



Canadian Bioinformatics Workshops

www.bioinformatics.ca

bioinformaticsdotca.github.io

Creative Commons

This page is available in the following languages:

Afrikaans Български Català Dansk Deutsch Ελληνικά English (CA) English (US) Esperanto
Castellano Castellano (AR) Español (CL) Castellano (CO) Español (Ecuador) Castellano (MX) Castellano (PE)
Euskara Suomiksi français français (CA) Galego മലബാറി ມາວລັກ ເംഗുഖ ພາສາ ພາສາ ພາສາ ພາສາ
Nederlands Norsk Sesotho sa Leboa polski Português română slovenski jezik српски srpski (latinica) Sotho svenska
中文 華語 (台灣) isiZulu



Attribution-Share Alike 2.5 Canada

You are free:



to Share — to copy, distribute and transmit the work



to Remix — to adapt the work



Under the following conditions:



Attribution. You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).



Share Alike. If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar licence to this one.

- For any reuse or distribution, you must make clear to others the licence terms of this work.
- Any of the above conditions can be waived if you get permission from the copyright holder.
- The author's moral rights are retained in this licence.

Disclaimer

Your fair dealing and other rights are in no way affected by the above.

This is a human-readable summary of the Legal Code (the full licence) available in the following languages:
[English](#) [French](#)

[Learn how to distribute your work using this licence](#)

Introduction to IGV

The Integrative Genomics Viewer



Malachi Griffith, Obi Griffith, Isabel Risch, Vida Talebian
RNA-seq Analysis 2024. June 17-19, 2024



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)

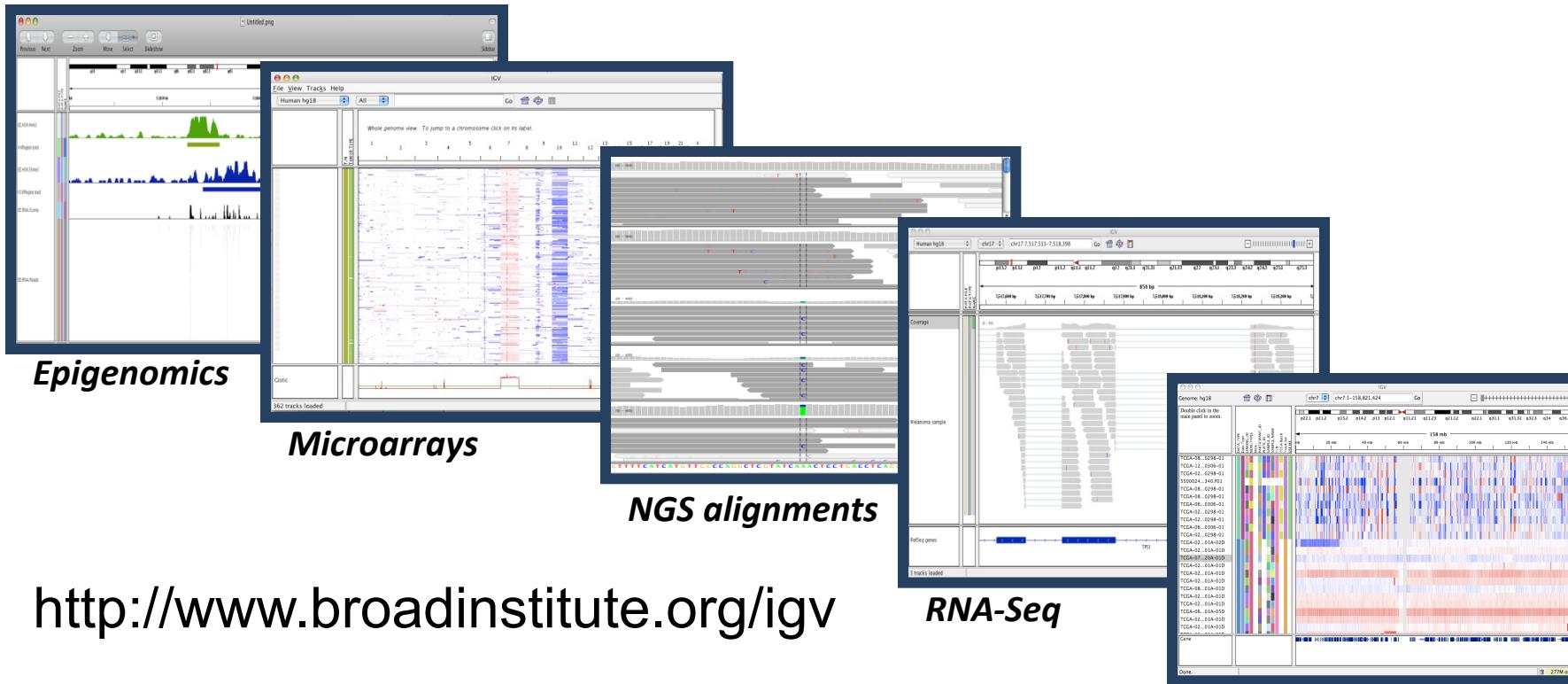


Savant
Genome
Browser

- 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>

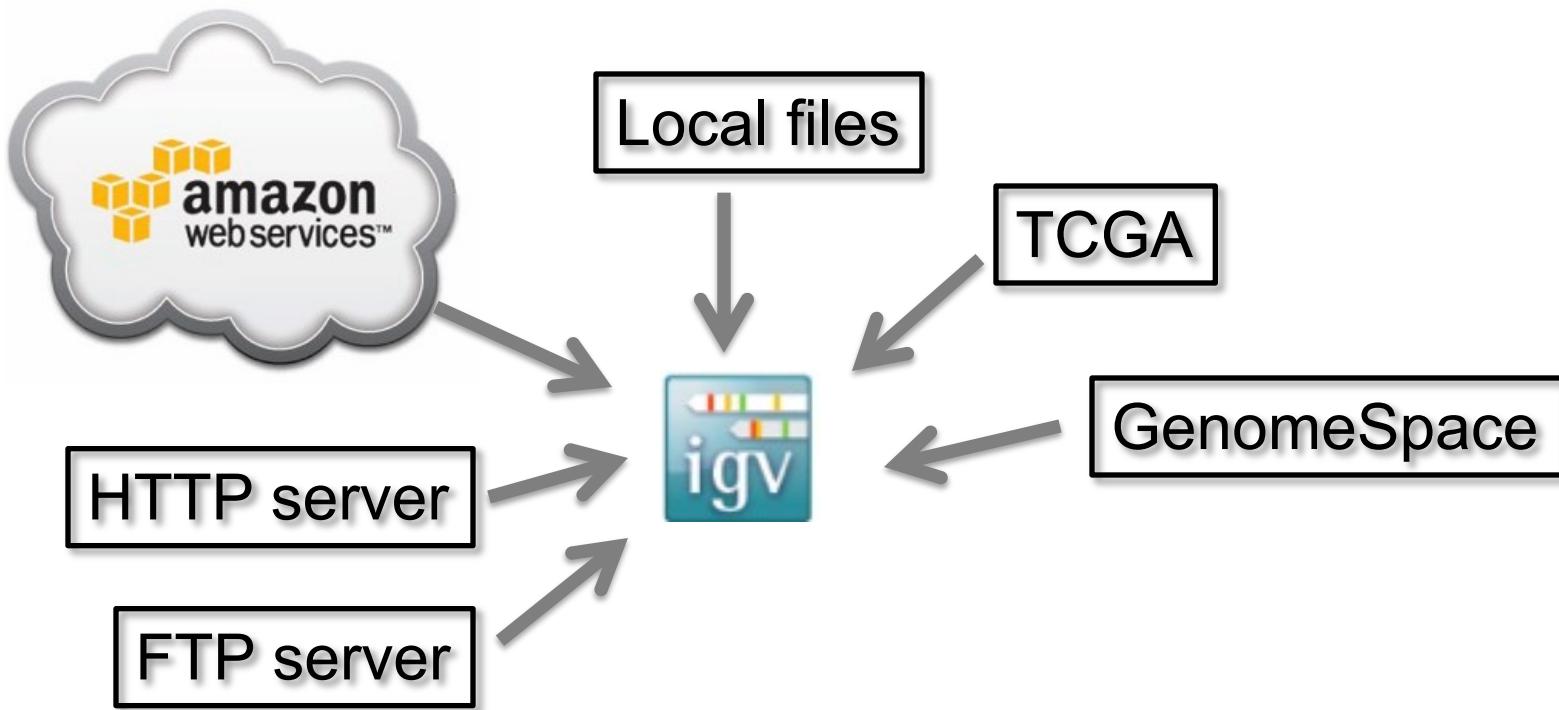
>85,000 registrations (2014)

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

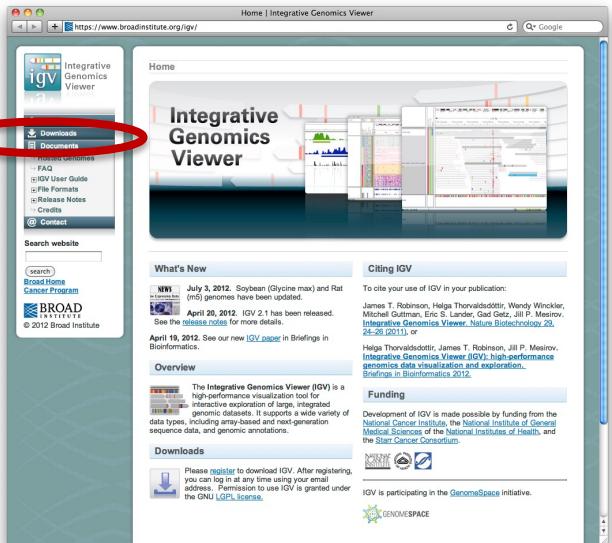


- View **local** files without uploading.
- View **remote** files without downloading the whole dataset.

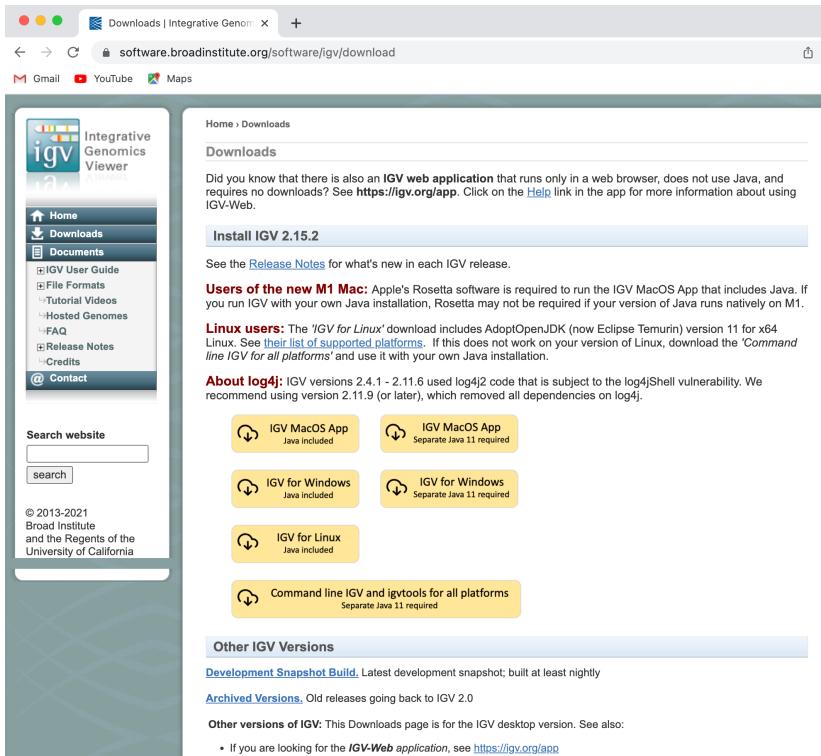
Using IGV: the basics

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

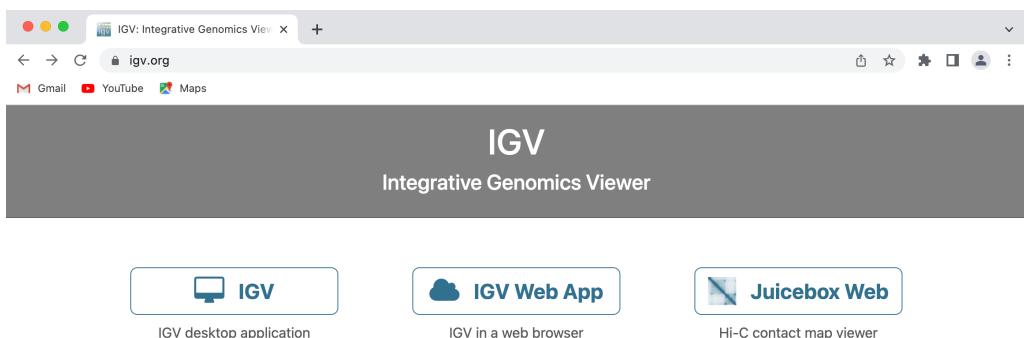
Launch IGV

The screenshot shows the IGV website's main page. A red circle highlights the 'Downloads' menu item in the top navigation bar. Below it, a sidebar lists various sections like 'Home', 'FAQ', 'B-IGV User Guide', 'B-File Formats', 'Release Notes', and 'Credits'. The main content area features a large image of the IGV interface and sections for 'What's New' and 'Citing IGV'. The 'Downloads' section is expanded, showing links for 'IGV Mac OS App' (Java included), 'IGV Mac OS App' (Separate Java 11 required), 'IGV for Windows' (Java included), 'IGV for Windows' (Separate Java 11 required), 'IGV for Linux' (Java included), and 'Command line IGV and igvtools for all platforms' (Separate Java 11 required).



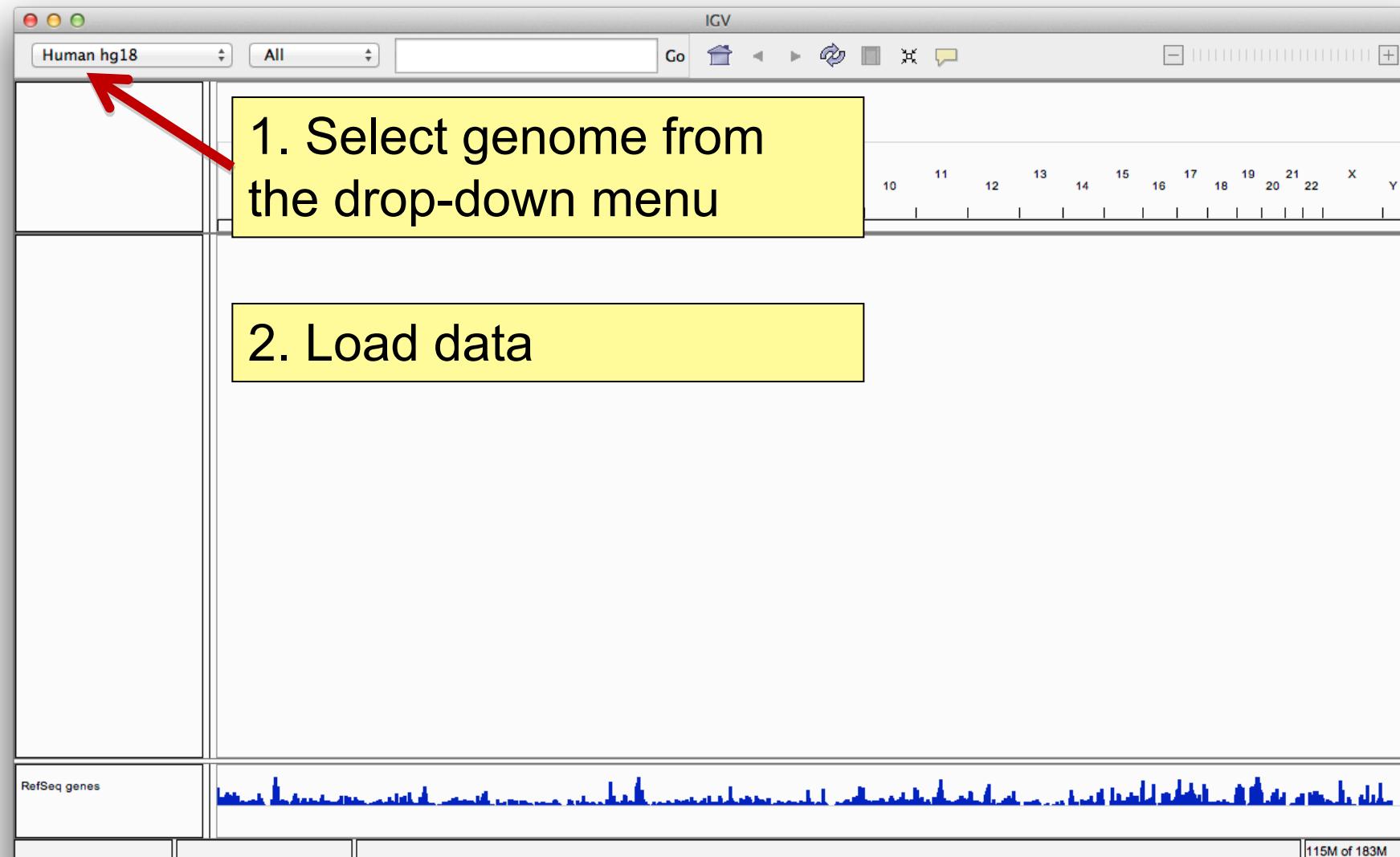
The screenshot shows the IGV Downloads page. It starts with a brief introduction about the IGV web application. The 'Install IGV 2.15.2' section is highlighted with a red circle. Below it, there's information for 'Users of the new M1 Mac', 'Linux users', and 'About log4j'. The page then lists download options for different platforms: 'IGV Mac OS App' (Java included), 'IGV Mac OS App' (Separate Java 11 required), 'IGV for Windows' (Java included), 'IGV for Windows' (Separate Java 11 required), 'IGV for Linux' (Java included), and 'Command line IGV and igvtools for all platforms' (Separate Java 11 required). At the bottom, there's a section for 'Other IGV Versions'.



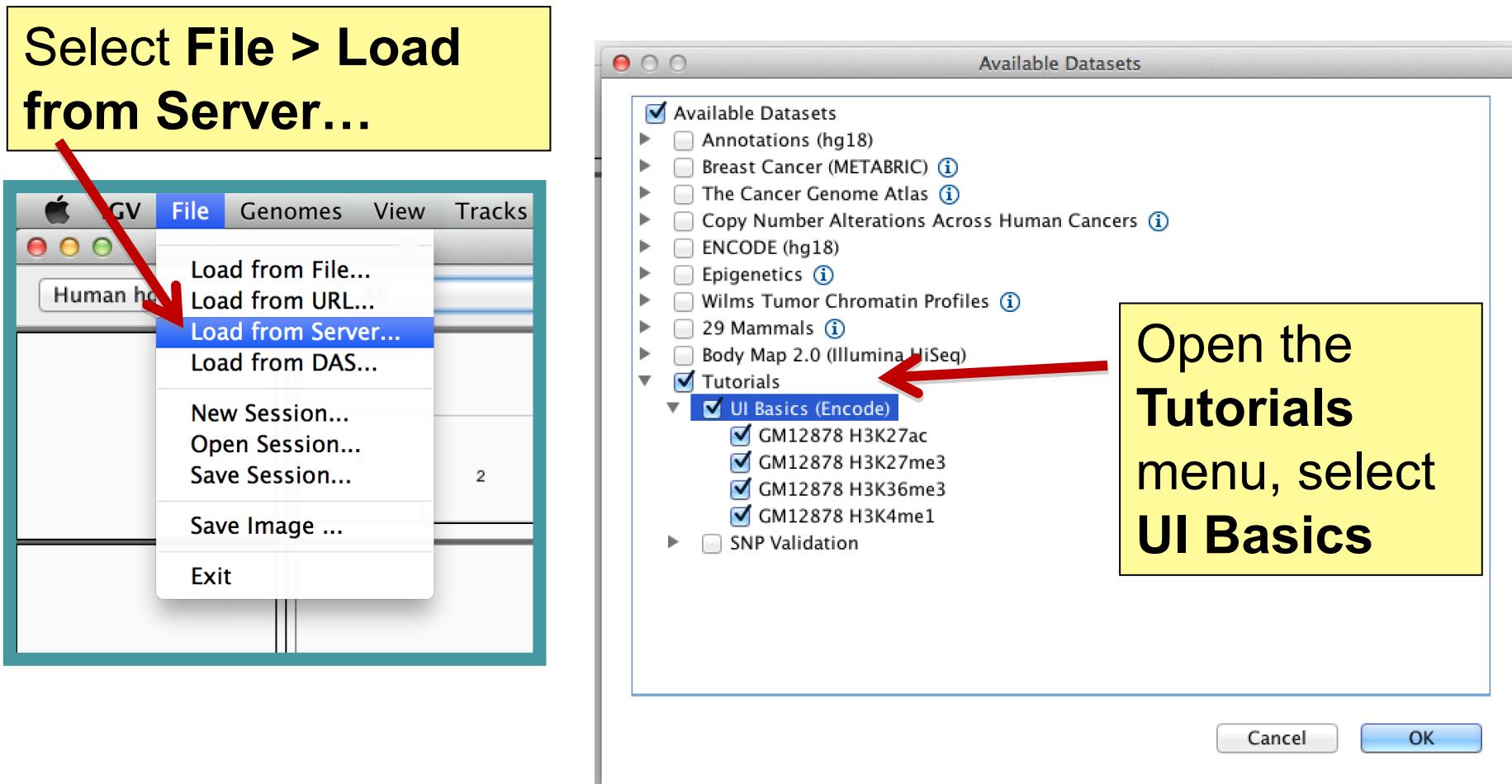
The screenshot shows the IGV desktop application's dock. It includes icons for the IGV desktop application, the IGV Web App, and the Juicebox Web viewer. The IGV desktop application icon is circled in red.



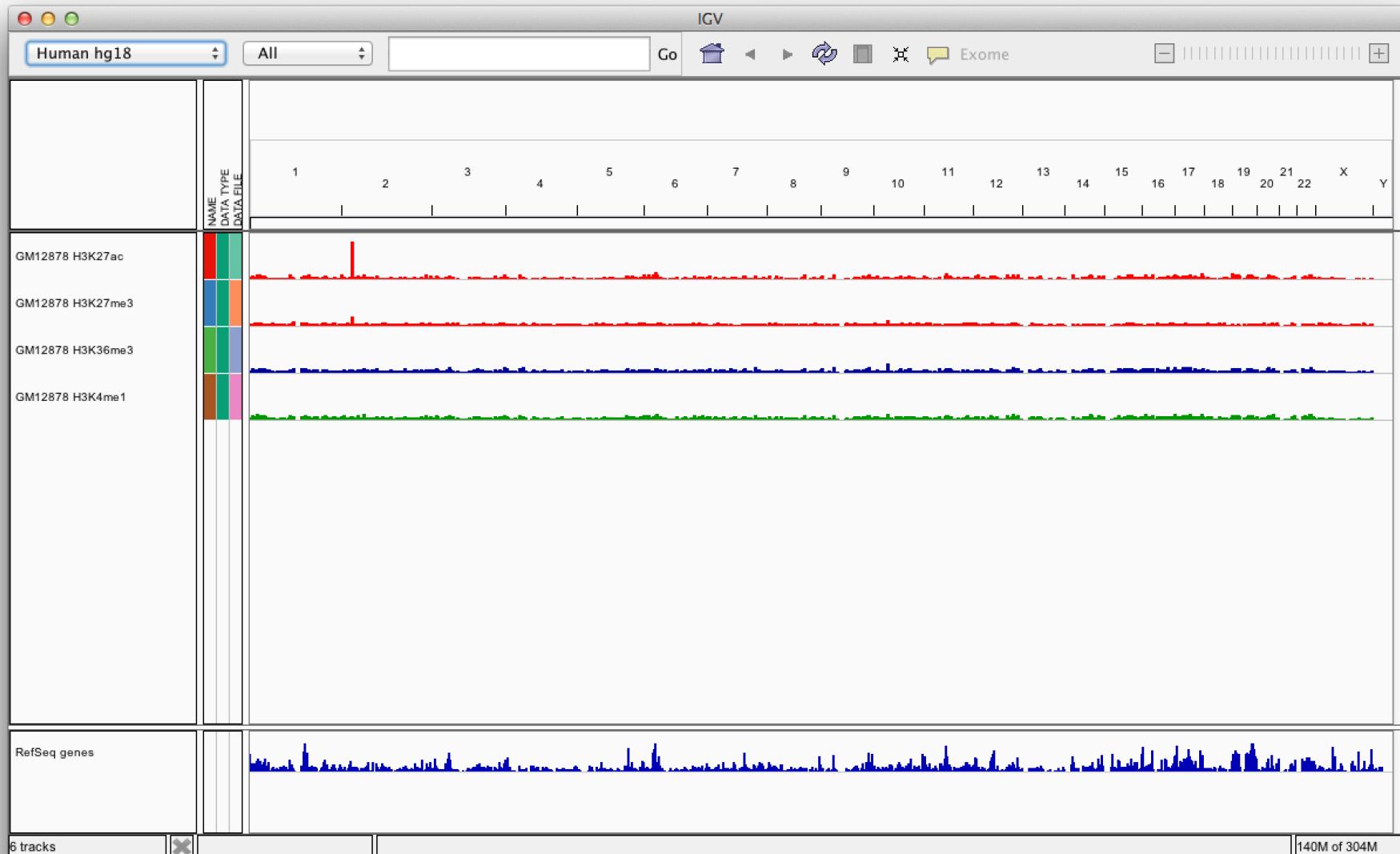
Select reference genome



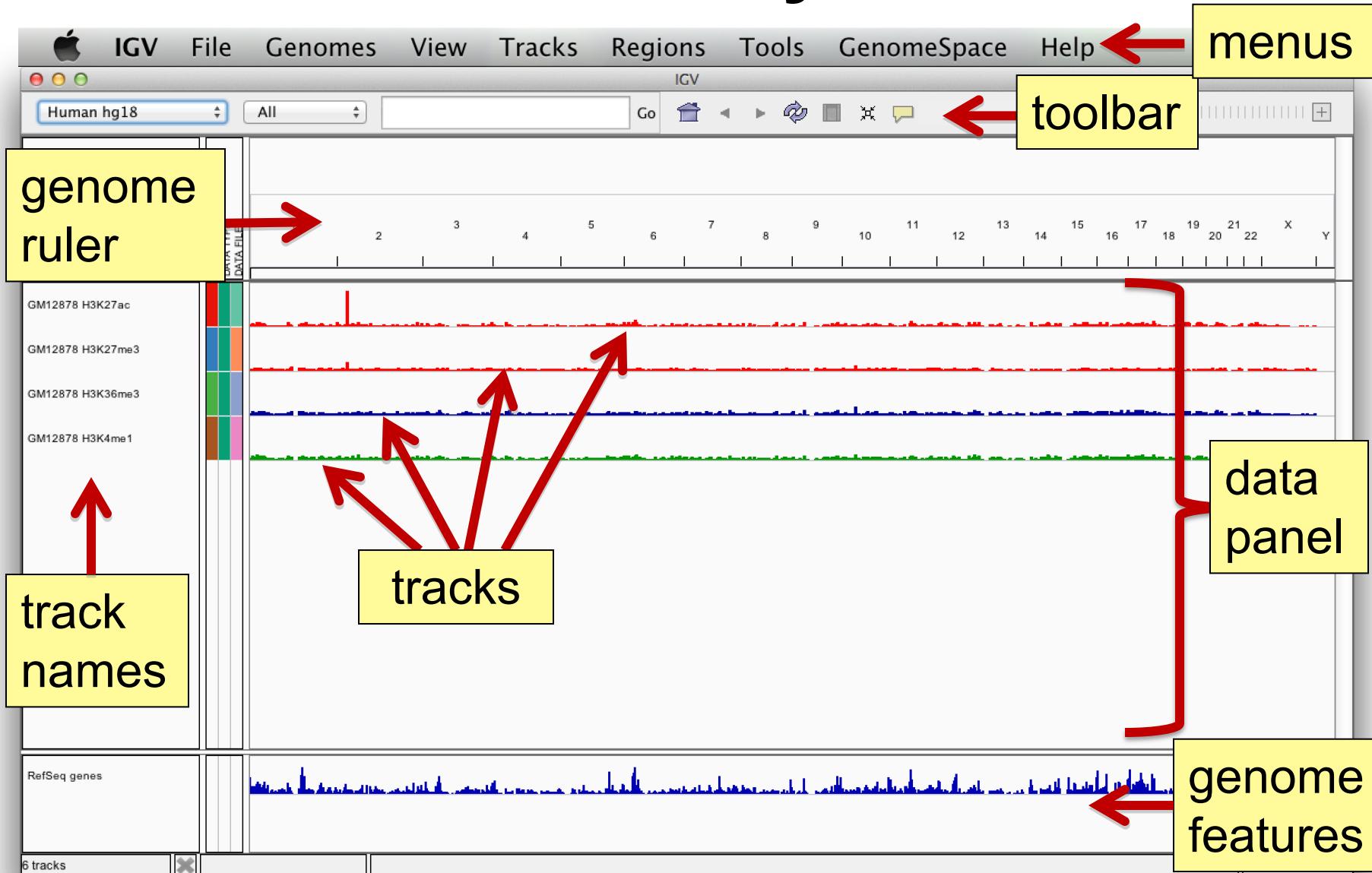
Load data



Screen layout



Screen layout



File formats and track types

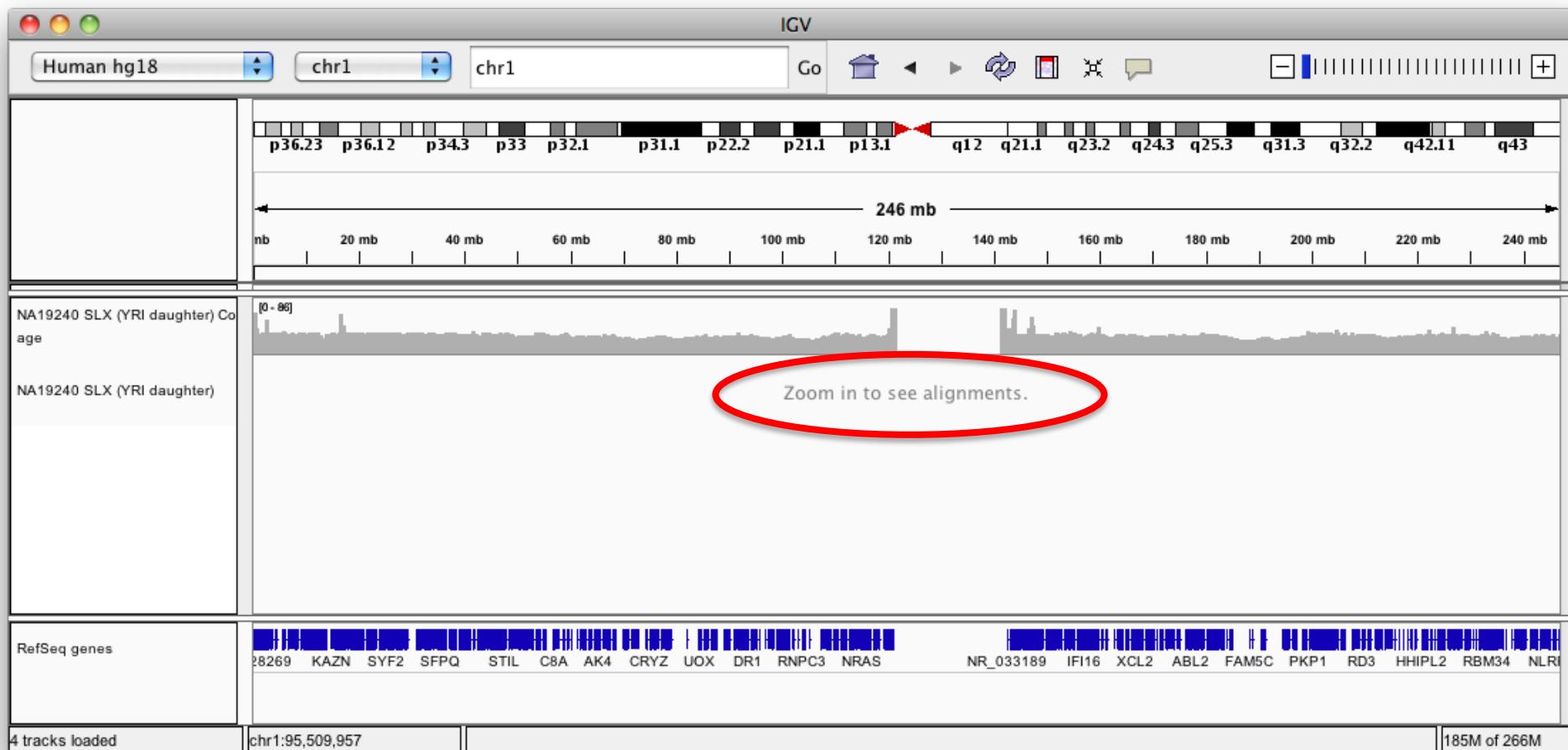
- The **file format** defines the track type.
- The **track type** determines the display options

- [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](#)

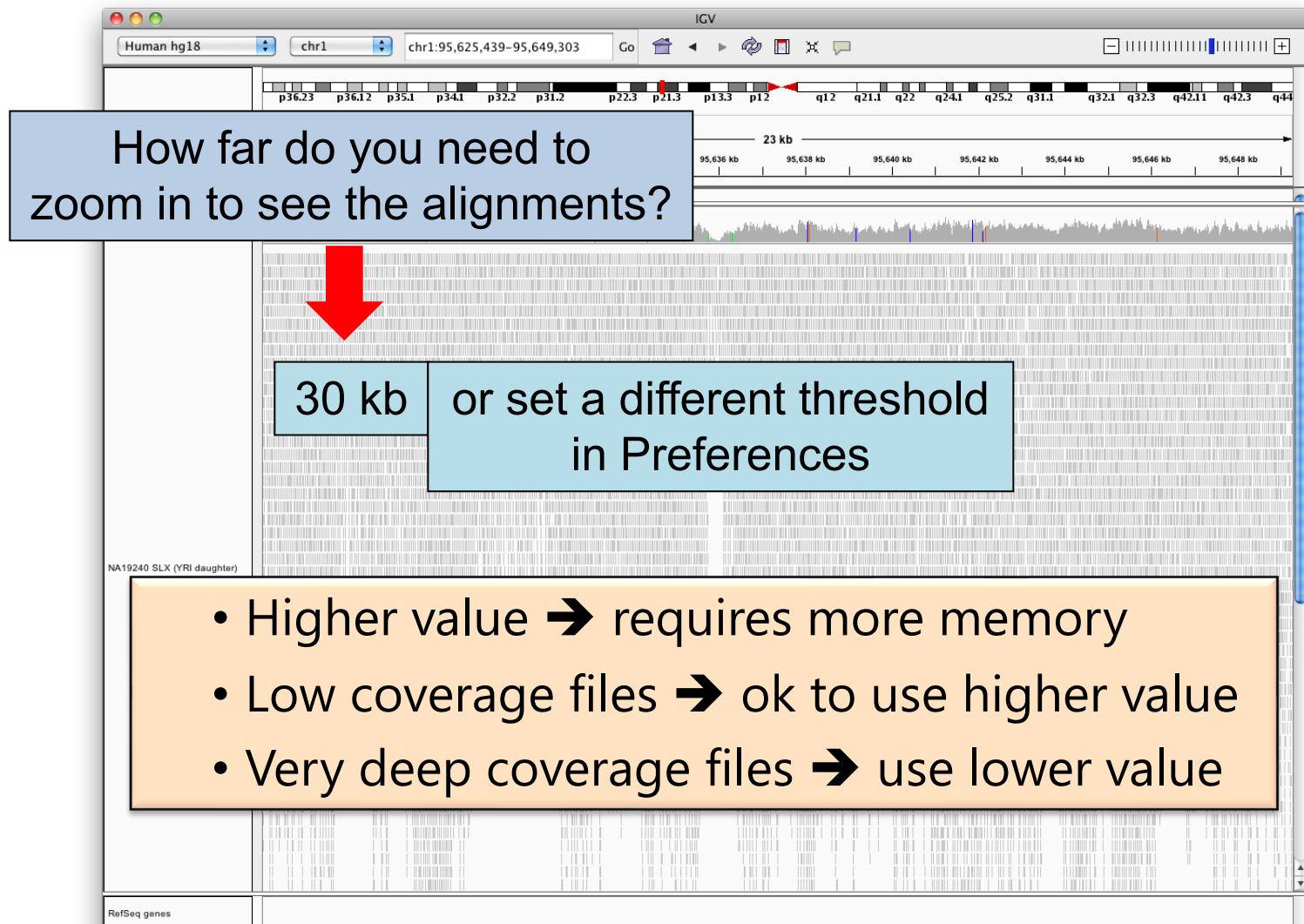
- For current list see: <https://software.broadinstitute.org/software/igv/FileFormats>

Viewing alignments

Whole chromosome view



Viewing alignments – Zoom in



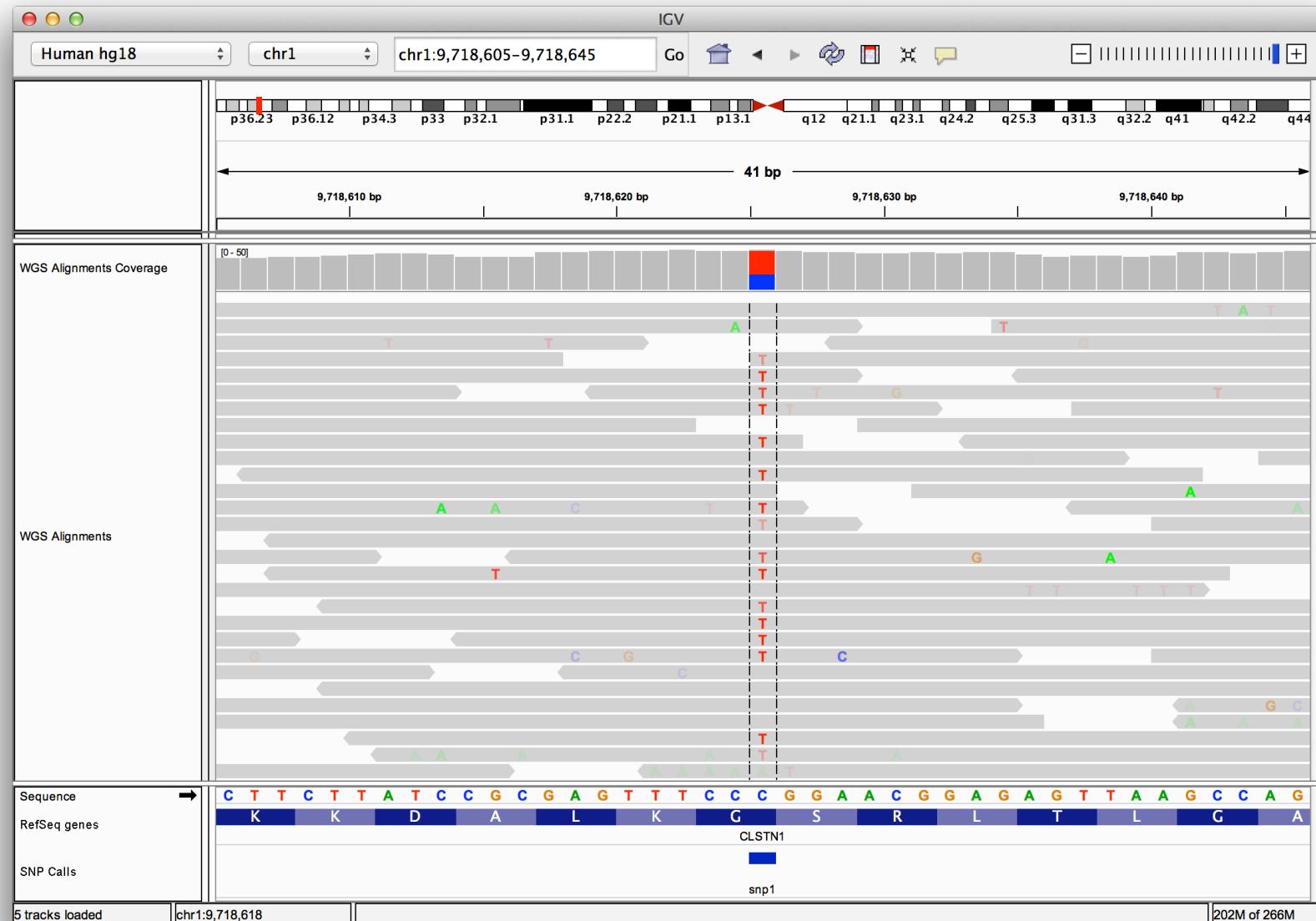
Viewing alignments – Zoom in



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



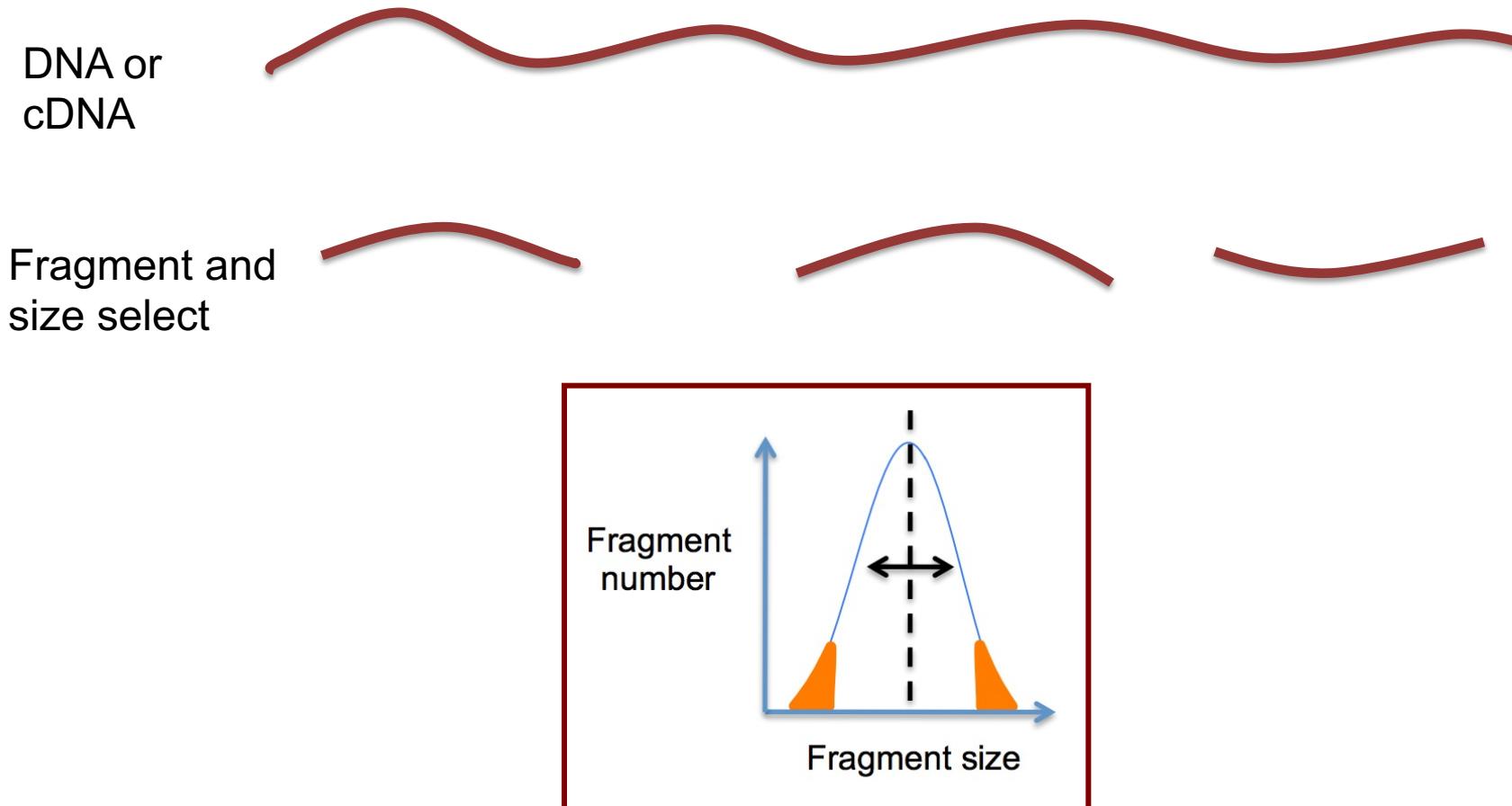
Viewing SNPs and SNVs



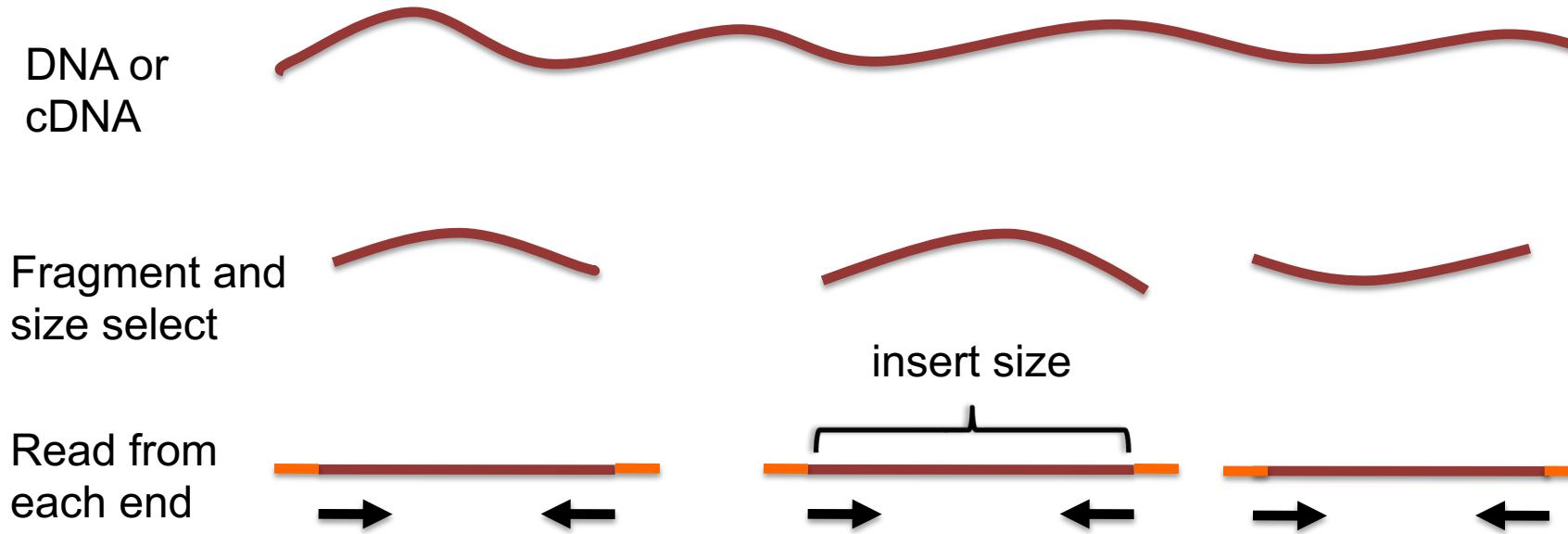
Viewing Structural Events

- 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)

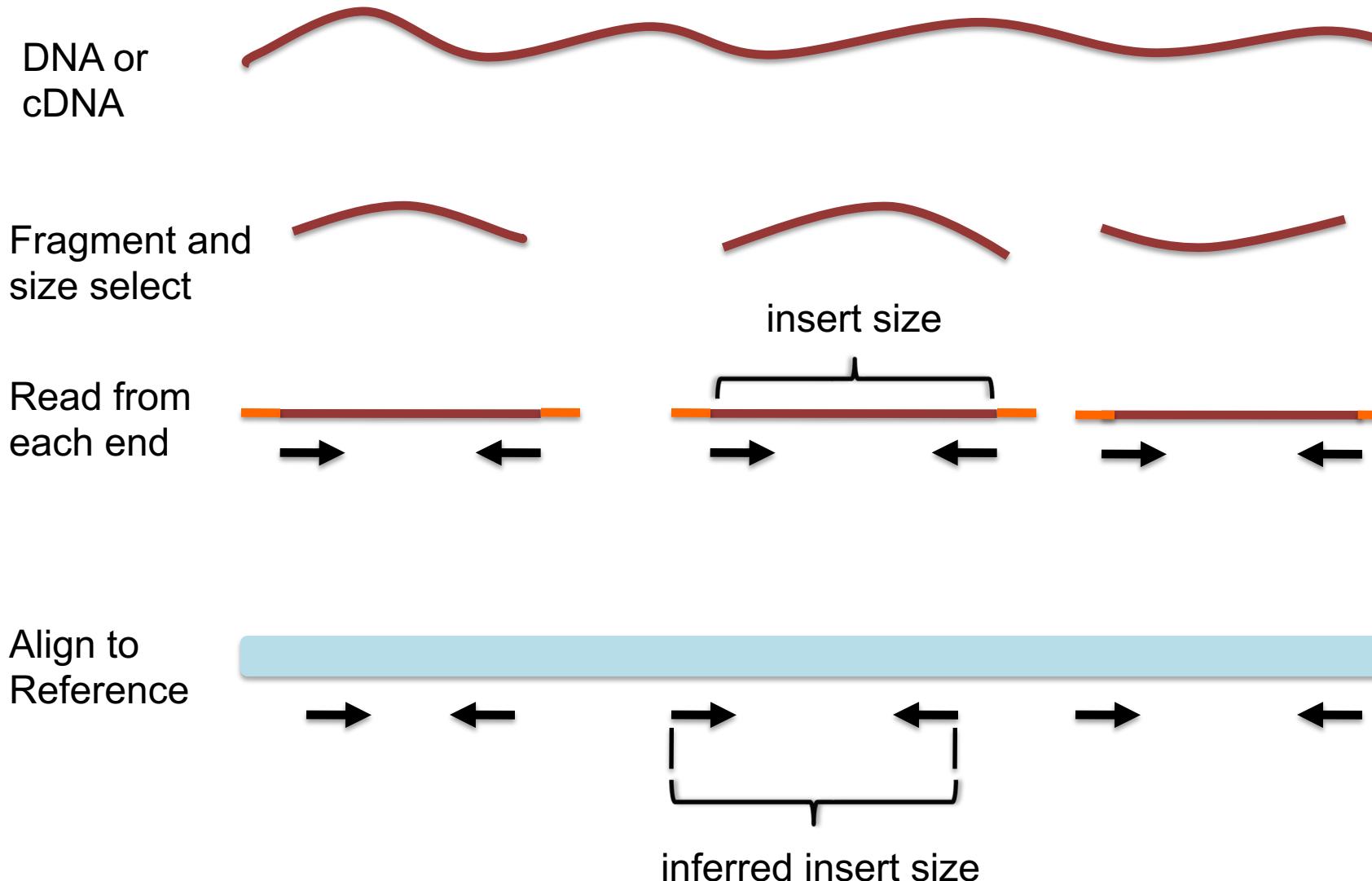
Paired-end sequencing



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

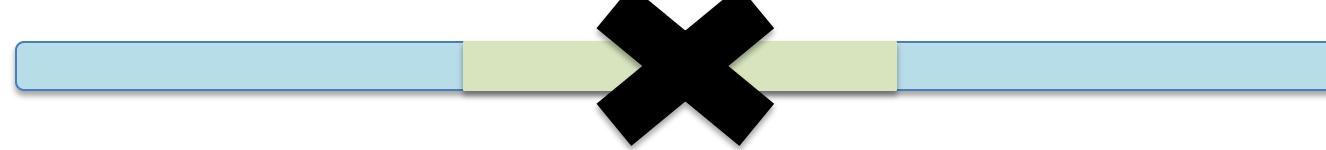
What is the effect of a deletion on inferred insert size?

Deletion

Reference
Genome



Subject



Deletion

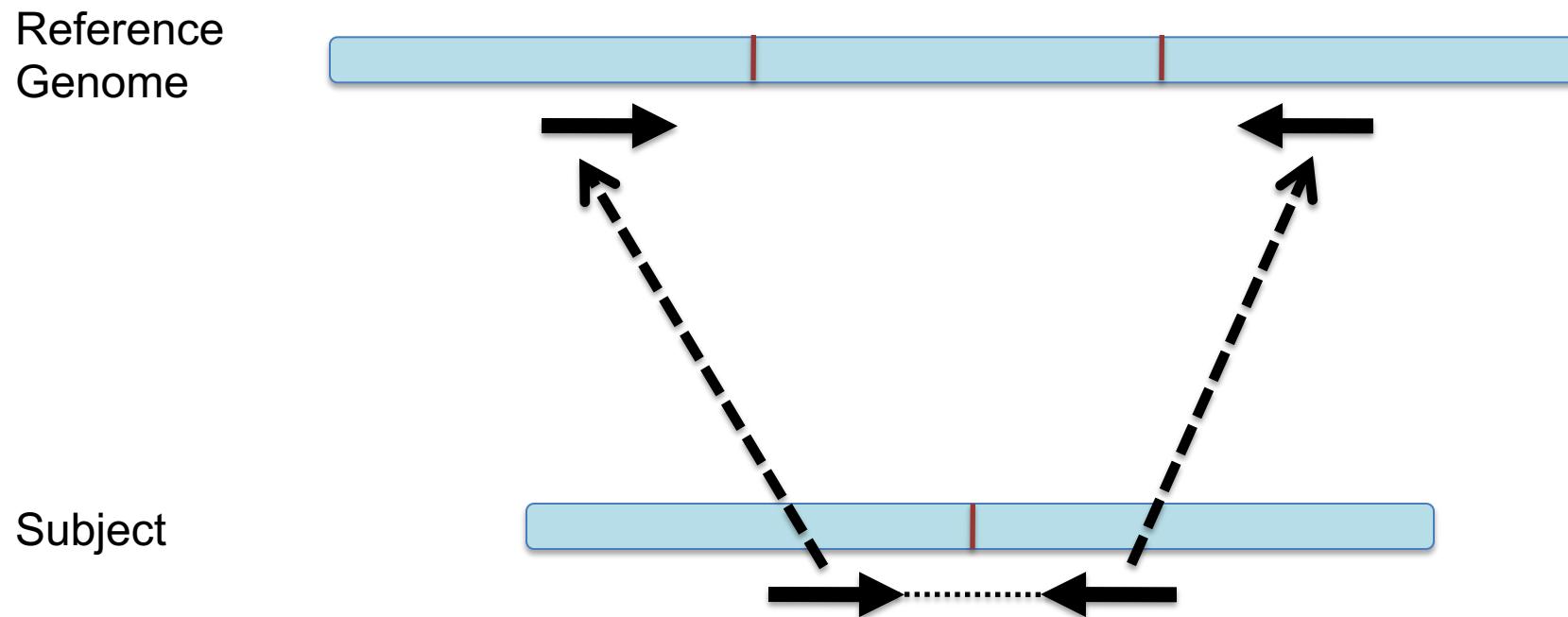
Reference
Genome



Subject



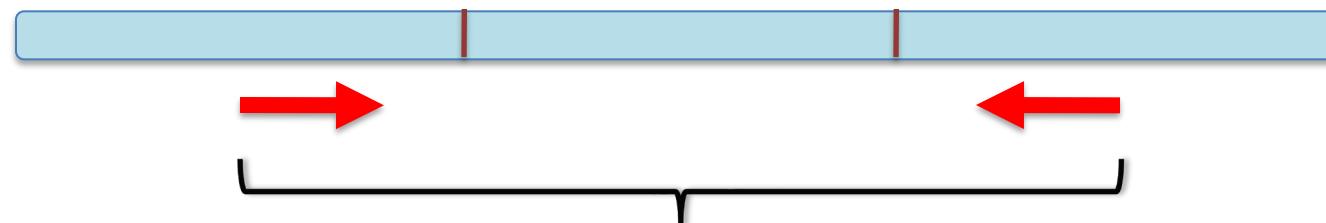
Deletion



Deletion

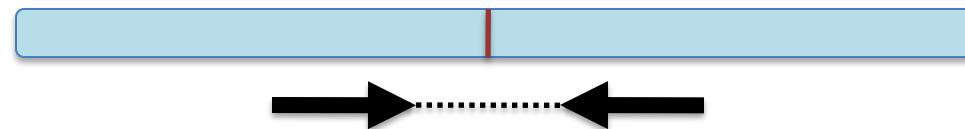
Inferred insert size is > expected value

Reference
Genome



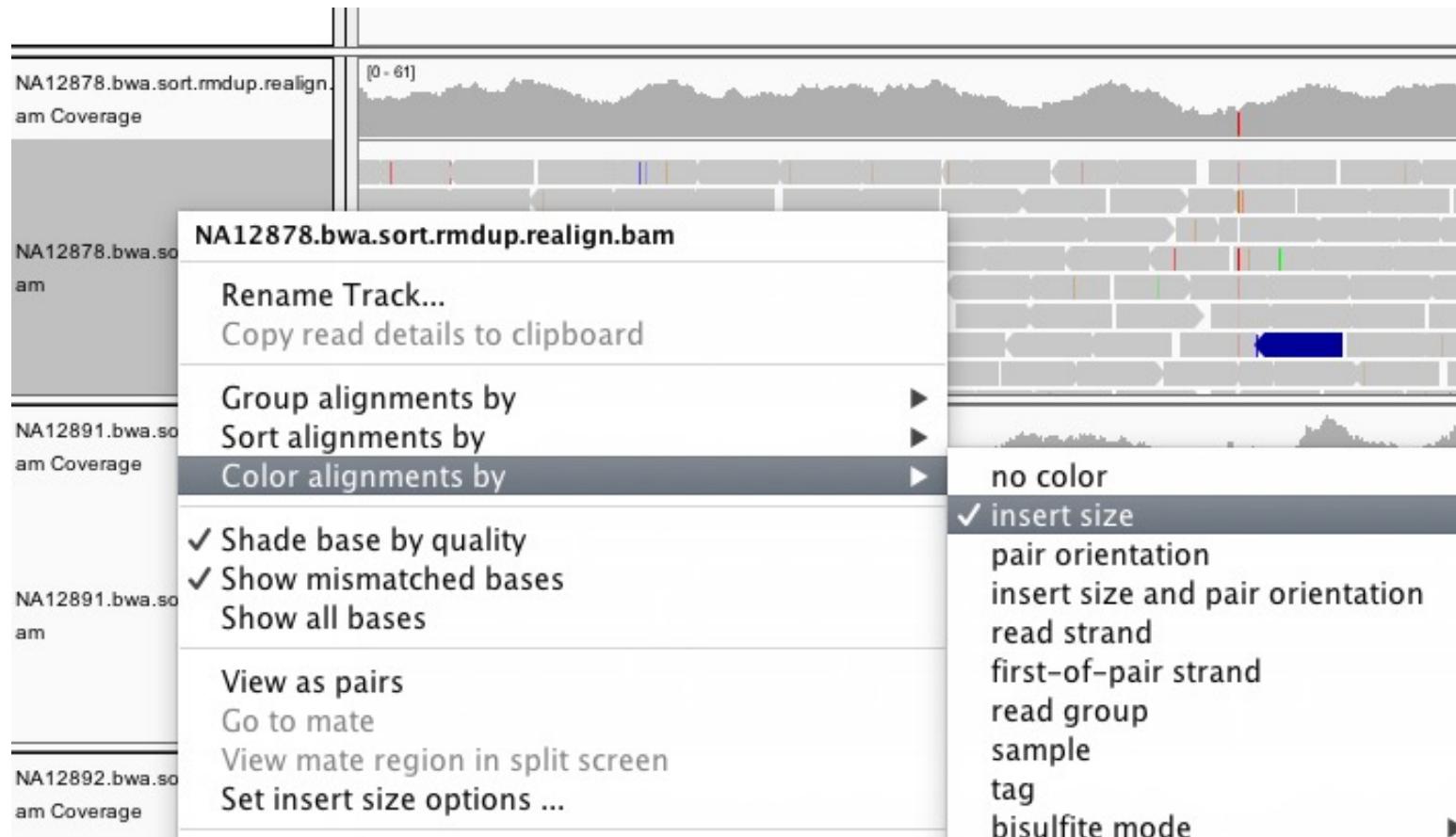
inferred insert size

Subject

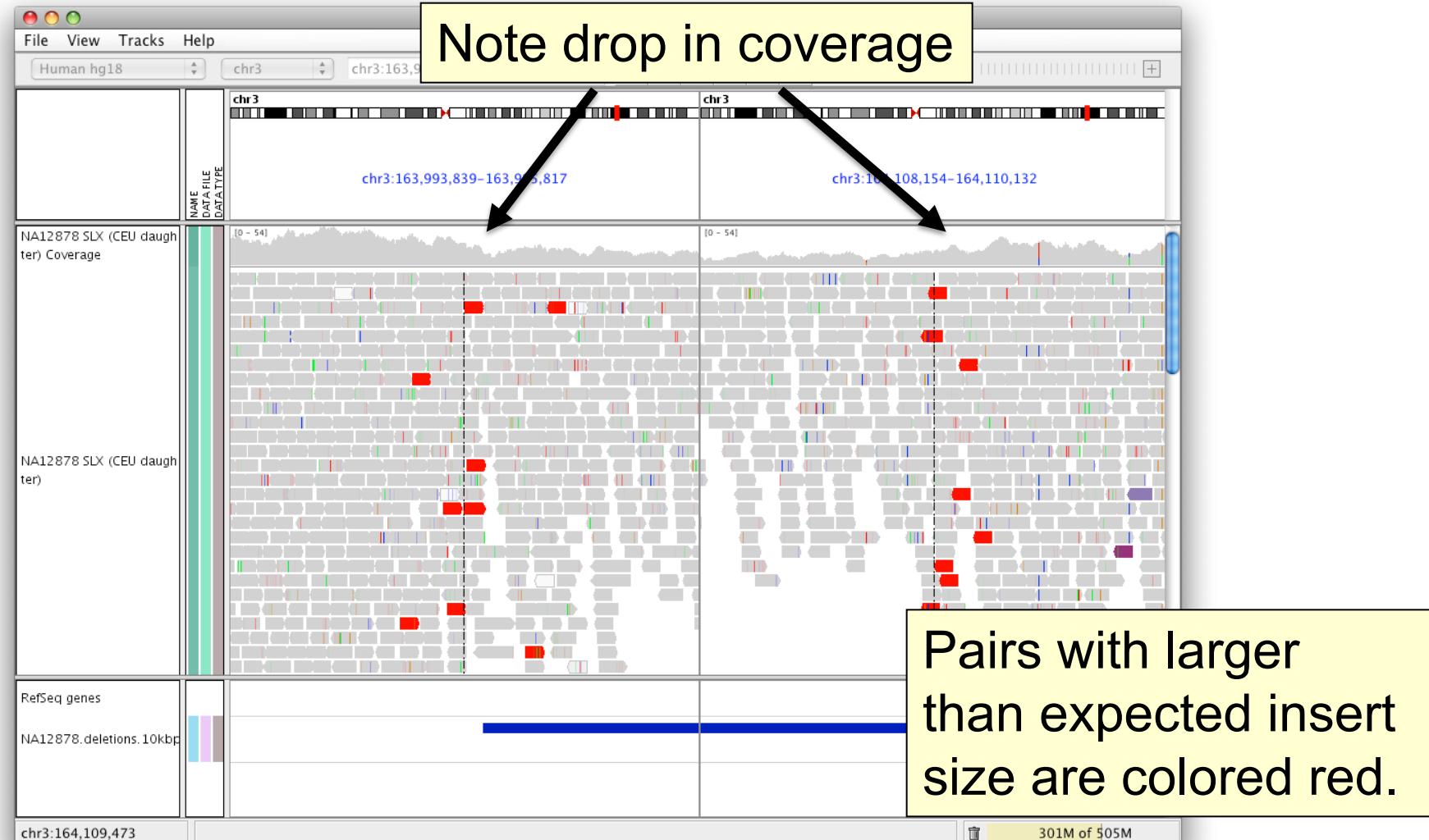


expected insert size

Color by insert size



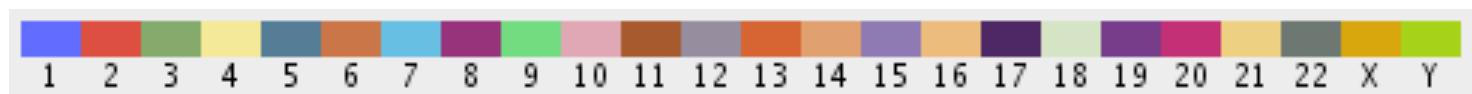
Deletion



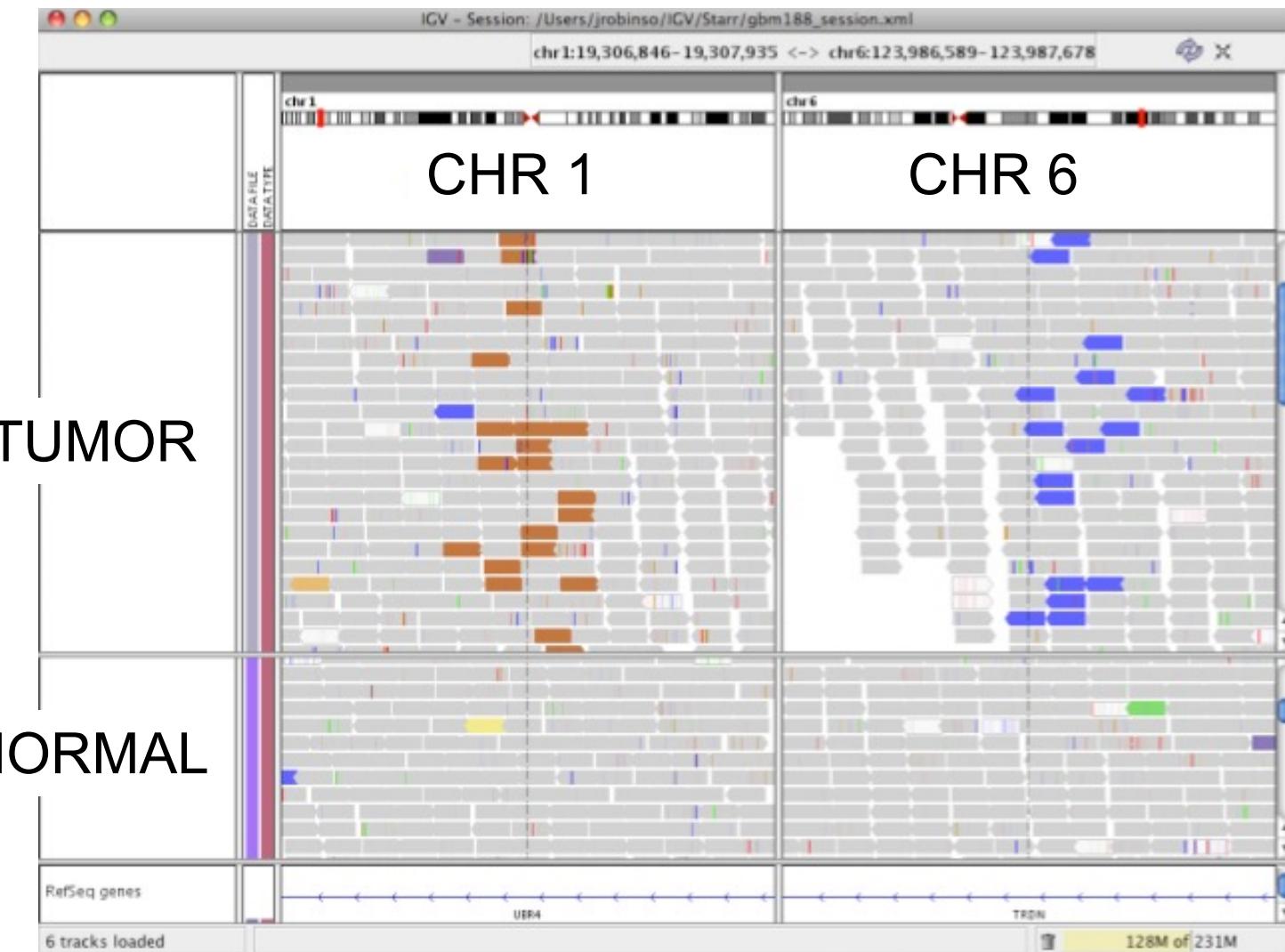
Insert size color scheme

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

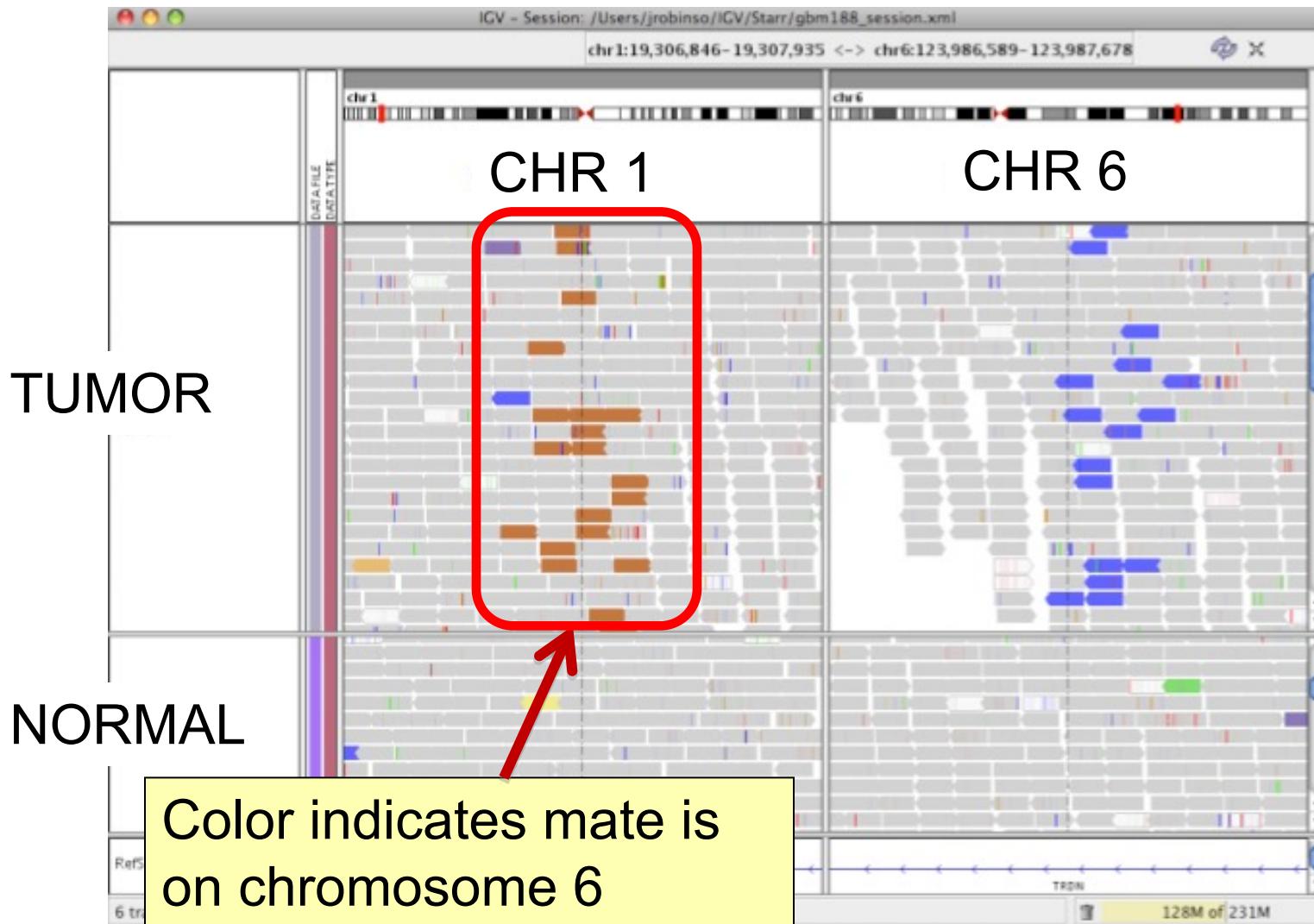
Each end colored by chromosome of its mate



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

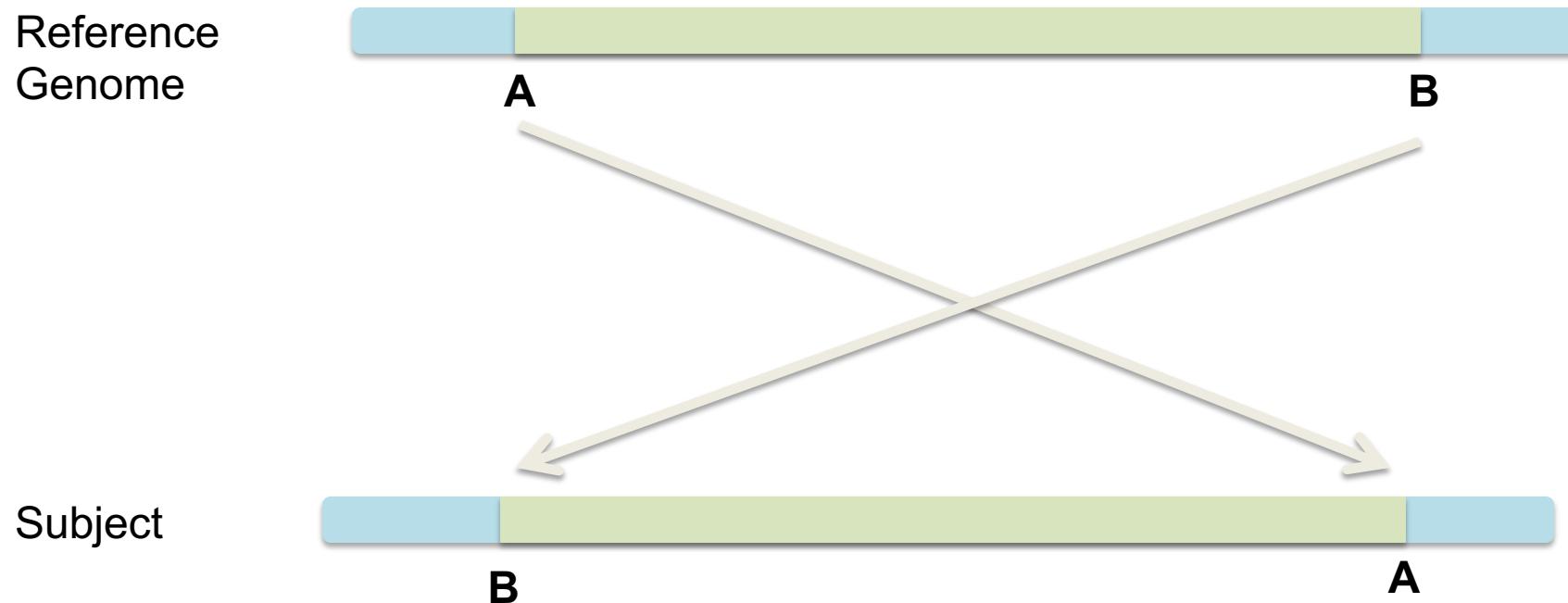
Reference
genome

Inversion

Reference
genome



Inversion

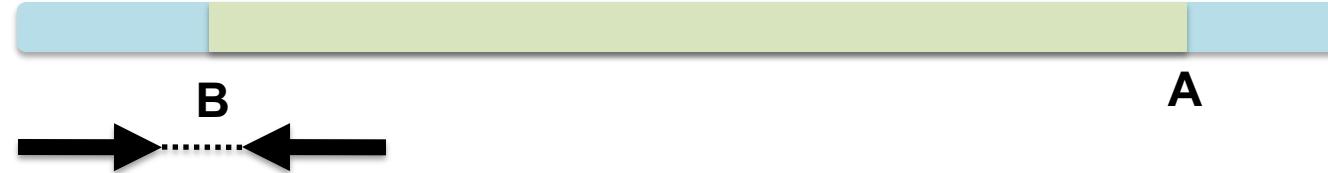


Inversion

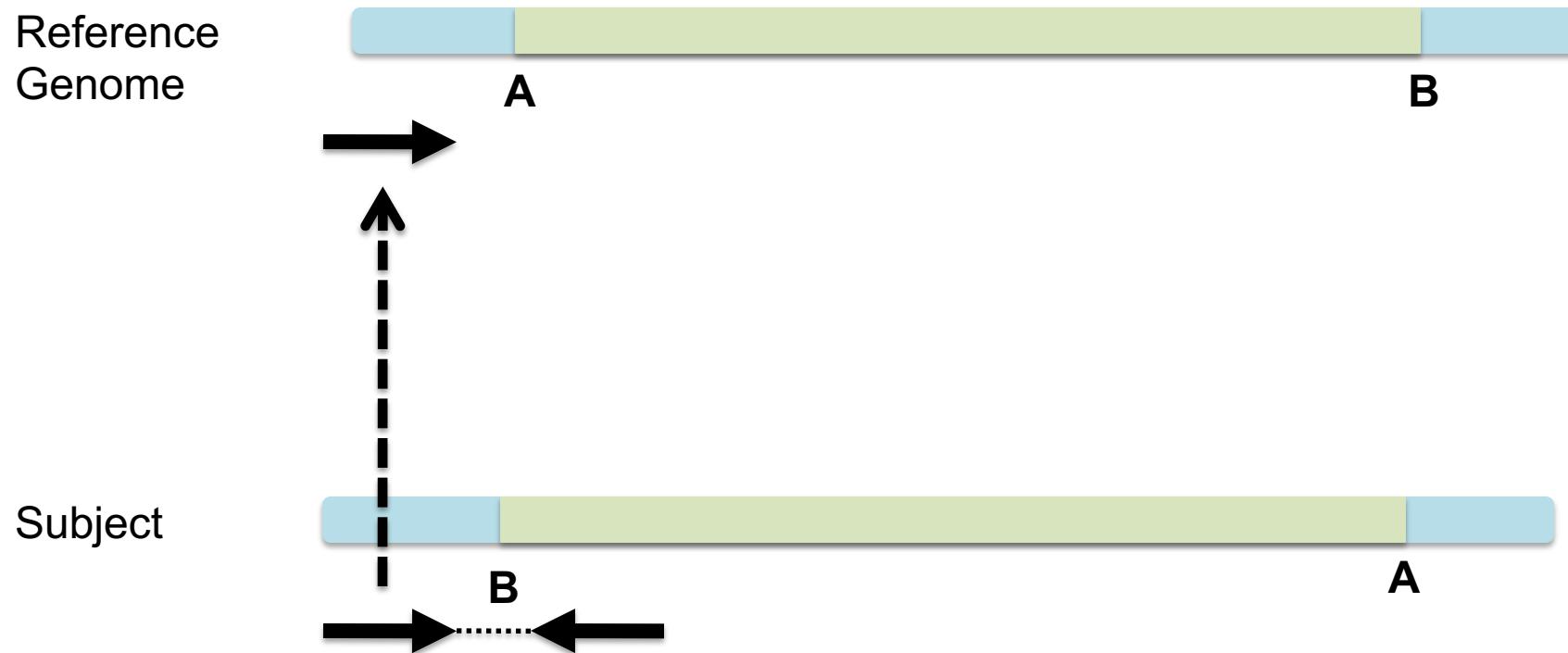
Reference
Genome



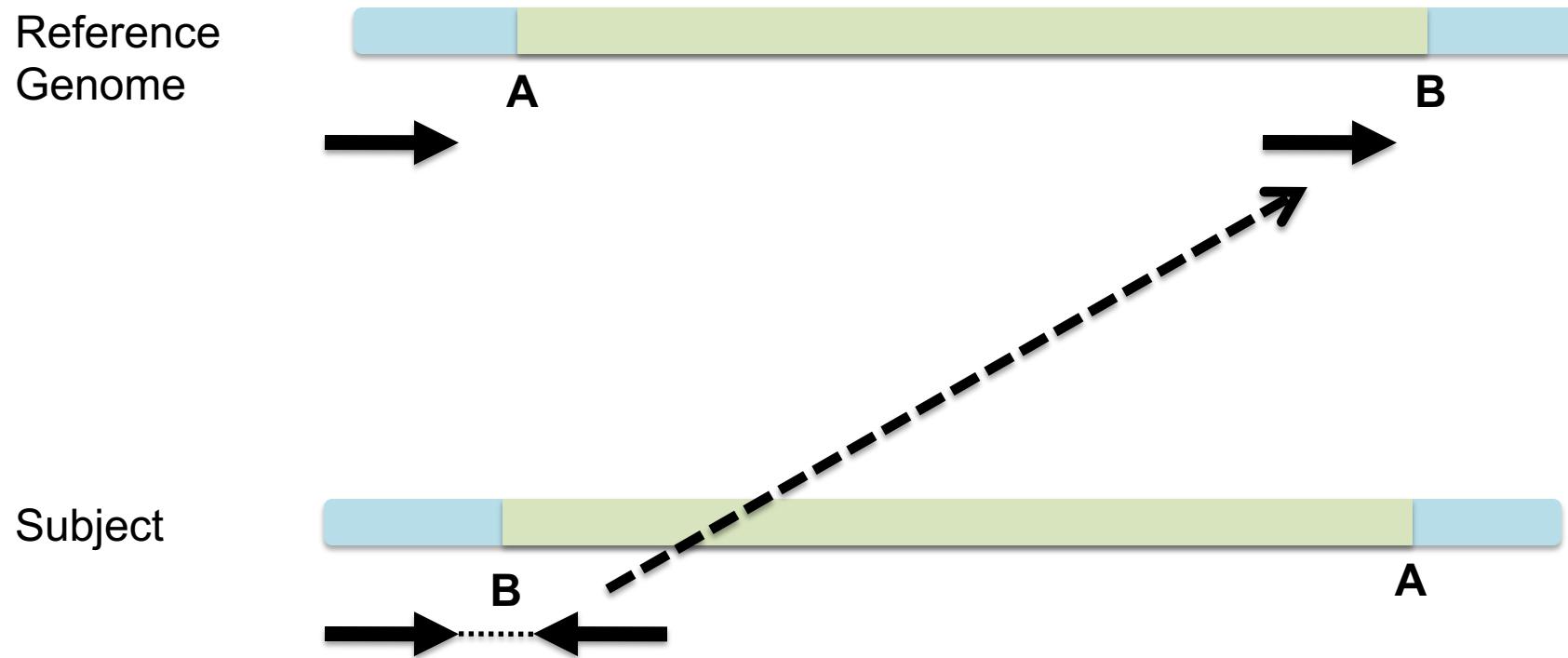
Subject



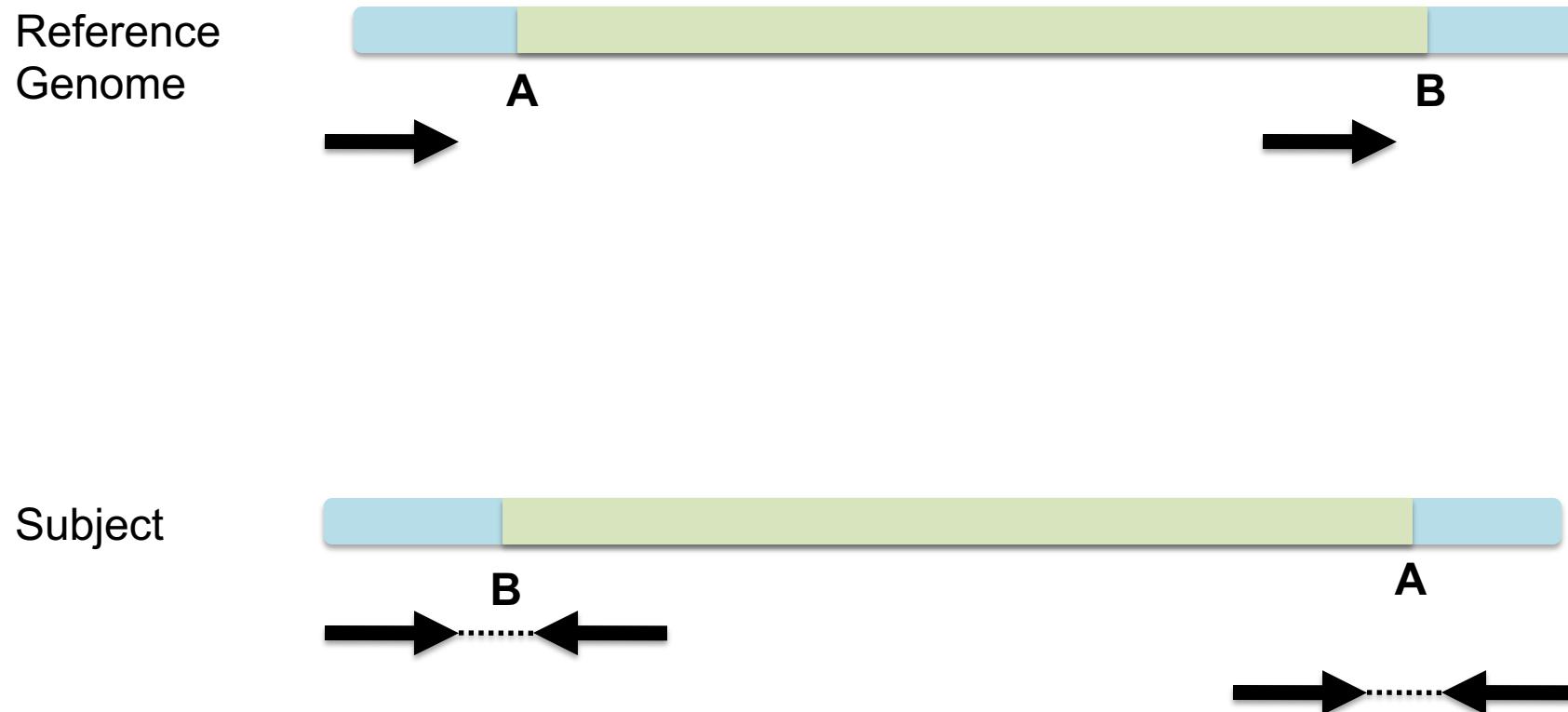
Inversion



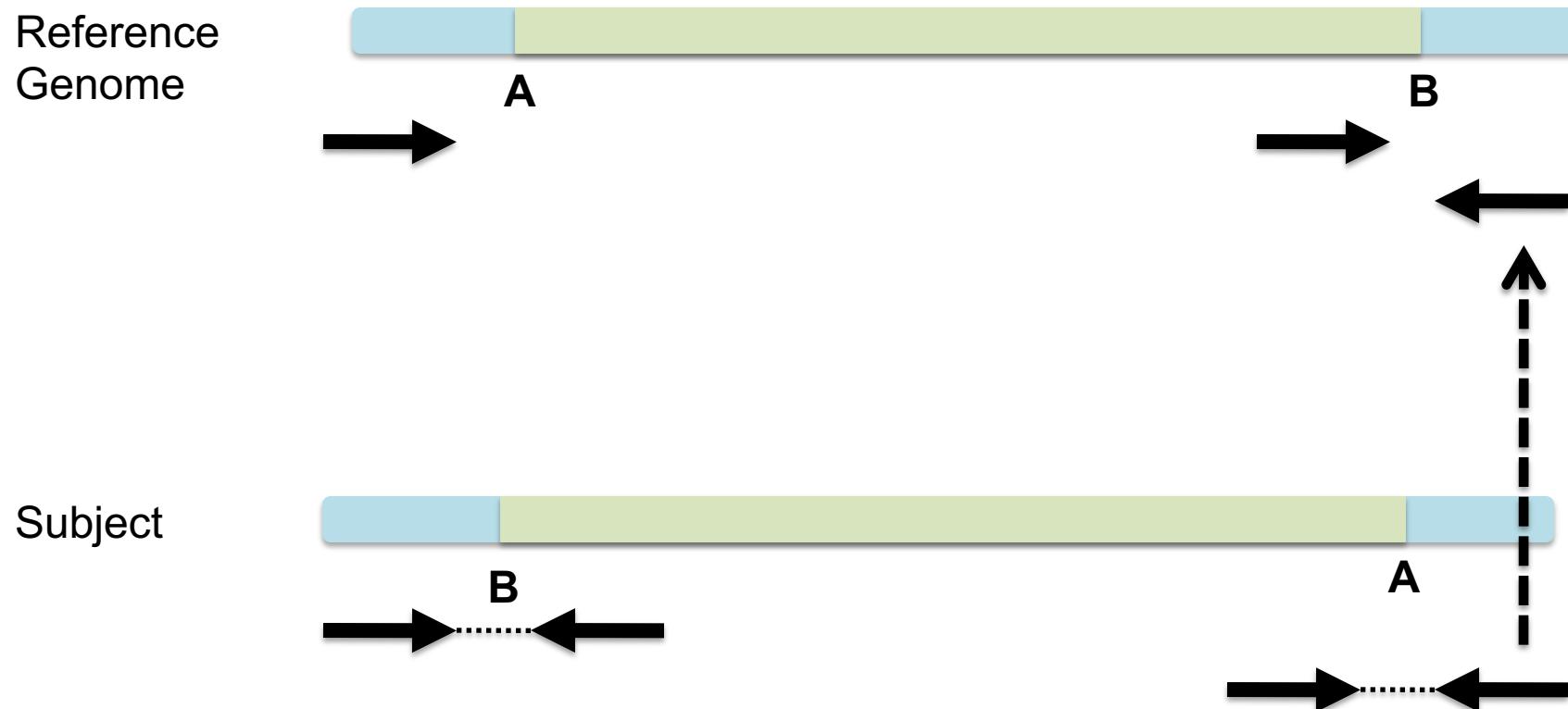
Inversion



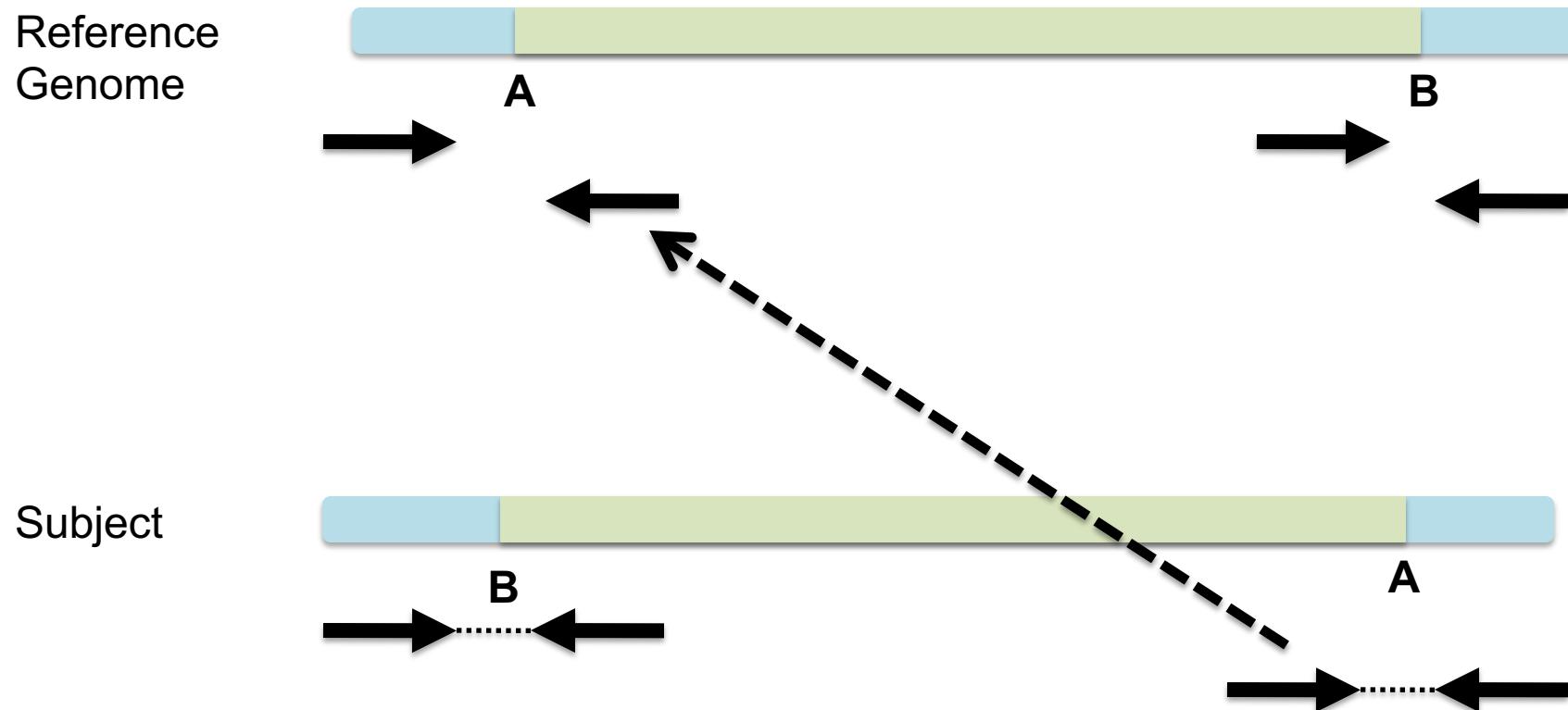
Inversion



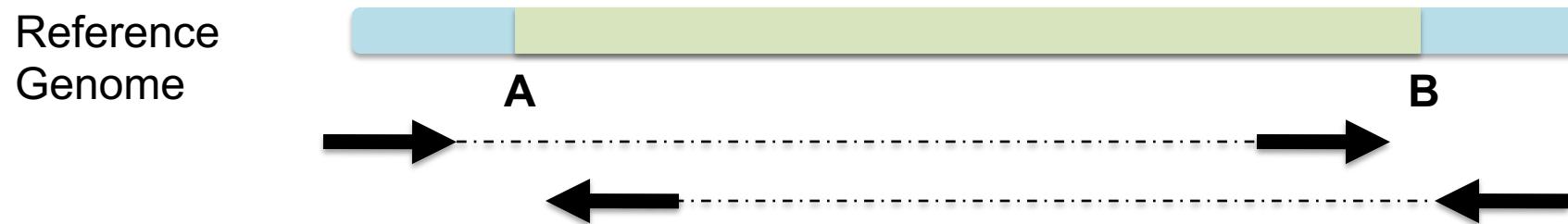
Inversion



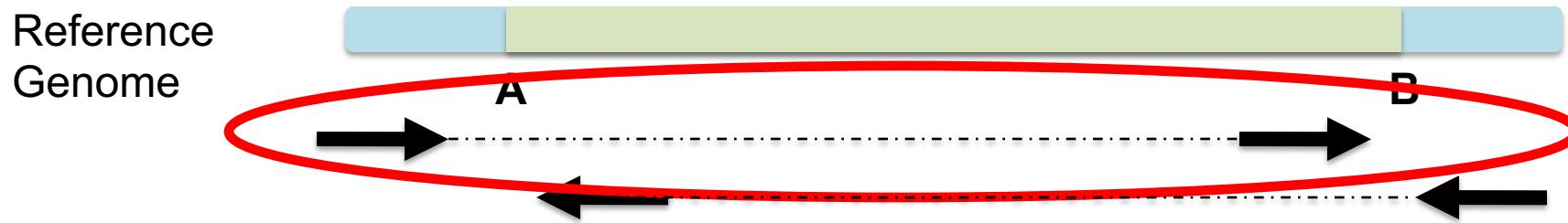
Inversion



Inversion

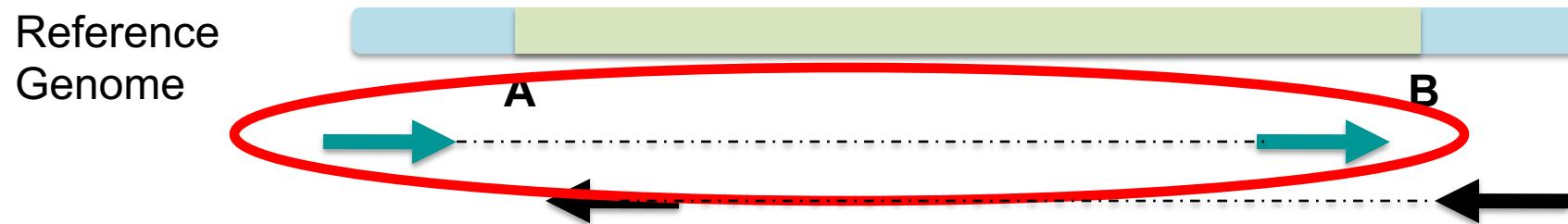


Inversion



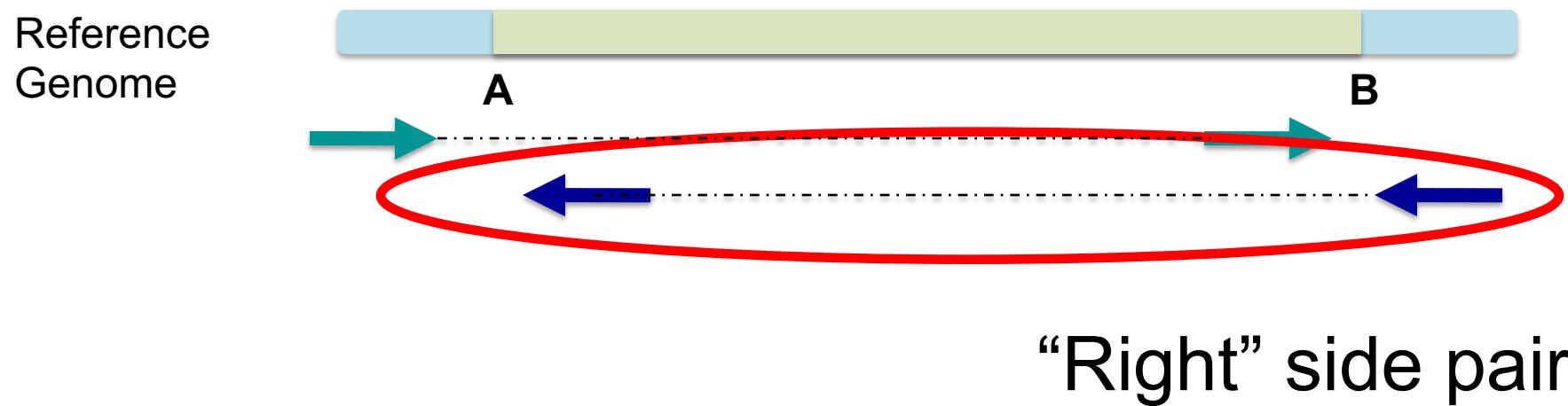
Anomaly: expected orientation of pair is inward facing (→ ←)

Inversion

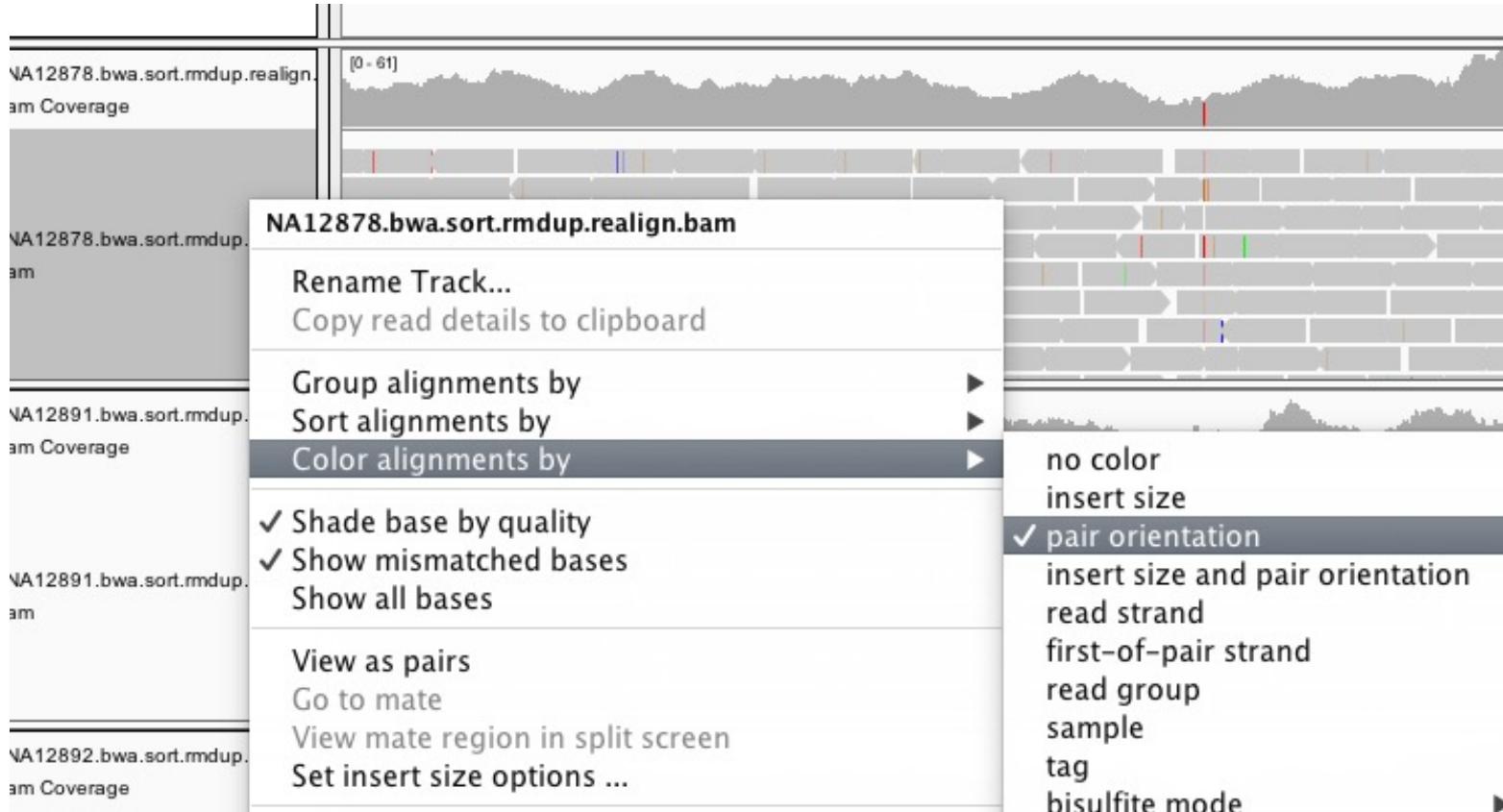


“Left” side pair

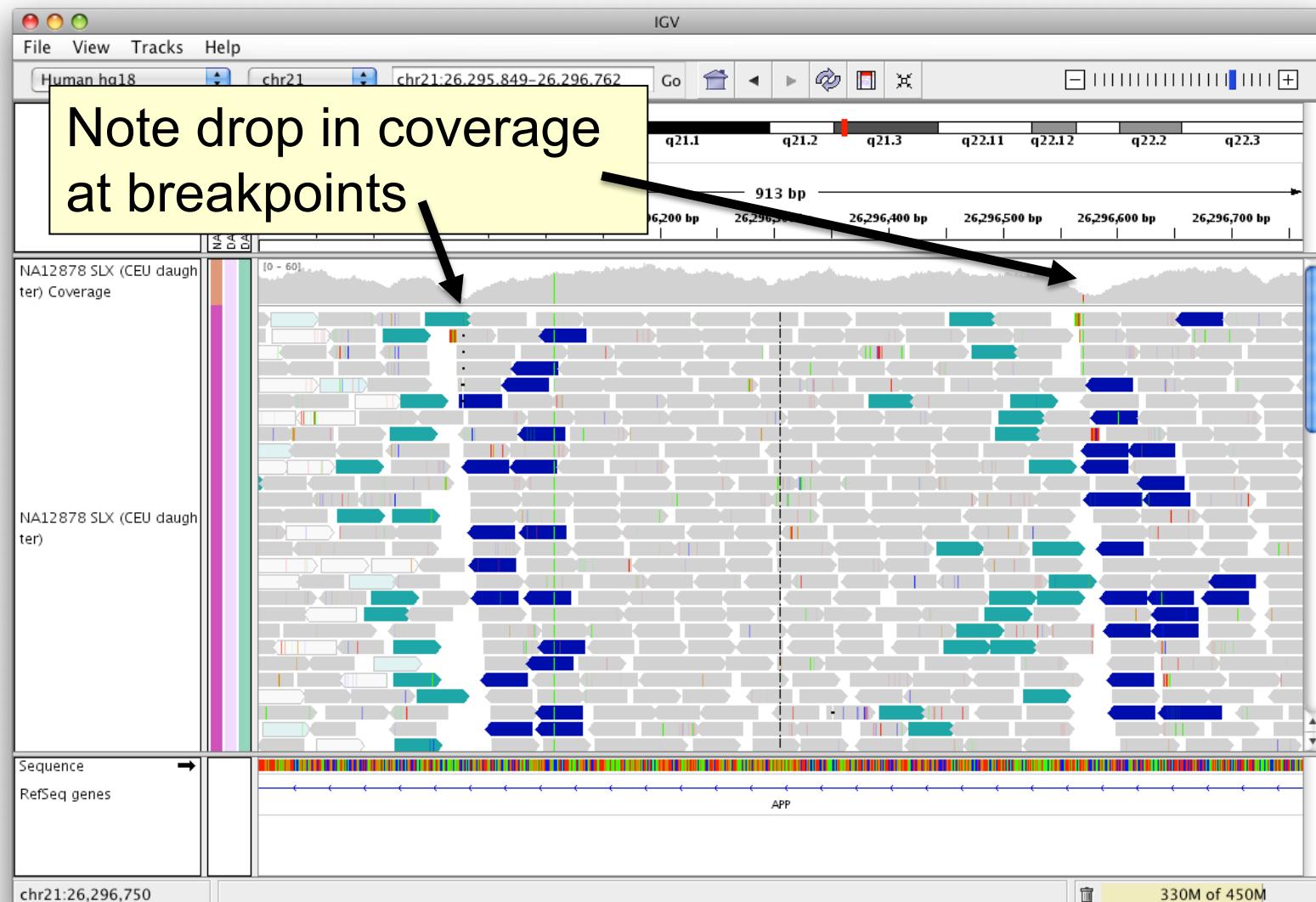
Inversion



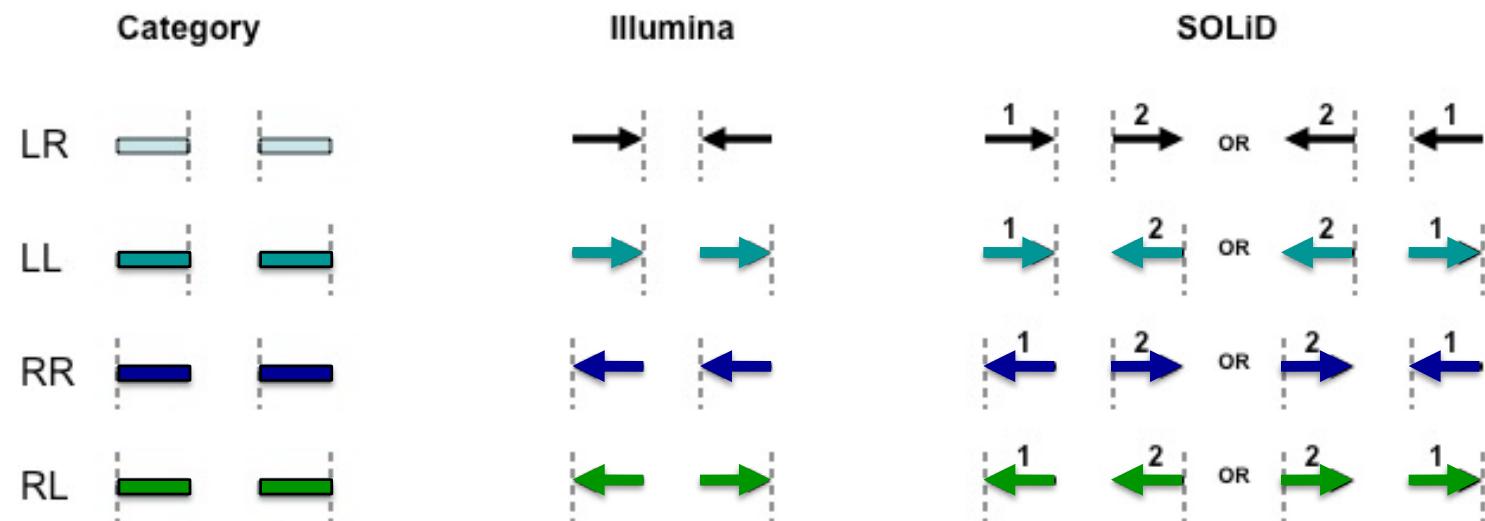
Color by pair orientation



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.

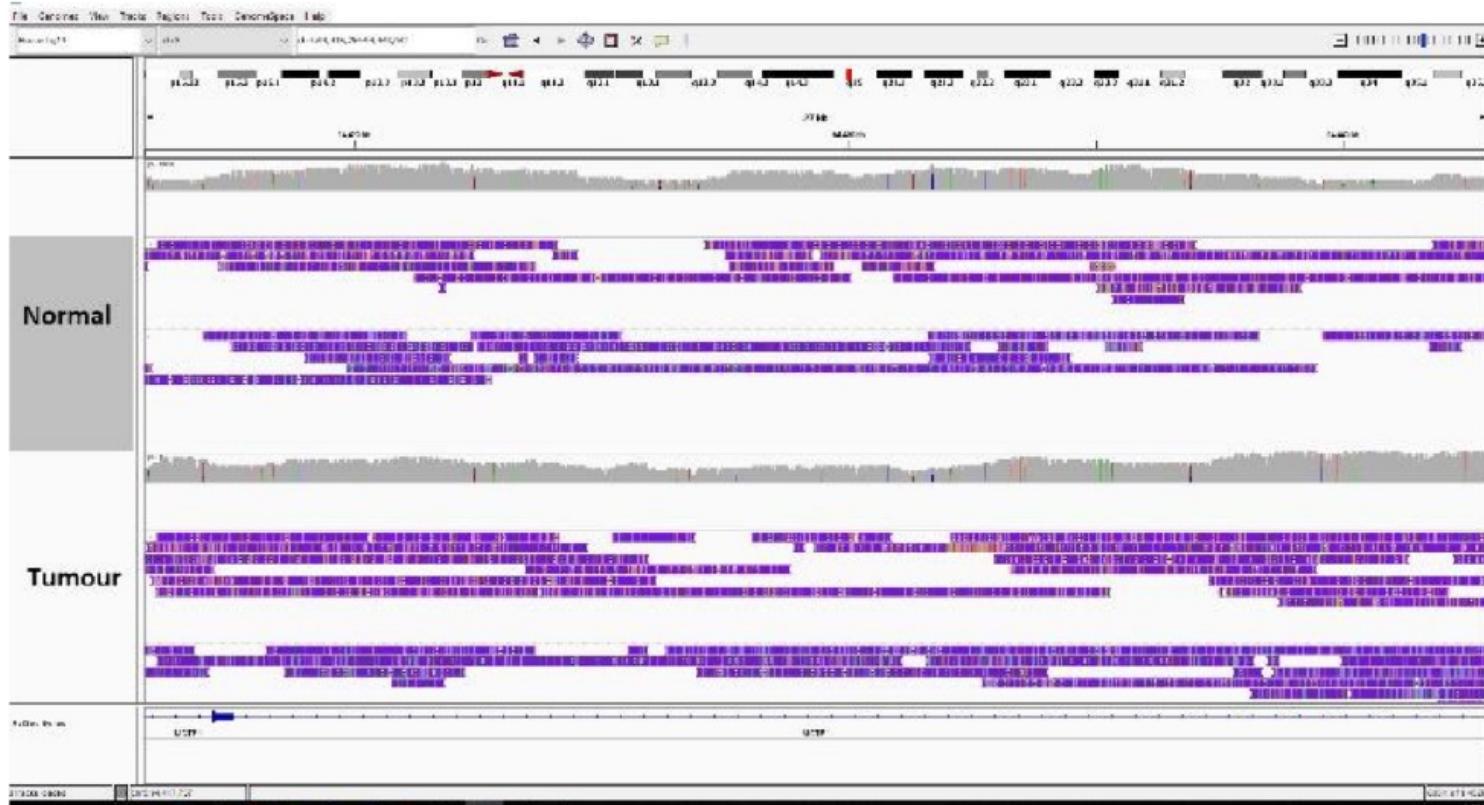
RL

Implies duplication or translocation with respect to reference.

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

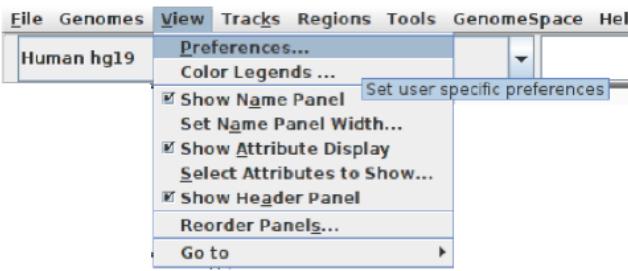
Figure courtesy of Bob Handsaker

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



Setting an indel threshold hides noise from small indels

The screenshot shows the 'View' menu open in the IGV software. The 'Set user specific preferences' option is highlighted with a yellow box and an arrow pointing to the 'Alignment Track Options' section of the 'Track Display Options' dialog. The dialog contains various settings for alignment tracks, including visibility thresholds, downsampling options, and filtering rules for indels and other features.

Track Display Options

On initial load show: Alignment Track Coverage Track Splice Junction Track

Alignment Track Options

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

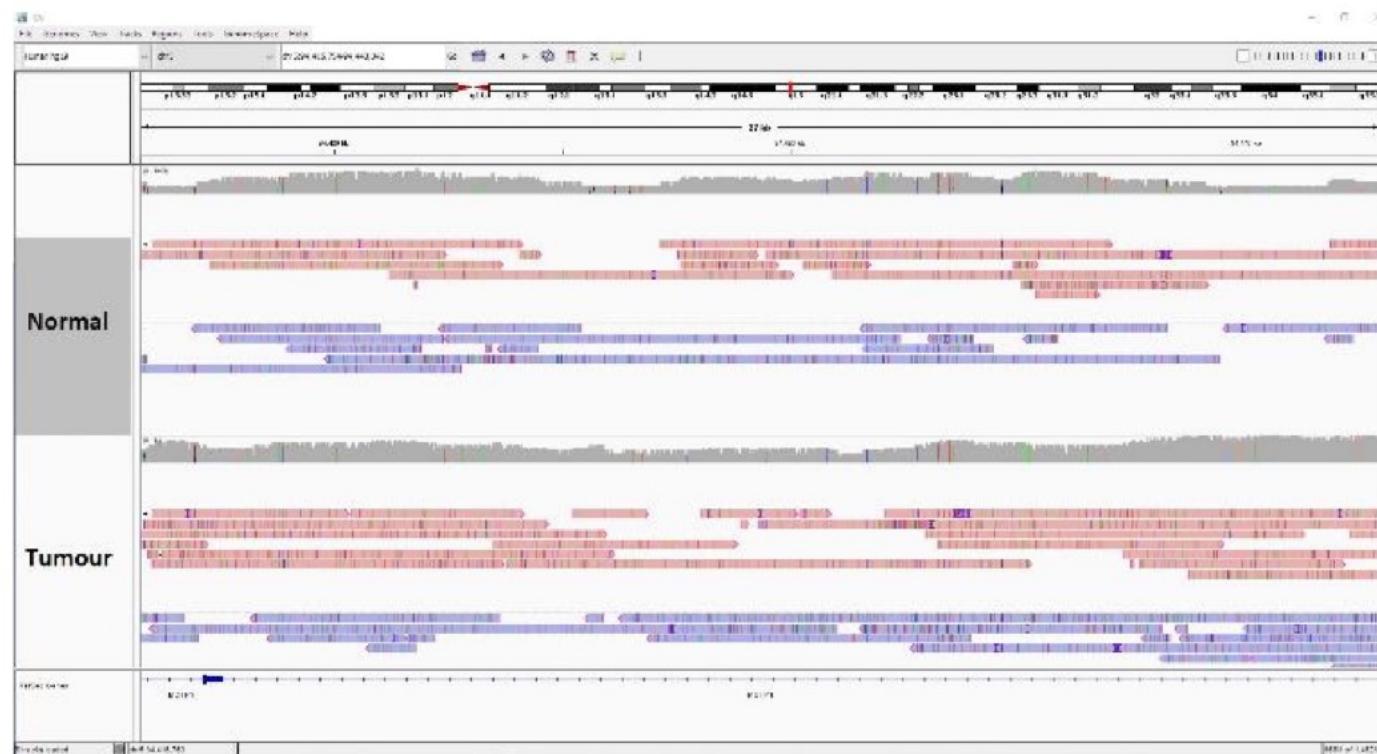
Insert Size Options

Defaults Minimum (bp): 50 Compute Minimum (percentile): 0.5

Maximum (bp): 1000 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

The screenshot shows the 'General' tab of the GATK Haplotype Caller Preferences dialog. A yellow callout box highlights the 'Show consensus mode' checkbox under the 'Alignment Track Options' section. An arrow points from this checkbox to a group of checkboxes labeled 'Quick consensus mode'.

General **Tracks** **Variants** **Charts** **Alignments** **Probes** **Proxy** **Advanced** **Cram**

Track Display Options

On initial load show: Alignment Track Coverage Track Splice Junction Track

Alignment Track Options

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 tags: 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

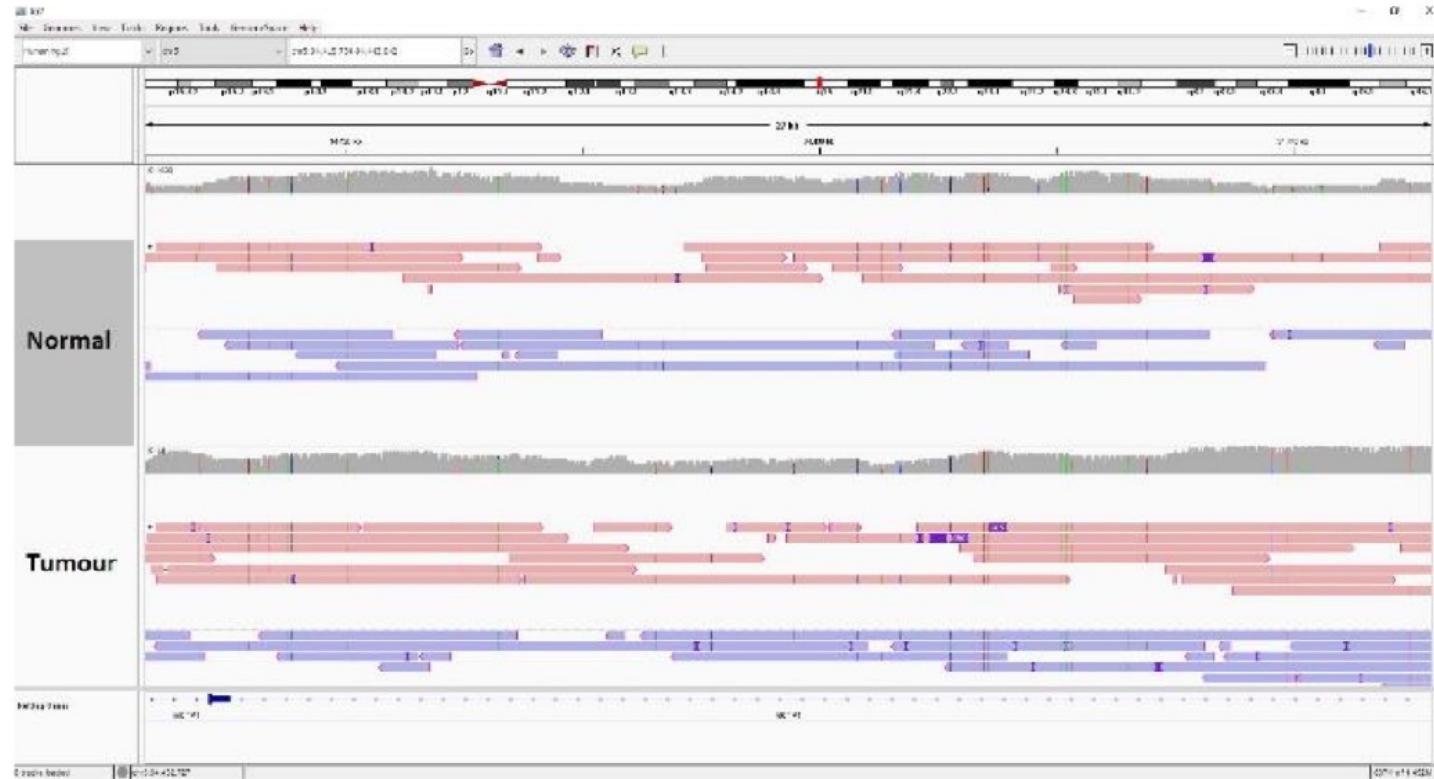
Defaults Minimum (bp): 50 Compute Minimum (percentile): 0.5

Maximum (bp): 1000 Maximum (percentile): 99.5

OK Cancel

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
inMedicine

Open

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

Erica K. Barnell, BS¹, Peter Ronning, BS¹, Katie M. Campbell, BS¹, Kilannin Krysiak, PhD^{1,2}, Benjamin J. Ainscough, PhD^{1,3}, Lana M. Sheta¹, Shahil P. Pema¹, Alina D. Schmidt, BS¹, Megan Richters, BS¹, Kelsy C. Cotto, BS¹, Arpad M. Danos, PhD¹, Cody Ramirez, BS¹, Zachary L. Skidmore, MEng¹, Nicholas C. Spies, BS¹, Jasreet Hundal, MS¹, Malik S. Sediqzad¹, Jason Kunisaki, BS¹, Felicia Gomez, PhD¹, Lee Trani, BS¹, Matthew Matlock, BS¹, Alex H. Wagner, PhD¹, 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.

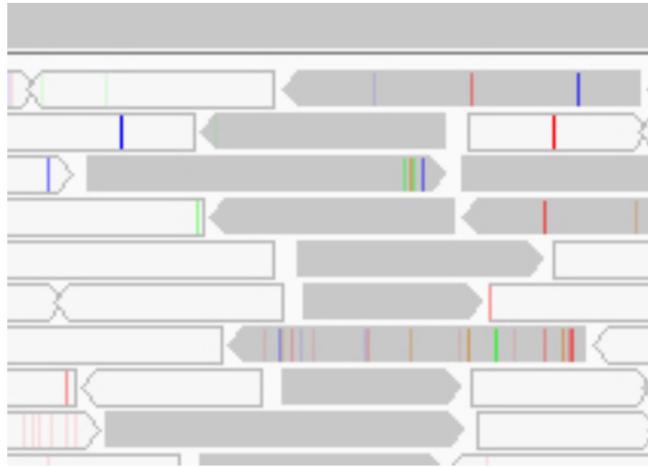
Results: After reading the SOP, average accuracy in somatic variant identification increased by 16.7% (p value = 0.0298) and average interreviewer agreement increased by 12.7% (p value < 0.001). Manual review conducted after reading the SOP did not significantly increase reviewer time.

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

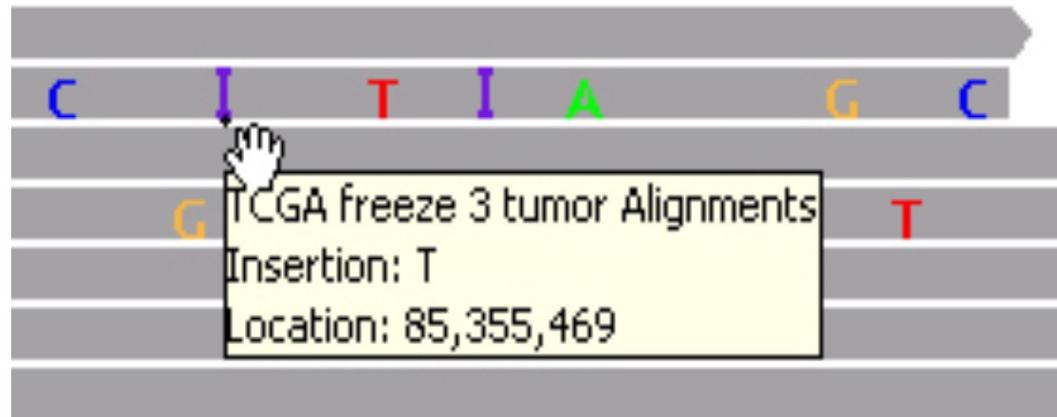
Genetics in Medicine (2018) <https://doi.org/10.1038/s41436-018-0278-z>

Keywords: somatic variant refinement; manual review

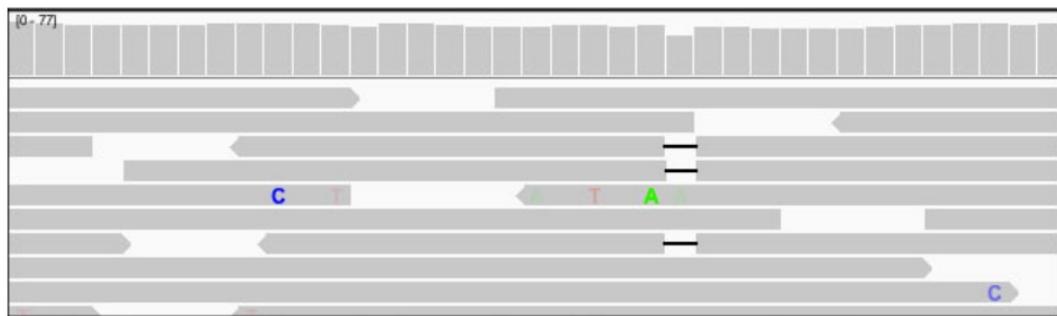
Other notes



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



Purple : Insertion



Gapped read/black bar: Deletion

We are on a Coffee Break & Networking Session

Workshop Sponsors:



Canadian Centre for
Computational
Genomics

