Copy Number and Structural Variation

Chris Miller

Some slides adapted from:

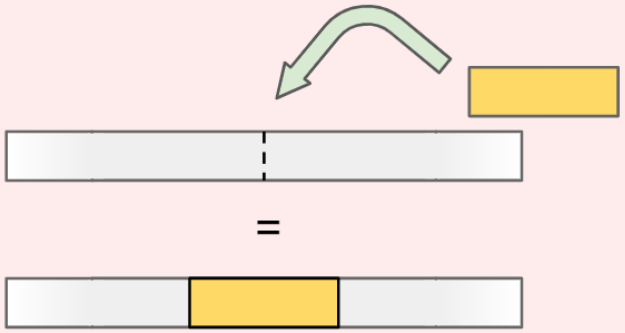
Malachi Griffith, Obi Griffith, Fouad Yousif
High-Throughput Biology: From Sequence to Networks
https://github.com/griffithlab/rnaseq_tutorial_wiki/blob/master/LectureFiles/cbw-cshl/2017/IGV_Tutorial_Brief.pptx

Tobias Rausch
Structural and copy-number variation analysis
https://www.ebi.ac.uk/training/materials/cancer-genomics-materials/structural-and-copy-number-variation-mutational-signatures/structural-and-copy-number-variation-analysis/

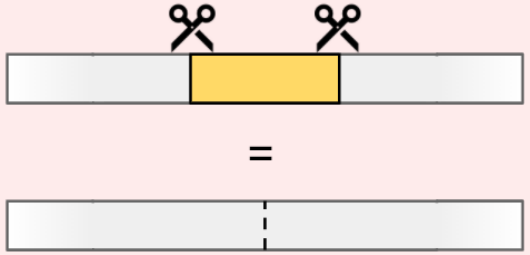
SV Types

Destructive (non-balanced)

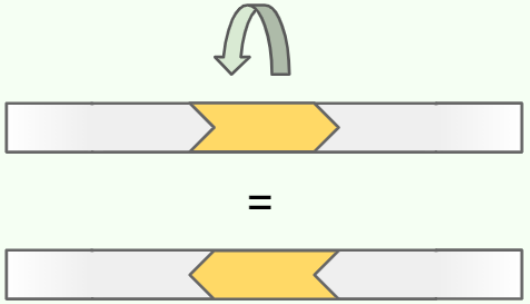
Non-destructive (balanced)



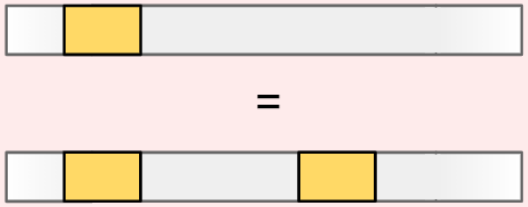
Insertion



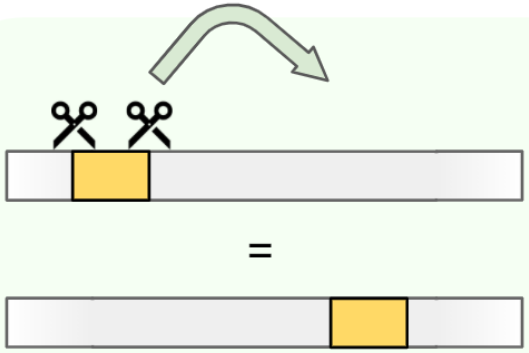
Deletion



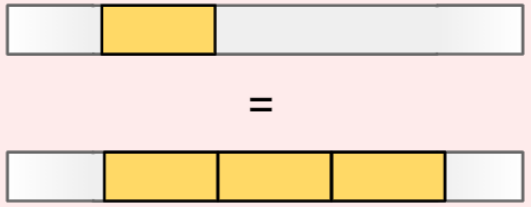
Inversion



Interspersed Duplication



Translocation
(inter- or intra- chromosomal)

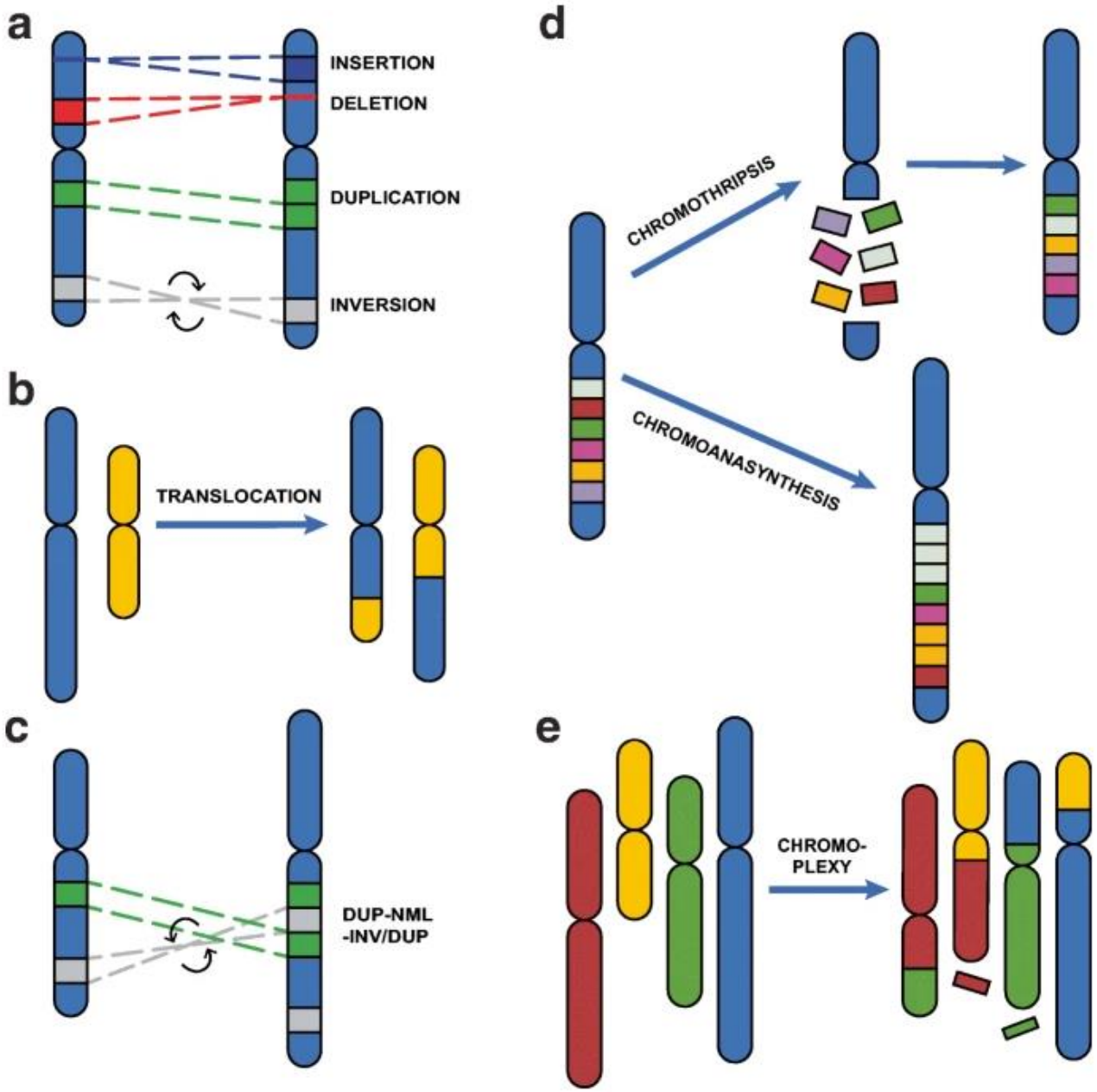


Tandem Duplication

https://www.melbournebioinformatics.org.au/tutorials/tutorials/longread_sv_calling/longread_sv_calling/

Types of Structural Variation

SVs can get complicated!



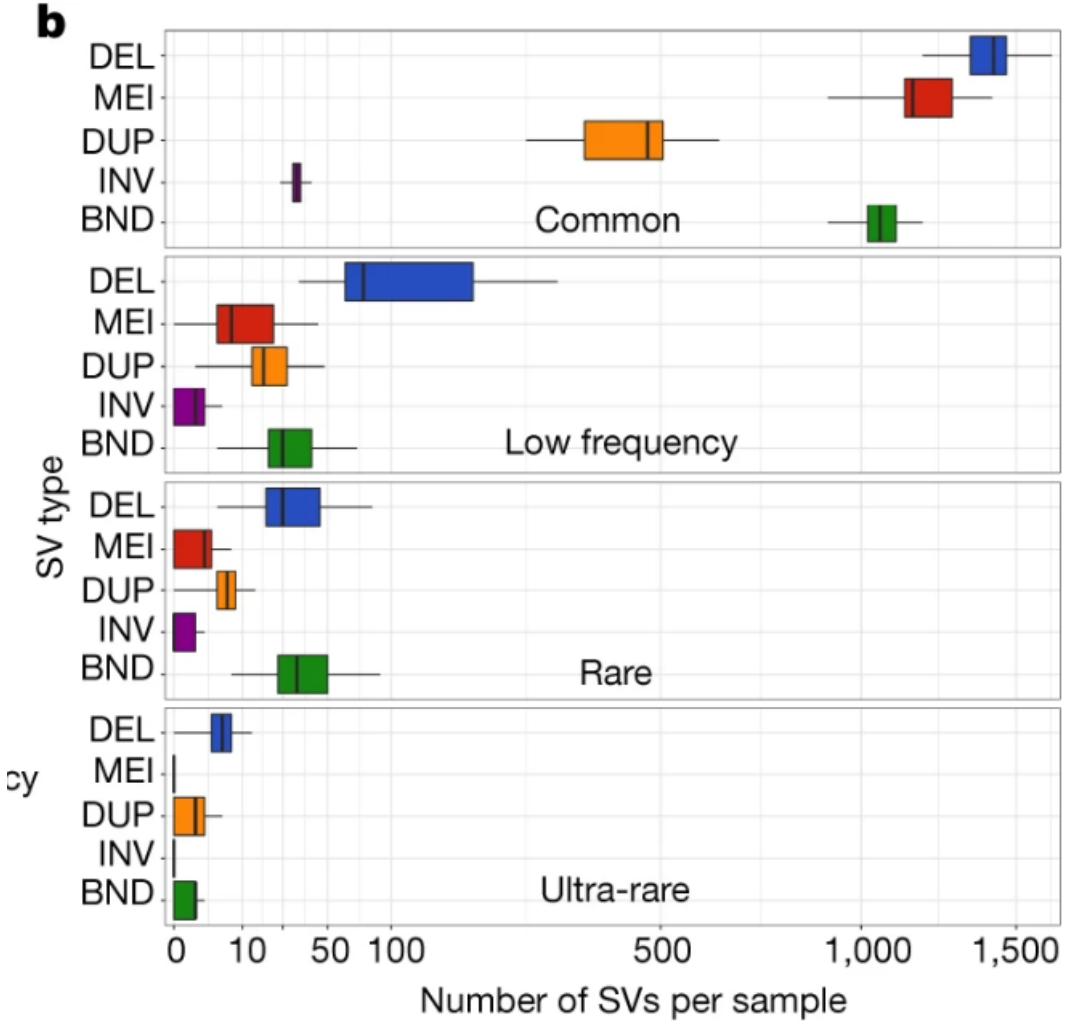
https://link.springer.com/article/10.1007/s10577-019-09623-z

Genomic diversity from SVs

- Underappreciated due to past limitations of technology

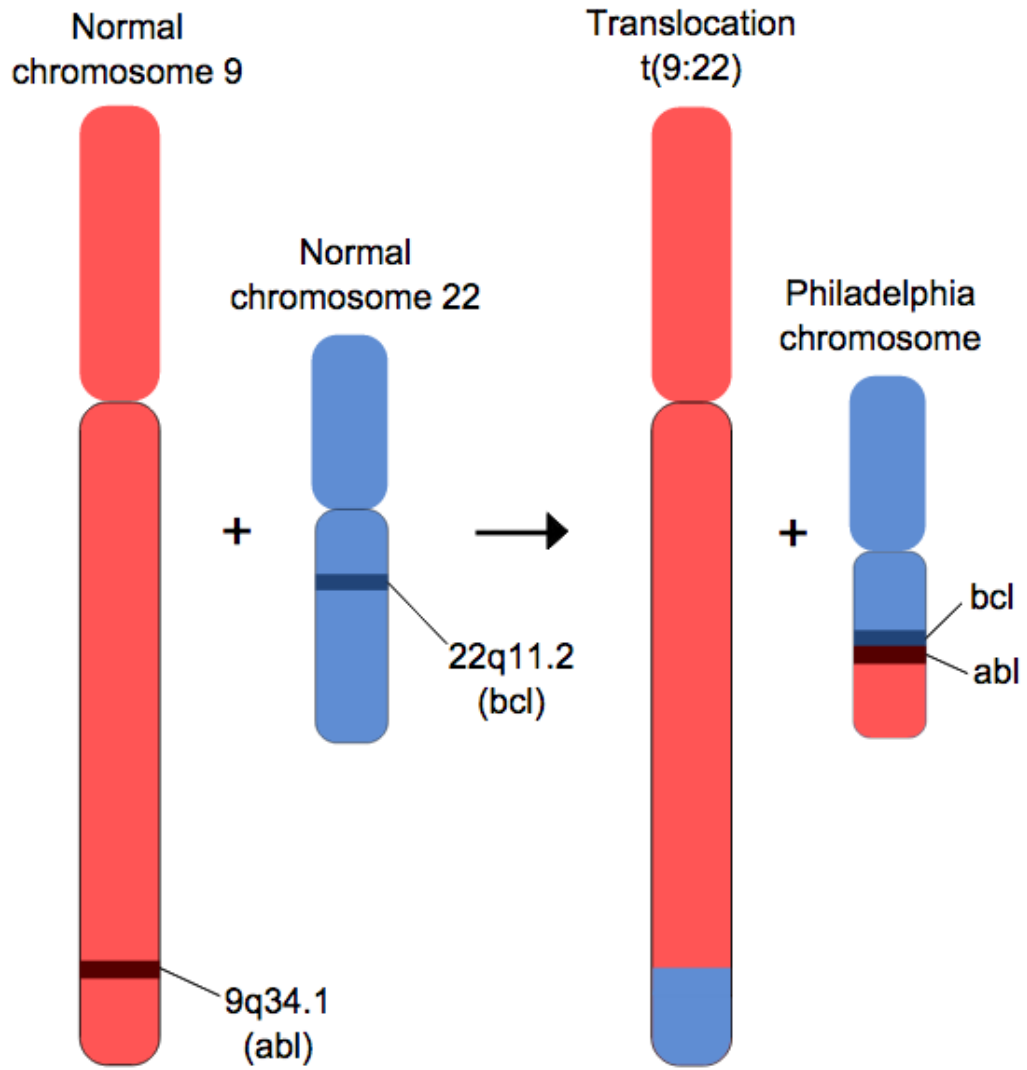
- probably about 1% of each genome (by bp) differs from the reference
  - only 0.1% different by SNPs

SV calls from WGS of 14,623 samples



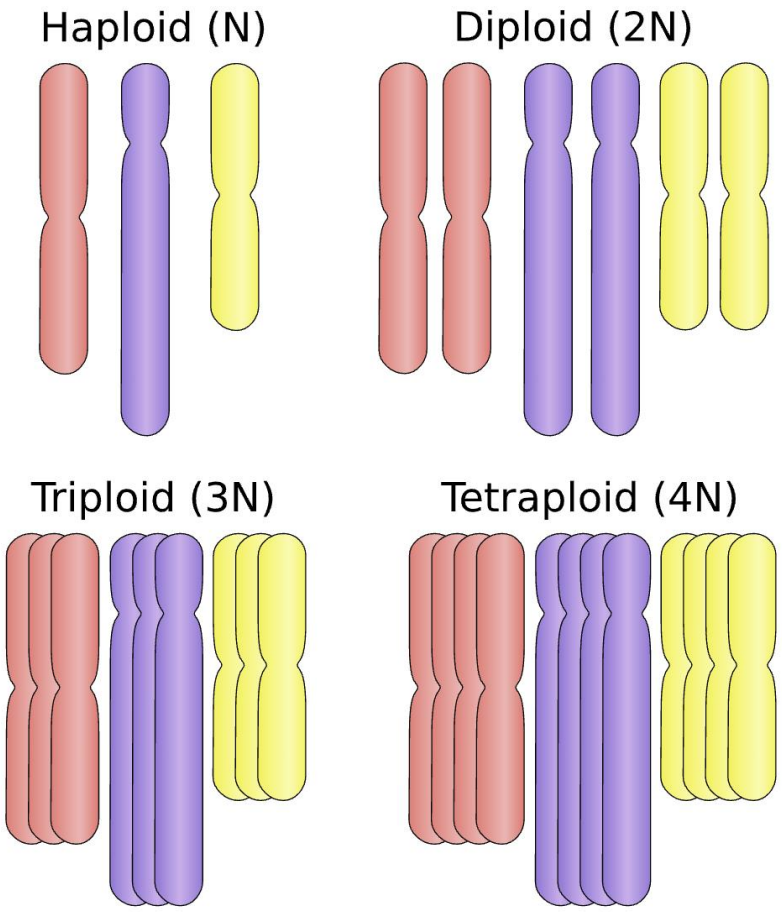
https://www.nature.com/articles/s41586-020-2371-0#Sec6

Somatic SVs are a frequent cause of cancer



https://commons.wikimedia.org/wiki/File:Schematic_of_the_Philadelphia_Chromosome.svg

Polyploidy



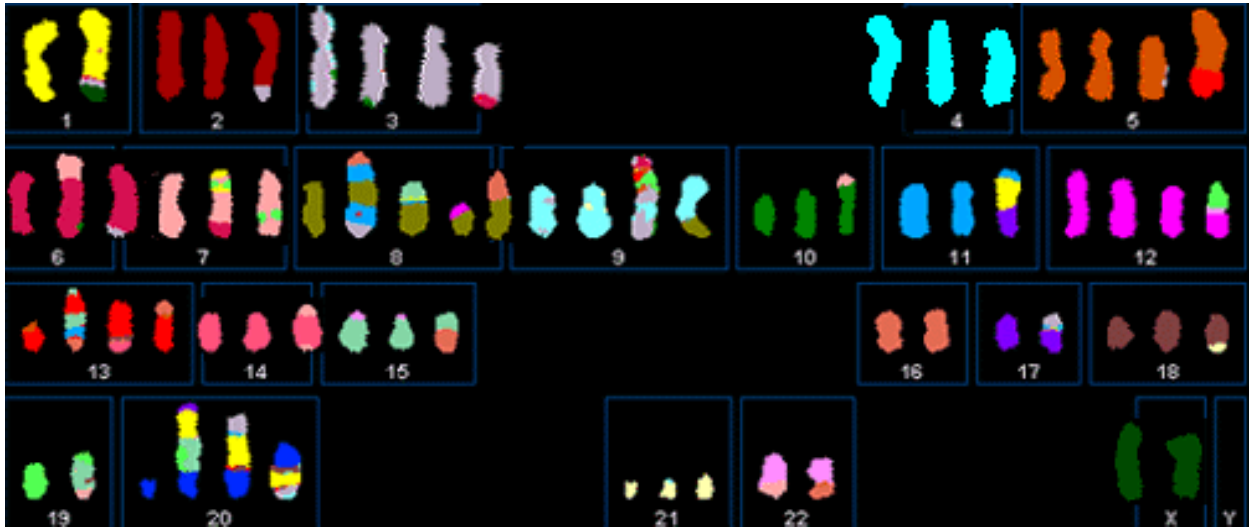
Whole Genome Doubling

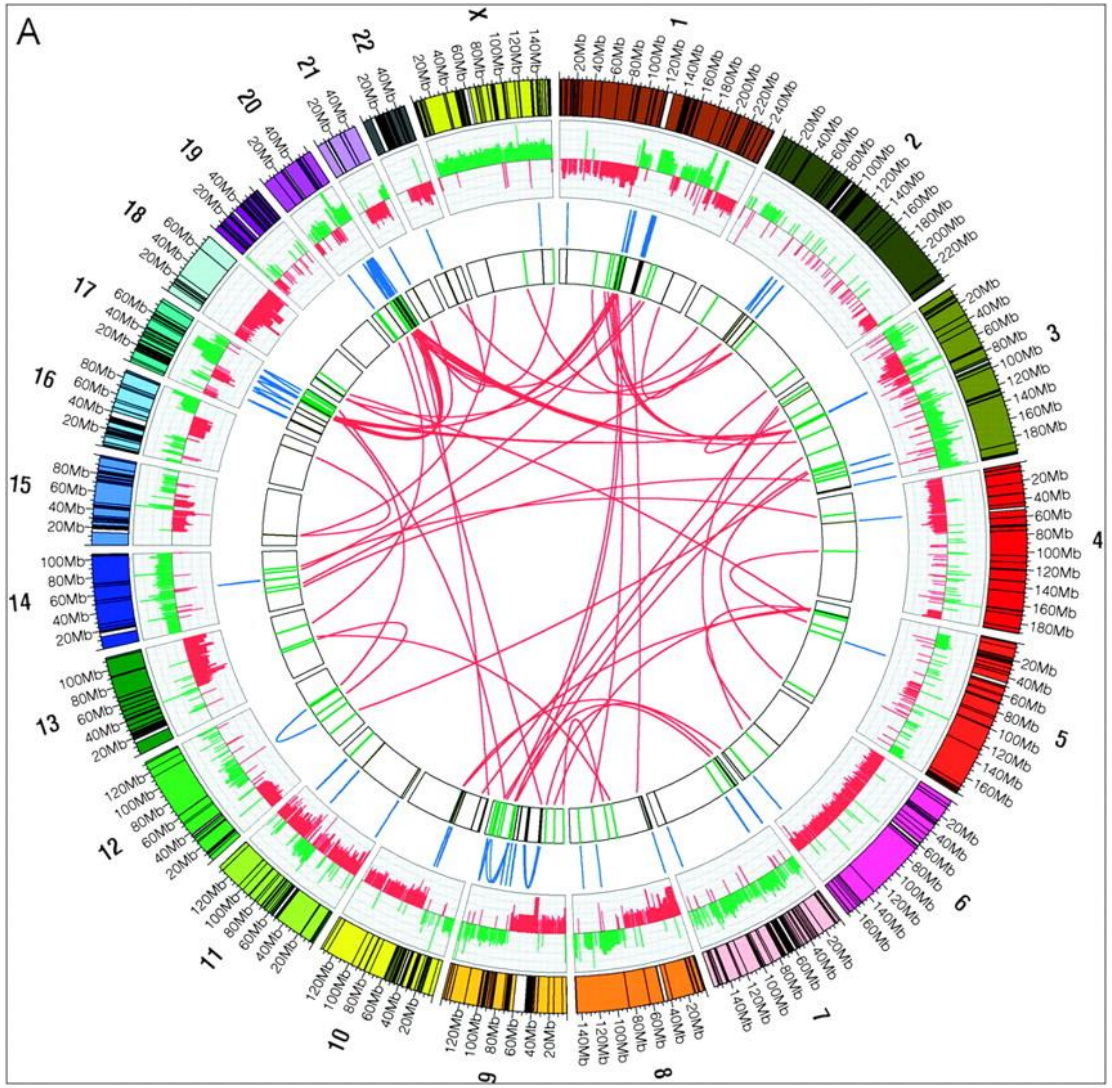
https://upload.wikimedia.org/wikipedia/commons/a/aa/Haploid%2C_diploid_%2Ctriploid_and_tetraploid.svg

https://www.mdpi.com/2073-4425/11/9/1046

Somatic SVs are a frequent cause of cancer

MCF7 Breast Cancer cell line

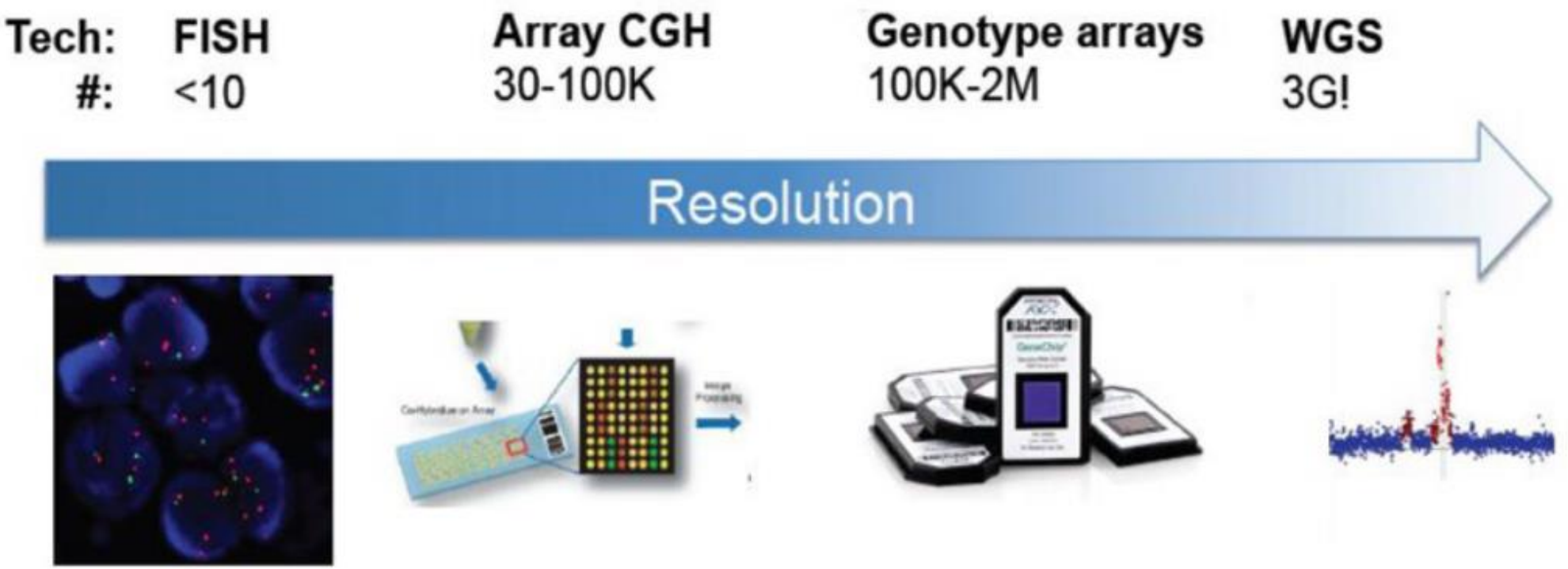




https://www.cellosaurus.org/pawefish/BreastCellLineDescriptions/mcf7.html

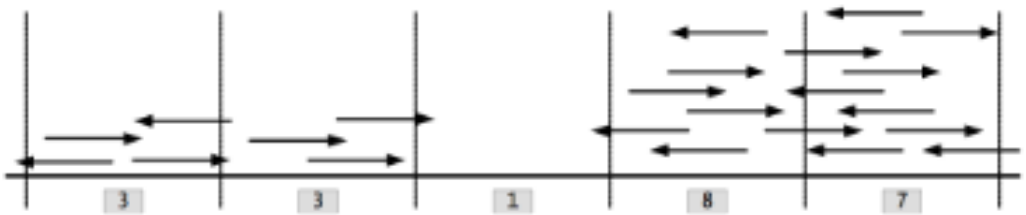
Hampton, et al. doi: 10.1101/gr.080259.108

Copy Number detection



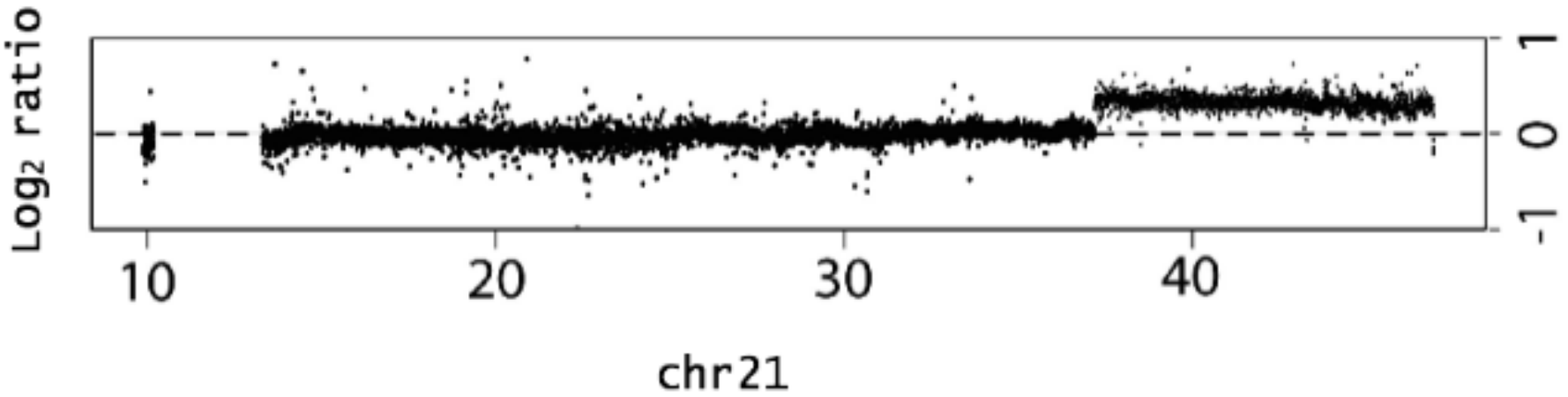
Copy Number detection

- Read counting in windows for tumor and normal data



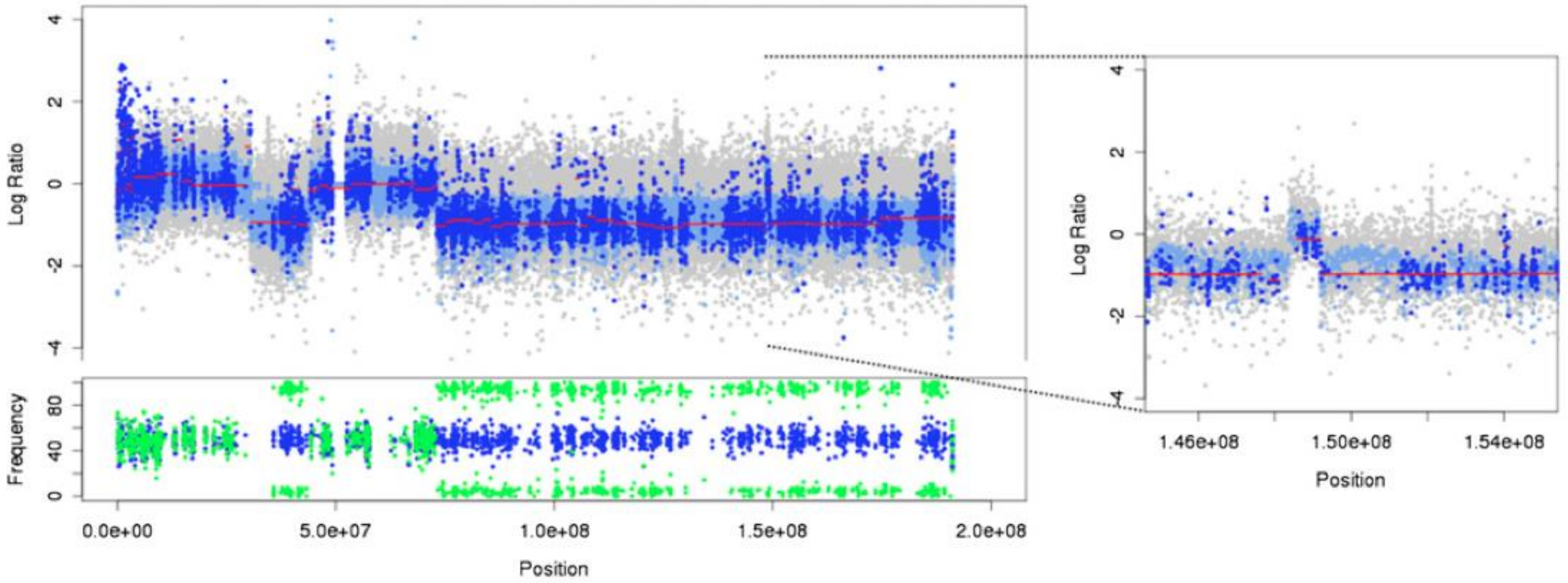
- Log2 ratio for each window
- Chromosome-wide plot

$$\log_2 \frac{\#\text{Reads}_{Disease}}{\#\text{Reads}_{Normal}}$$



Copy Number detection

- Gets more noisy with targeted sequencing, but still works!



- B-allele frequency for CN-neutral Loss of Heterozygosity

Copy Number detection

- Other factors:
  - Sample prep
  - GC-bias
  - Probe affinities
  - Sample Purity
  - Subclonal populations

- Cleaner data, deeper data = higher resolution
  - even a 0.5x low-pass WGS experiment can pick up large events!

Copy Number detection

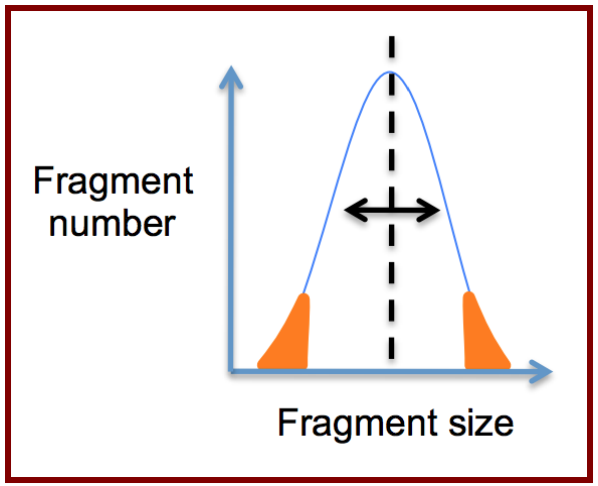
- There are few decent packages for doing this

- CNVkit is my go-to algorithm these days

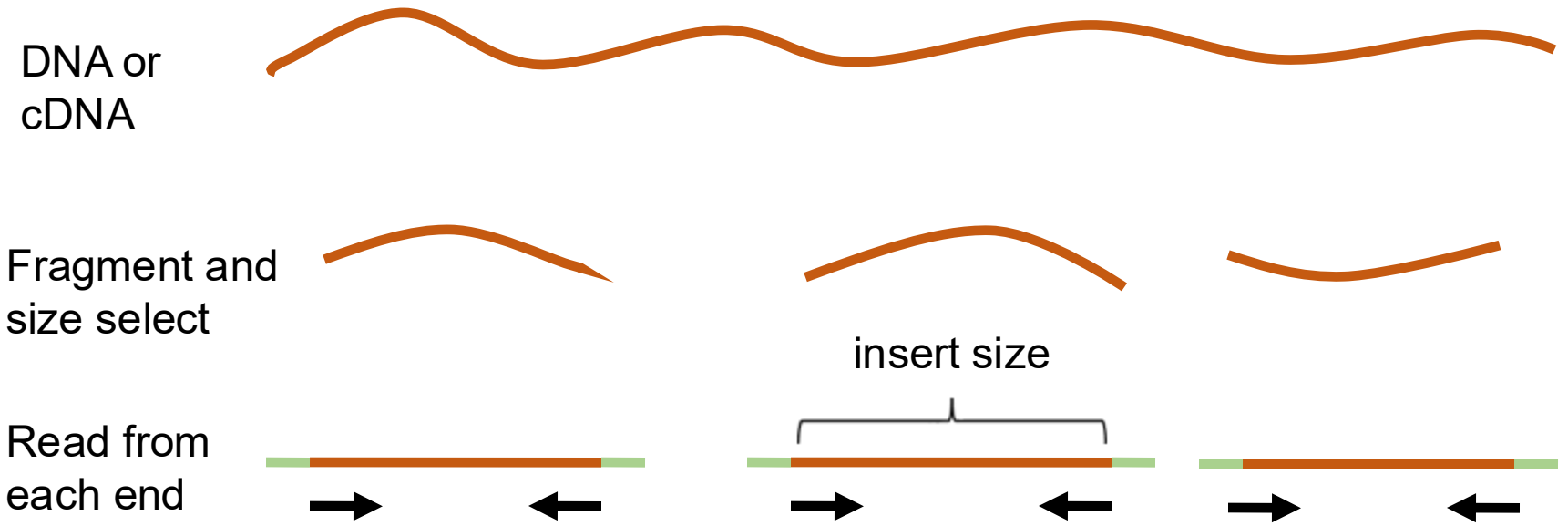
Paired-end sequencing

DNA or
cDNA

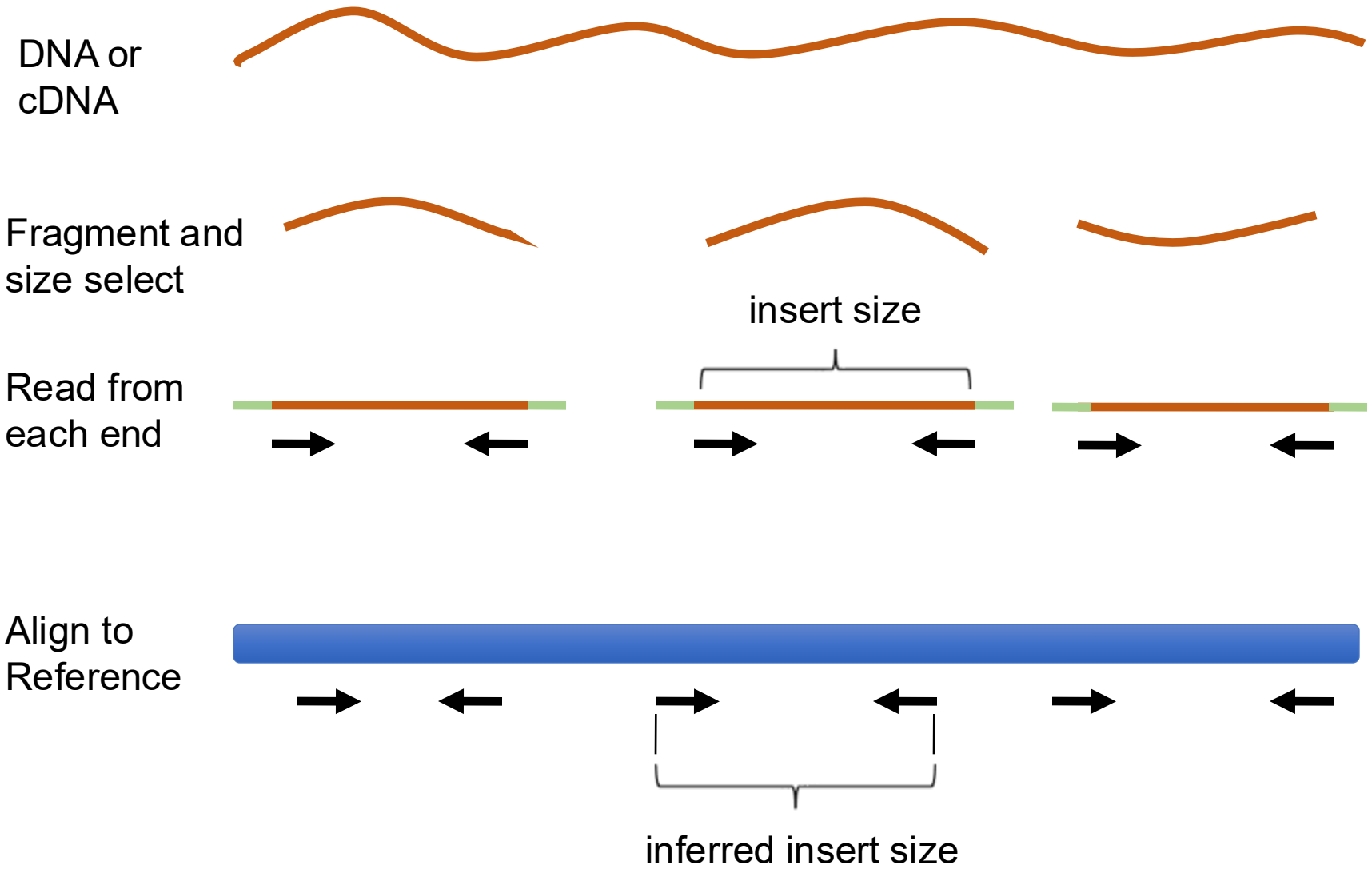
Fragment and
size select



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)

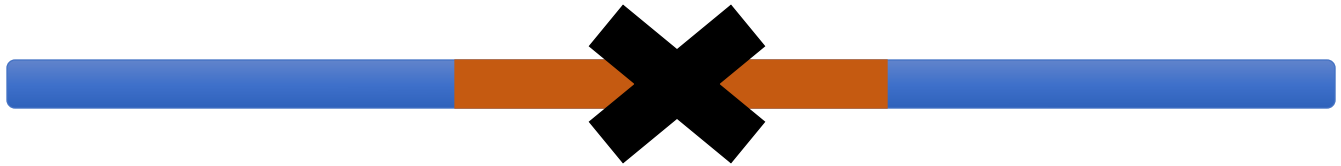
Deletion

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

Deletion

Reference Genome

Subject

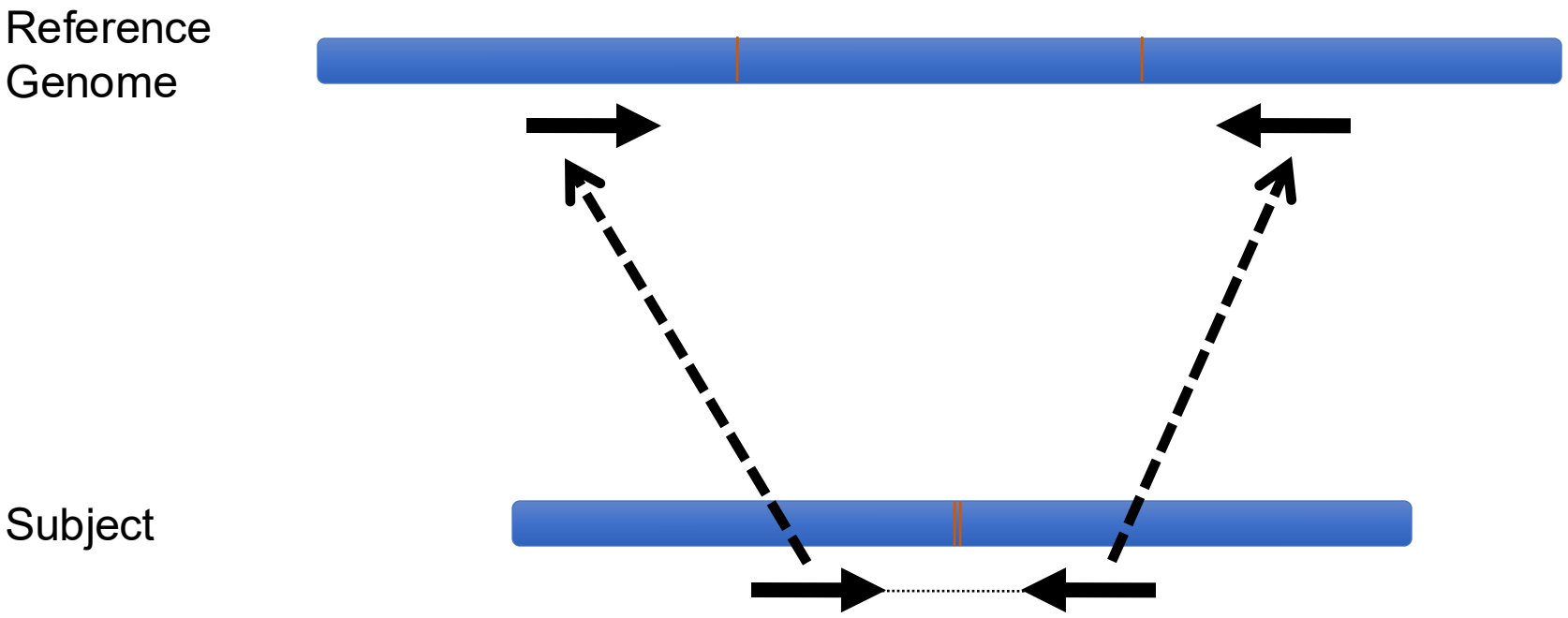


Deletion

Reference Genome

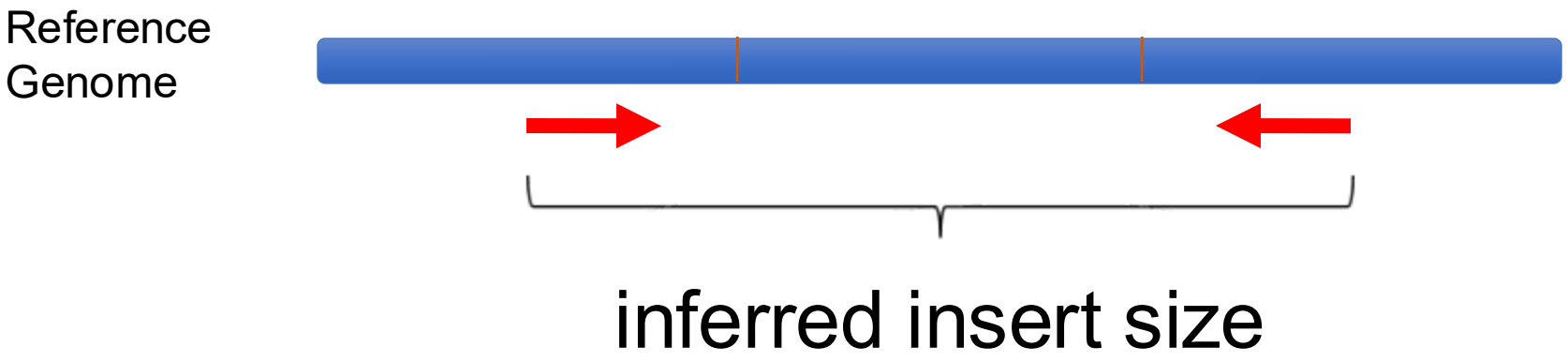
Subject

Deletion

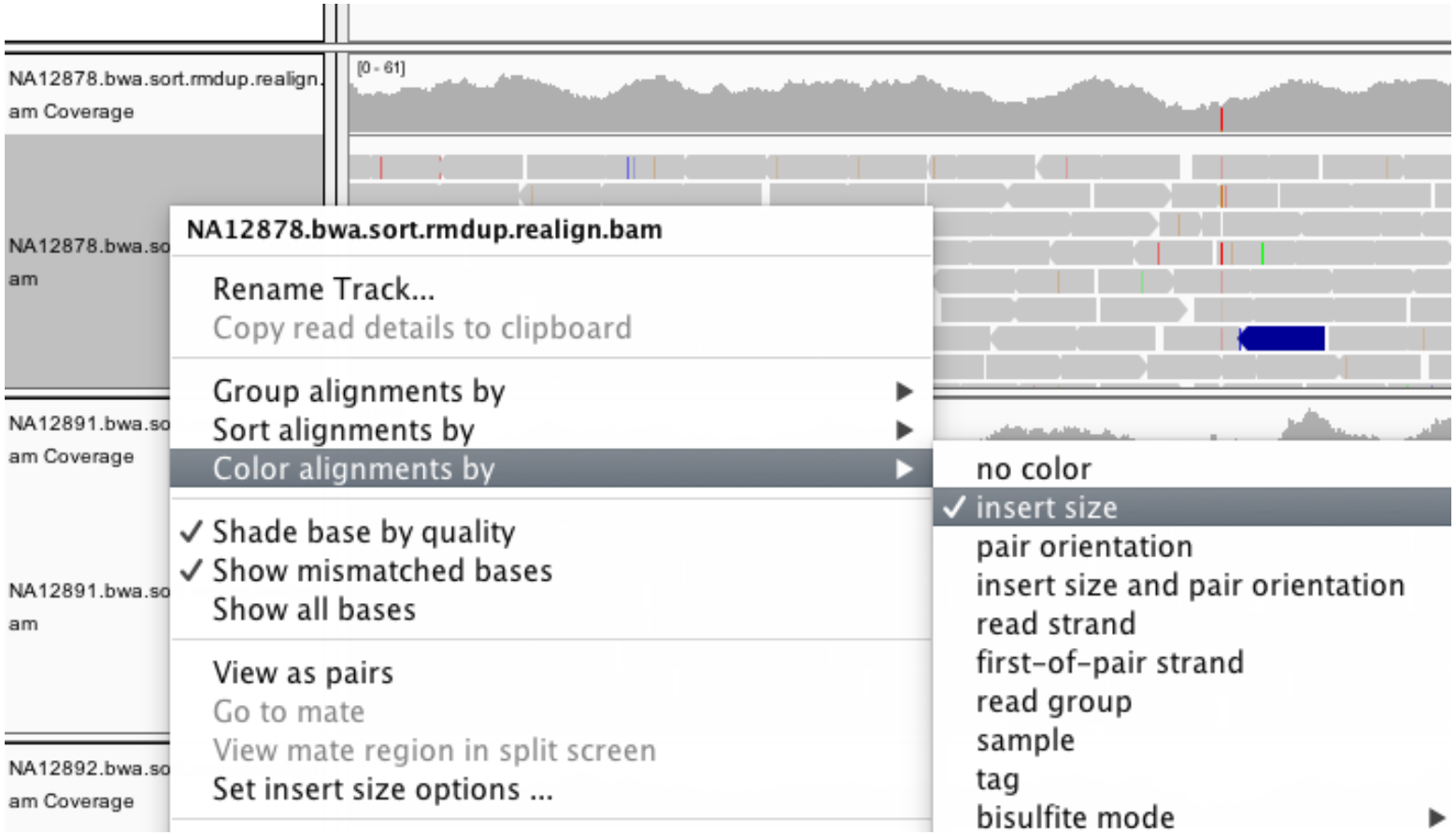


Deletion

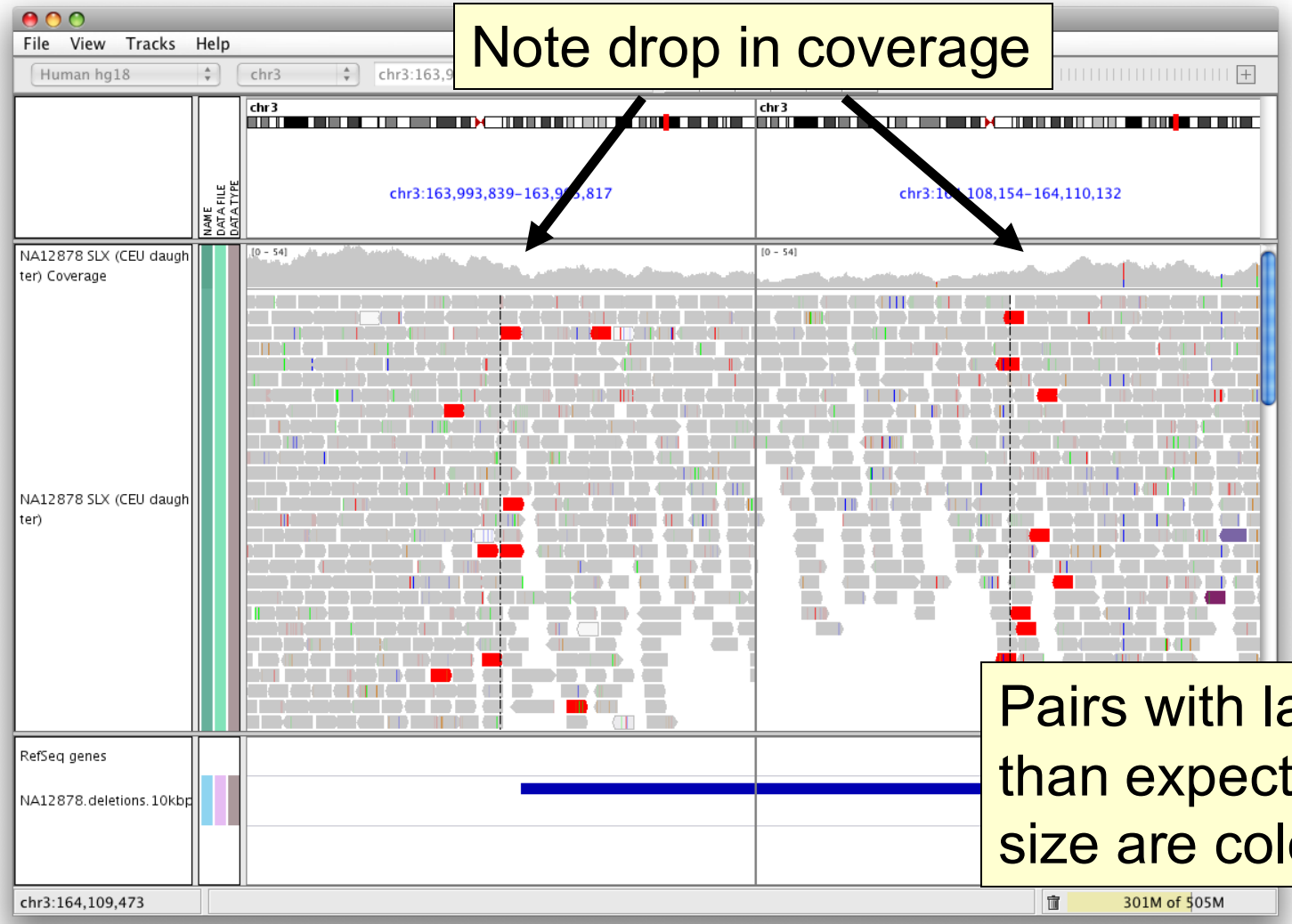
Inferred insert size is > expected value



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
,817
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.

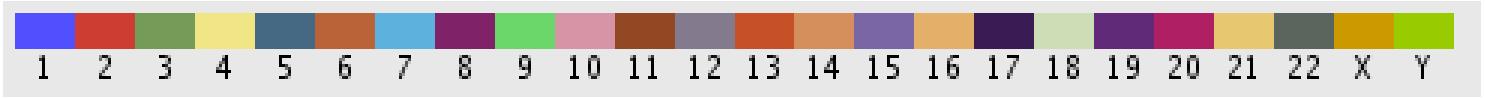
Insert size color scheme

- Smaller than expected insert size:

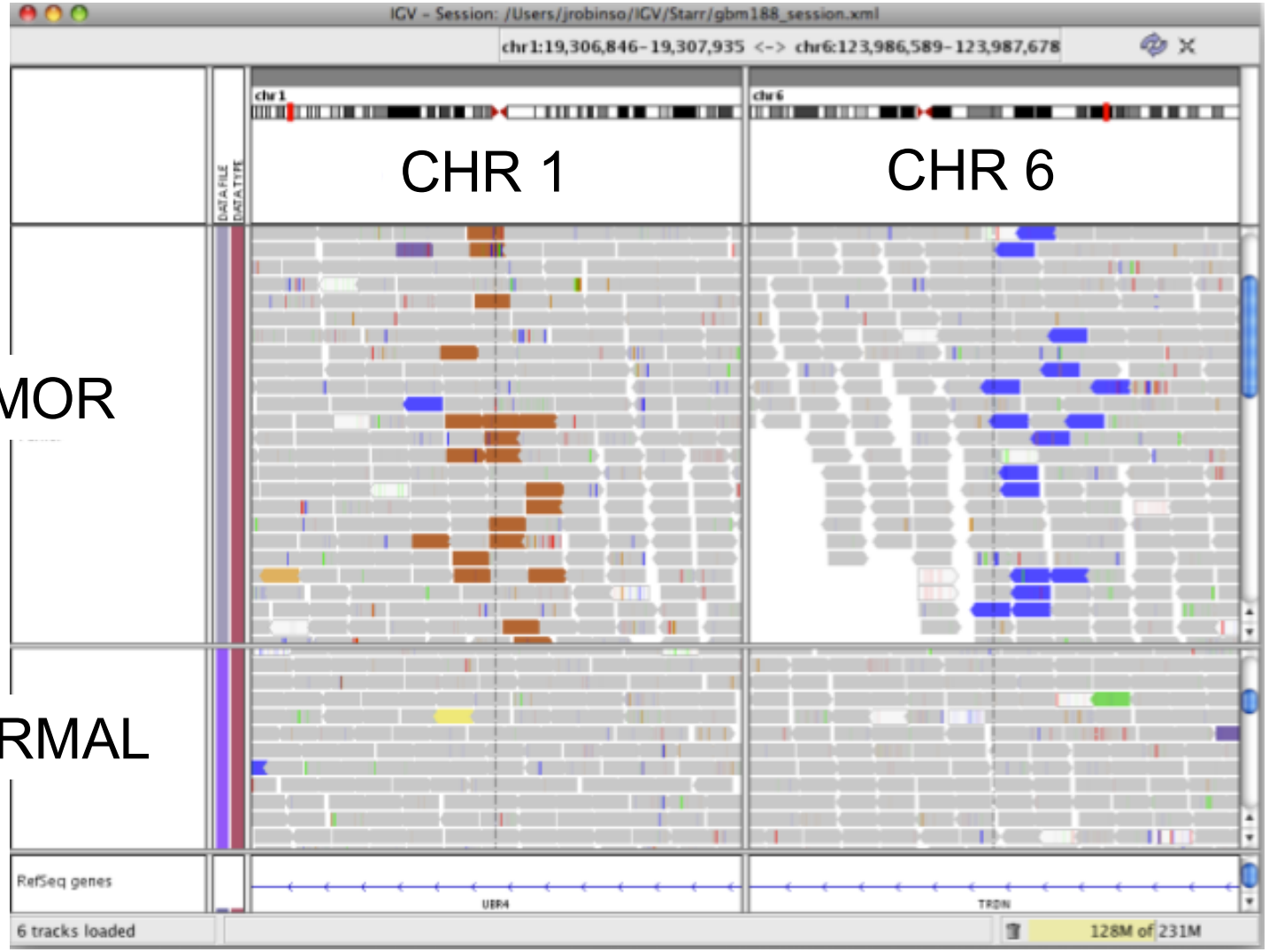
- Larger than expected insert size:

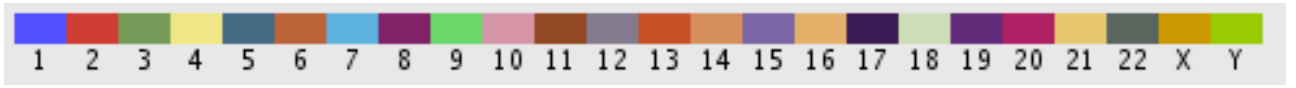
- Pairs on different chromosomes

Each end colored by chromosome of its mate

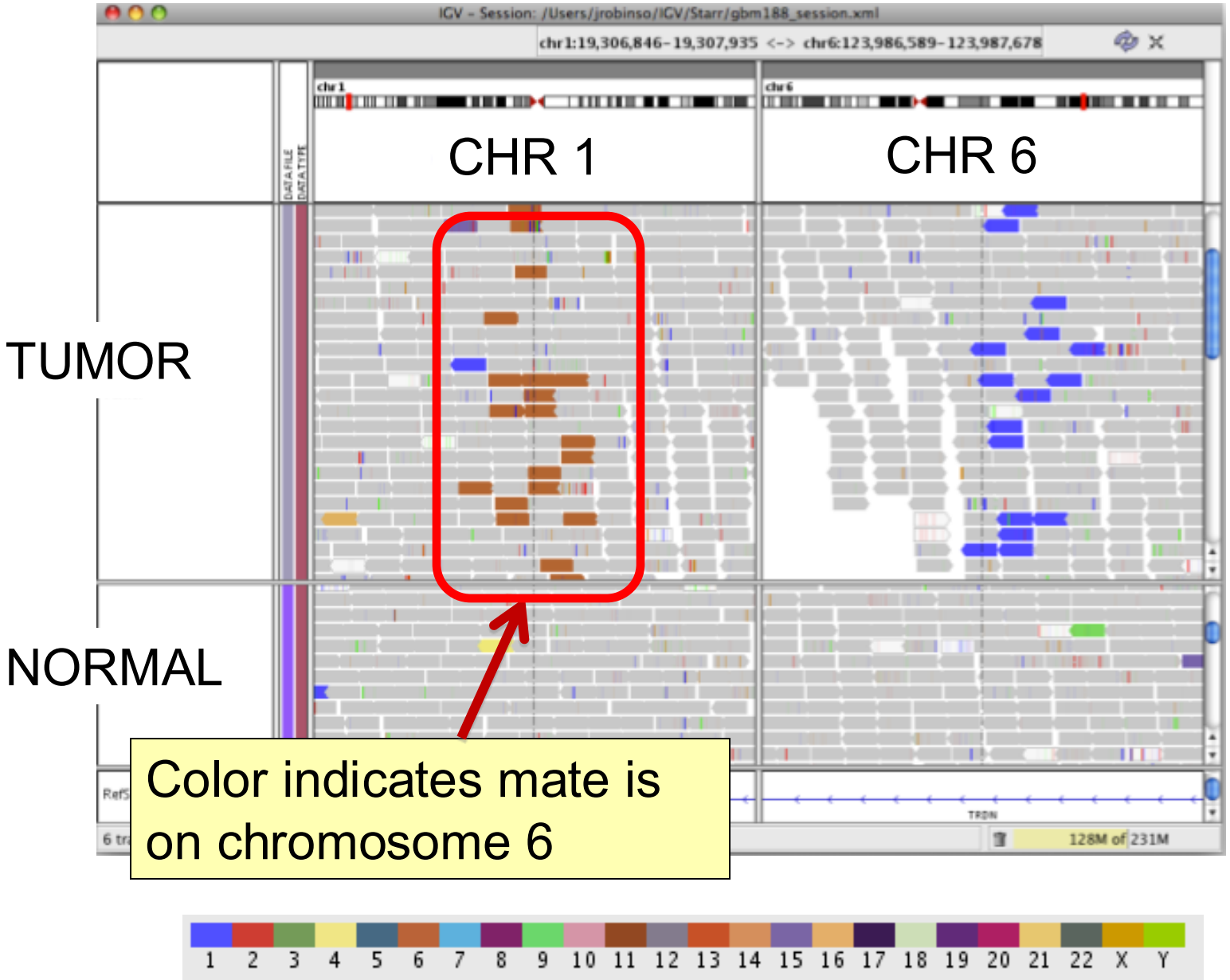


Rearrangement





Rearrangement



Interpreting Read-Pair Orientations

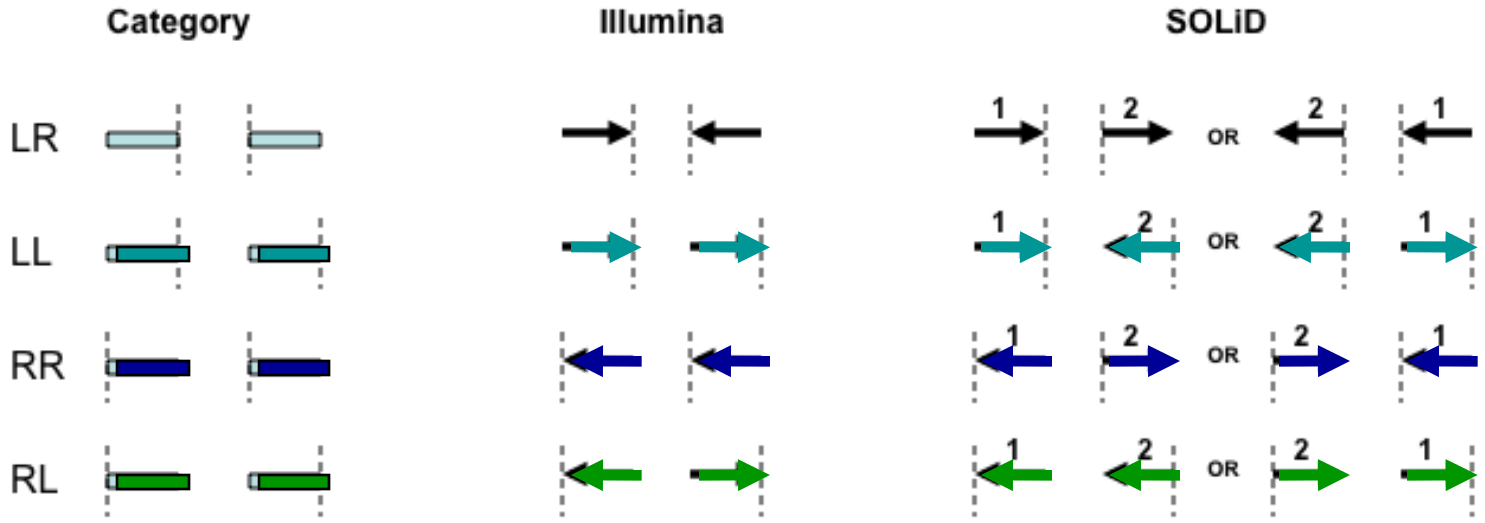
Orientation of paired reads can reveal structural events:

- Inversions
- Duplications
- Translocations
- Complex rearrangements

Orientation is defined in terms of

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

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

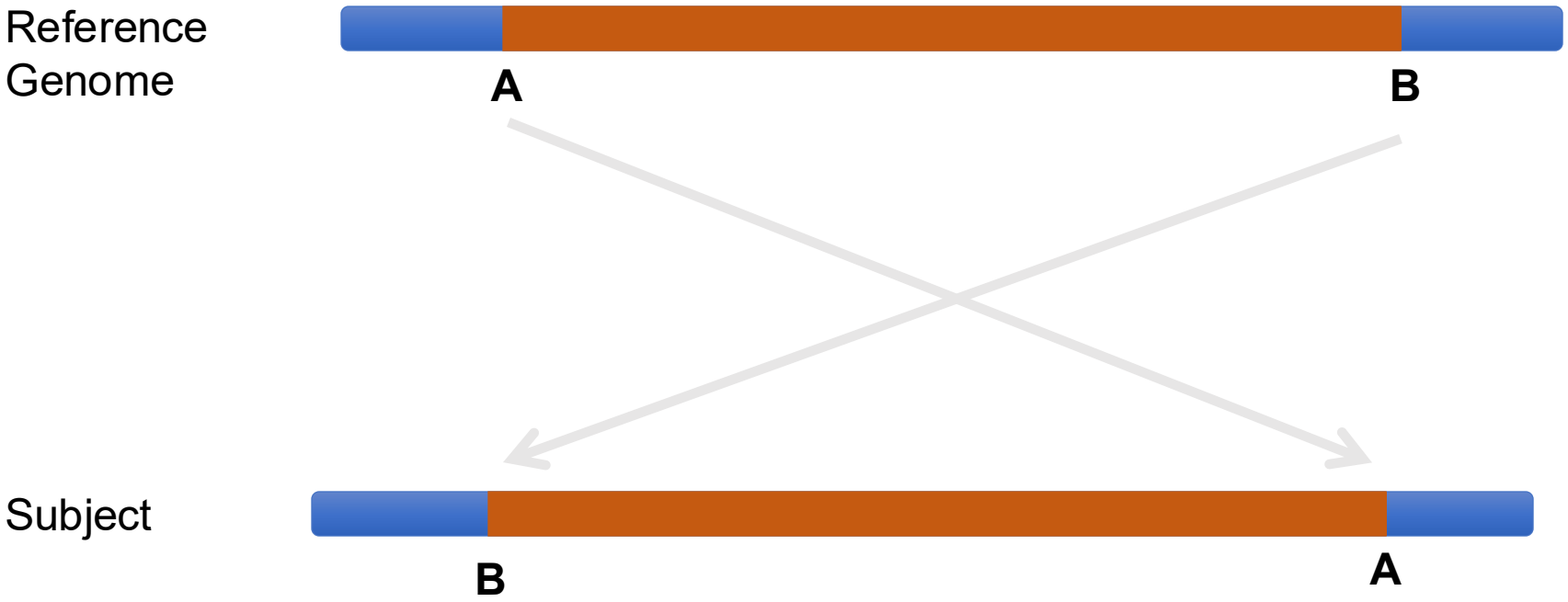
Inversion

Reference genome

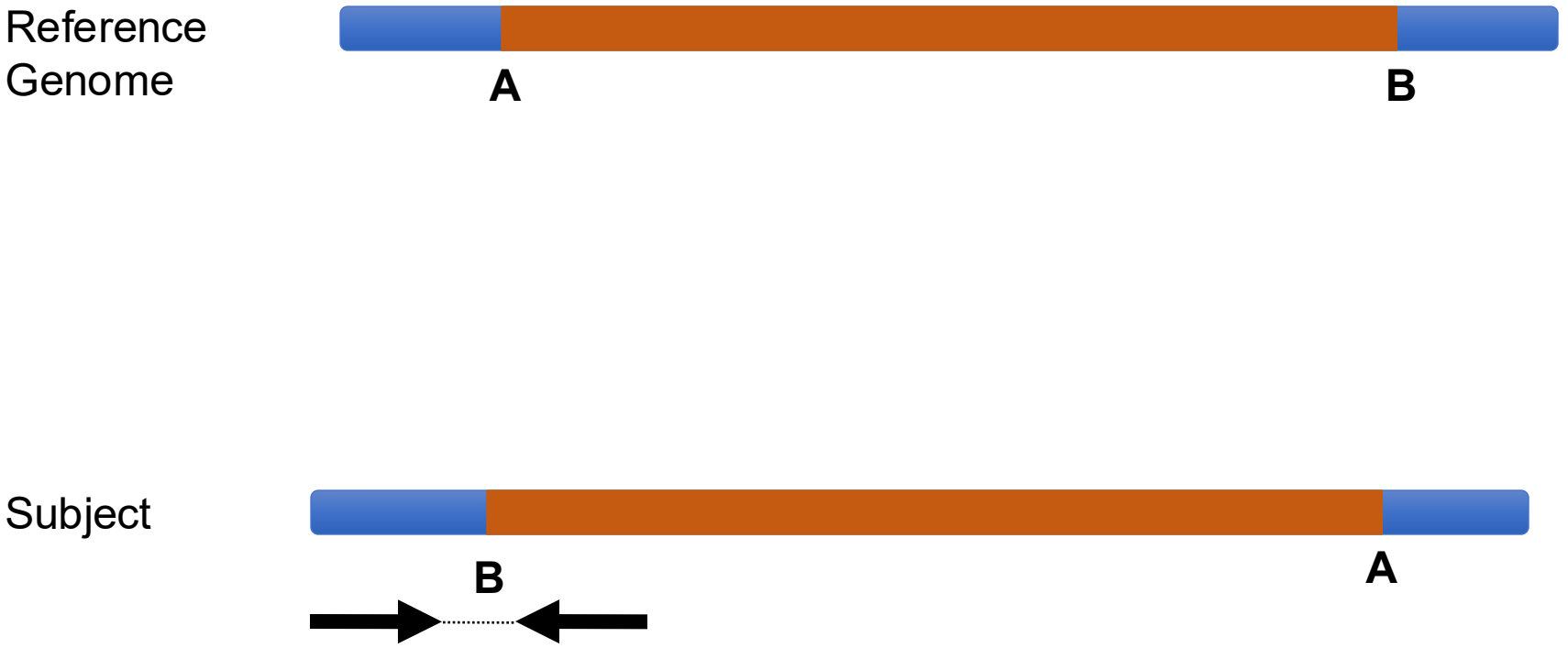
Inversion

Reference genome

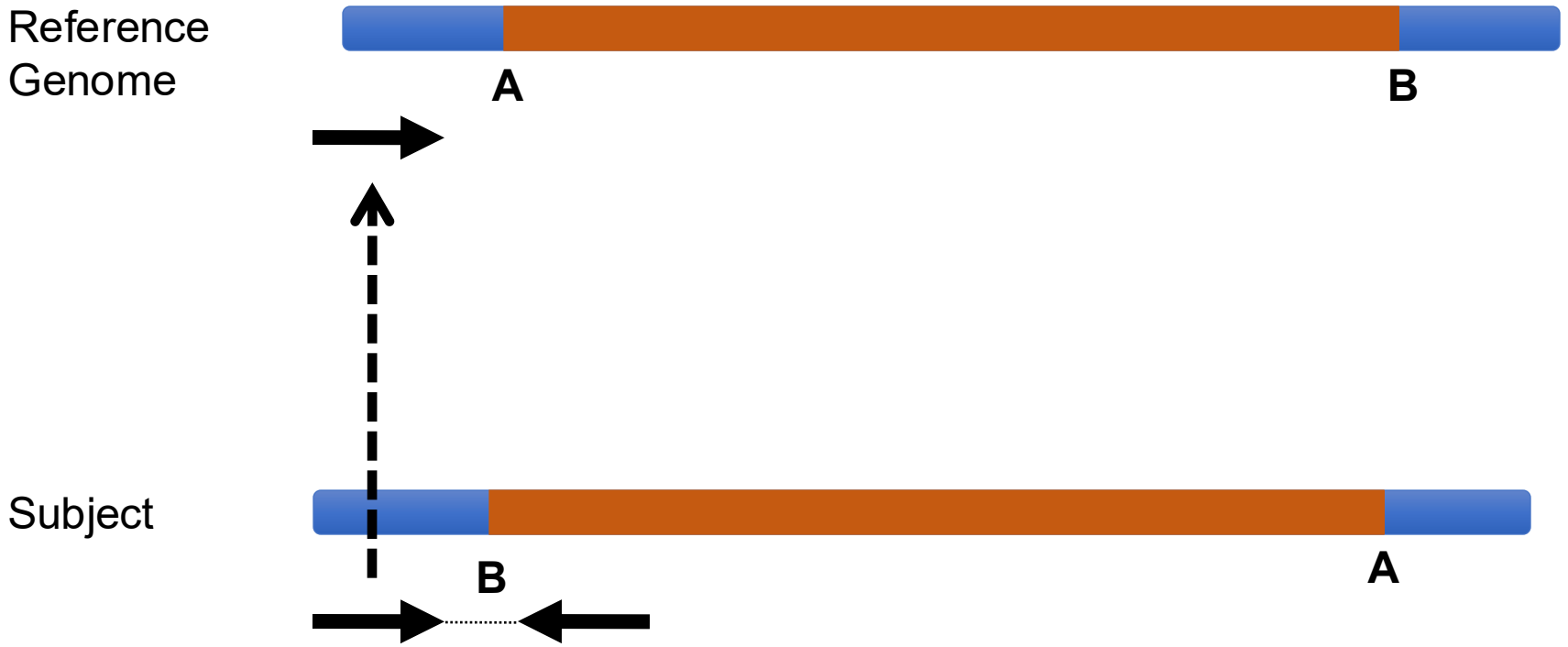
Inversion



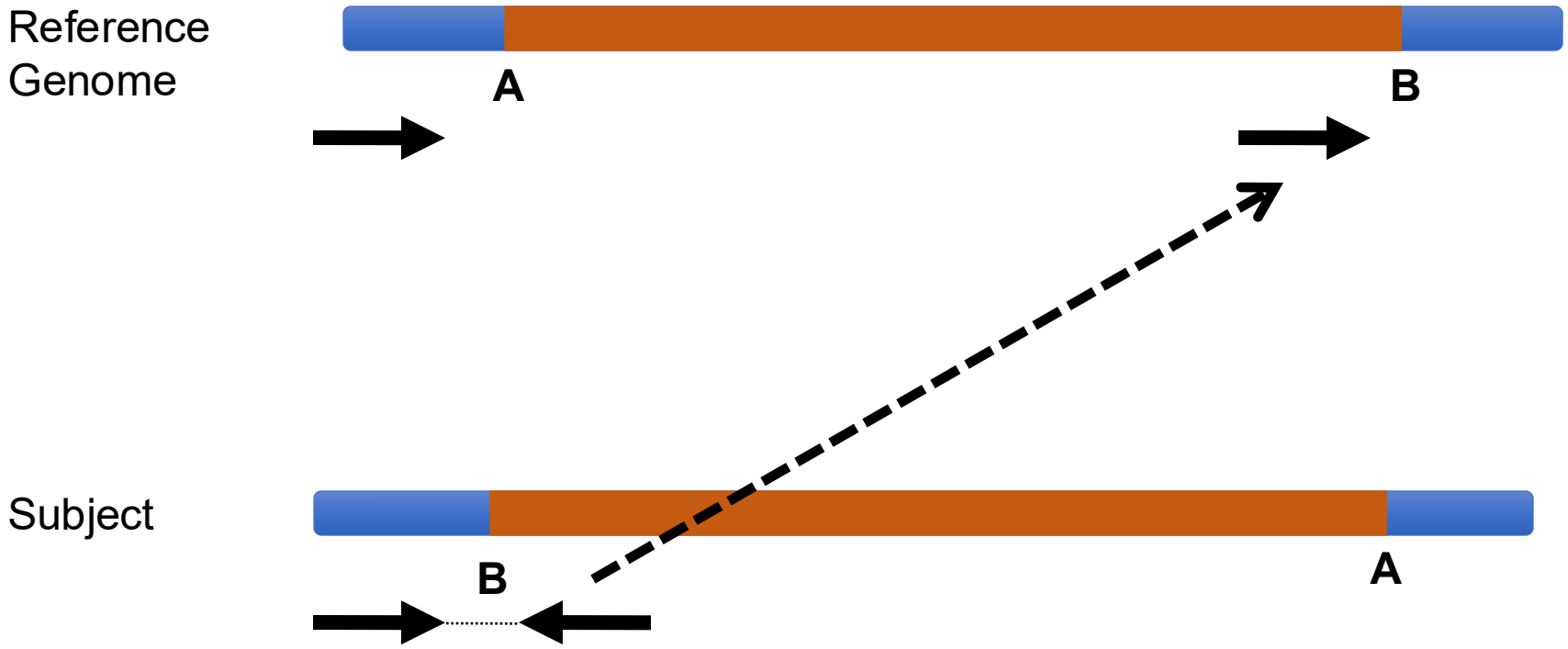
Inversion



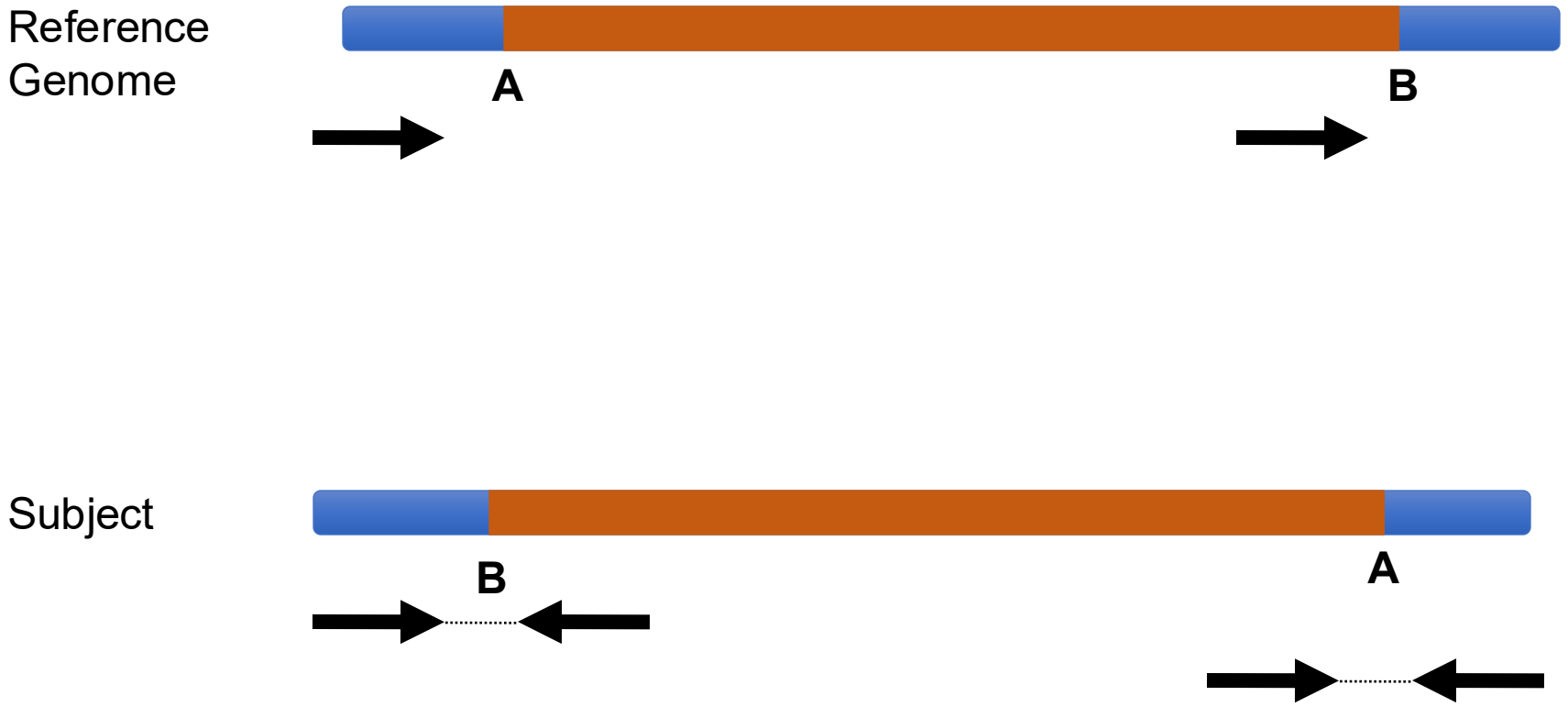
Inversion



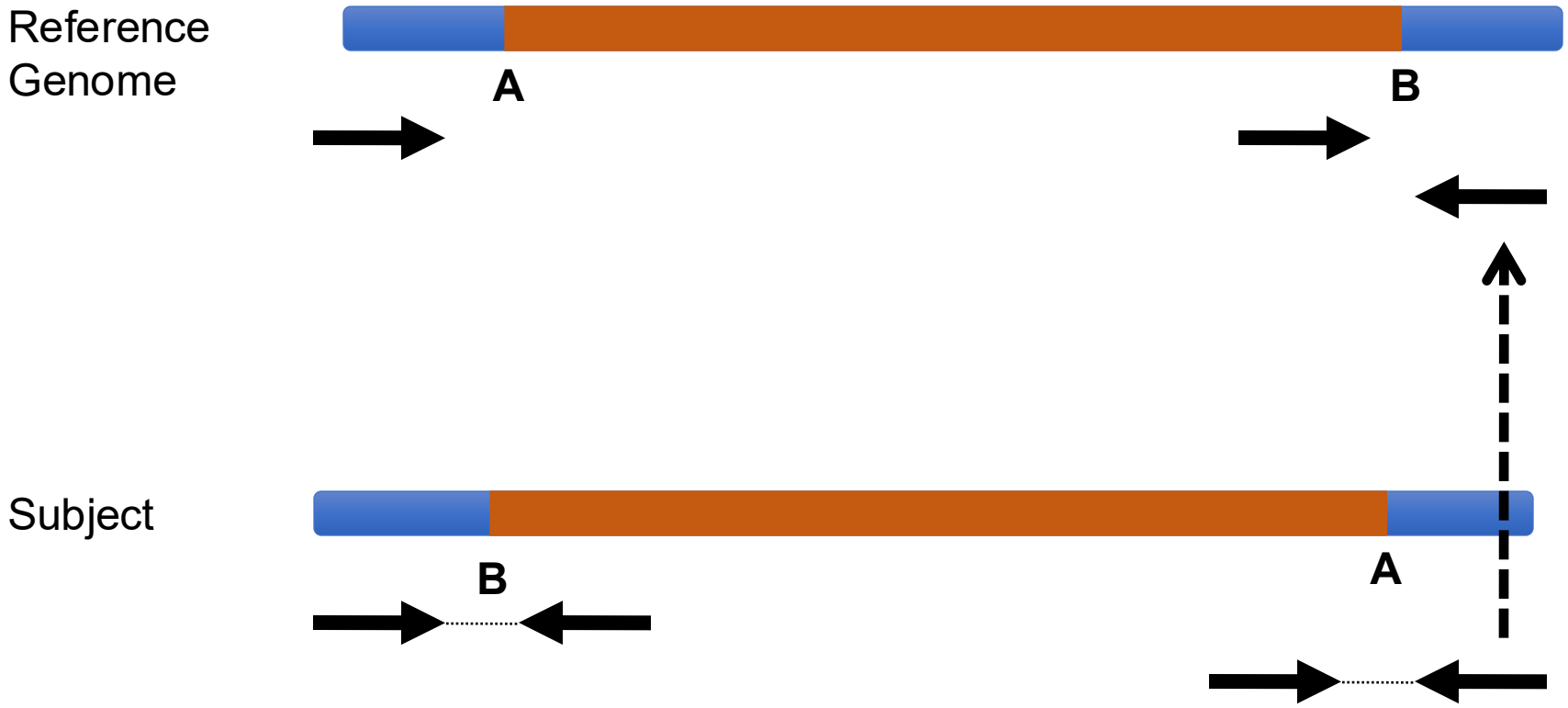
Inversion



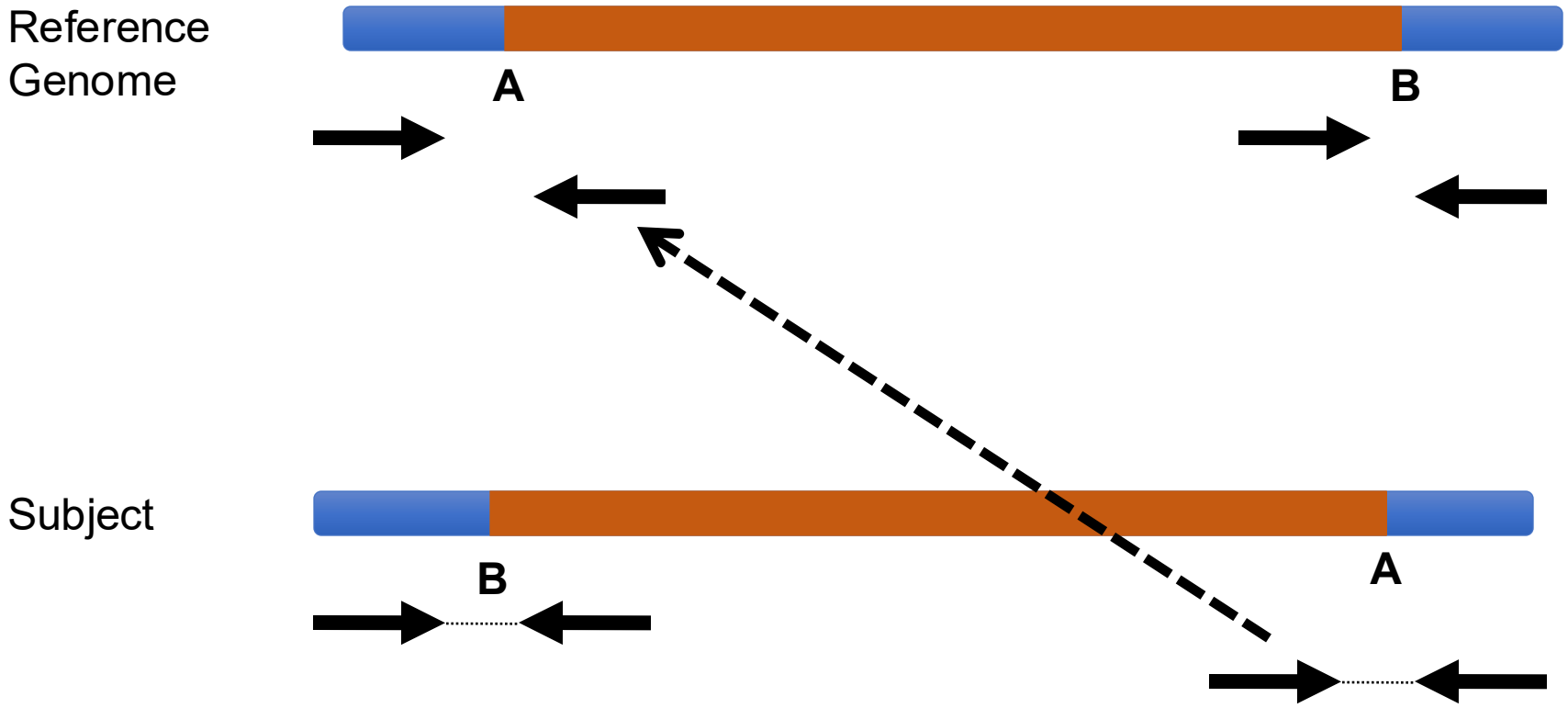
Inversion



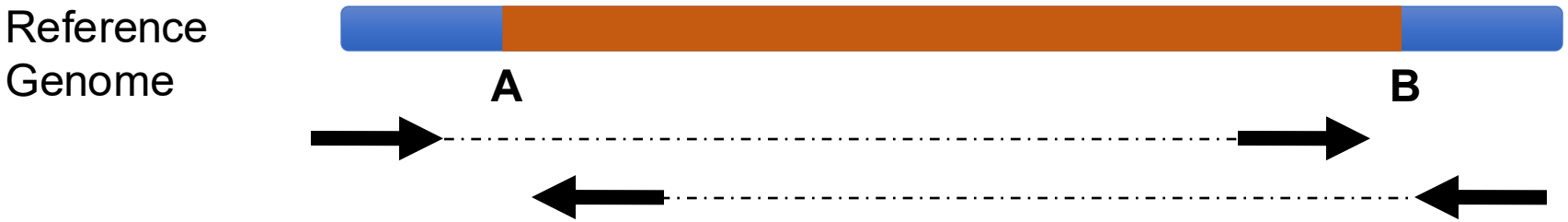
Inversion



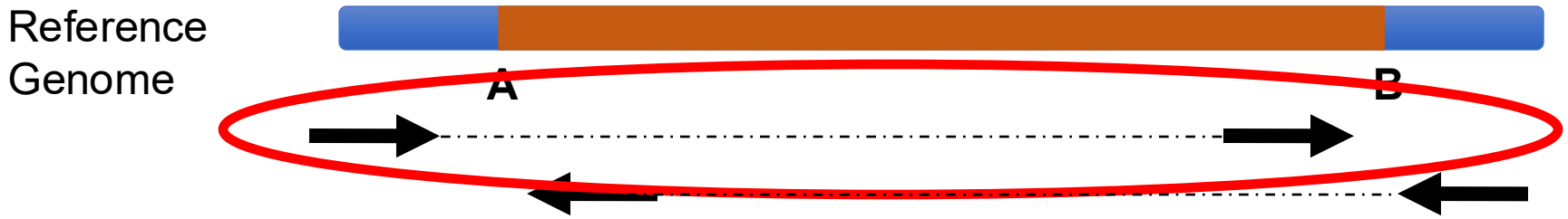
Inversion



Inversion



Inversion

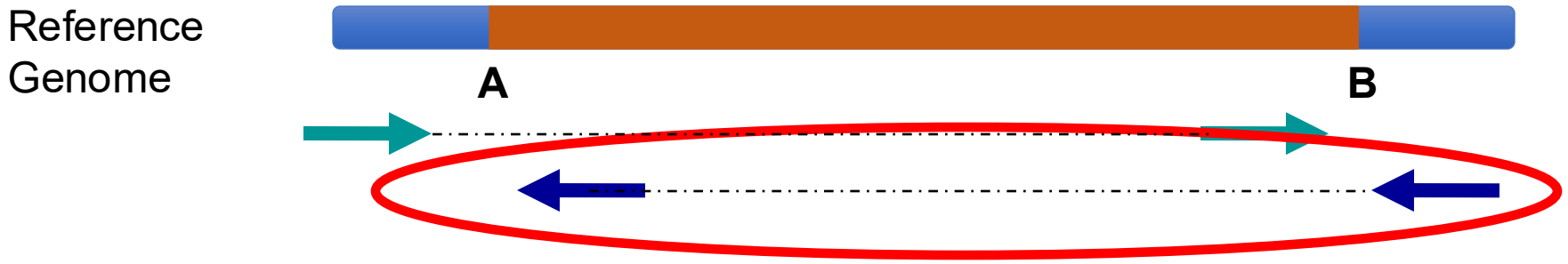


Anomaly: expected orientation of pair is inward facing (→ ←)

Inversion

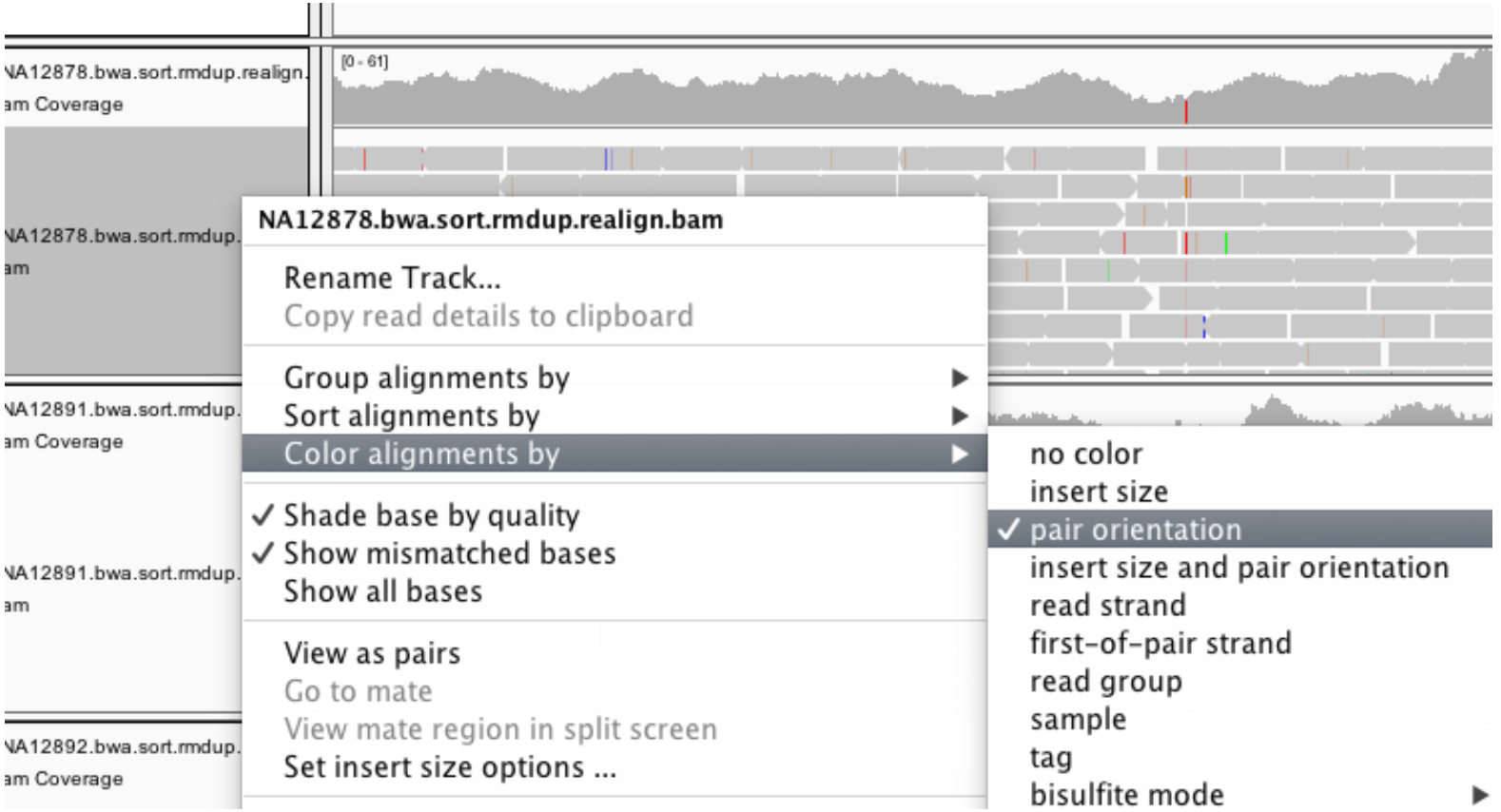
“Left” side pair

Inversion

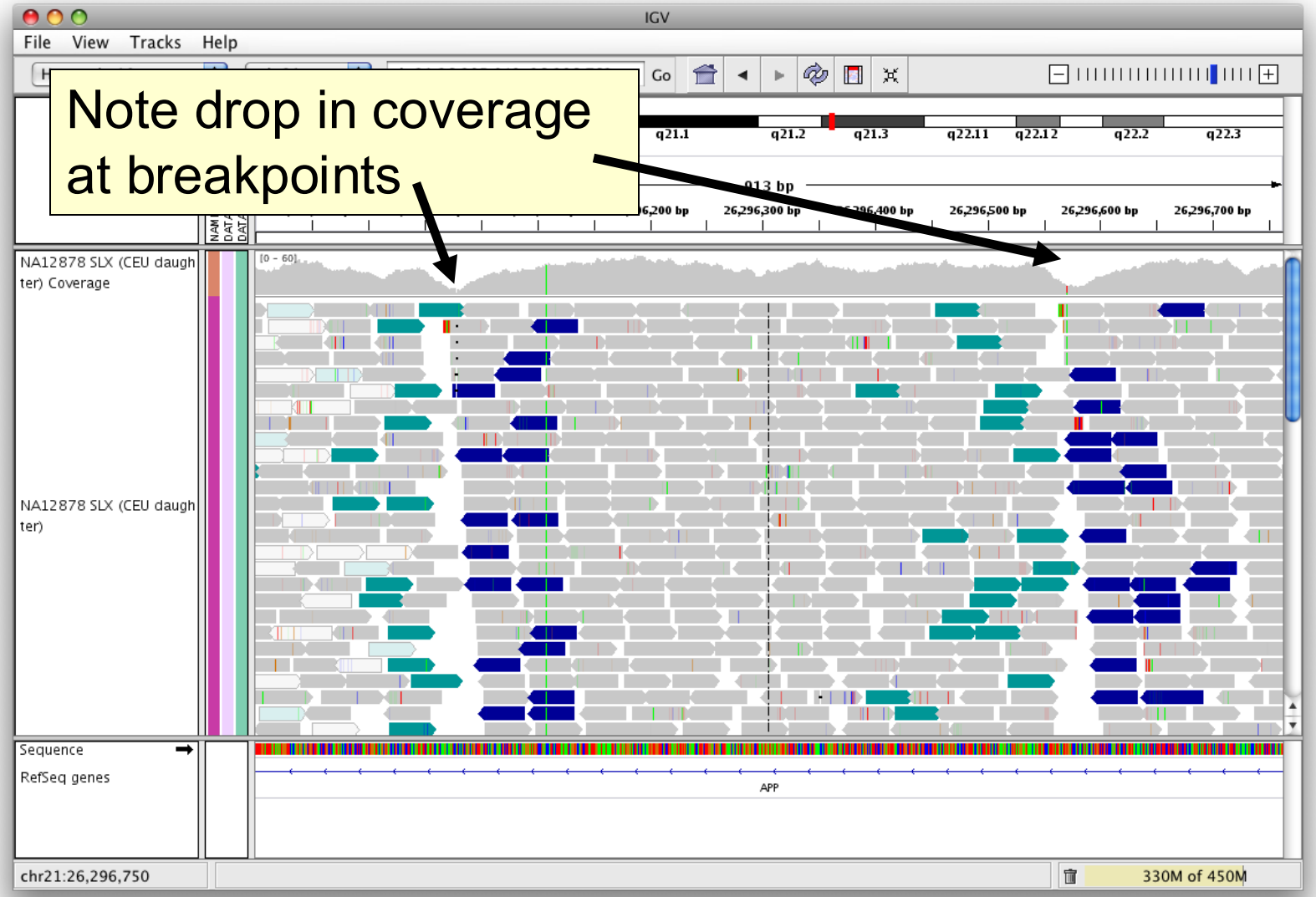


“Right” side pair

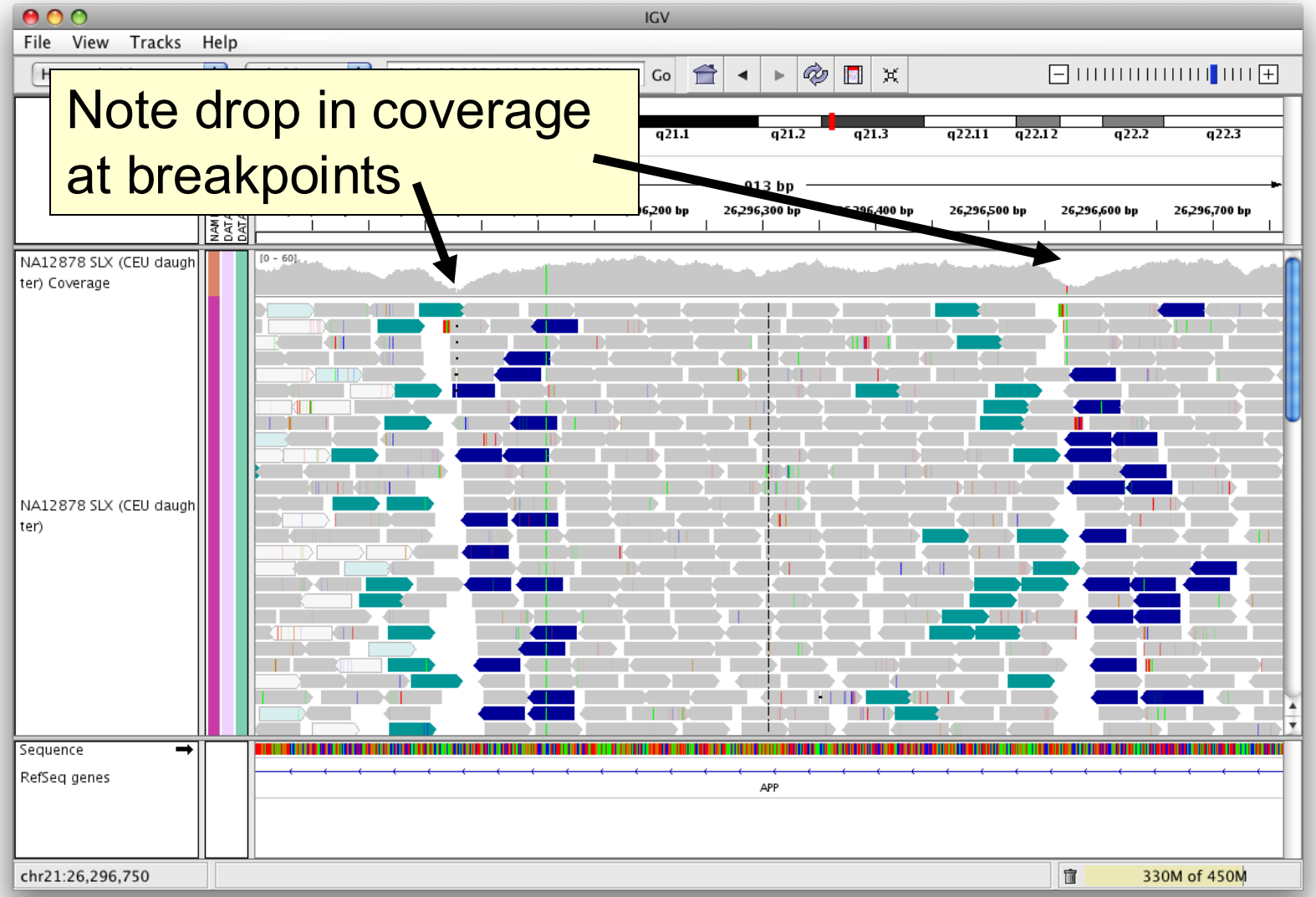
Color by pair orientation



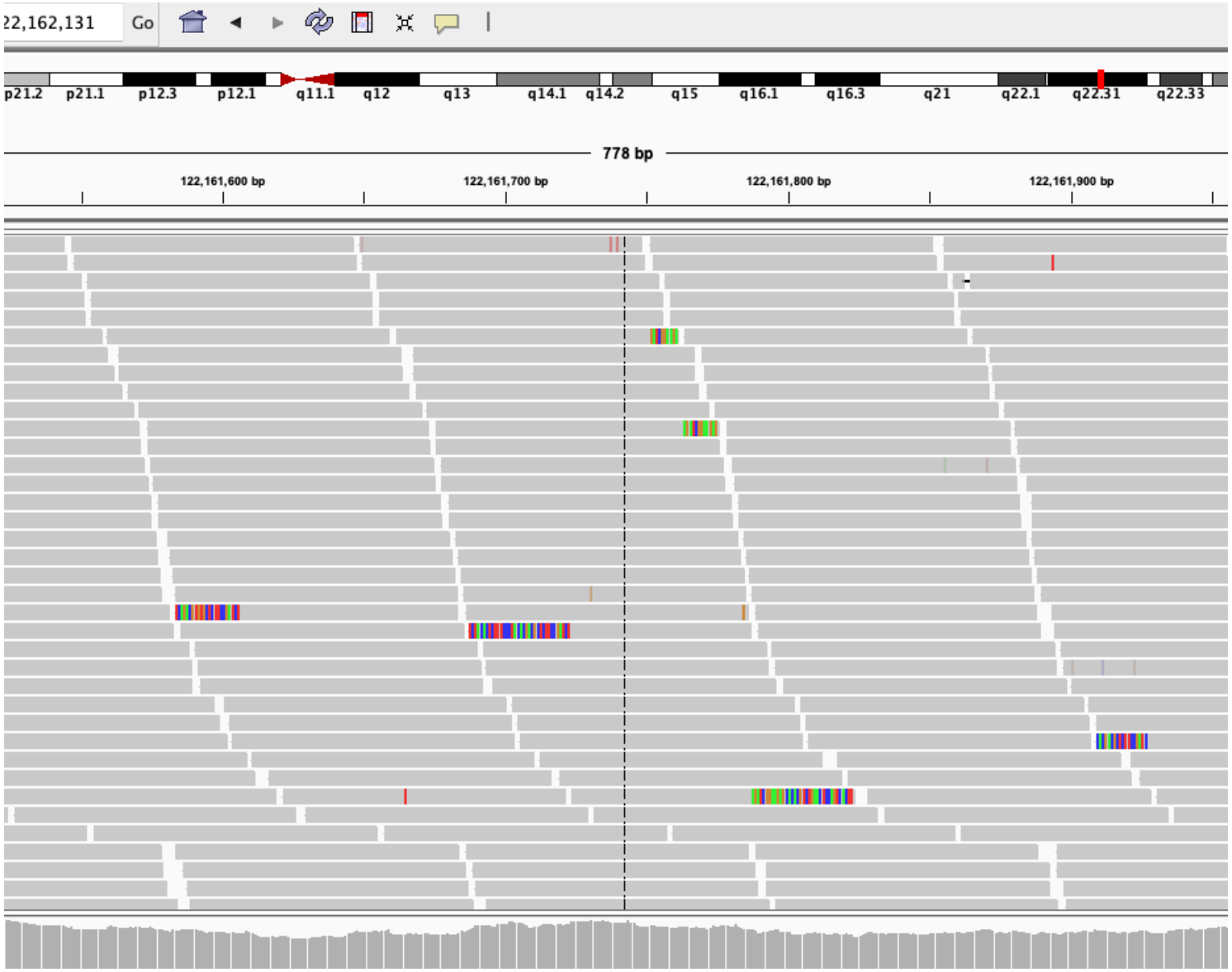
Inversion


IGV
File View Tracks Help
Go
Note drop in coverage at breakpoints
q21.1 q21.2 q21.3 q22.11 q22.12 q22.2 q22.3
26,296,200 bp 26,296,300 bp 26,296,400 bp 26,296,500 bp 26,296,600 bp 26,296,700 bp
NA12878 SLX (CEU daughter) Coverage
[0 - 60]
NA12878 SLX (CEU daughter)
Sequence
RefSeq genes
APP
chr21:26,296,750
330M of 450M

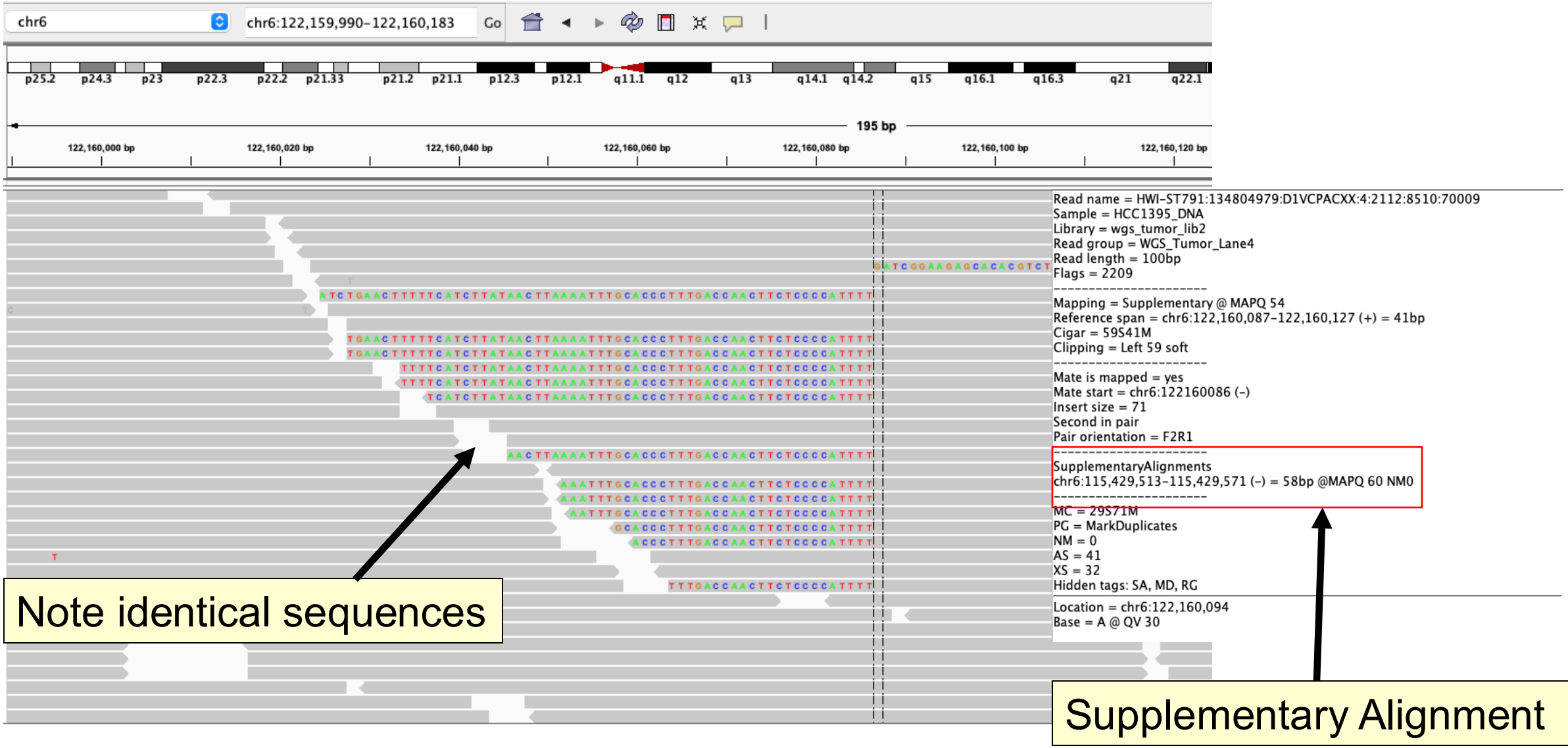
Inversion


IGV
File View Tracks Help
Go
Note drop in coverage at breakpoints
q21.1 q21.2 q21.3 q22.11 q22.12 q22.2 q22.3
26,296,200 bp 26,296,300 bp 26,296,400 bp 26,296,500 bp 26,296,600 bp 26,296,700 bp
NA12878 SLX (CEU daughter) Coverage
[0 - 60]
NA12878 SLX (CEU daughter)
Sequence
RefSeq genes
APP
chr21:26,296,750
330M of 450M

Soft-clipping



Soft-clipping



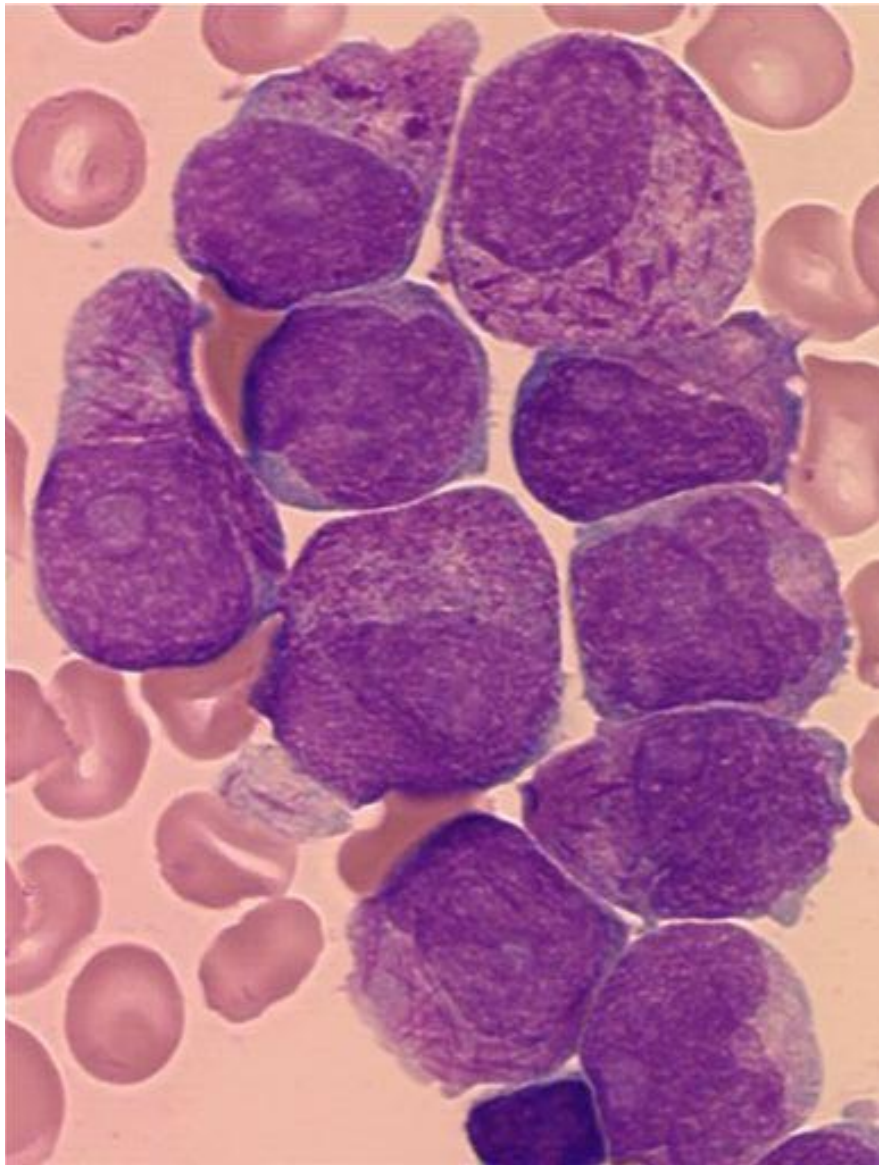
Note identical sequences

Supplementary Alignment

Assignment

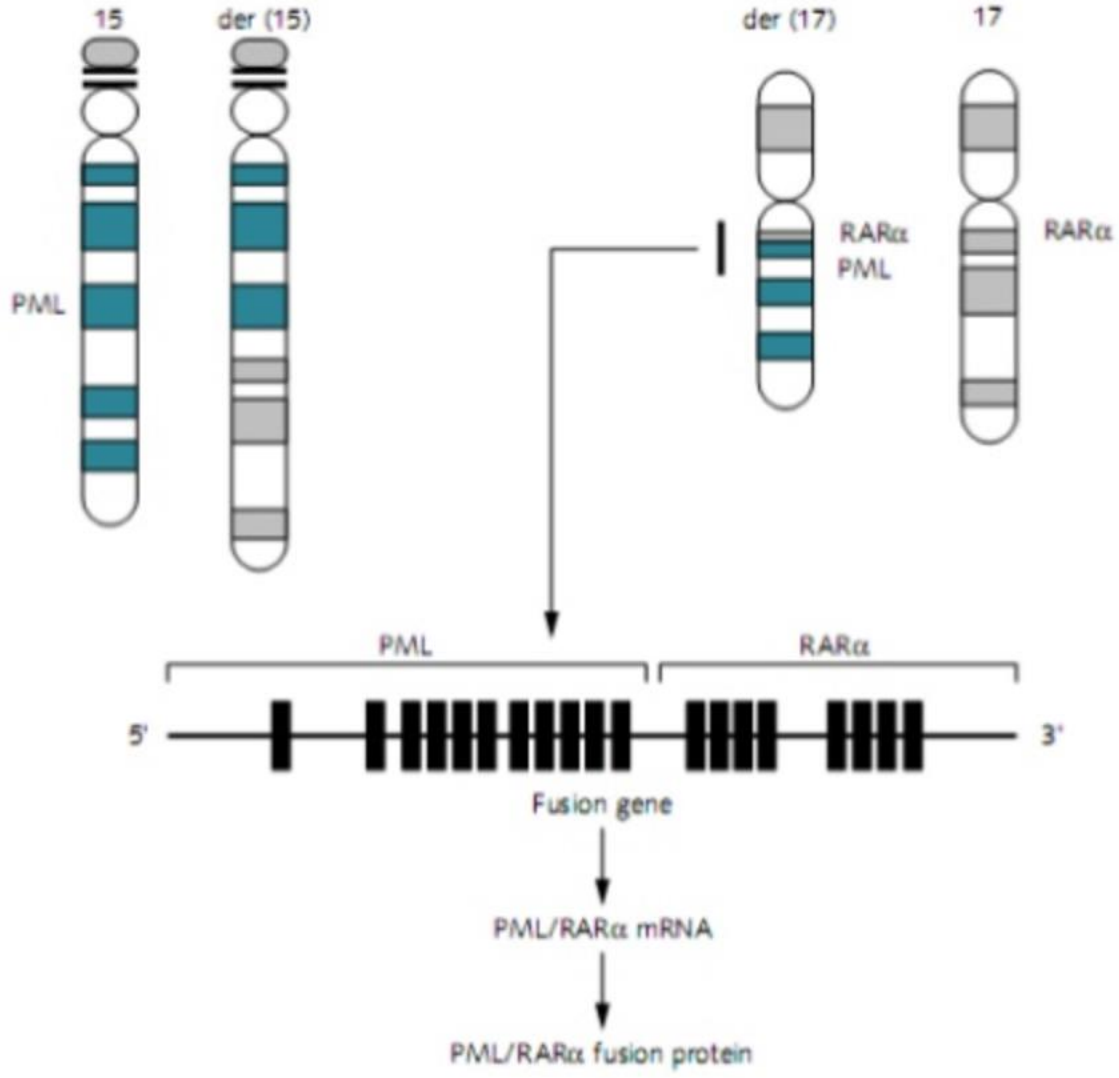
https://gist.github.com/chrisamiller/1150bcdb1a269b6c32d1f2a77dccb9aa

AML52: An atypical M3 AML

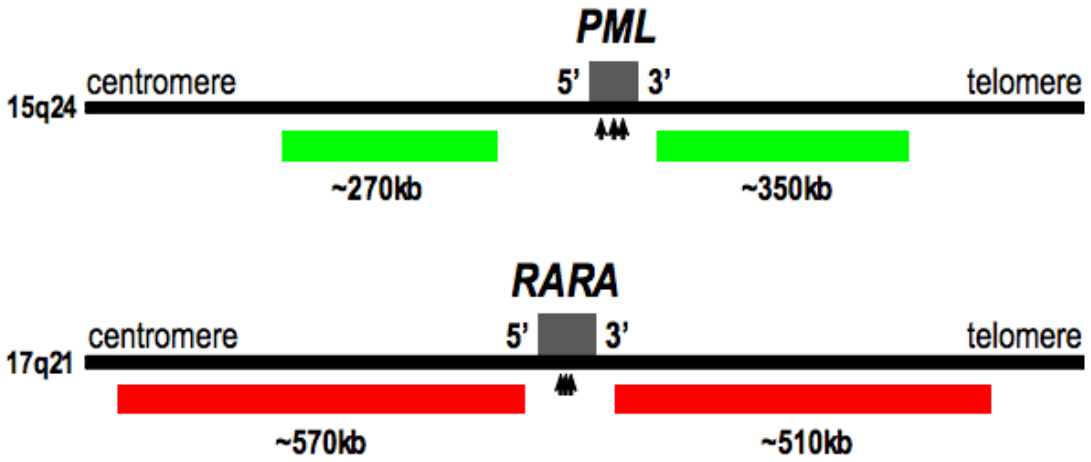


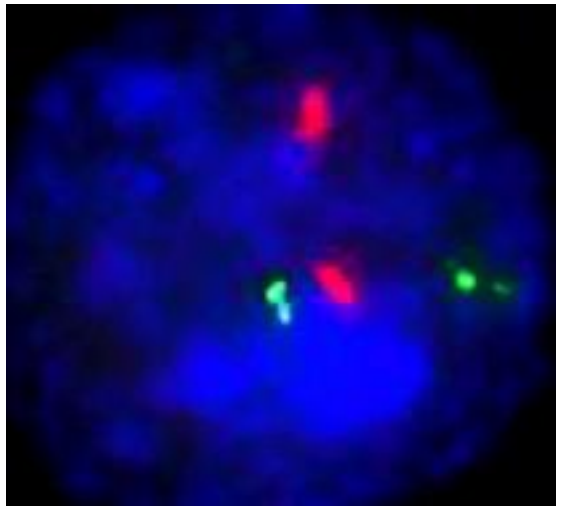
37 y.o. female with AML;
M3 morphology

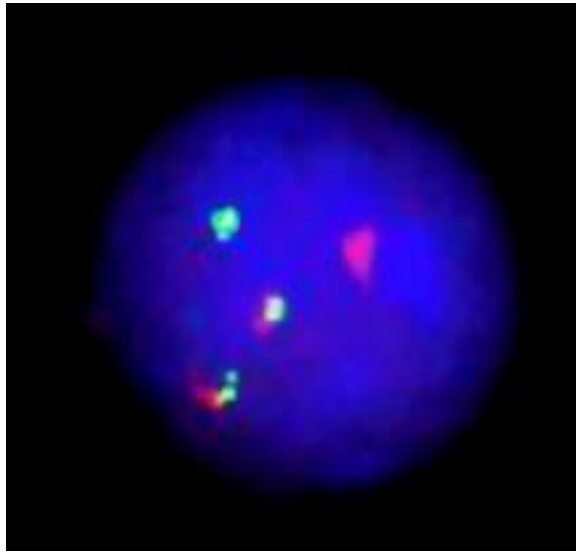
PML-RARA fusion



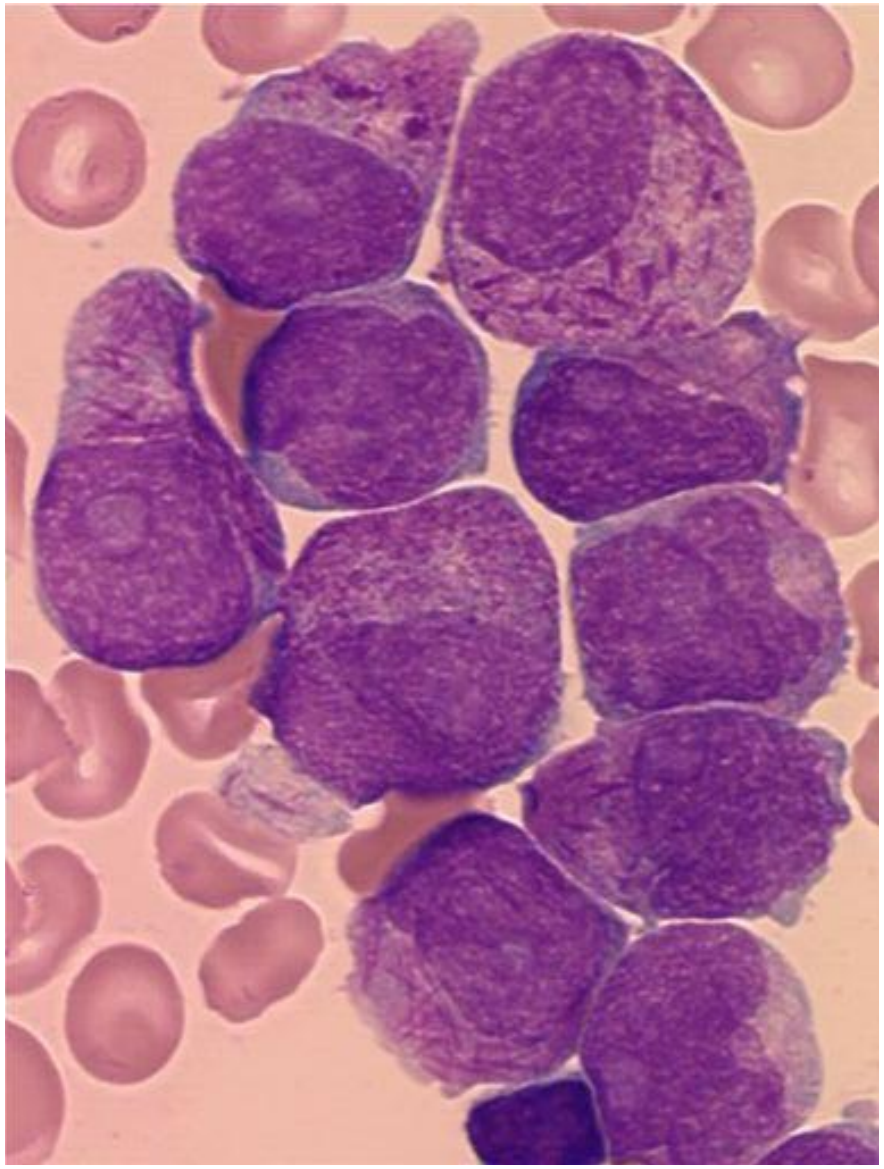
https://image.slidesharecdn.com/sinhaematology2012-121128014851-phpapp01/95/genetics-in-haematology2012-28-


PML
15q24
centromere
5'
3'
telomere
~270kb
~350kb
RARA
17q21
centromere
5'
3'
telomere
~570kb
~510kb





AML52: An atypical M3 AML

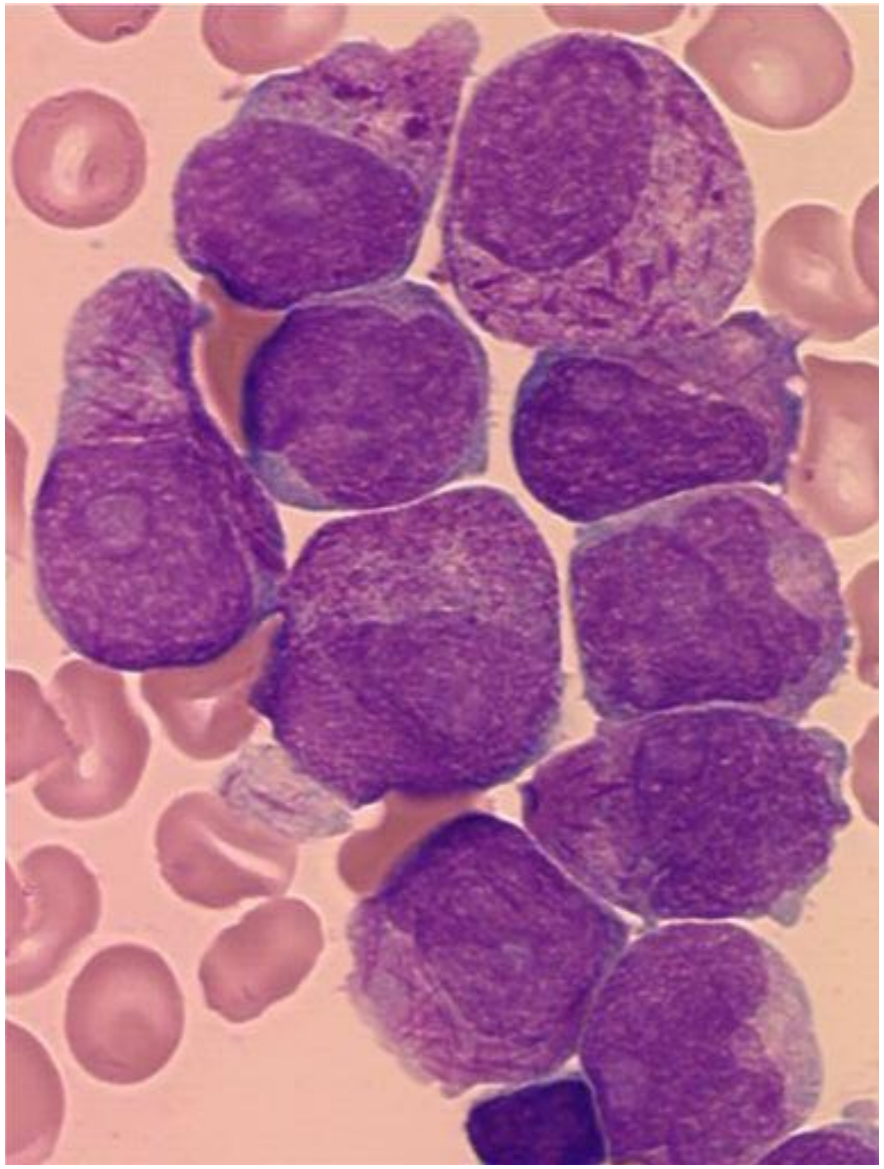


37 y.o. female with AML;
M3 morphology

↓

Chemo + ATRA

AML52: An atypical M3 AML



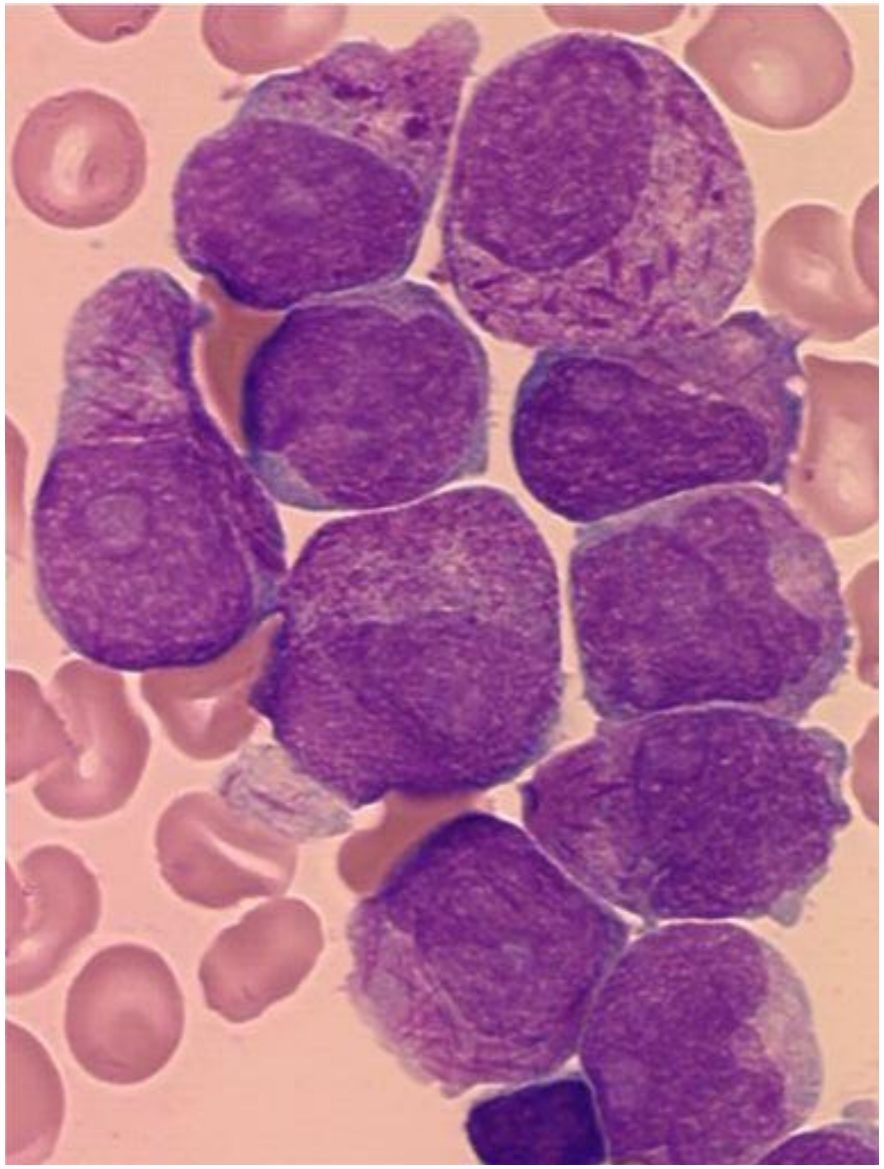
37 y.o. female with AML;
M3 morphology

↓ Chemo + ATRA

Complex cytogenetics,
negative for PML-RARA

↓ Chemo only

AML52: An atypical M3 AML



37 y.o. female with AML;
M3 morphology

Chemo + ATRA

Complex cytogenetics,
negative for PML-RARA

Chemo only

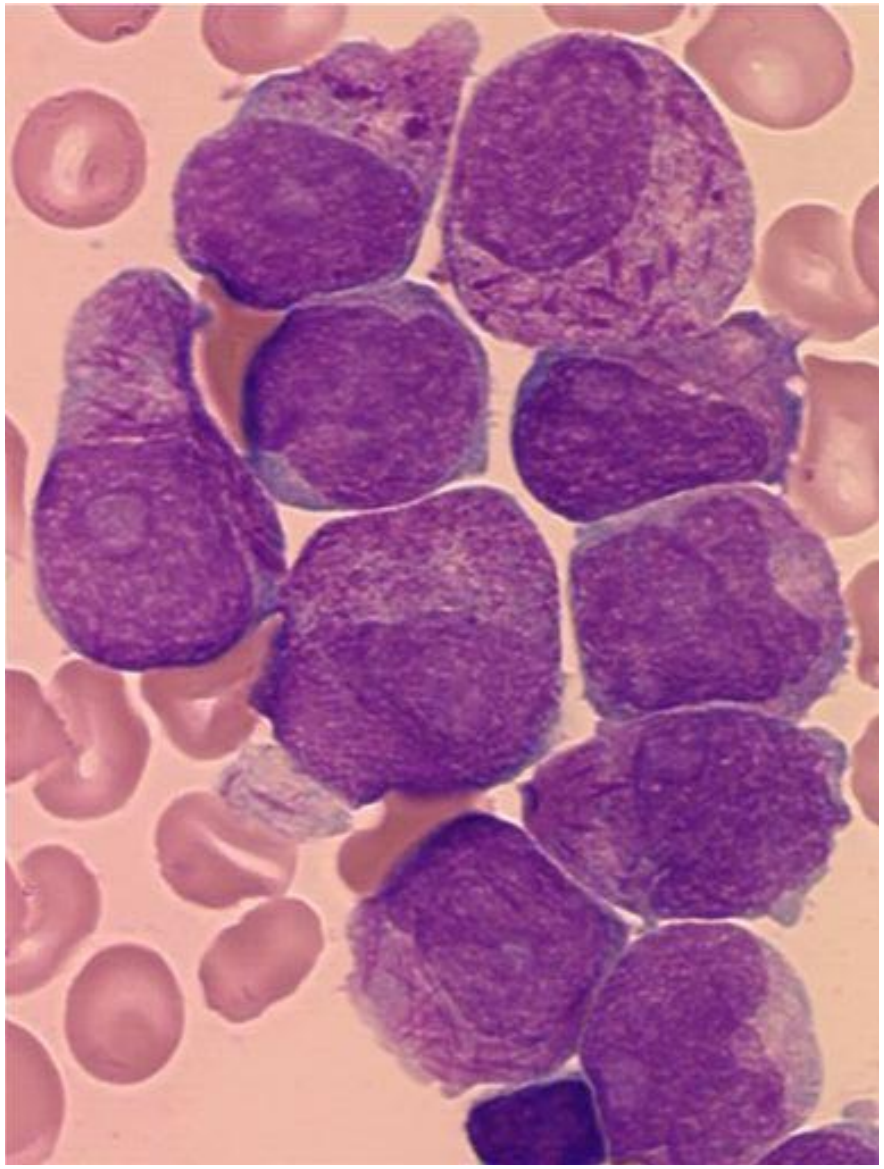
Achieved remission, but faced decision
Referred to WUSTL

???

Allogeneic
SCT

Consolidation
+ ATRA

AML52: An atypical M3 AML



37 y.o. female with AML;
M3 morphology

Chemo + ATRA

Complex cytogenetics,
negative for PML-RARA

Chemo only

Achieved remission, but faced decision
Referred to WUSTL

???

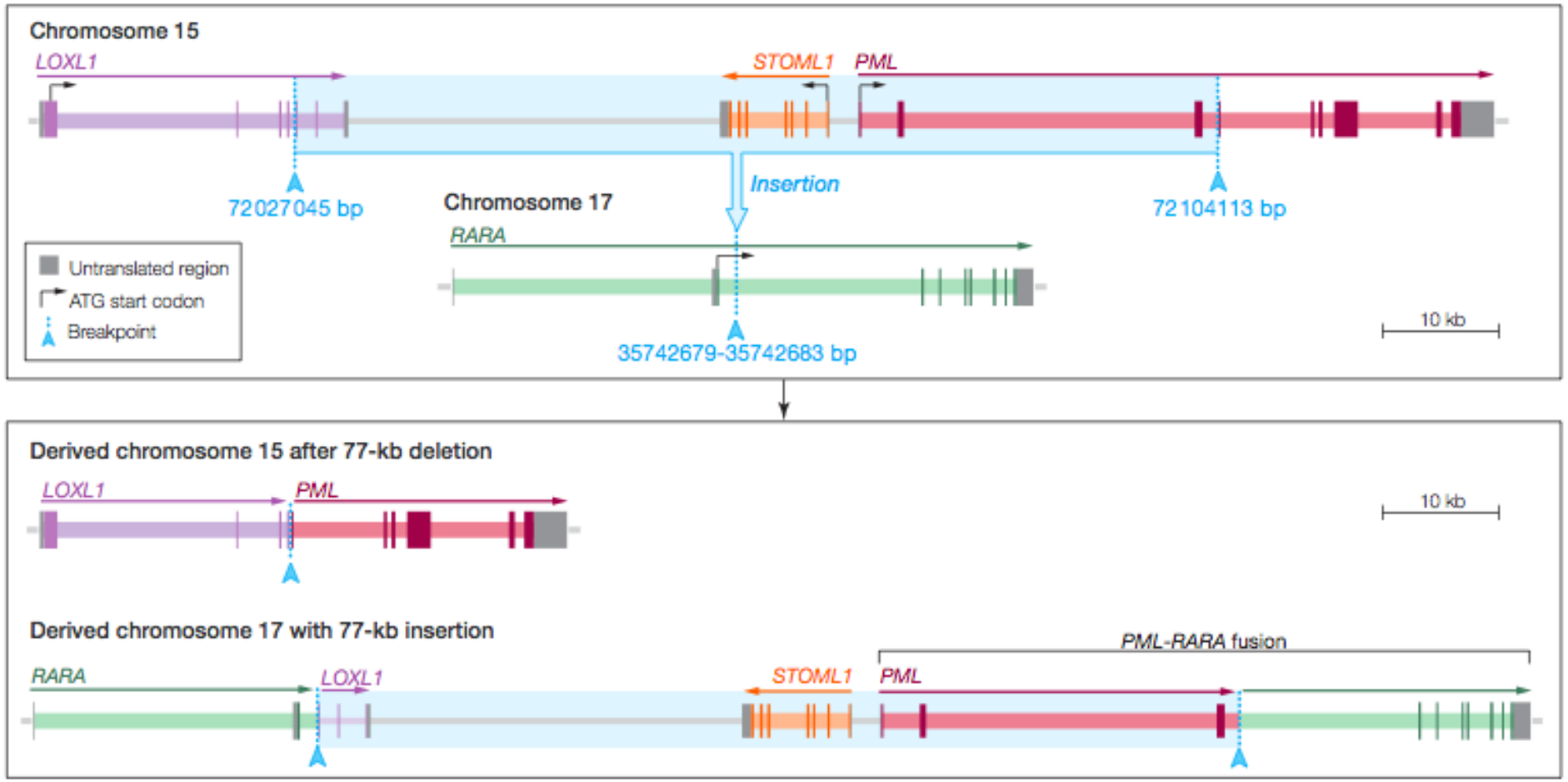
Allogeneic
SCT

Consolidation
+ ATRA

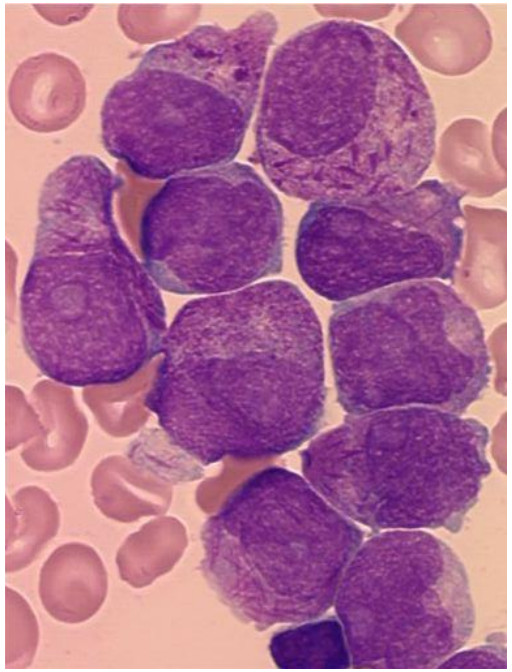
JAMA
The Journal of the American Medical Association
JAMA. 2011;305(15):1577-1584. doi: 10.1001/jama.2011.497

Use of Whole-Genome Sequencing to Diagnose a Cryptic Fusion Oncogene

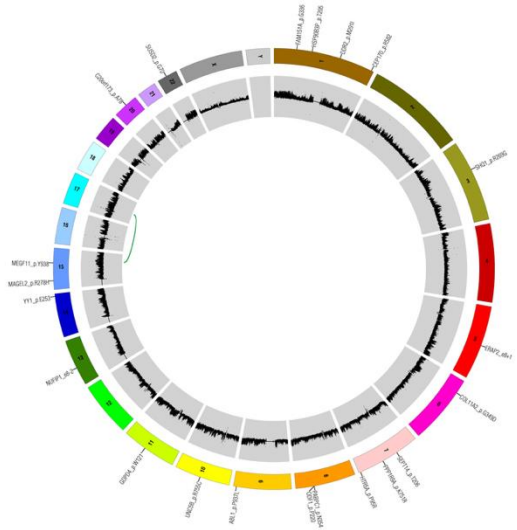
A Breakpoints in chromosomes 15 and 17 resulting in *PML-RARA* fusion



AML52: An atypical M3 AML



37 y.o. female with *de novo* AML, M3 morphology, CTG, no PML-RARA. Referred to WUSM for SCT.



Detection of PML-RARA fusion by WGS. Confirmed by FISH, RT-PCR

Consolidation: Chemo + ATRA

Sustained remission

Additional cryptic M3 AMLs