AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

Bioinformatics and Genome Analyses Course

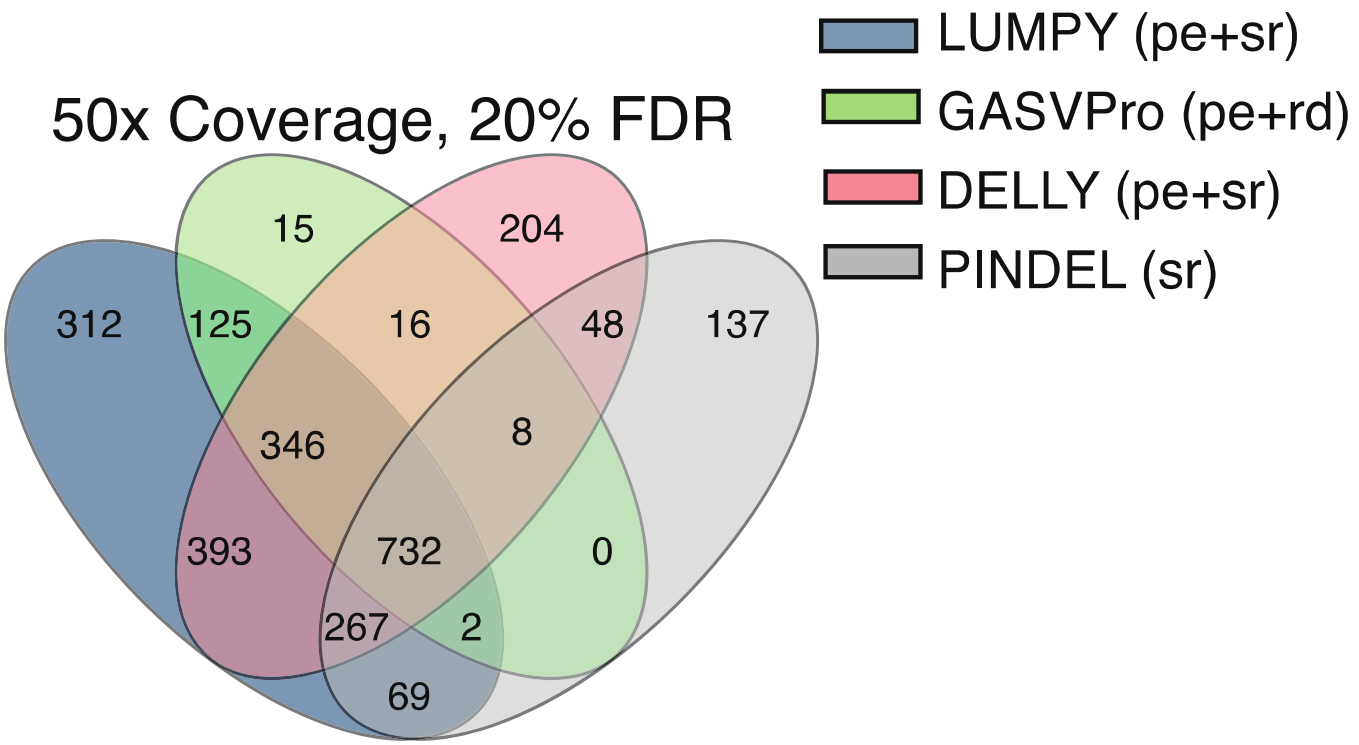
SV Calling

Stephan Pabinger
stephan.pabinger@ait.ac.at

http://pabinger.site44.com
Slides curtsy of Fritz Sedlazeck
https://www.nature.com/articles/s41592-018-0001-7

Challenges: Pursuing the diploid genome

Accurate prediction of SVs ?

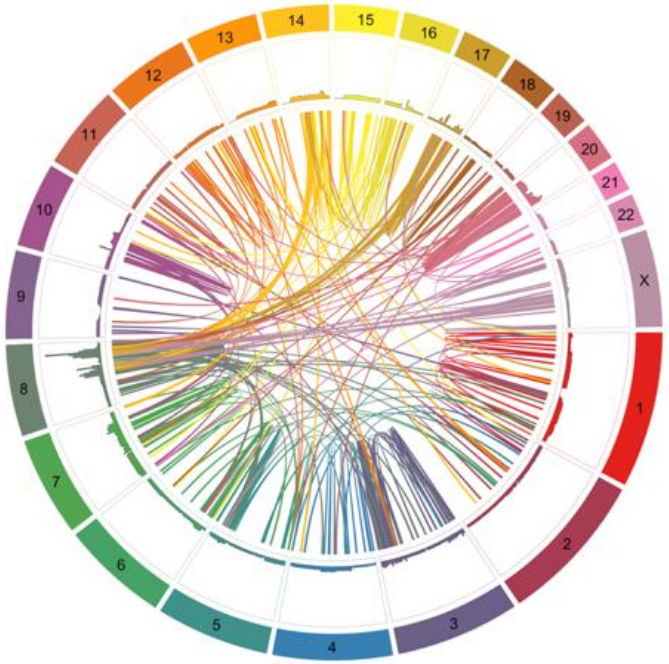


→ **Use long read sequencing**

Layer et.al. (2014)

How can we fully leverage long reads?

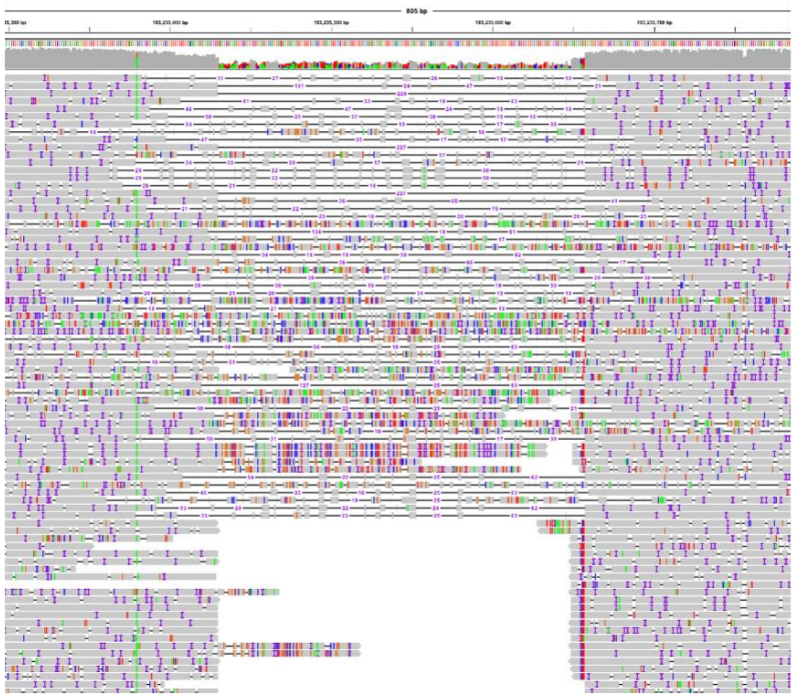
1. Improvement in mapping (NextGenMap-LR)
2. Improvement in SV calling (Sniffles)
3. Improvements over real data



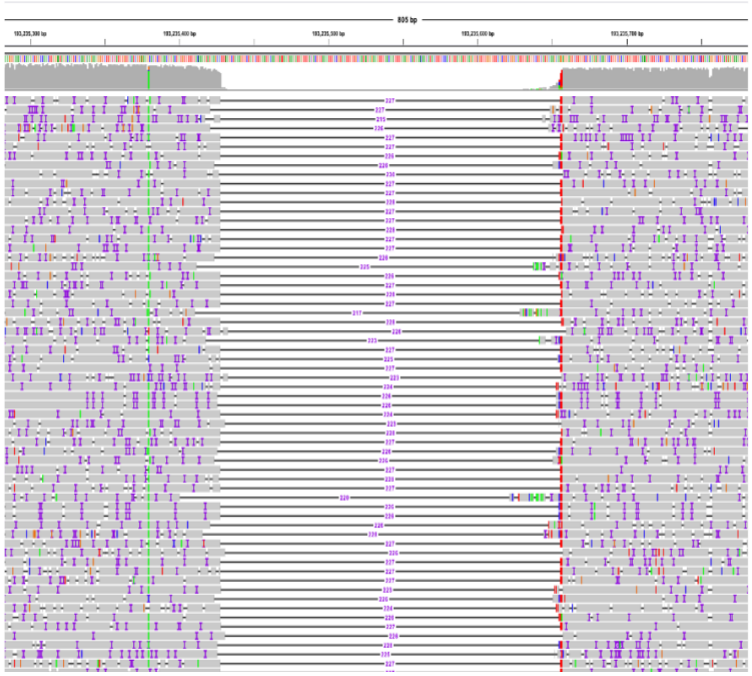
Why another mapper?

AUSTRIAN INSTITUTE OF TECHNOLOGY

BWA-MEM:



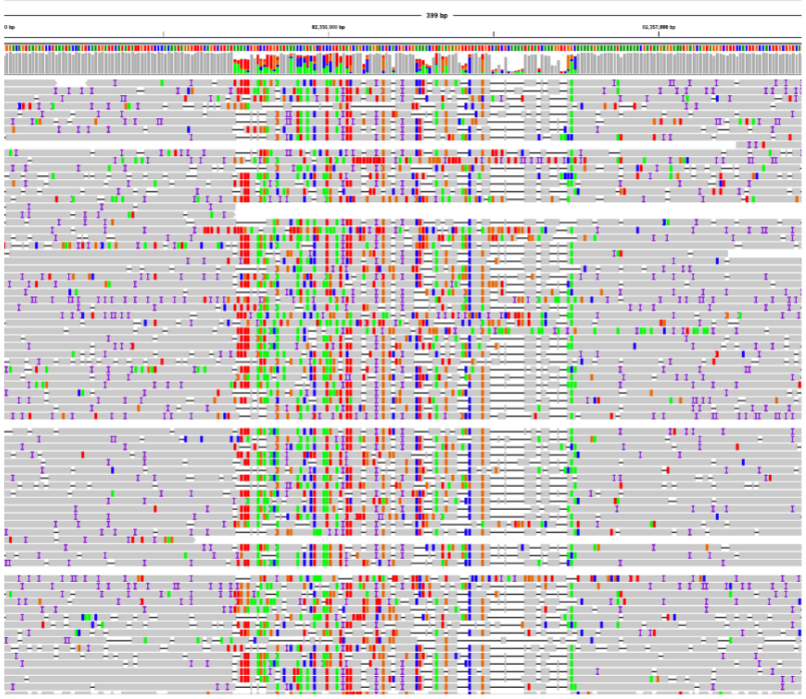
NGMLR:



Why another mapper?

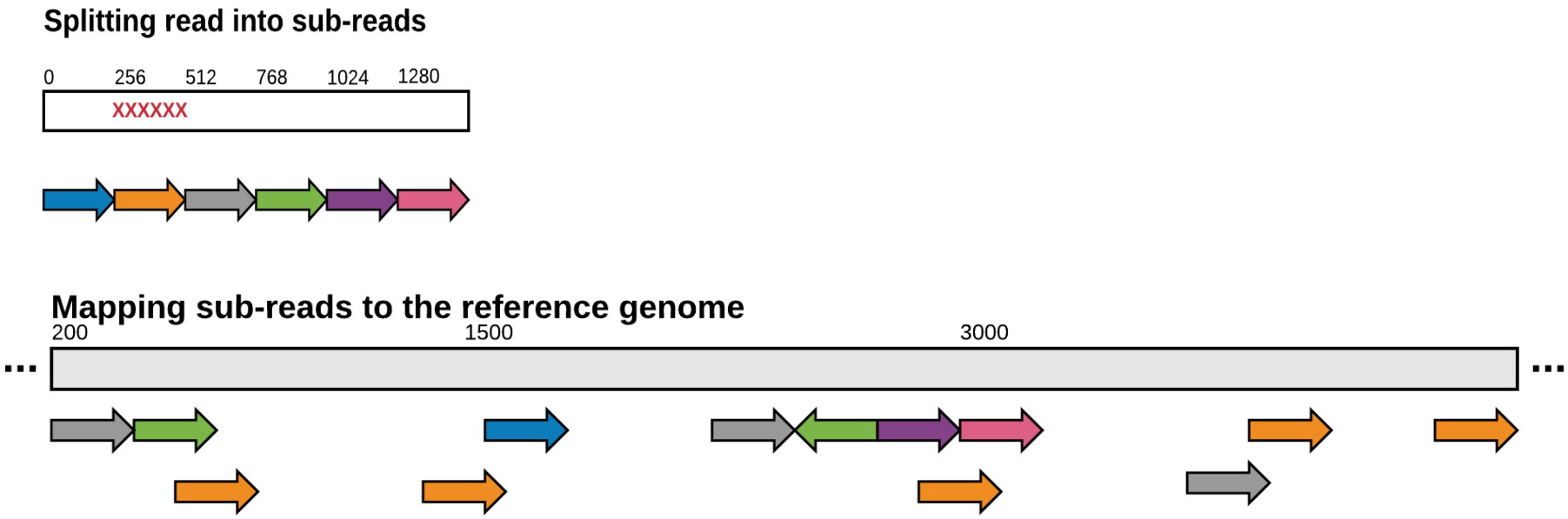
AUSTRIAN INSTITUTE OF TECHNOLOGY

BWA-MEM:

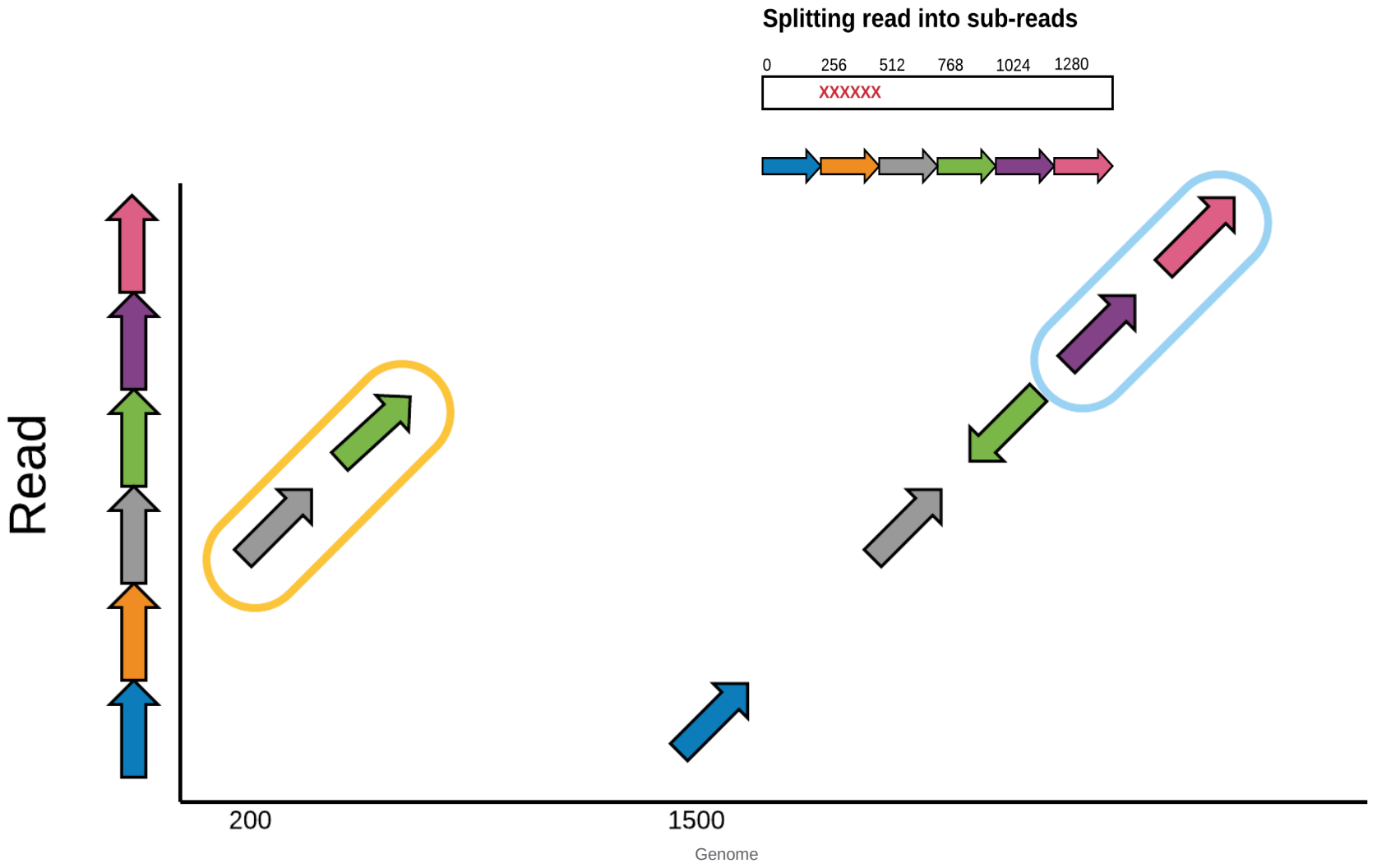


NGMLR:

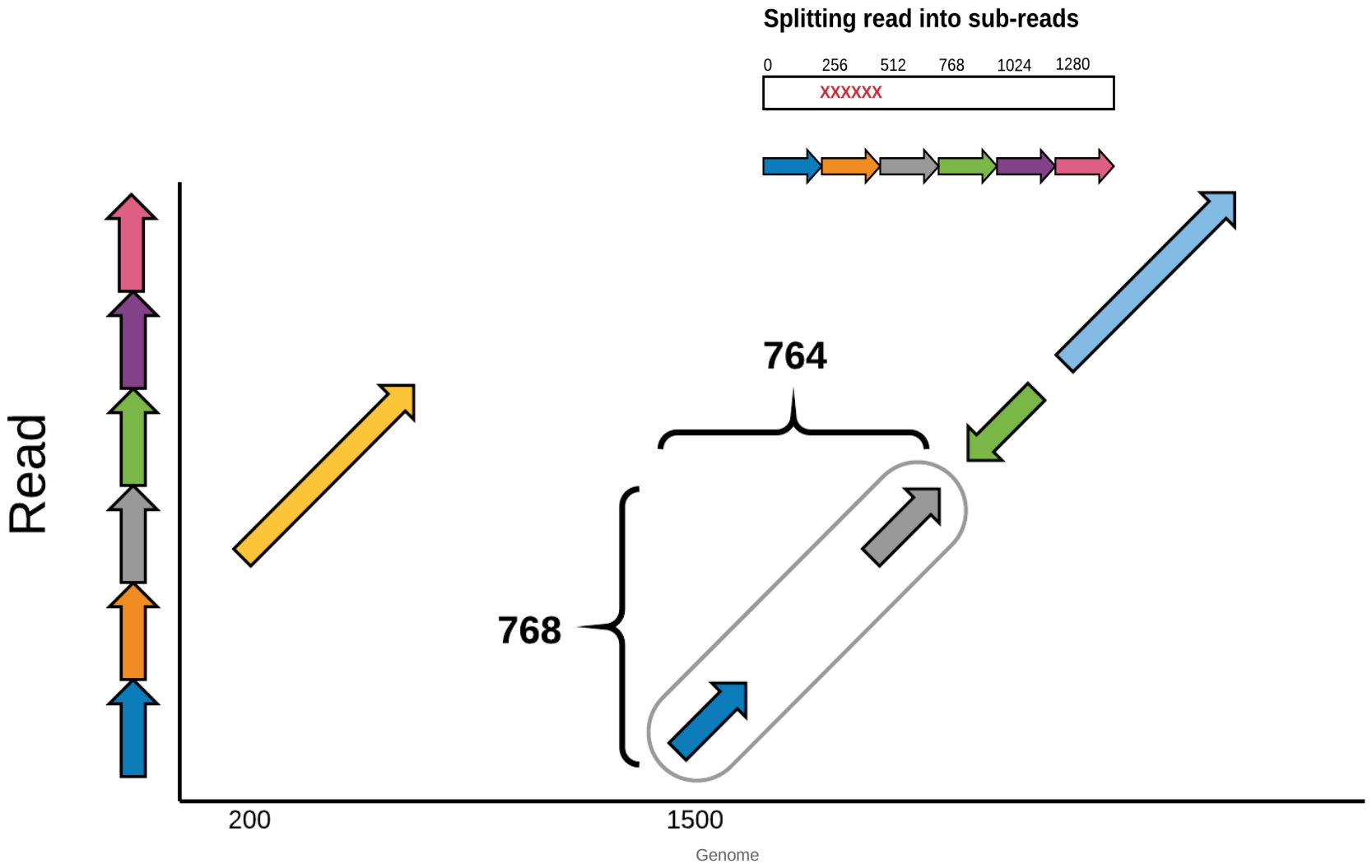
NGMLR: Split reads



NGMLR: Split reads

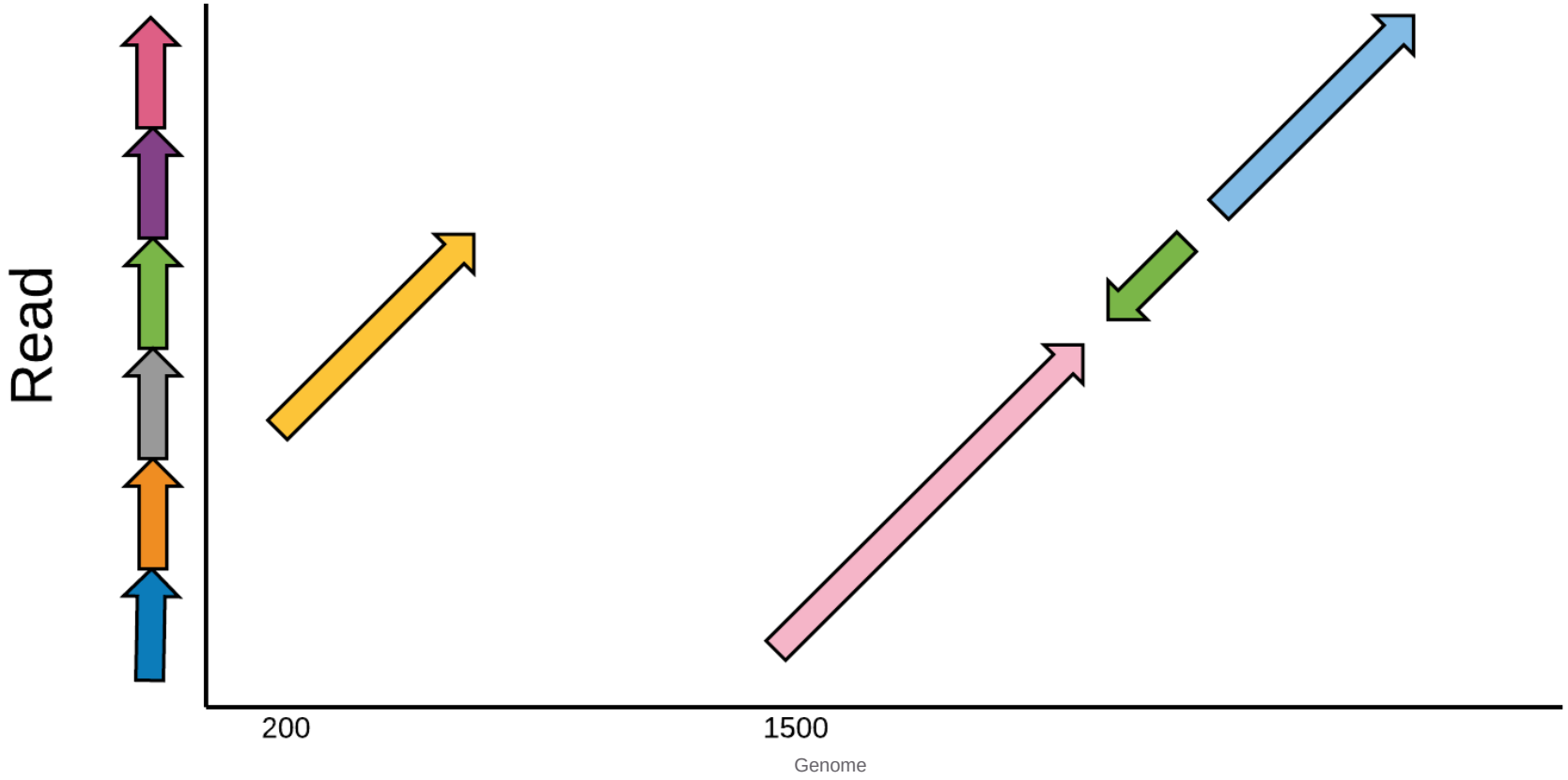


NGMLR: Split reads



NGMLR: Split reads

With the convex scoring model of NGMLR, extension of an indel is penalized proportionally less the longer the indel is



Sniffles – SV variant caller

AUSTRIAN INSTITUTE OF TECHNOLOGY

- Detecting all SV types
- Parameter estimation
- Detect sequencing artifacts

- Optional:
  - Genotype estimation
  - Clustering/phasing of SVs

Sniffles

- Estimates the parameters to adapt itself to the underlying dataset, such as the distribution in alignment scores and distances between indels and mismatches on the read, as well as the ratios of the best and second-best alignment scores

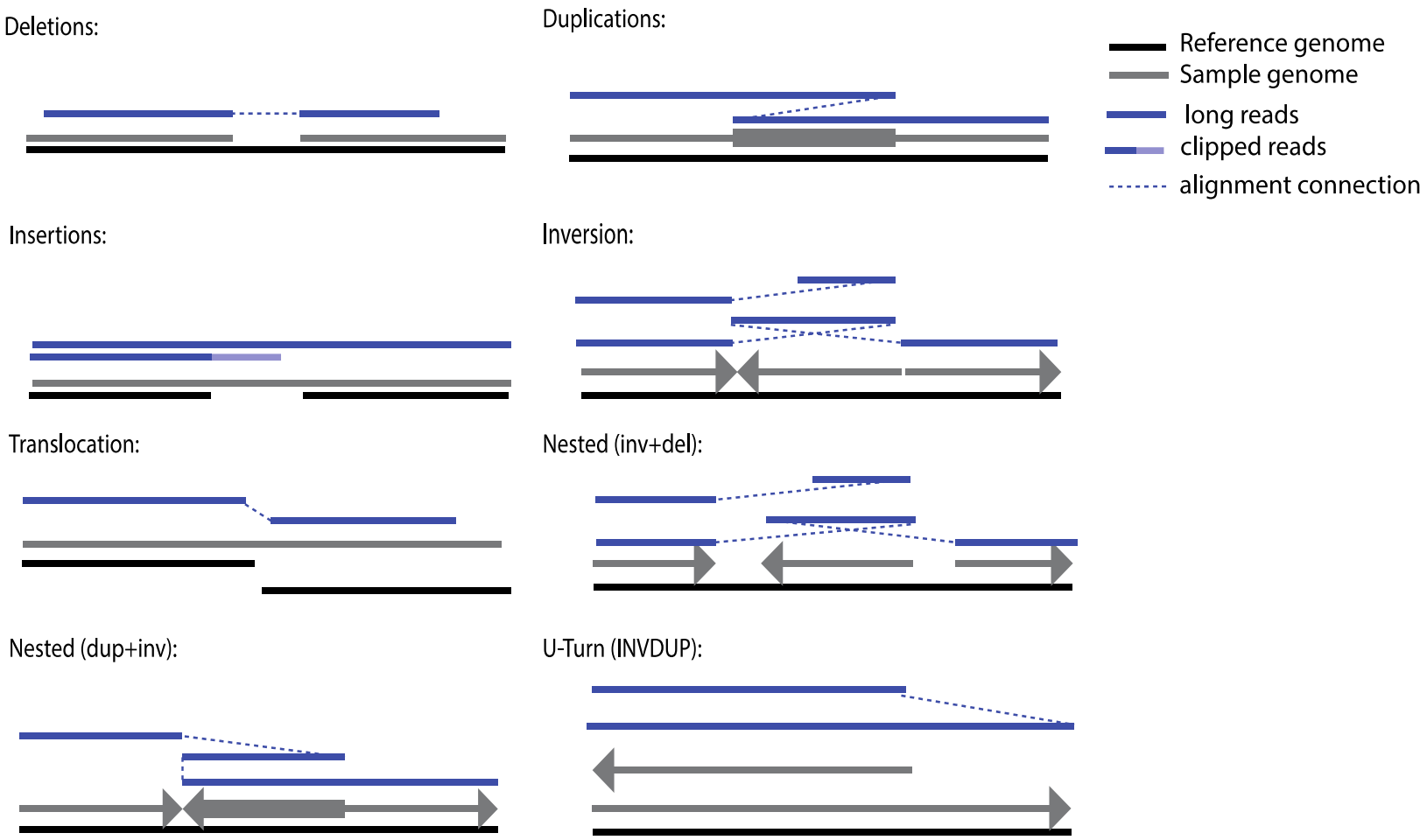
- Sniffles then scans for potential SVs

- Putative SVs are clustered and scored according to the number of supporting reads, the type and length of the SV, the consistency of the SV composition, and other features

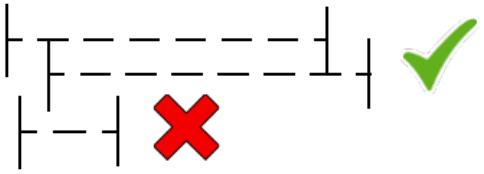
- Genotypes the variant calls → homozygous and heterozygous SVs

Sniffles: Detection of SVs

Split the reads



Sniffles: Clustering of SVs



When are two events the same?

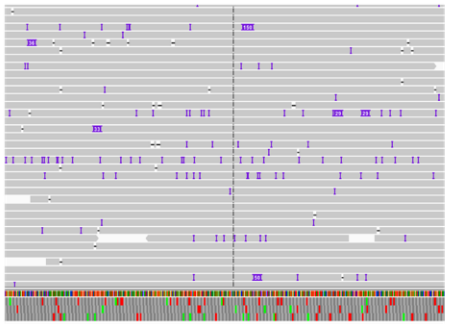
- Allowed distance depending on the size of the event

Detecting clustering of noise?

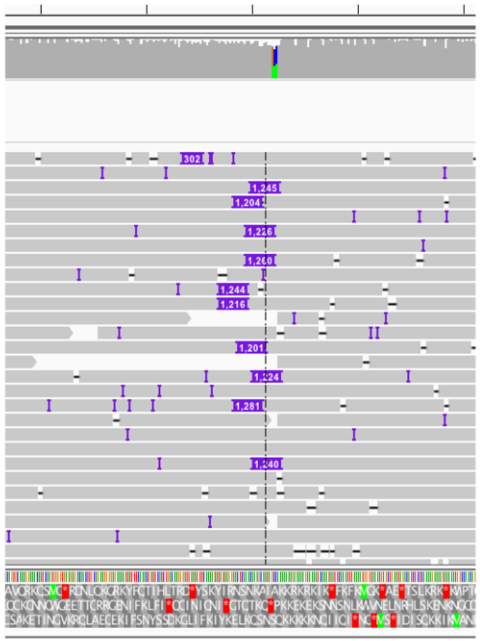
- Random appearance of Insertions (5-100bp)

- Standard deviation
  - Higher noise -> more likely artifact

Phantom insertion events:



Scattered events:



NA12878

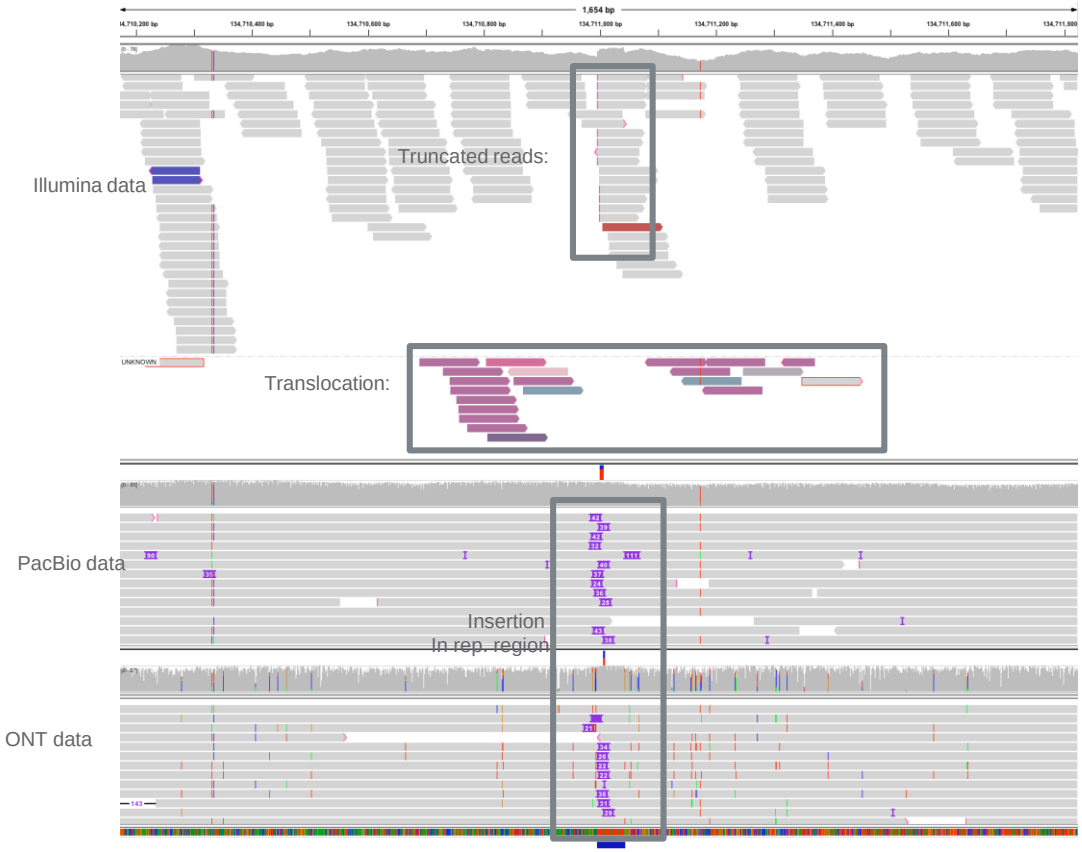
- Healthy female
- Gold standard in genomics
- Sequenced with many technologies independently:
  - Illumina, PacBio, Oxford Nanopore

3.2 NA12878: Deletion calling

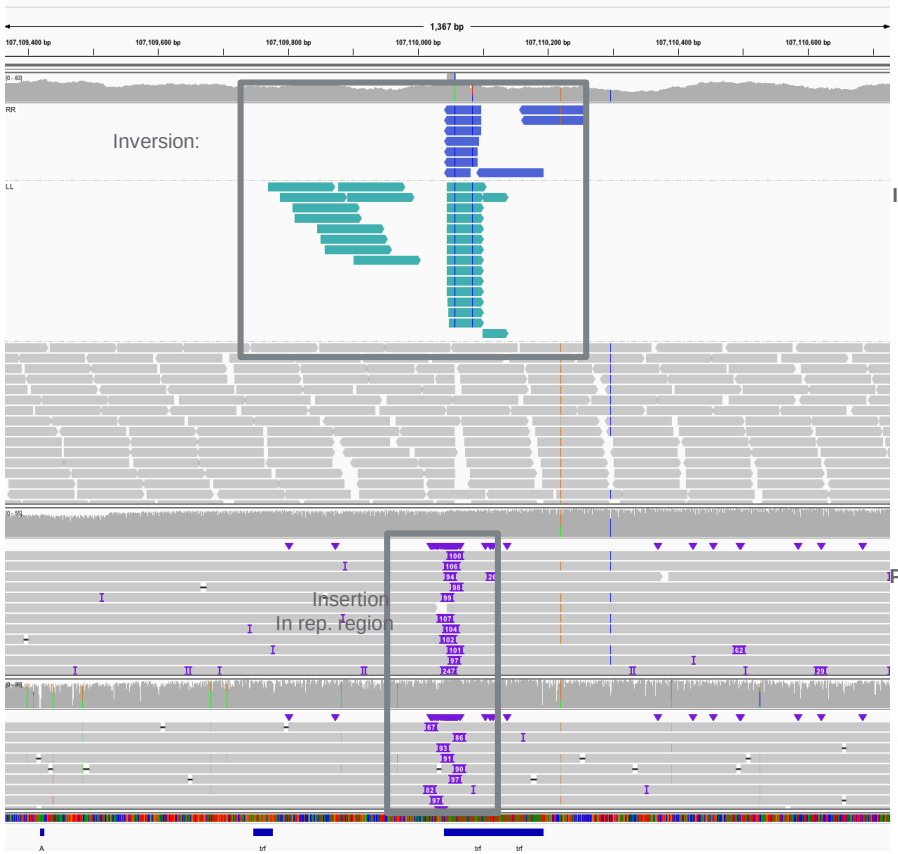
Tech.	Cov.	Avg len	SVs	DEL	DUP	INV	INS	TRA
PacBio	55x	4,334	22,877	9,933	162	611	12,052	119
Oxford Nanopore	28x	6,432	32,409	**27,147**	87	323	4,809	43
Illumina	50x	2 x 101	7,275	3,744	731	553	0	2,247

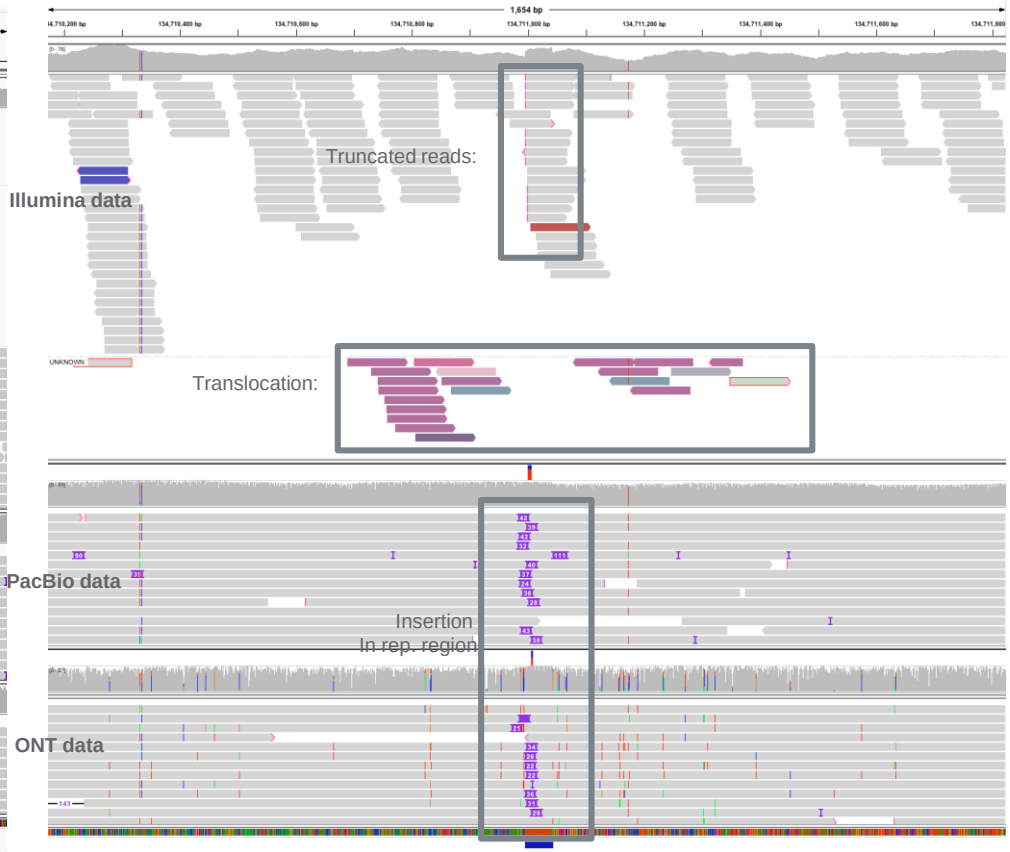
NA12878: check 2,247 vs 119 TRA

Overlap	Illumina TRA(%)
Translocations	7.74
Insertions	53.05
Deletions	12.06
Duplications	0.57
Nested	0.31
High coverage	1.87
Low complexity	9.79
Explained	**85.40**



NA12878: check 2,247 vs 119 TRA





Practical