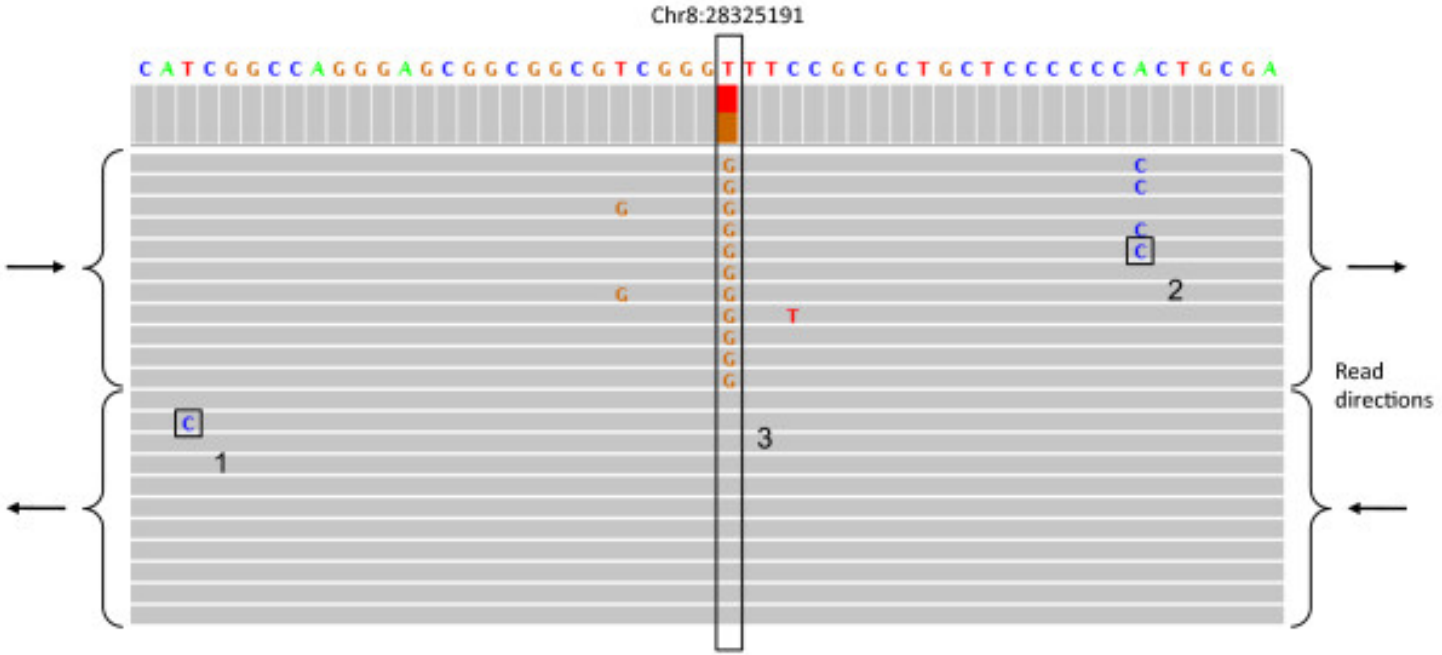
Variant calling a structural variation detection with short and long reads

23.3.2021

Monika Čechová

@biomonika
biomonika.org

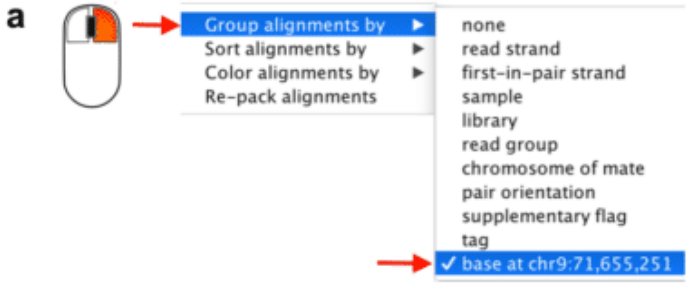
Recognizing sequencing errors from true variants

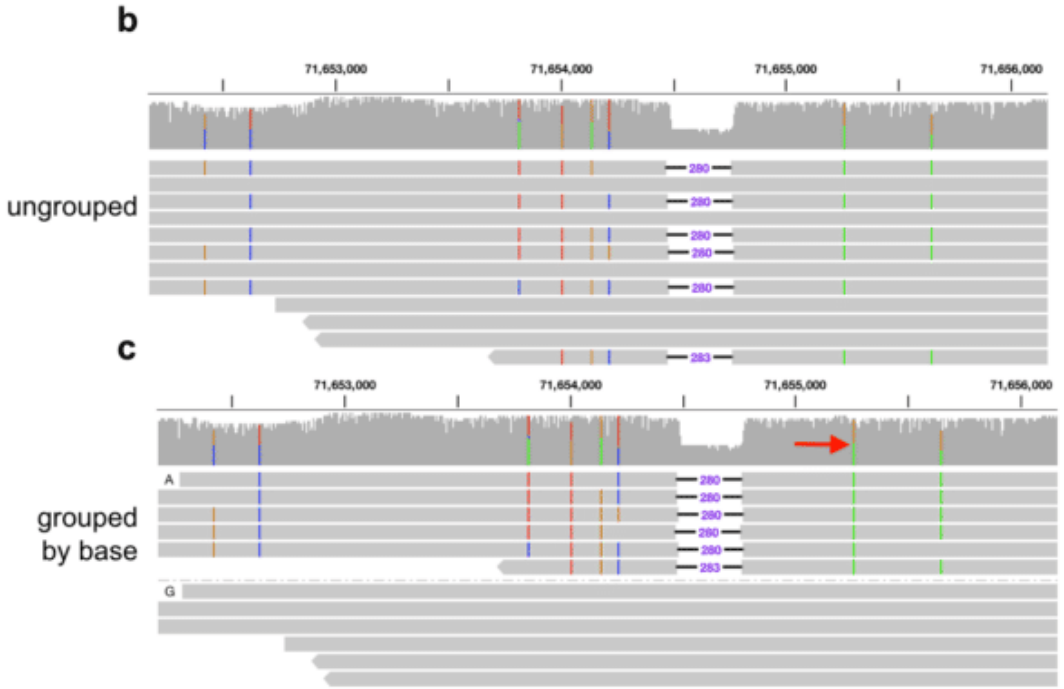


If possible, validate with an independent dataset.

https://slideplayer.com/slide/4393651/

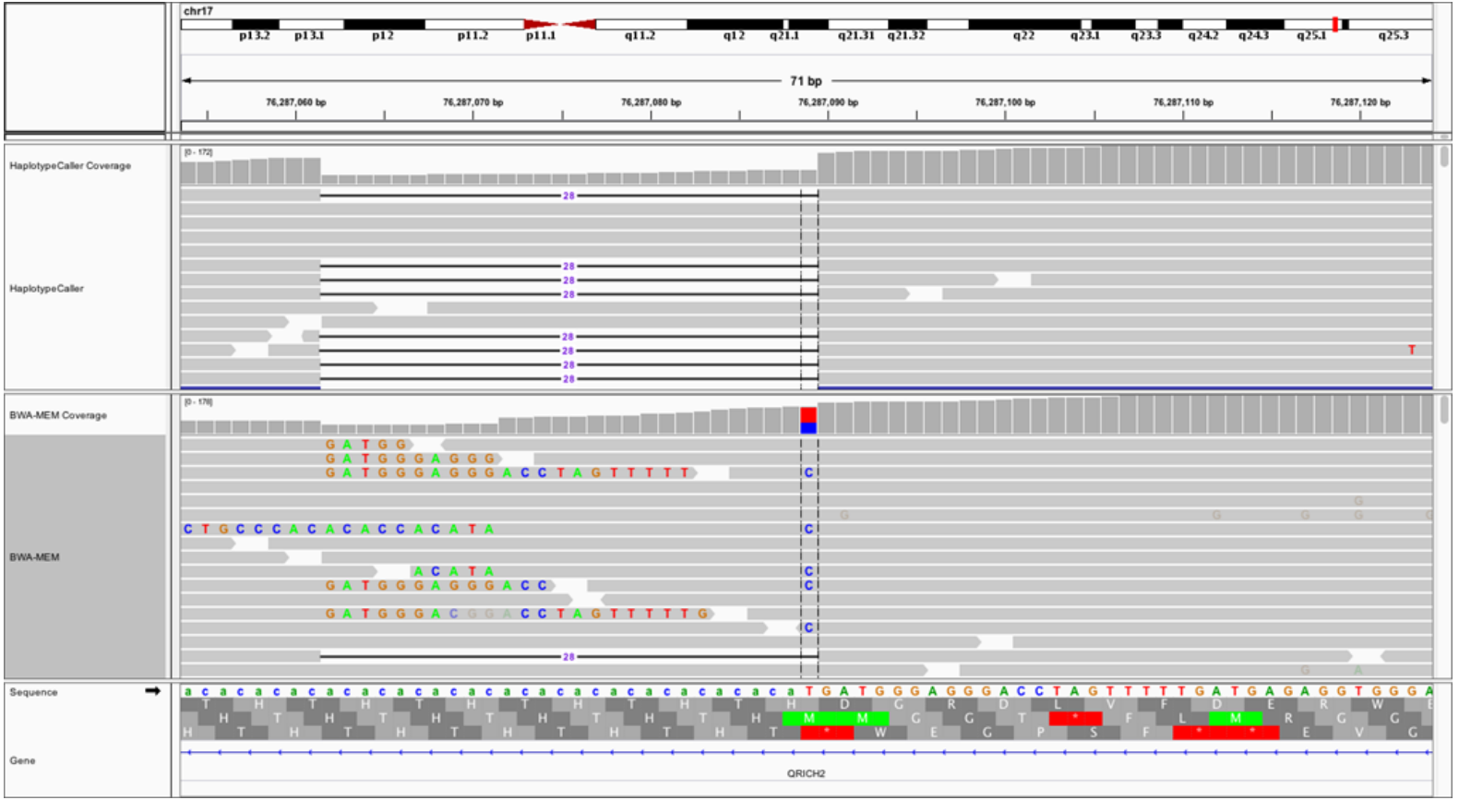
Grouping



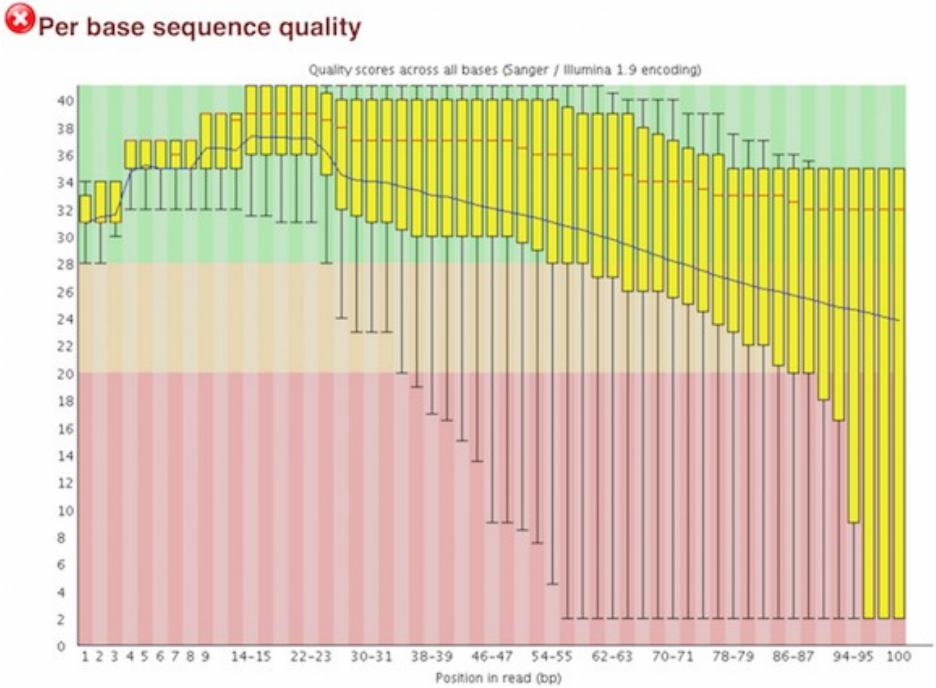


https://www.pacb.com/blog/igv-3-improves-support-pacbio-long-reads/

Indel re-alignment



Soft-clipping versus hard-clipping



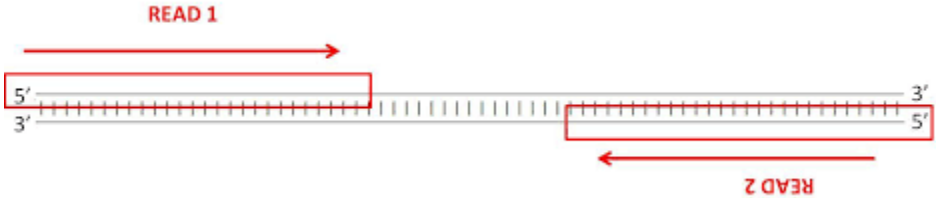


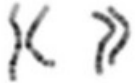
Image from: http://www.cureffi.org/2012/12/19/forward-and-reverse-reads-in-paired-end-sequencing/

https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/02_assessing_quality.html

Family trio

https://twitter.com/infoecho/status/1369518026257698822

MORE COMPLETE VARIANT DETECTION EXPLAINS MORE RARE DISEASE CASES



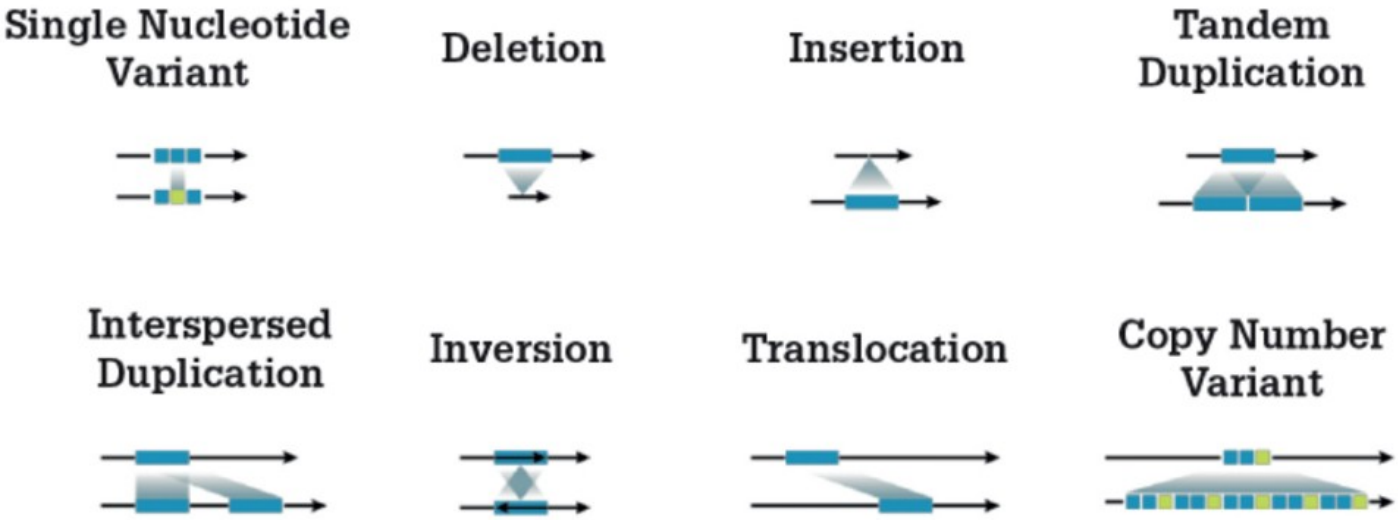

Karyotyping	Microarrays	Short-read Sequencing		Long-read Sequencing
		Exome	Genome	HiFi Genome
Chromosomal abnormalities	Copy-number variants >50kb	SNVs & indels, some large exonic variants	SNVs, indels, some large variants	**SNVs, indels, SVs, CNVs, phasing, translocations, inversions, repeat expansions**
~5% explanation rate	~10%	~30%	~40%	**up to 67%**
Phelan Proc. of Greenwood Genetics Center 1996	De Vries AJHG 2008	De Ligt NEJM 2012	Gilissen Nature 2014	

Recognizing sequencing errors from true variants

- Sequencing coverage (is it uneven? Is there a drop? are stacked read endings present?)
- Mapping quality / multi-mapping reads
- Base quality (Phred score)
- Where is the variant located within the read? (indel realignment)
- Phasing/haplotype informations
- The sequence context (transposable elements/satellite repeats/multi-copy genes)
- Expected allele/variant frequency (ploidy, copy number for the “collapsed regions”)
- PCR amplification error (consider strands!)
- Somatic mutations

Simulations can help estimate how ‘likely’ are ‘unlikely’ events.

Structural variants



https://thepathologist.com/diagnostics/smrt-long-read-sequencing-solves-genetic-mysteries

SV detection in IGV

Variant detection

Manually find the following variants using IGV and Illumina reads:

- SNP 1bp
- ins 1bp
- del 1bp
- ins 50bp
- del 50bp
- ins 250bp
- translocation 1000bp

Please share your results with everyone in the chat window.

Variant detection

Now additionally load PacBio and Nanopore track. Discuss what you see.

SAM FORMAT

https://genome.sph.umich.edu/wiki/SAM

BAM + CRAM

SV detection

- Sniffles
  https://github.com/fritzsedlazeck/Sniffles
- SVanalyzer
  https://svanalyzer.readthedocs.io/en/latest/
- Pepper
  https://github.com/kishwarshafin/pepper

Homework (3 points)

Compare variants called with freebayes and samtools (e.g. using a venn diagram). You can use vcf-compare from vcftools or any other tool for this comparison.

Now set the parameters of freebayes in a way that makes you confident in your results. Discuss the results in the 3 generated VCF files.

We will start the next lecture by each group presenting their results.