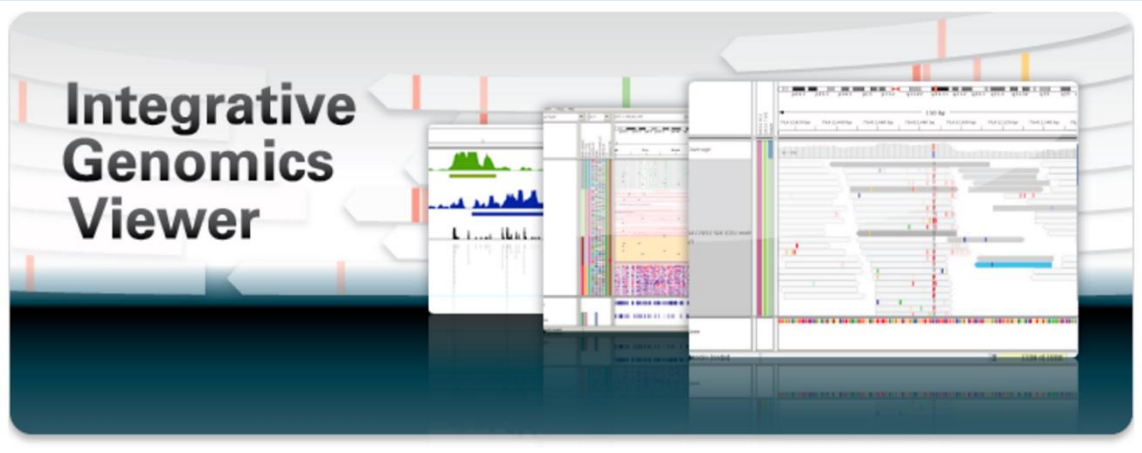
Cold Spring Harbor Laboratory

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 11-22, 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

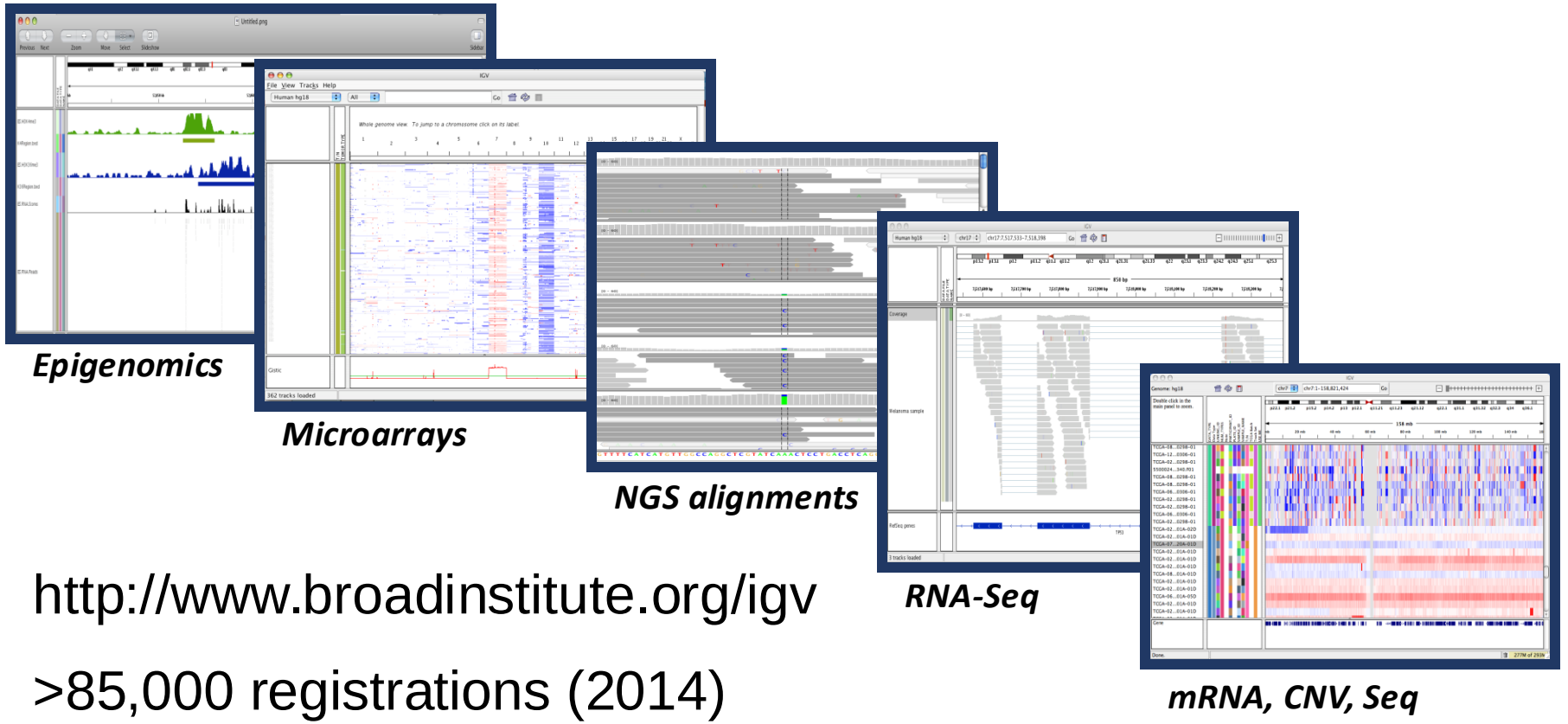
Trackster (part of Galaxy)

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



http://www.broadinstitute.org/igv

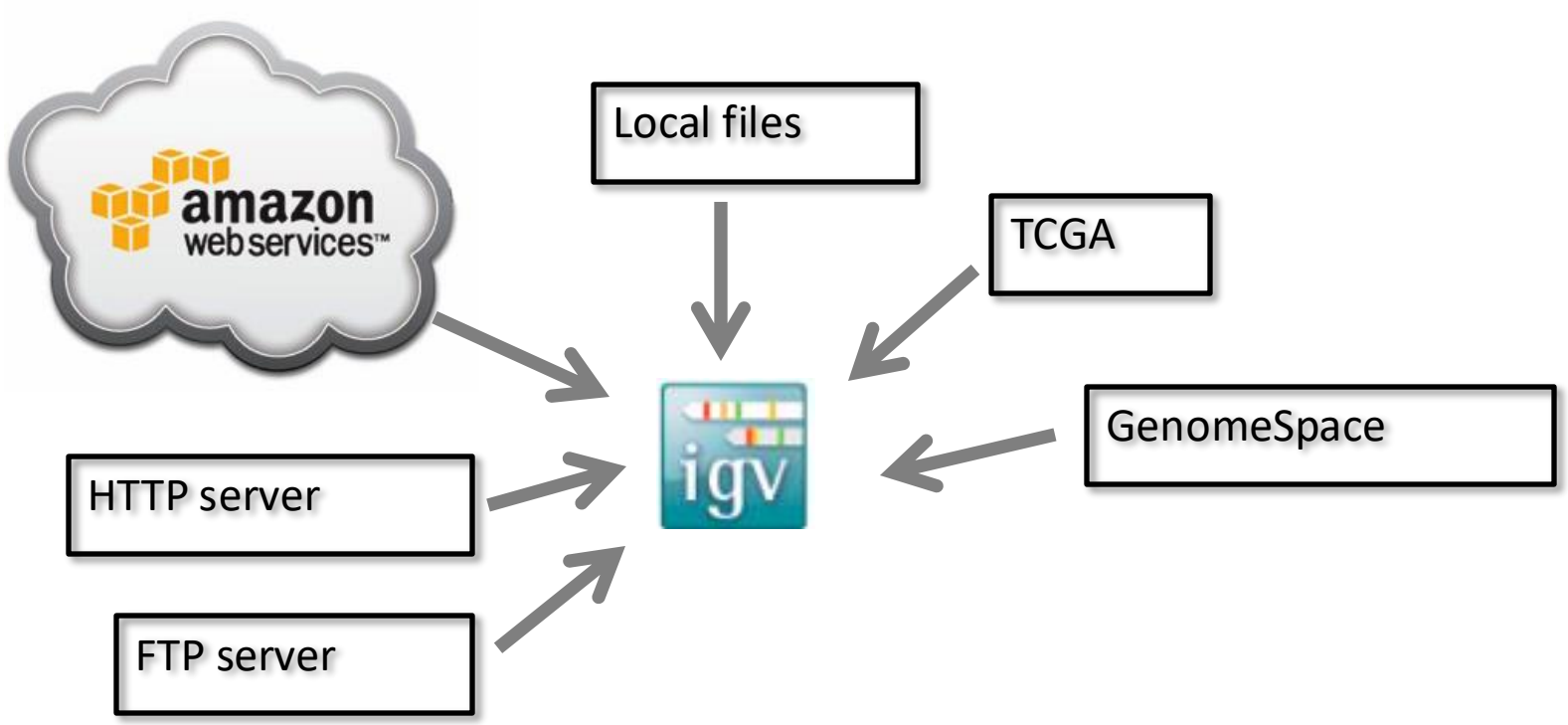
>85,000 registrations (2014)

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
    - Single Nucleotide Variants (SNVs)
    - Structural Variations (SVs)

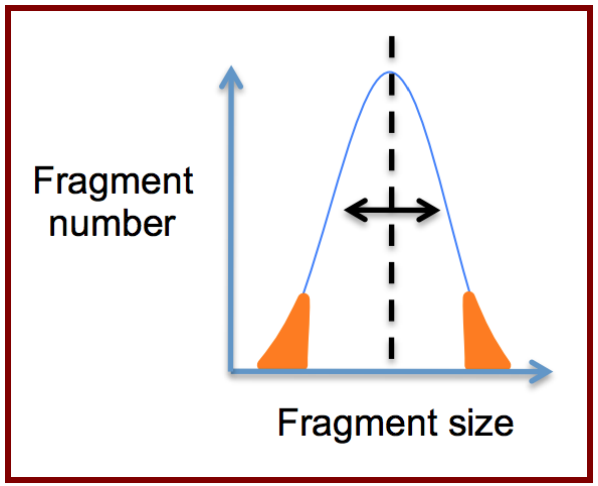
SNVs and Structural variations

- 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

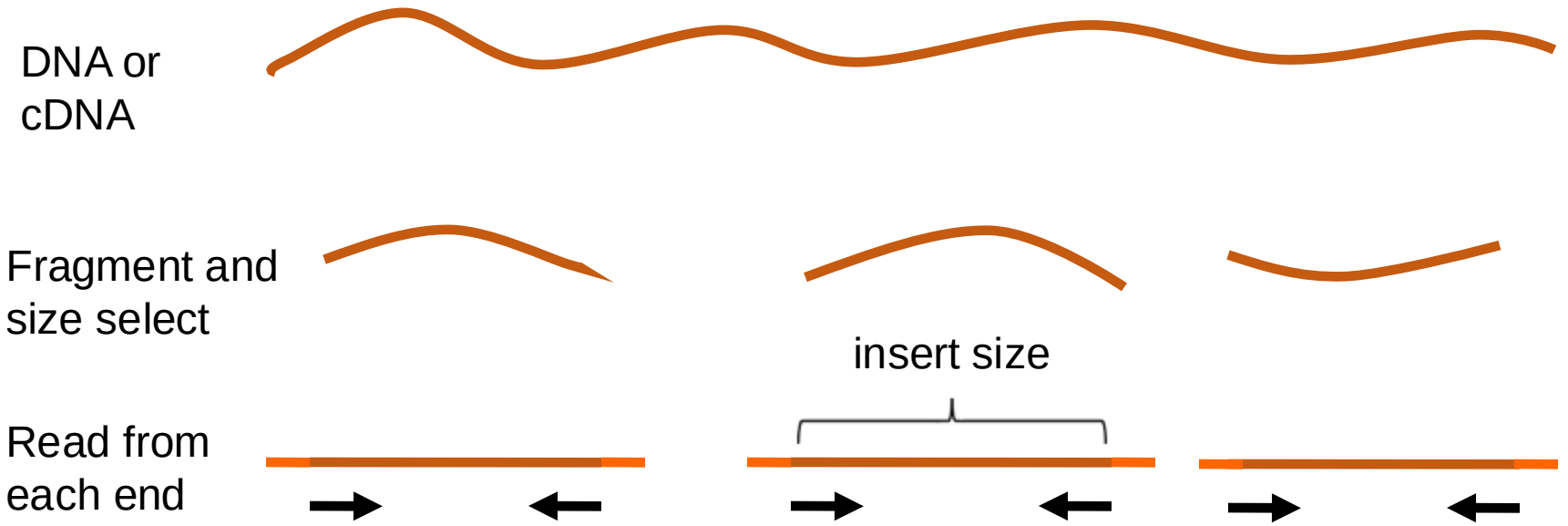
Paired-end sequencing

DNA or cDNA

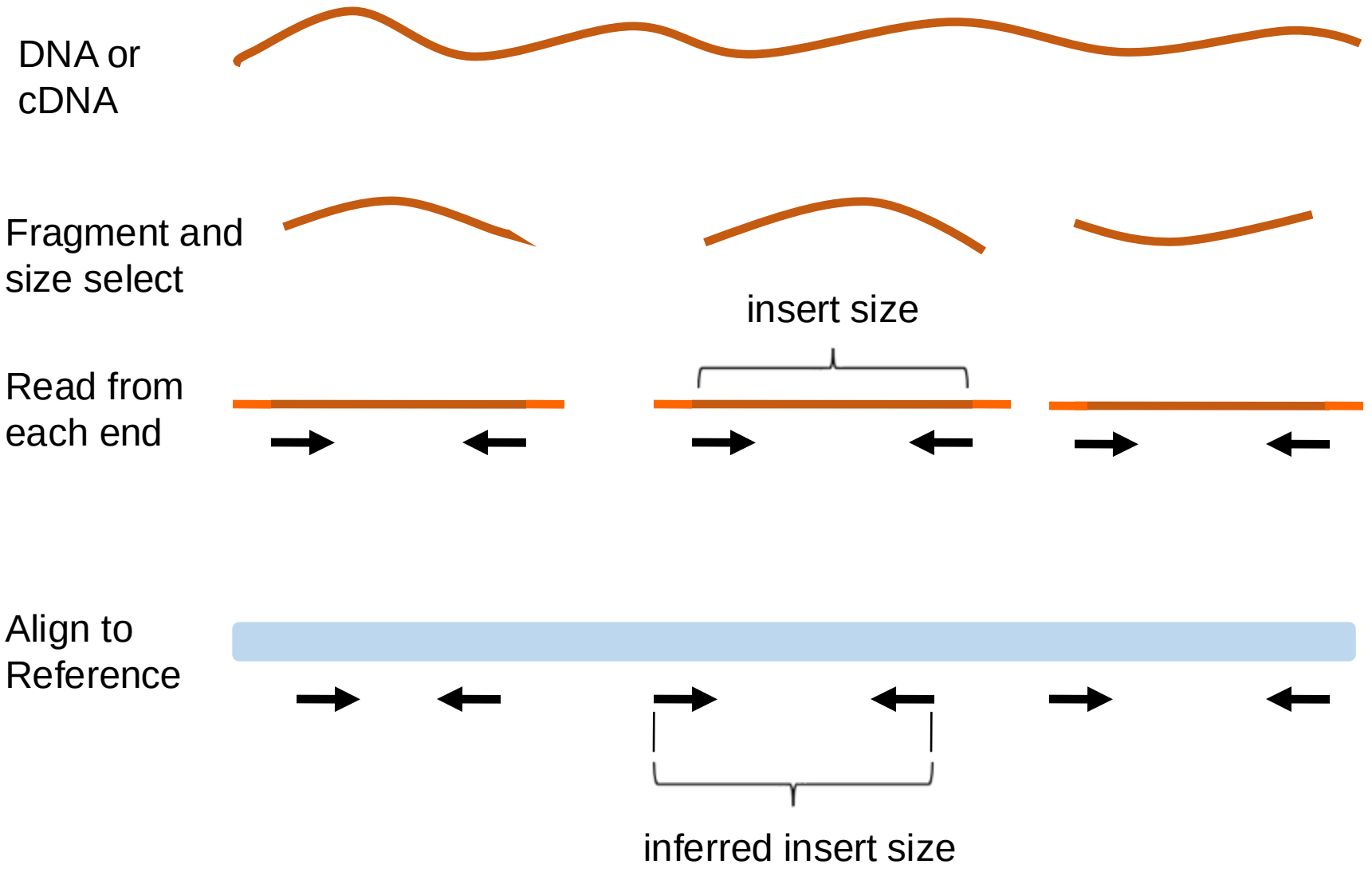
Fragment and size select



Paired-end sequencing



Paired-end sequencing



Interpreting inferred insert size

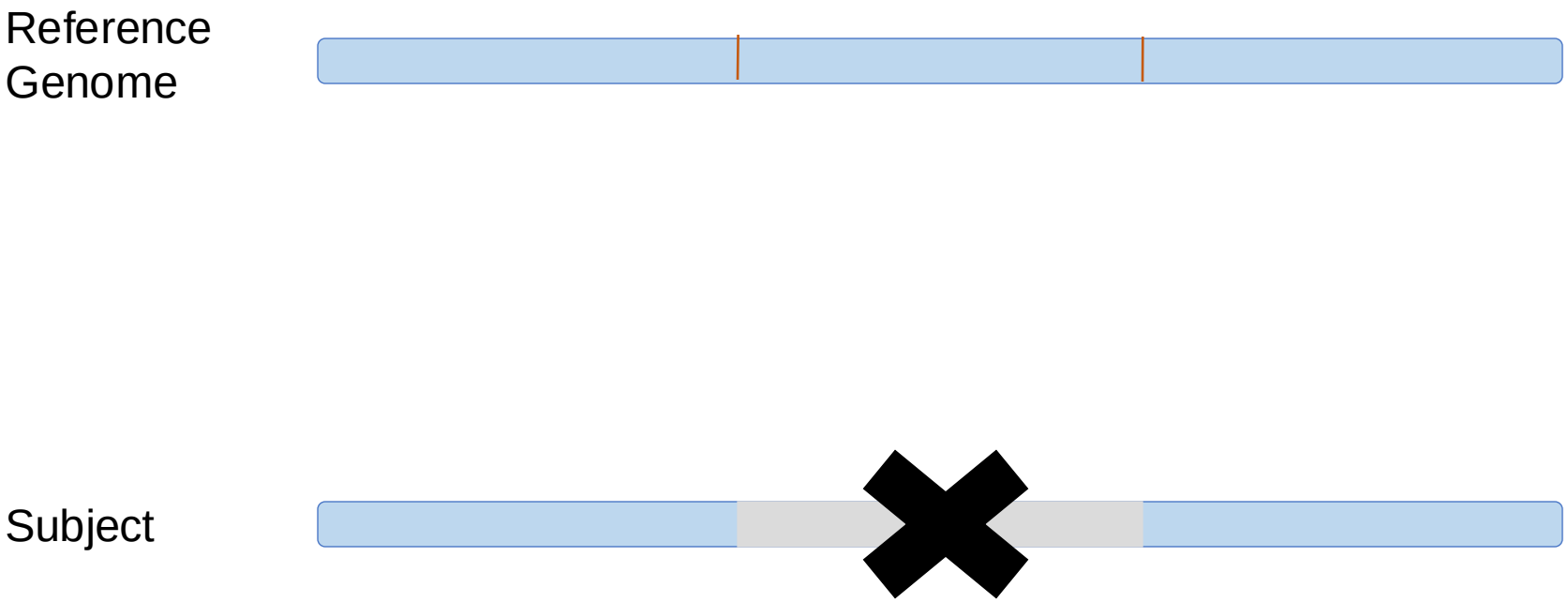
The “inferred insert size” can be used to detect structural variants including

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

Deletion

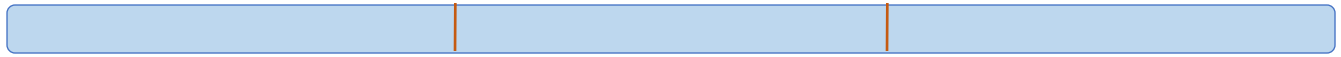
What is the effect of a deletion on inferred insert size?

Deletion



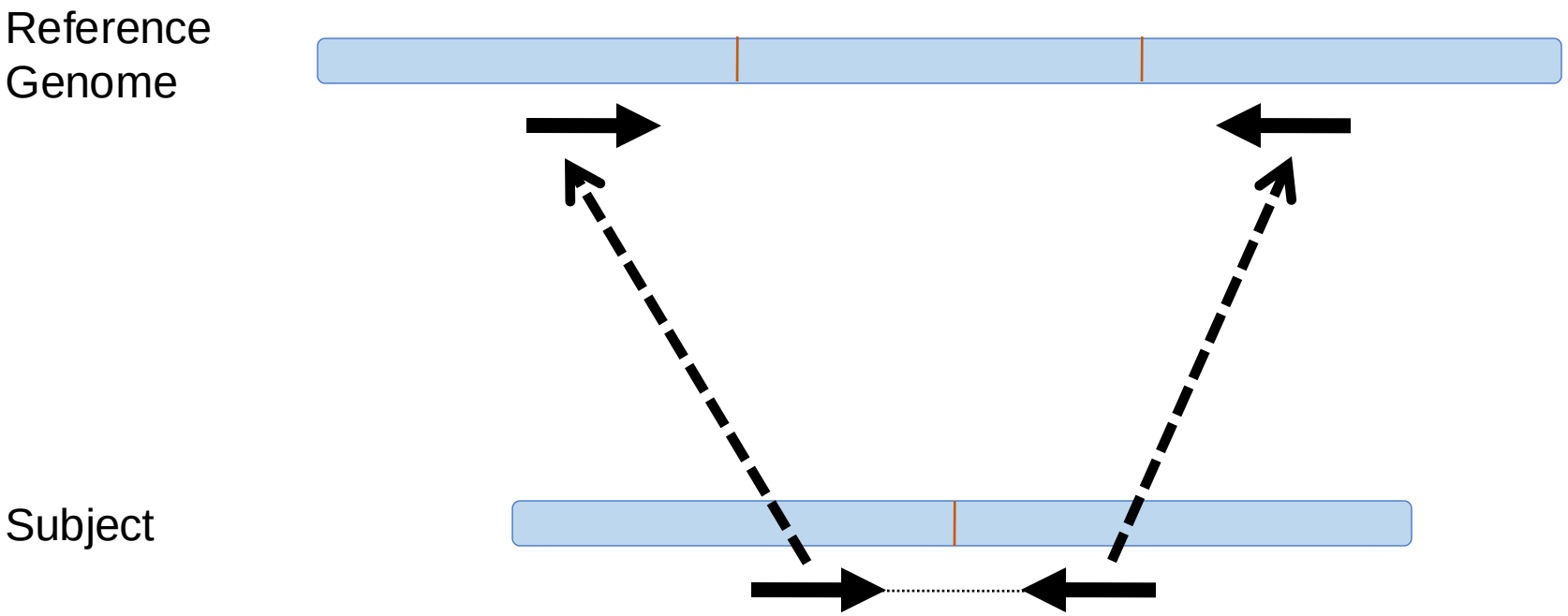
Deletion

Reference Genome



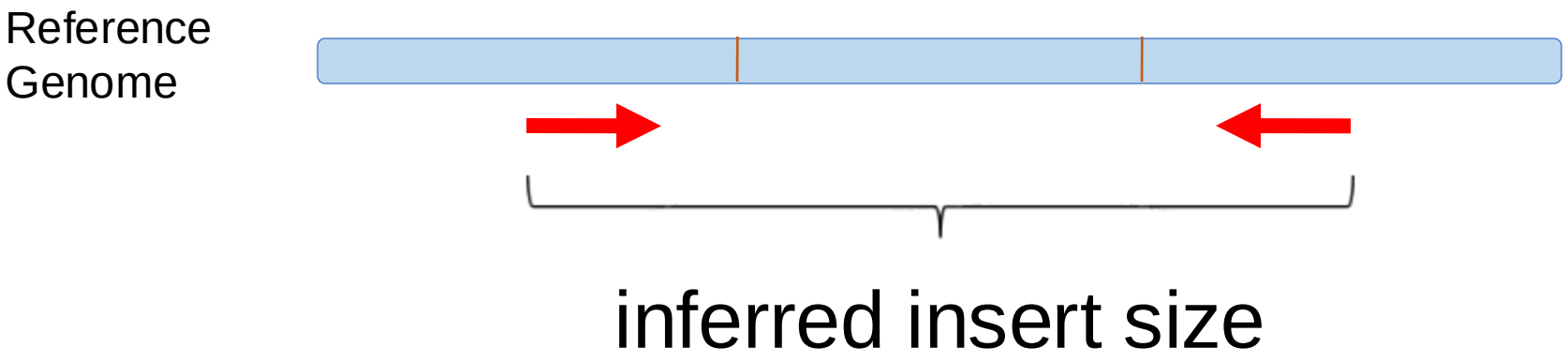
Subject

Deletion



Deletion

Inferred insert size is > expected value



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

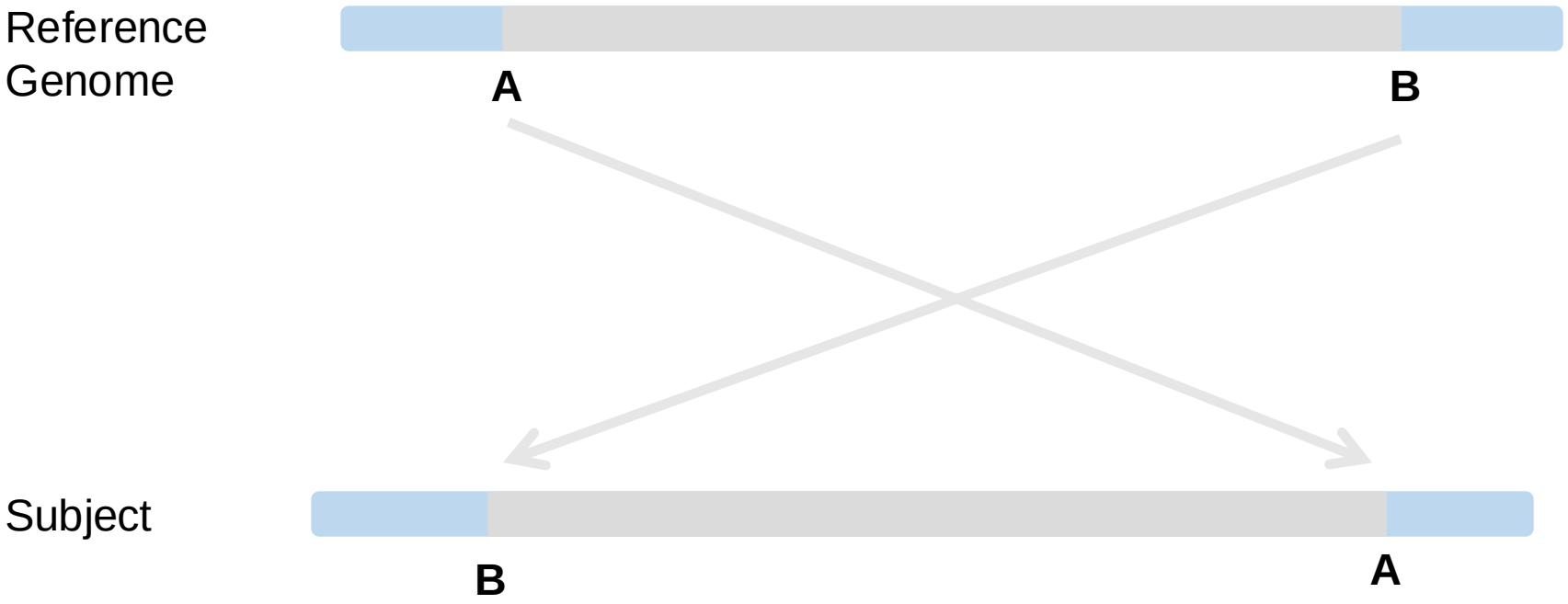
Inversion

Reference genome

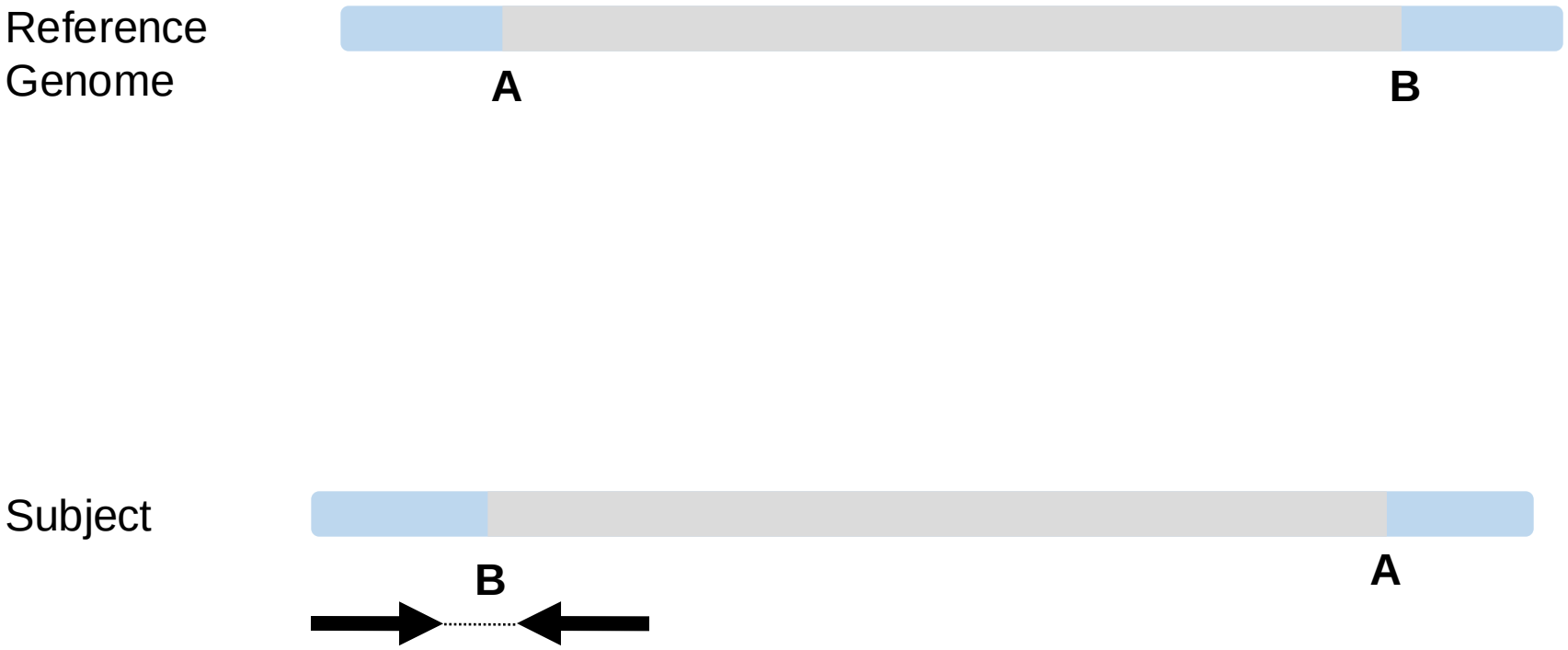
Inversion

Reference genome

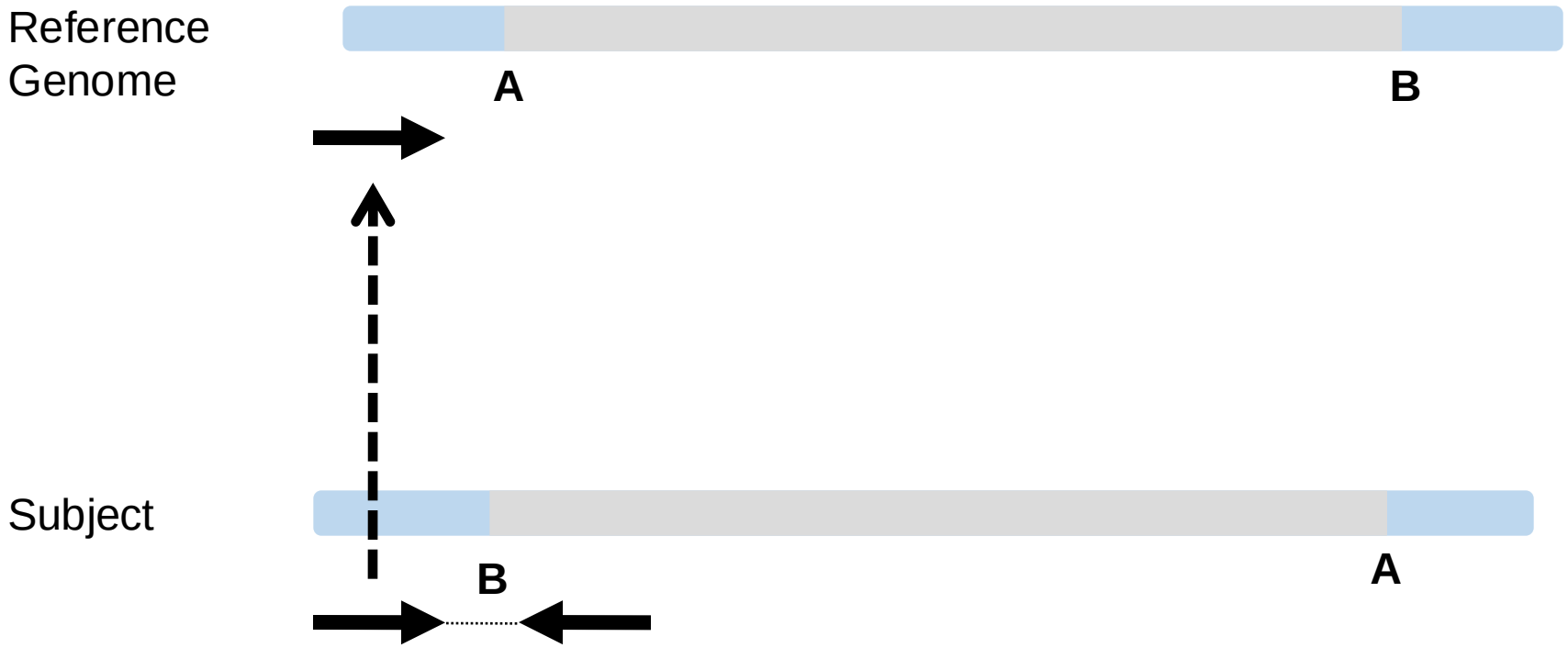
Inversion



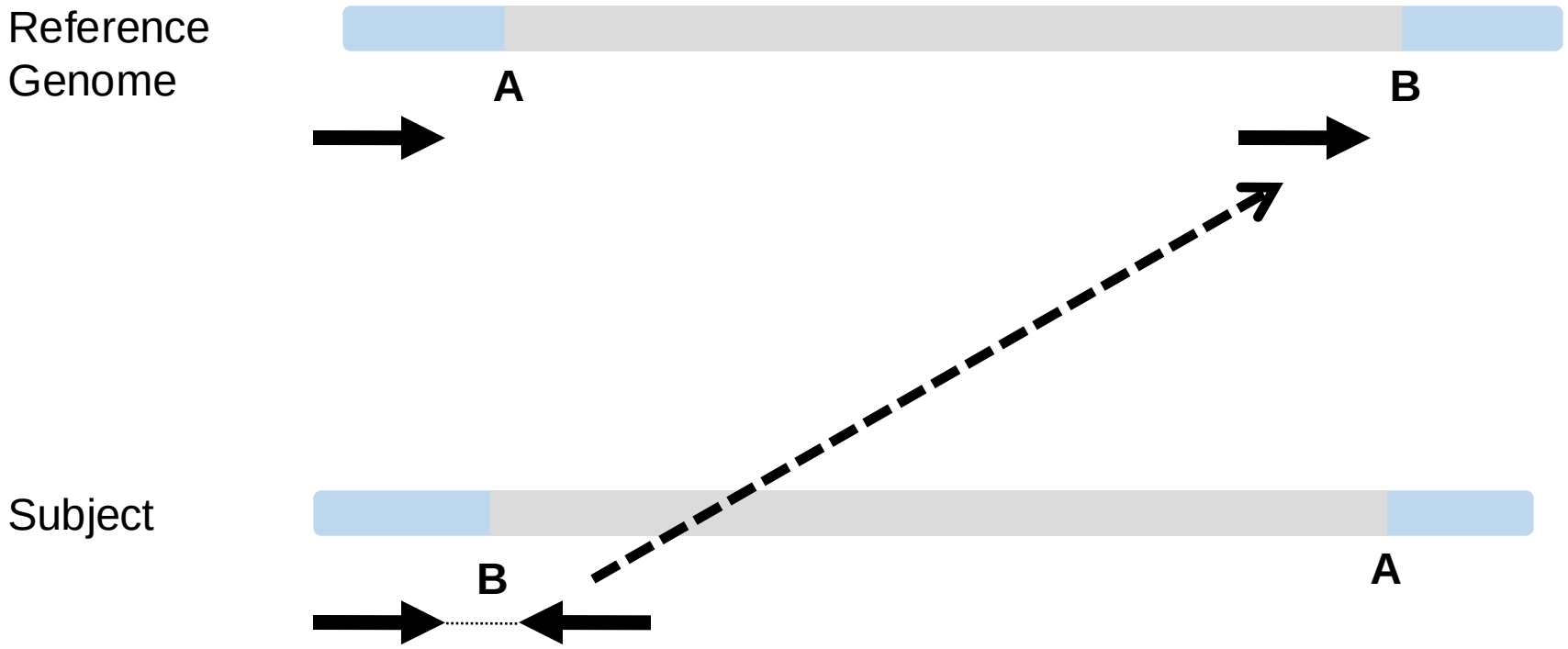
Inversion



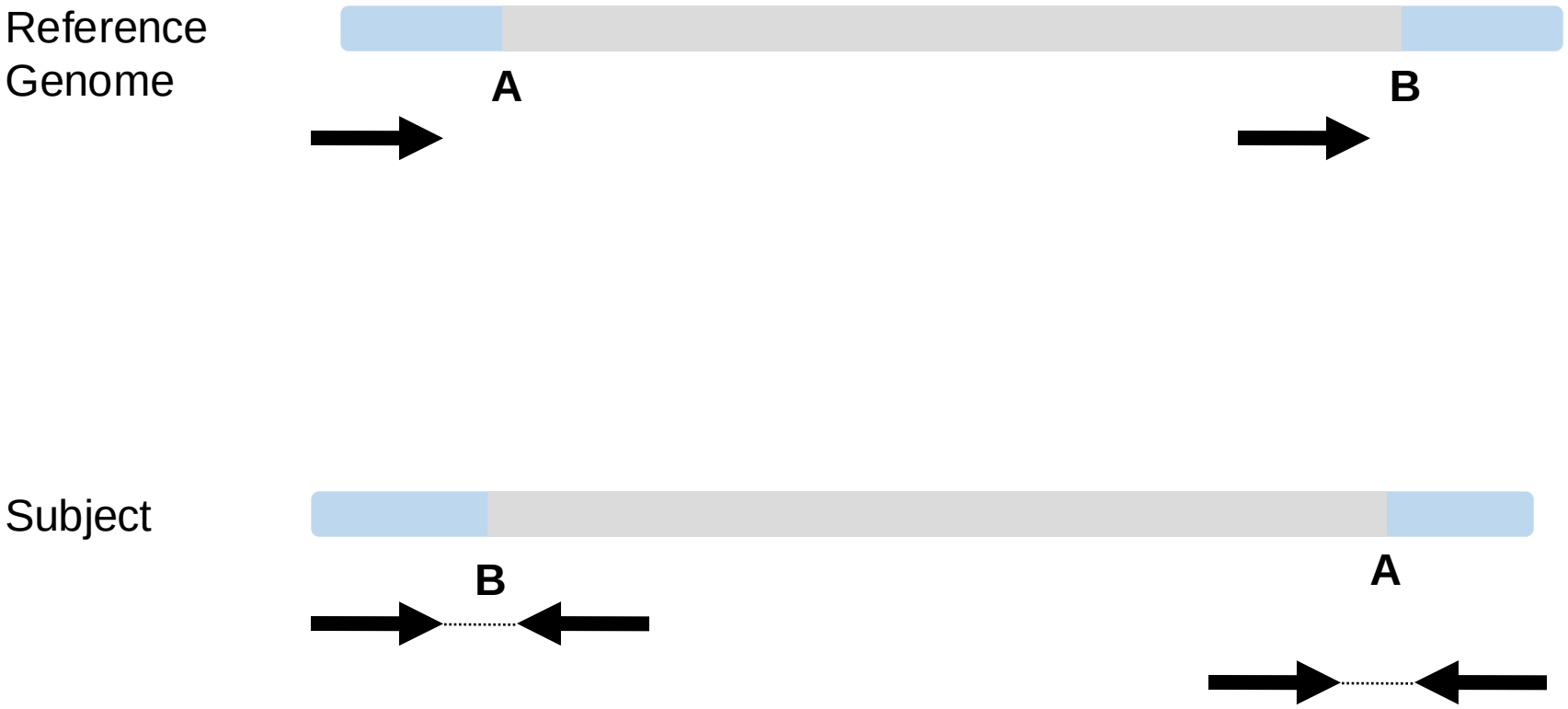
Inversion



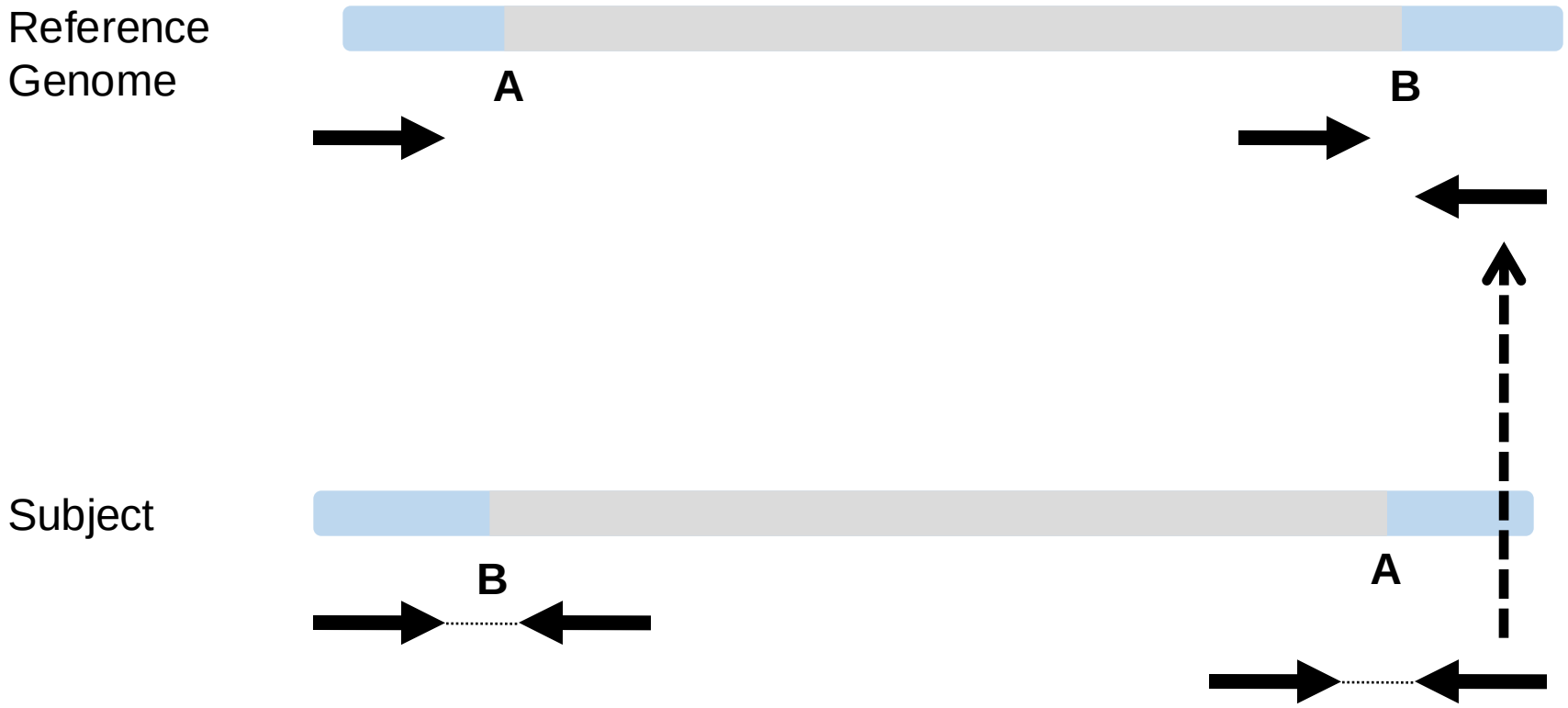
Inversion



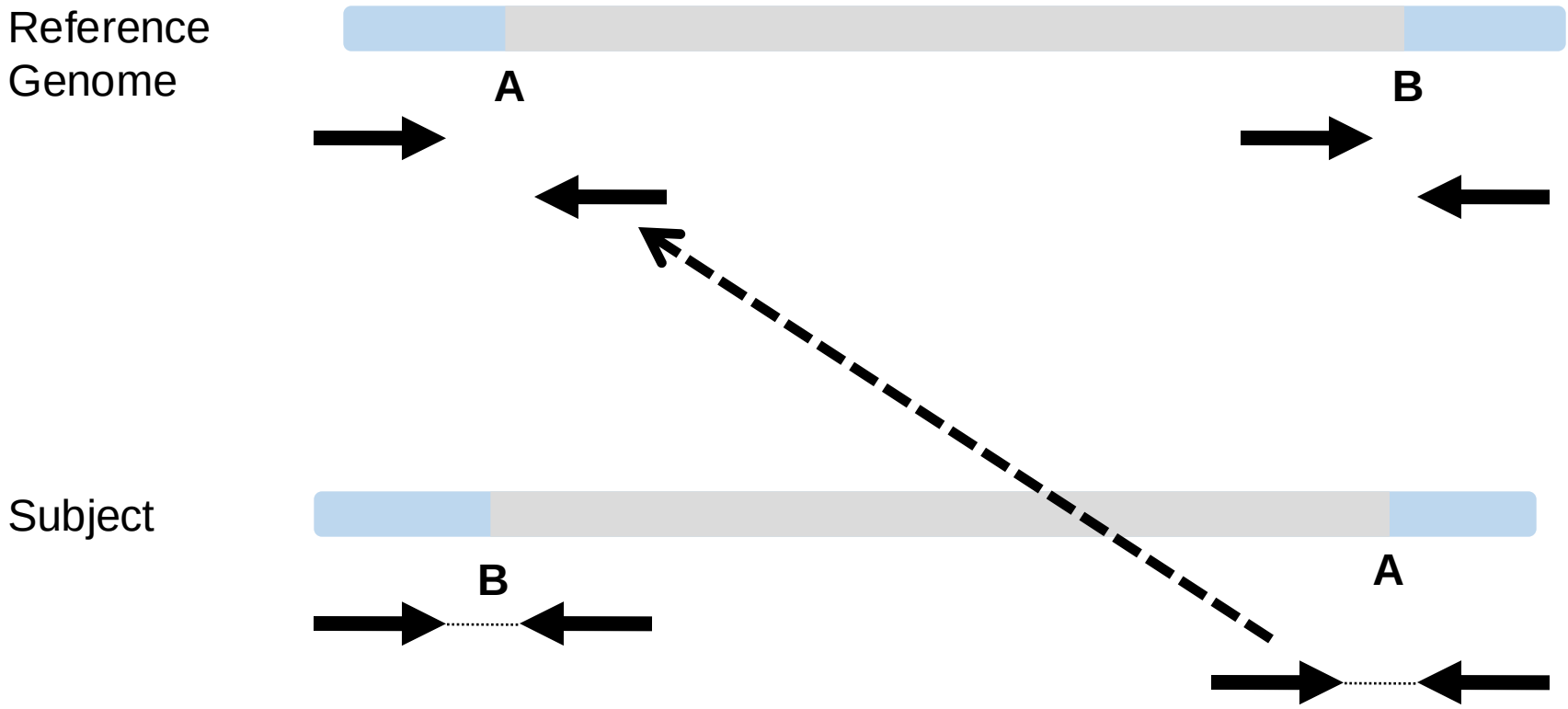
Inversion



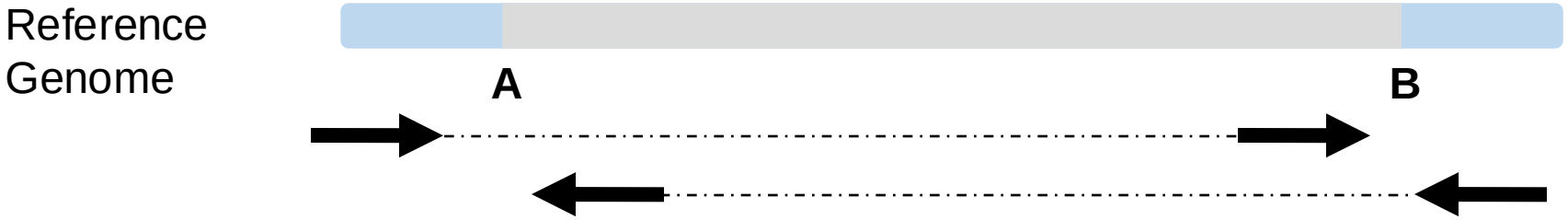
Inversion



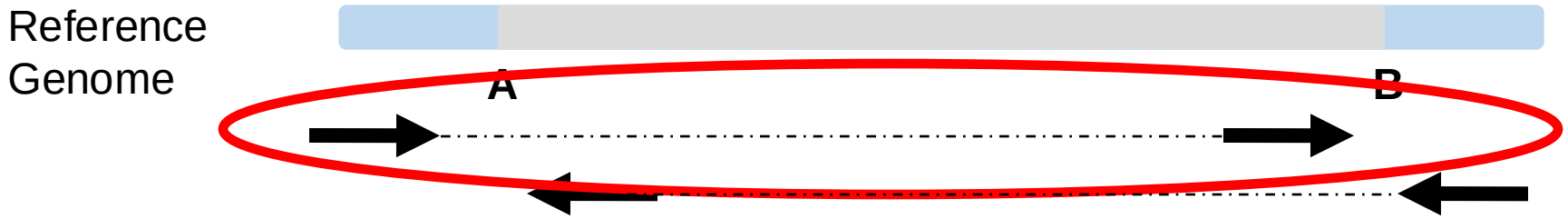
Inversion



Inversion

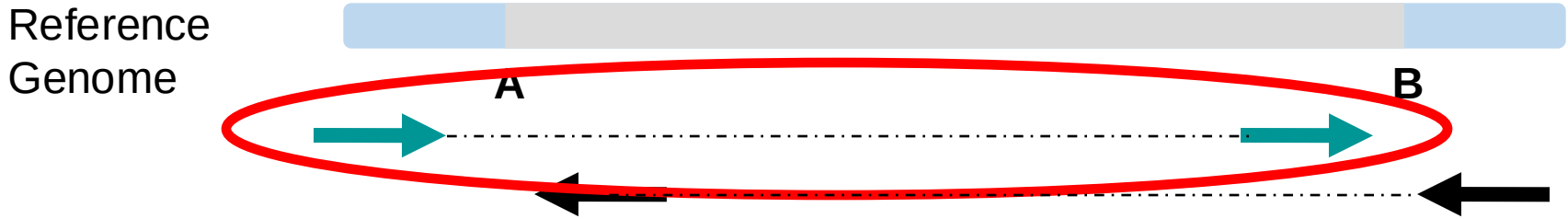


Inversion



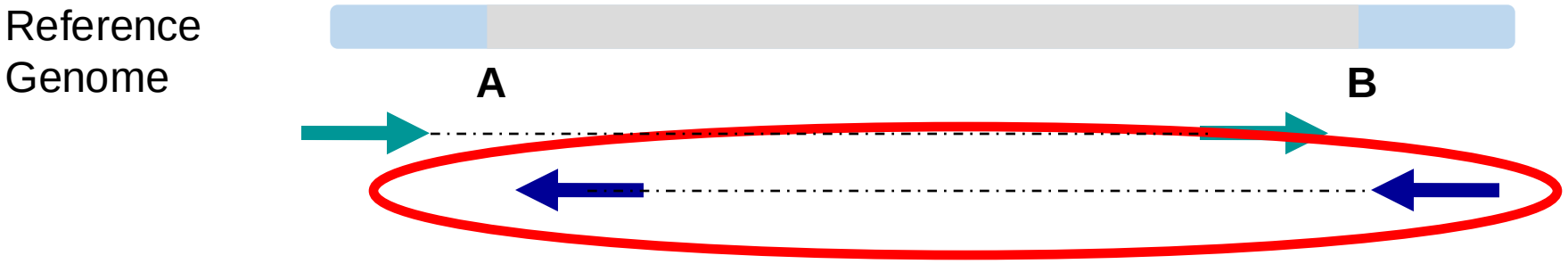
Anomaly: expected orientation of pair is inward facing (→ ←)

Inversion



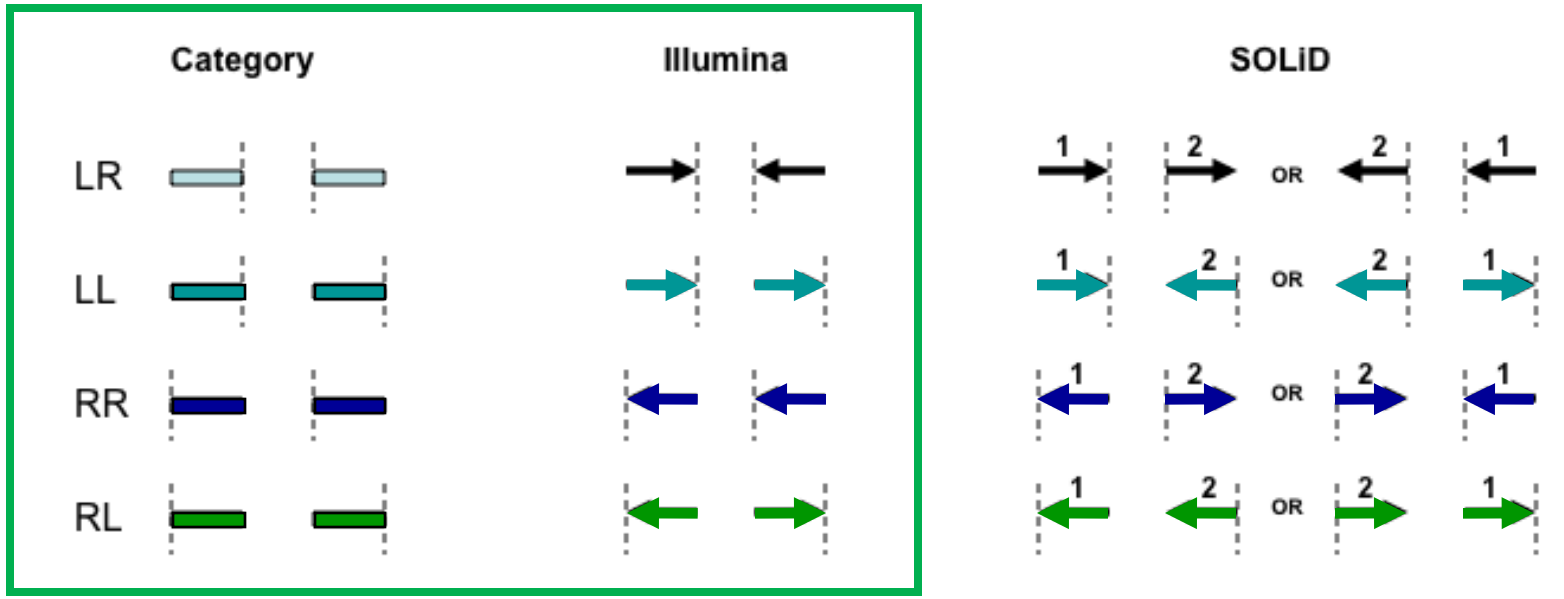
“Left” side pair

Inversion



“Right” side pair

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.

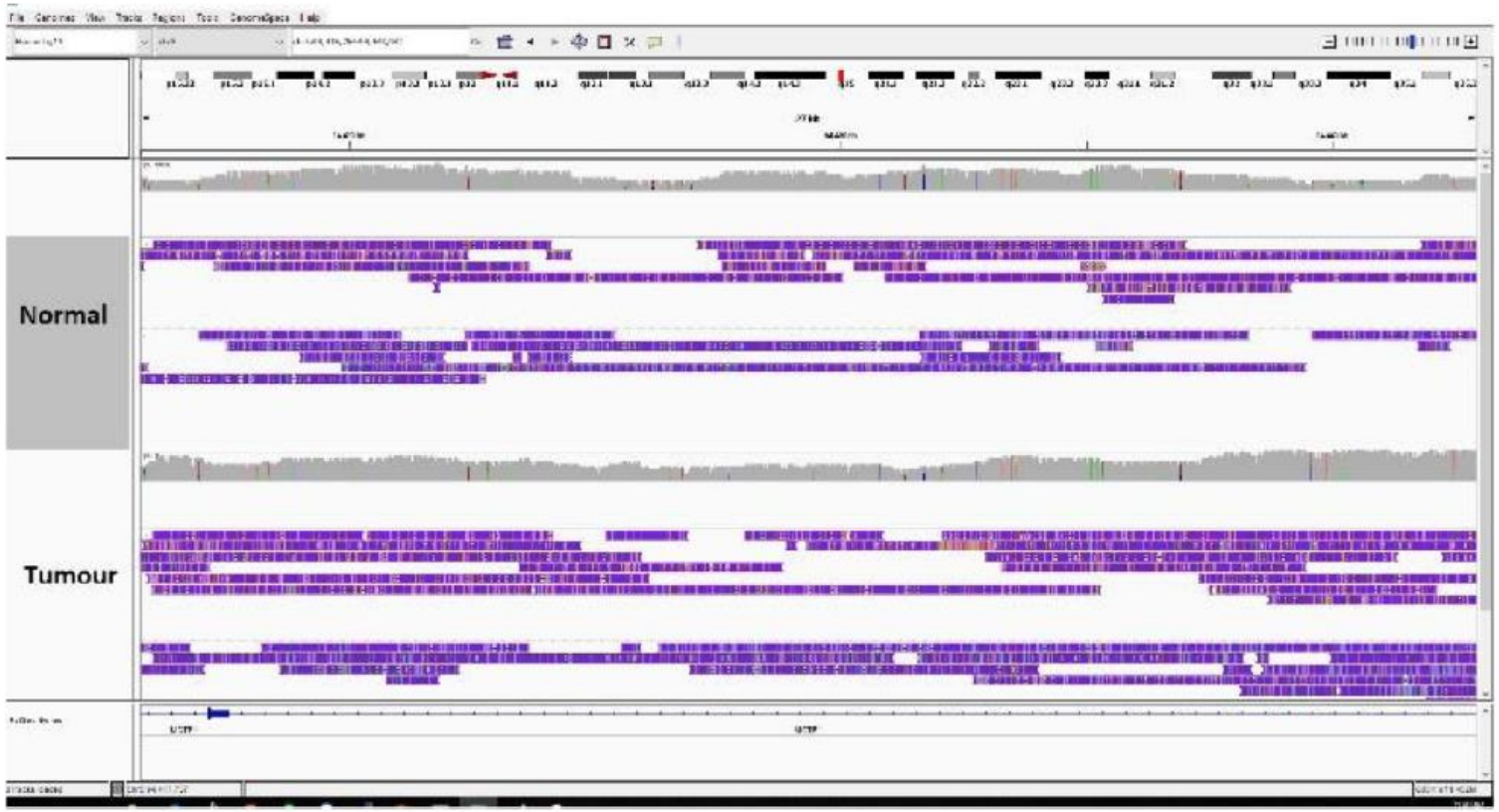
RL Implies duplication or translocation with respect to reference.

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

Figure courtesy of Bob Handsaker

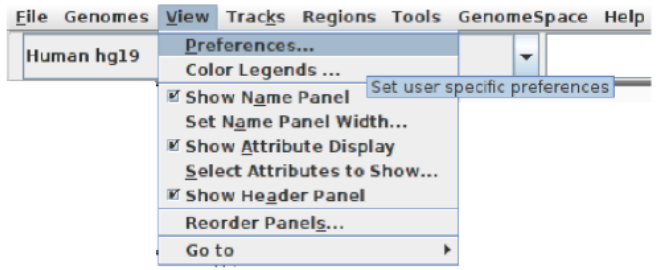
Let’s view some genomes!

Long read considerations

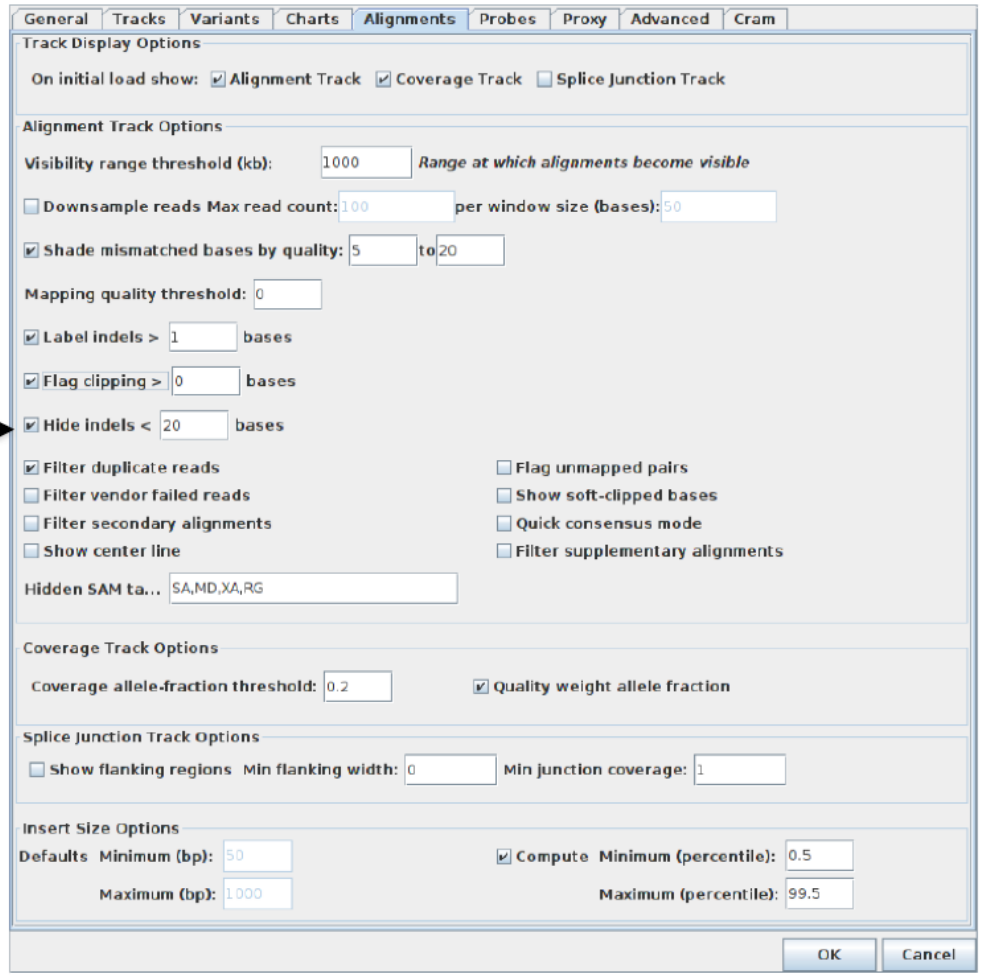


- Commonly see lots of small indels and single base errors that are simply noise
- Can be removed to be able to view the data more cleanly

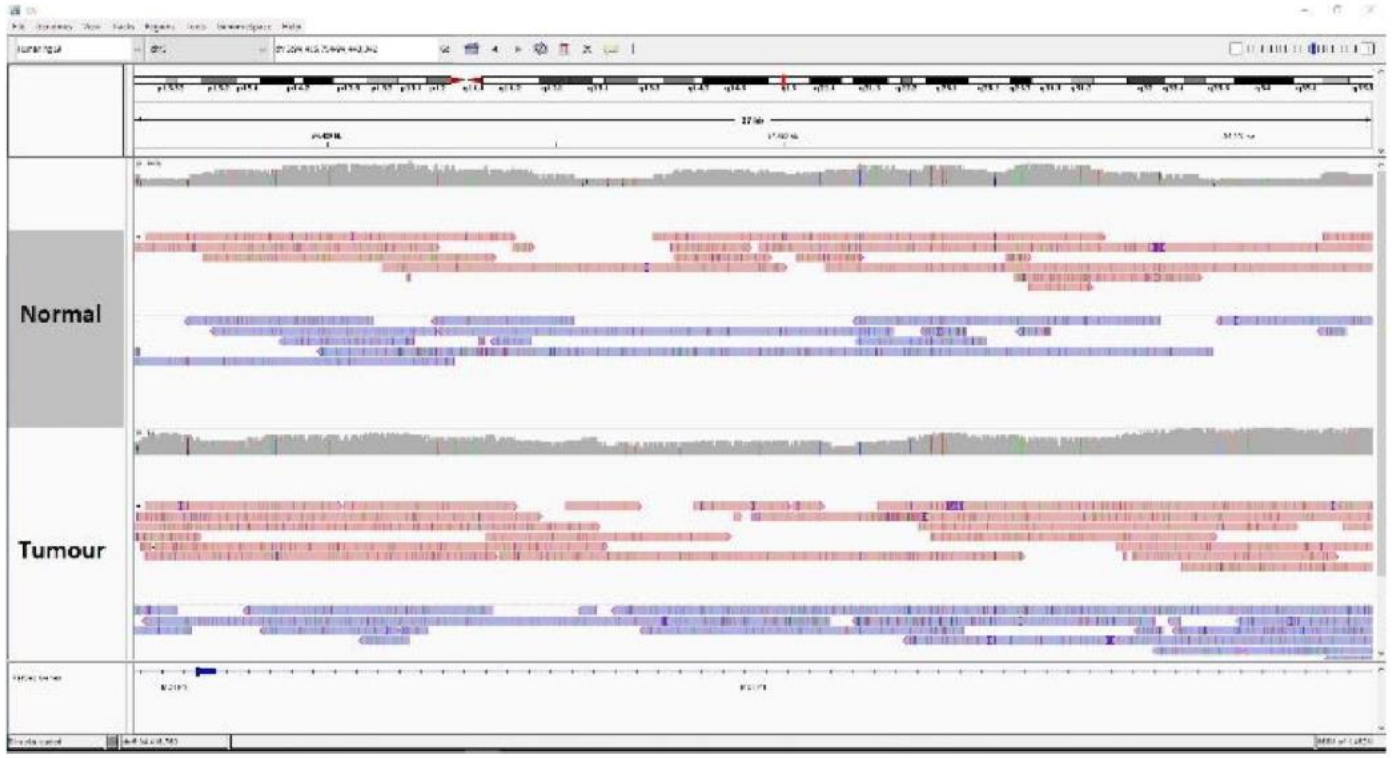
Long read considerations



Setting an indel threshold hides noise from small indels

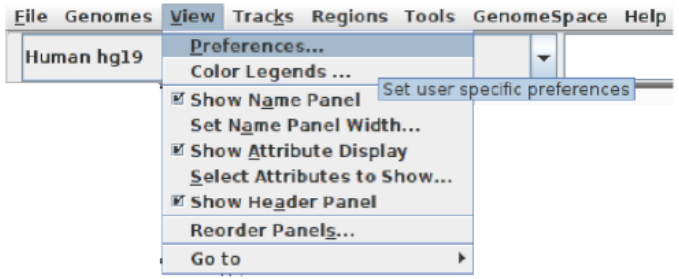


Long read considerations

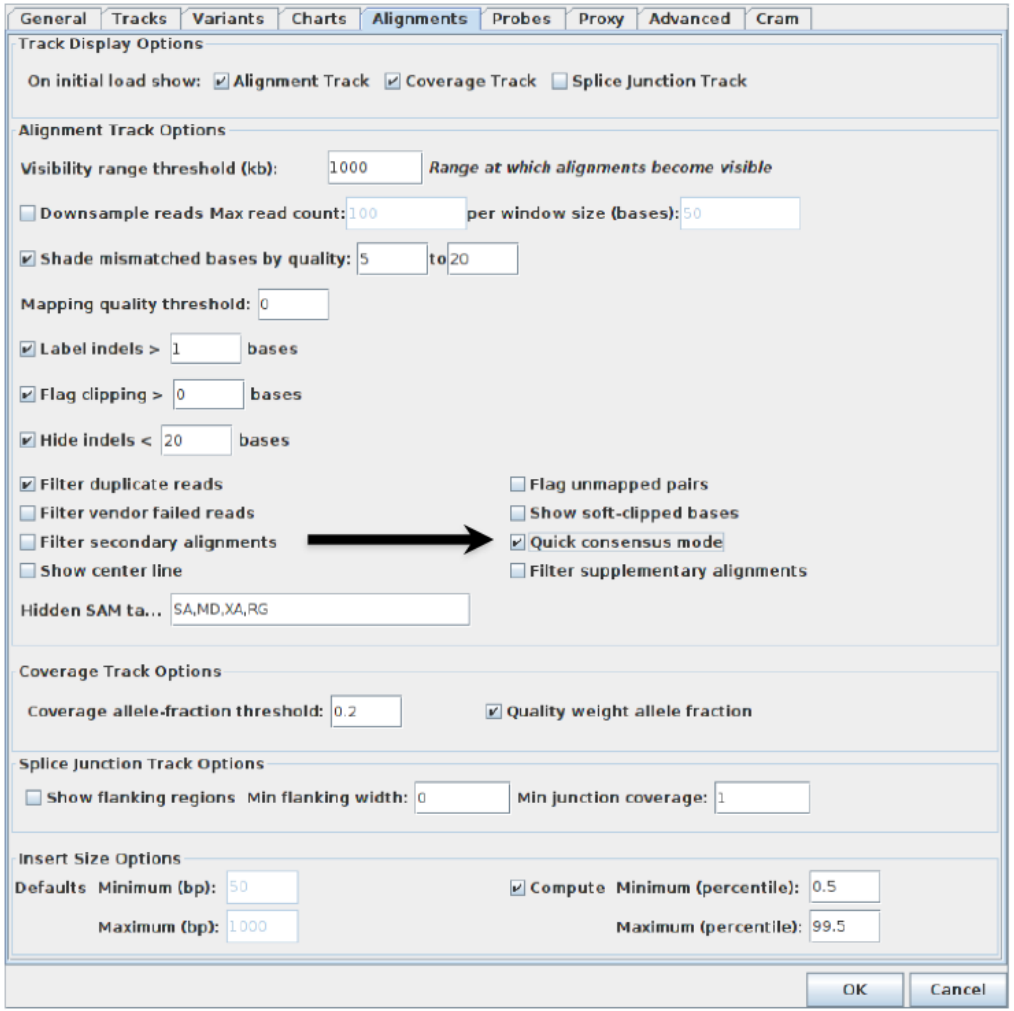


- Reads are not all purple dashes
- Next step would be to call a consensus at each position

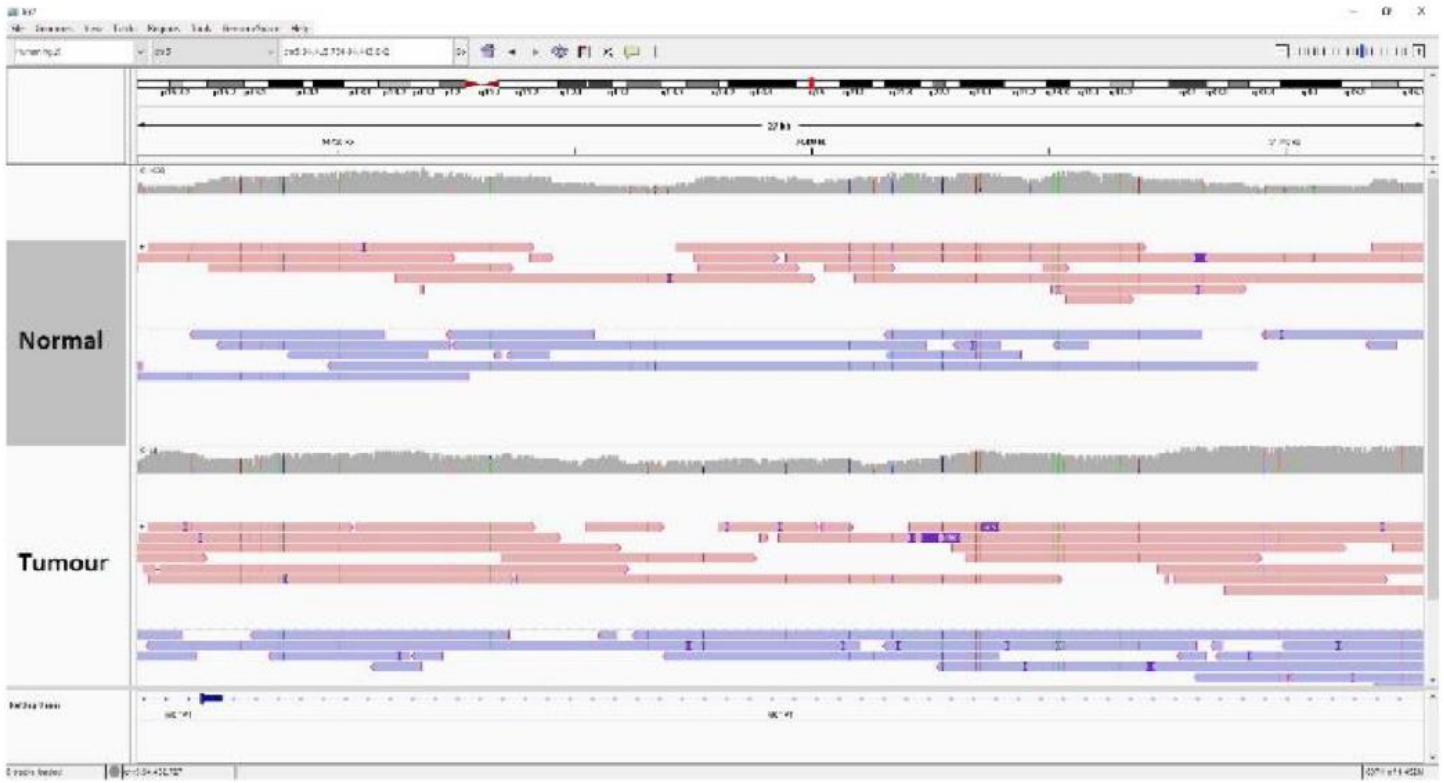
Long read considerations



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.

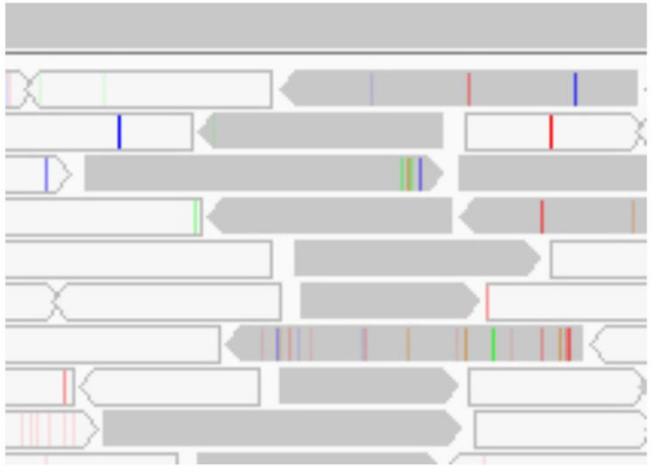
Results: After reading the SOP, average accuracy in somatic variant identification increased by 16.7% (p value = 0.0298) and average interreviewer agreement increased by 12.7% (p value < 0.001). Manual review conducted after reading the SOP did not significantly increase reviewer time.

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

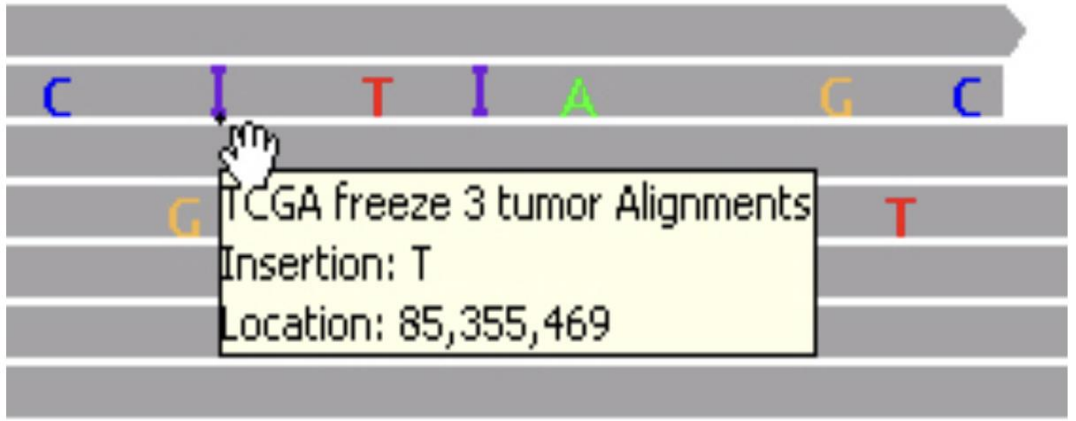
Genetics in Medicine (2018) https://doi.org/10.1038/s41436-018-0278-z

Keywords: somatic variant refinement; manual review

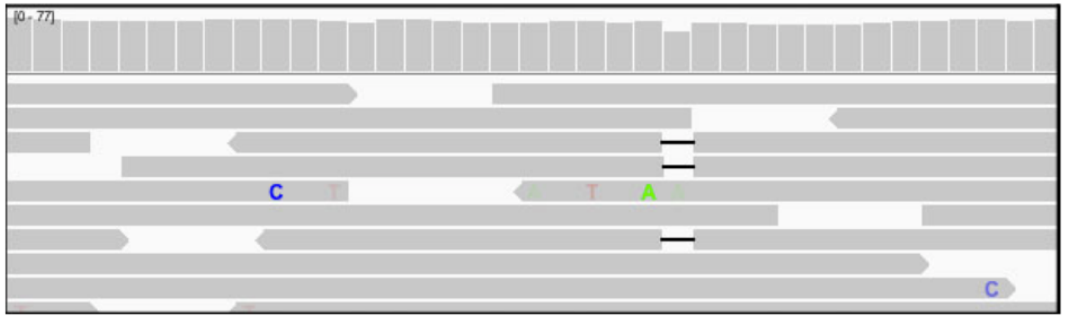
Other notes



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