



CEITEC

Central European Institute of Technology
BRNO | CZECH REPUBLIC

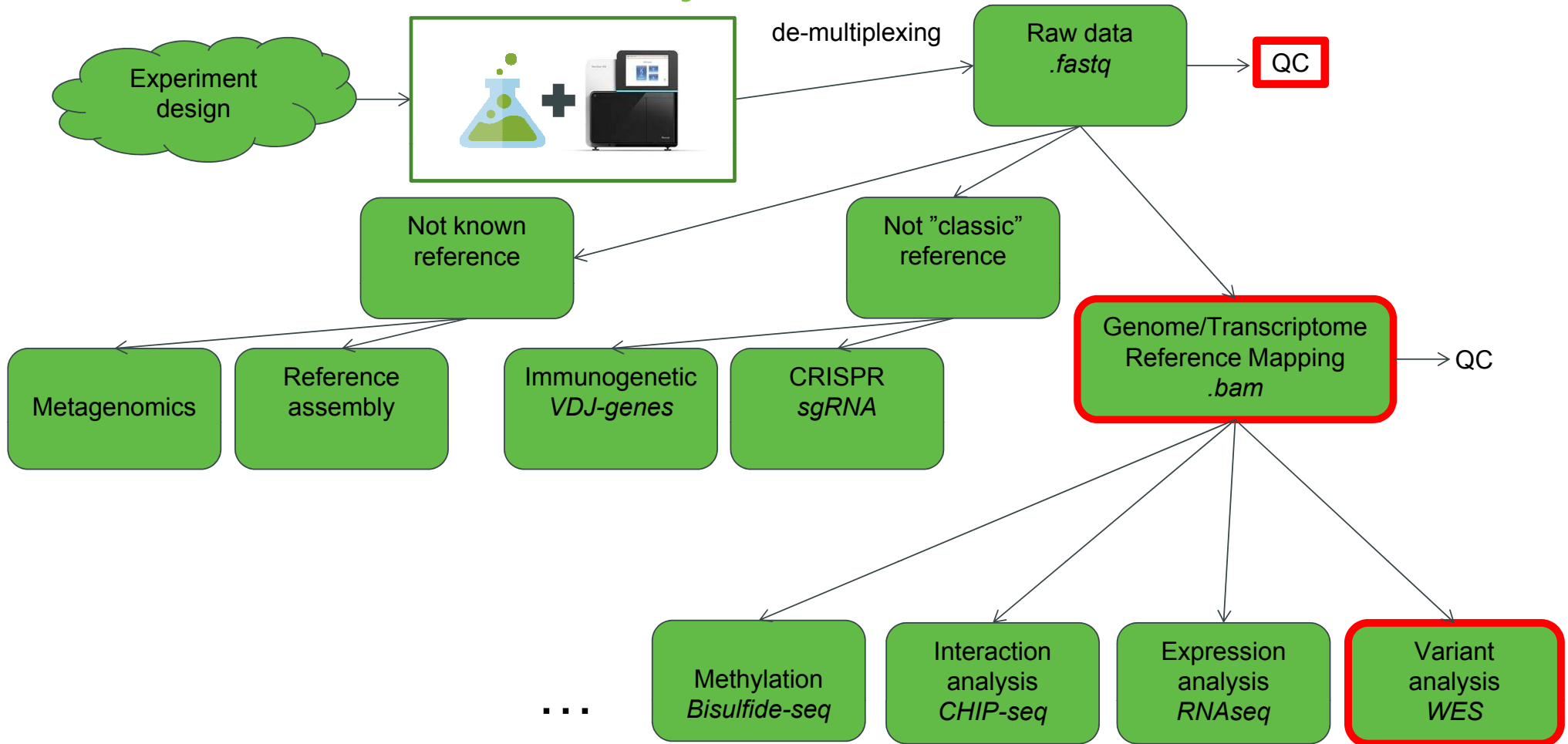


Modern Genomic Technologies (LF:DSMGT01)

Lecture 2 : DNA re-sequencing

Vojta Bystry
vojtech.bystry@ceitec.muni.cz

NGS data analysis



DNA re-sequencing

- Variant Calling
- Medical purposes
- Cancer genomics

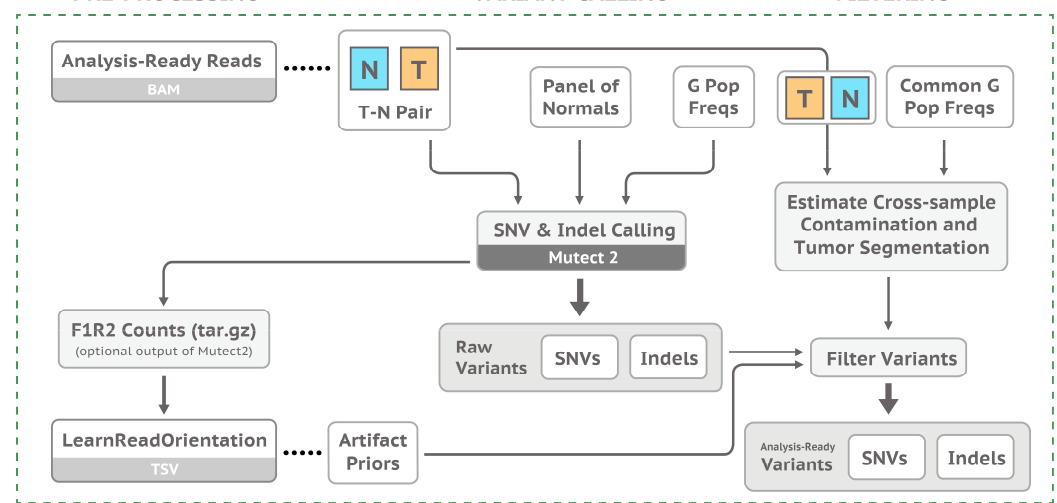
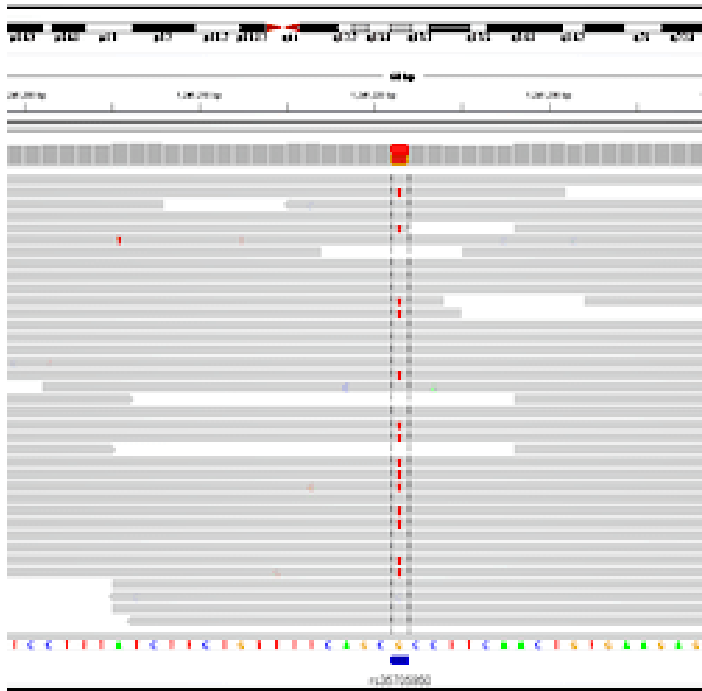
- Small variants (SNV + small indels) vs. Structural Variants

- Germline vs. Somatic

Mapping

- Computationally most demanding
- More or less standardized
- Output .bam
 - .bam = binary (zipped) .sam
 - .sam = Sequence Alignment Map DNA re-sequencing
- Tools
 - BWA - DNA
 - STAR - RNA

Small Variant calling



Mapping QC

General Statistics

[Copy table](#)
[Configure Columns](#)
[Plot](#)
 Showing 12/12 rows and 16/24 columns.

K Reads Mapped	% GC	Ins. size	≥ 100X	≥ 500X	≥ 20X	≥ 30X	Median cov	Mean cov	% Aligned	Fold Enrichment	Target Bases 30X	% Dups	% Dups	% GC	K Seqs
100 827.9	48%	176	43.3%	0.8%	93.2%	88.7%	89.0X	111.8X	99.6%	43	83%				
Dups												4.7%			
													26.8%	47%	50 603.8
													25.4%	47%	50 603.8
100 523.1	48%	178	42.8%	0.8%	93.2%	88.8%	88.0X	111.2X	99.6%	43	84%				
Dups												4.6%			
													26.7%	47%	50 460.3
													25.5%	47%	50 460.3
84 081.9	48%	172	33.7%	0.5%	92.1%	86.4%	75.0X	94.4X	99.6%	44	80%				
Dups												4.5%			
													24.4%	47%	42 202.7
													23.3%	47%	42 202.7

Mapping QC

Qualimap Report: BAM QualiMap

Summary

Globals

Reference size	3,101,804,739
Number of reads	84,405,388
Mapped reads	84,038,132 / 99.56%
Unmapped reads	367,256 / 0.44%
Mapped paired reads	84,038,132 / 99.56%
Mapped reads, first in pair	42,129,277 / 49.91%
Mapped reads, second in pair	41,908,855 / 49.65%
Mapped reads, both in pair	83,774,794 / 99.25%
Mapped reads, singletons	263,338 / 0.31%
Secondary alignments	0

Secondary alignments	0
Supplementary alignments	7,807 / 0.01%
Read min/max/mean length	30 / 80 / 80.02
Clipped reads	2,065,102 / 2.45%

Globals (inside of regions)

Regions size/percentage of reference	45,326,818 / 1.46%
Mapped reads	63,363,519 / 75.07%
Mapped reads, only first in pair	31,877,600 / 37.77%
Mapped reads, only second in pair	31,485,919 / 37.3%
Mapped reads, both in pair	63,167,455 / 74.84%
Mapped reads, singletons	196,064 / 0.23%
Correct strand reads	0 / 0%
Clipped reads	2,065,102 / 2.45%
Duplicated reads (flagged)	2,968,557 / 4.68%

ACGT Content (inside of regions)

Number/percentage of A's	1,090,175,822 / 25.48%
Number/percentage of C's	1,048,730,118 / 24.52%
Number/percentage of T's	1,108,474,060 / 25.91%
Number/percentage of G's	1,030,171,088 / 24.08%
Number/percentage of N's	237,846 / 0.01%
GC Percentage	48.6%

Coverage (inside of regions)

Mean	94.3822
Standard Deviation	97.2737

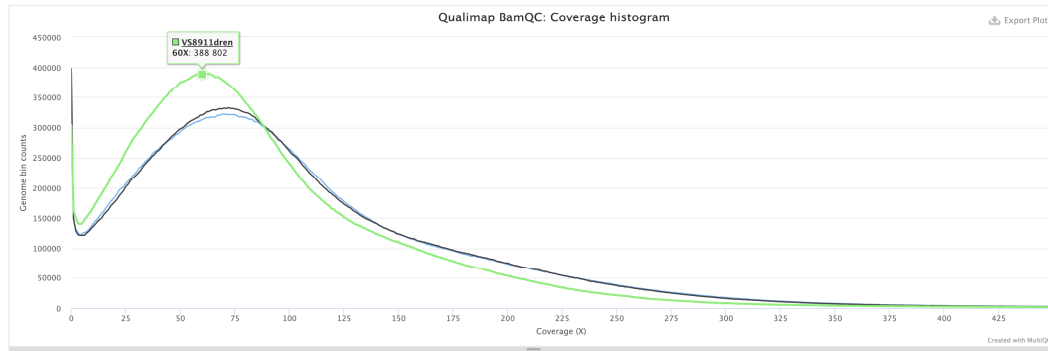
Mapping QC - coverage

Coverage histogram

Distribution of the number of locations in the reference genome with a given depth of coverage.

Help

Y-Limits: on

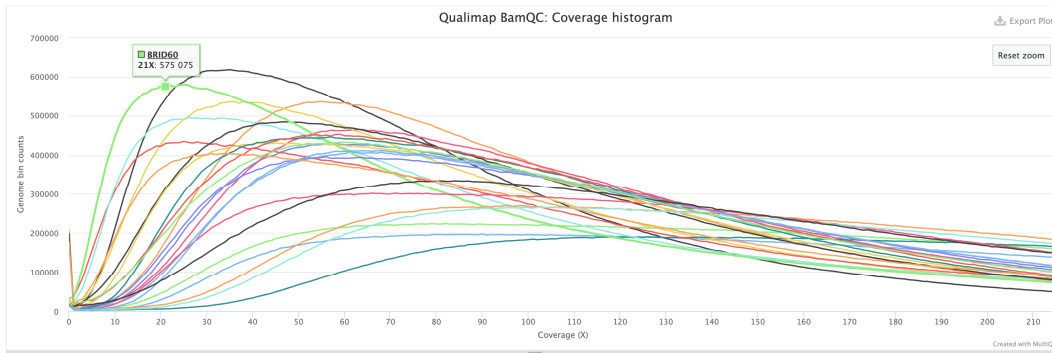


Coverage histogram

Distribution of the number of locations in the reference genome with a given depth of coverage.

Help

Y-Limits: on

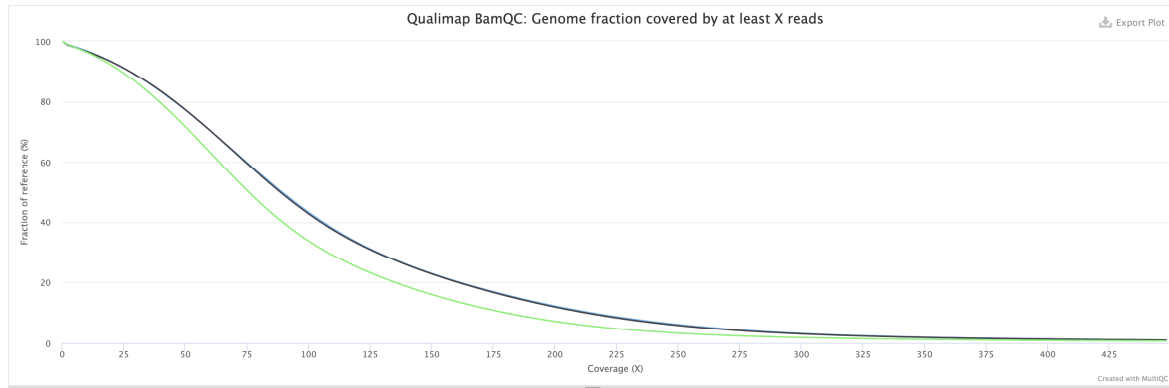


Mapping QC – cumulative coverage

Cumulative genome coverage

Percentage of the reference genome with at least the given depth of coverage.

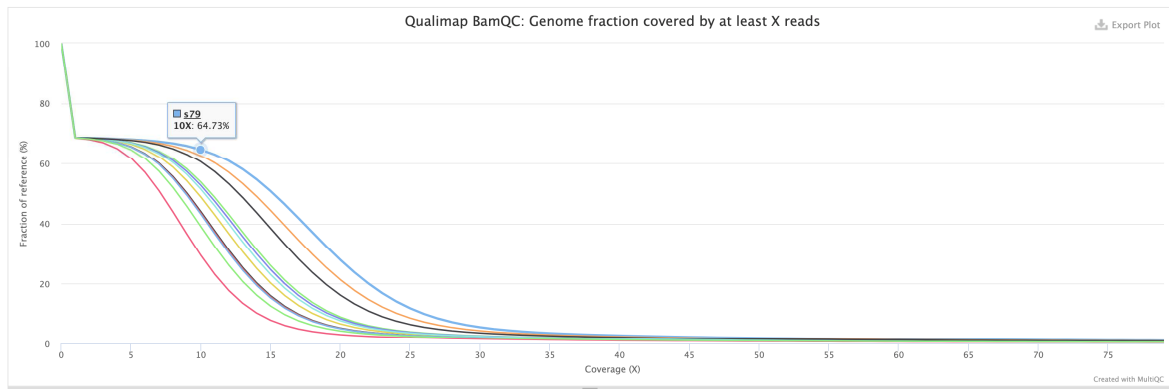
Help
Y-Limits: on



Cumulative genome coverage

Percentage of the reference genome with at least the given depth of coverage.

Help
Y-Limits: on



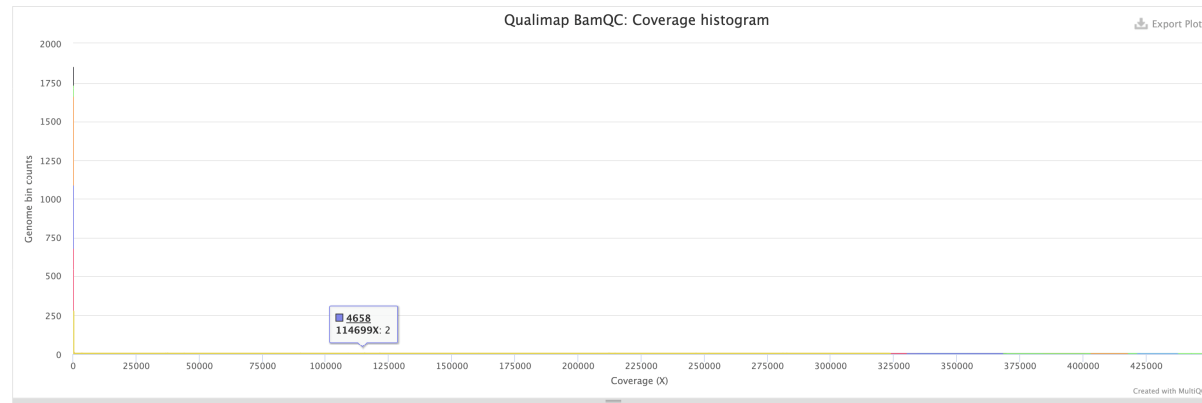
Mapping QC

Coverage histogram

Distribution of the number of locations in the reference genome with a given depth of coverage.

Help

Y-Limits: on

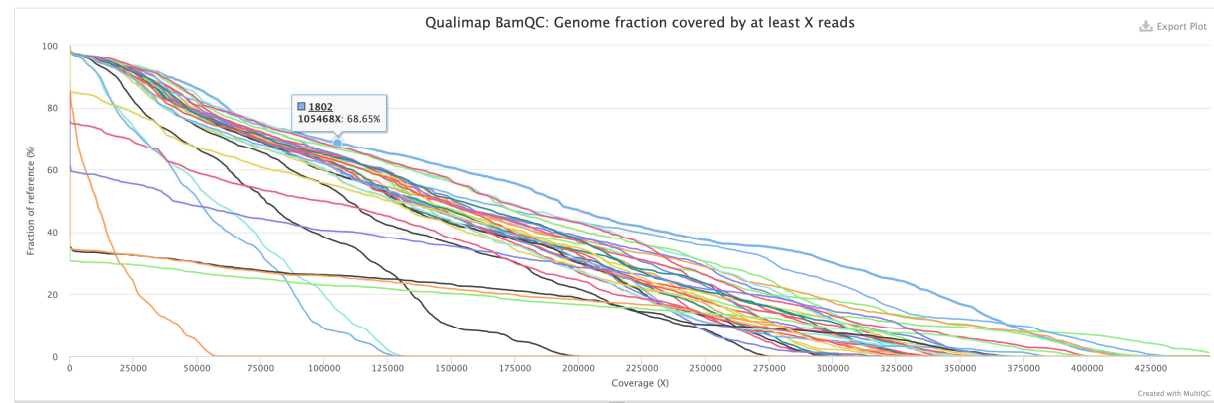


Cumulative genome coverage

Percentage of the reference genome with at least the given depth of coverage.

Help

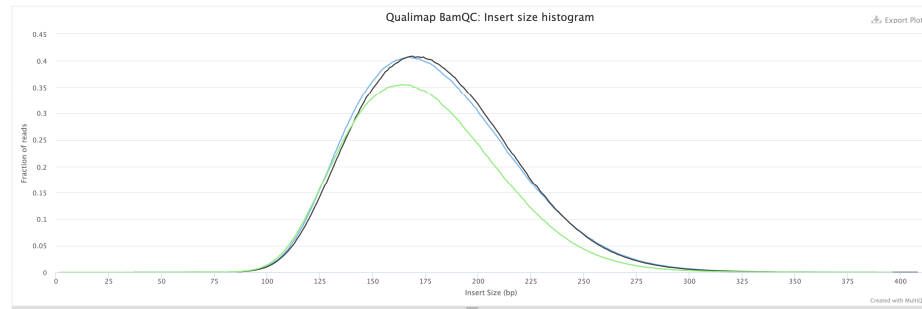
Y-Limits: on



Mapping QC

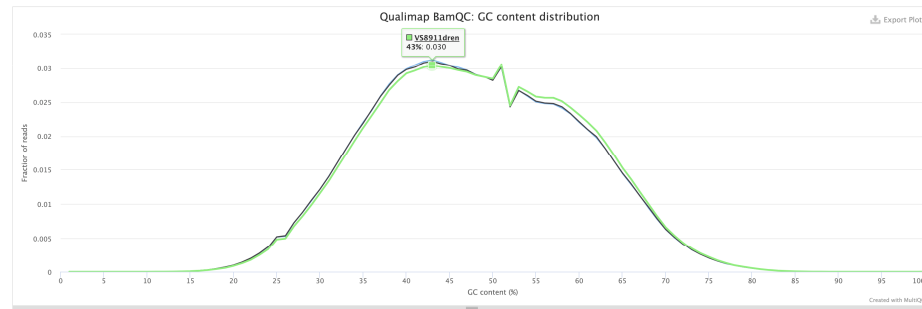
Insert size histogram

Distribution of estimated insert sizes of mapped reads.



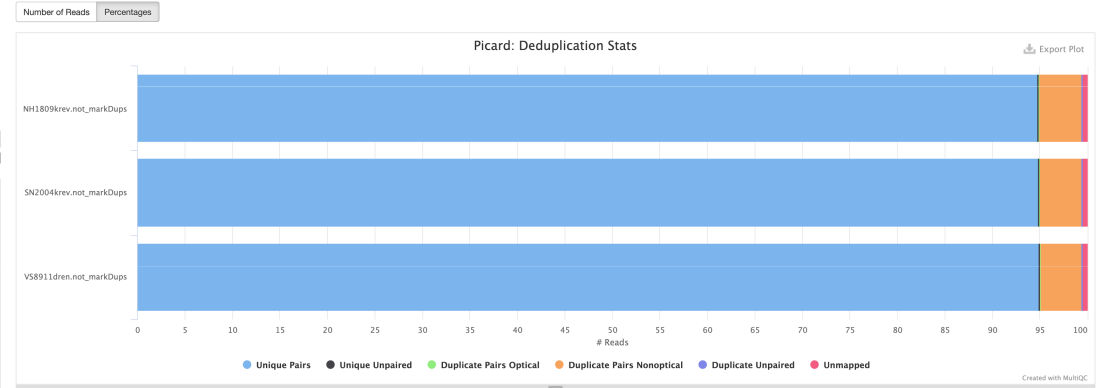
GC content distribution

Each solid line represents the distribution of GC content of mapped reads for a given sample.

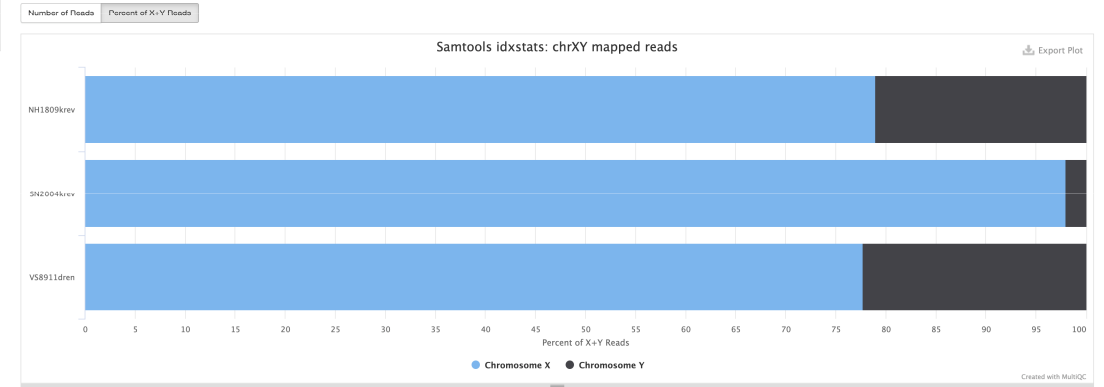


Mark Duplicates

Number of reads, categorised by duplication state. **Pair counts are doubled** - see help text for details.

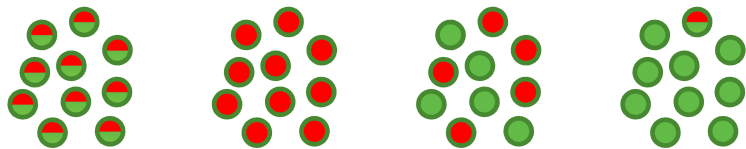
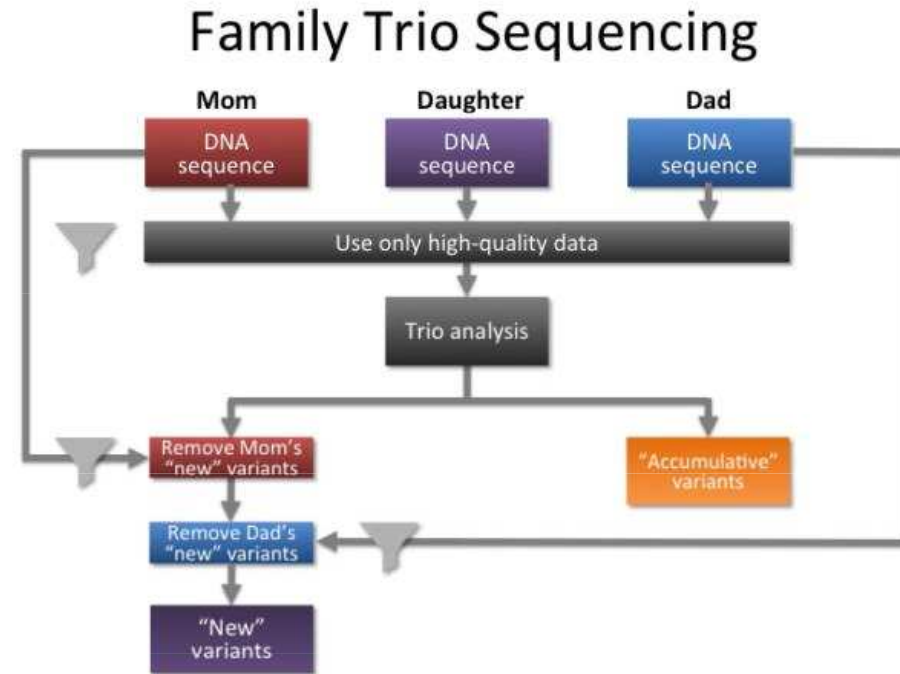


XY counts



Variant Calling - Germline

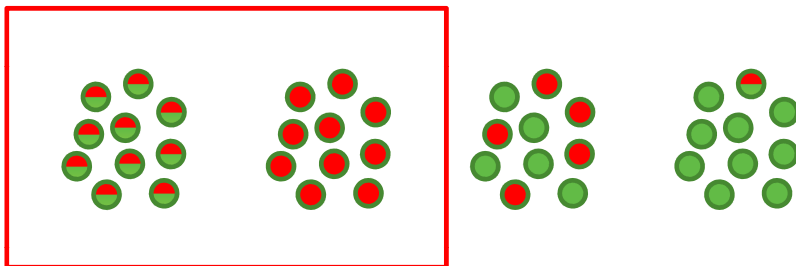
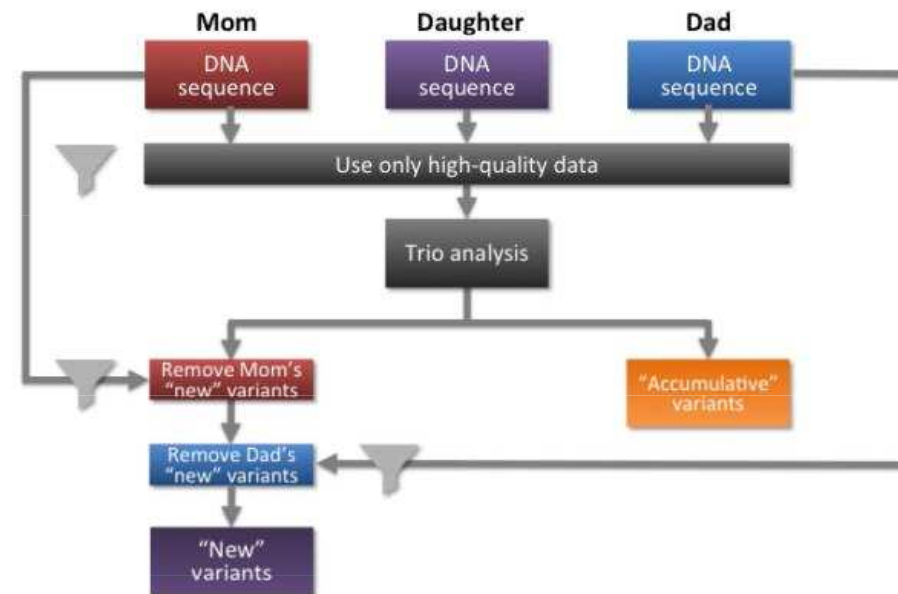
- What you have from birth
- Family trio sequencing
- Predispositions



Variant Calling - Germline

- What you have from birth
- Family trio sequencing
- Predispositions

Family Trio Sequencing

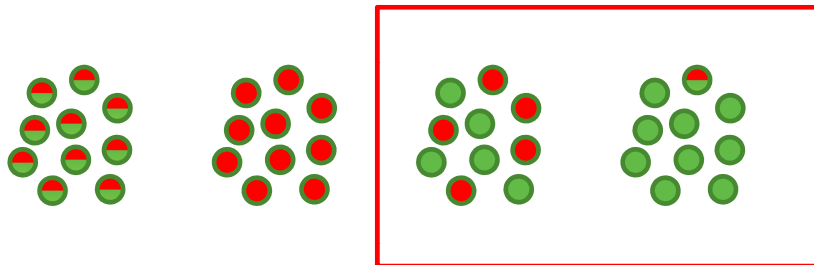


Variant Calling - Germline

- Tools:

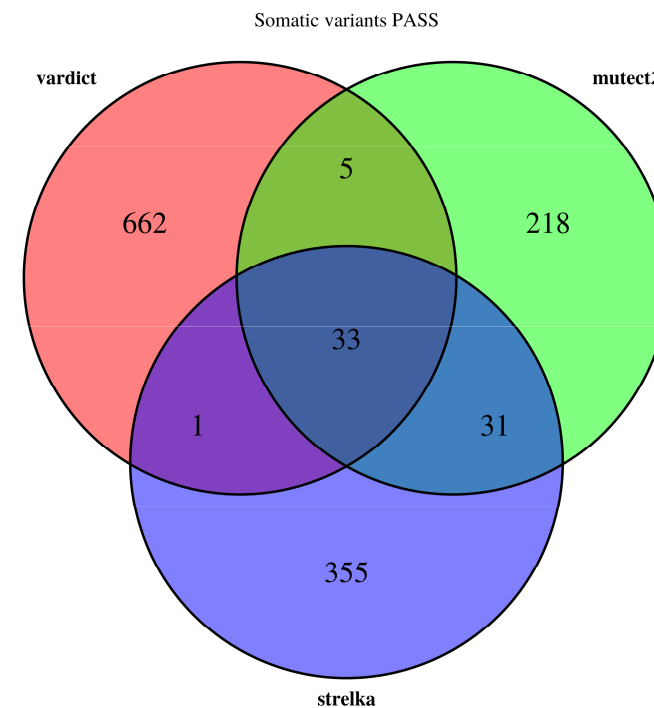
Variant Calling - Somatic

- Diagnostics / prognostic / therapy decision
- Tumor – normal paired
 - Somatic variant calling without normal needs high coverage
- Expected variant heterogeneity
- Indirectly correlates to the necessary coverage



Variant Calling - Somatic

- Multiple tools:
 - strelka2, verdict, mutect2, somaticsniper, lofreq, muse, varscan
- Ensemble caller
 - SomaticSeq
 - Use machine learning to detect TP from FP
- Sensitivity vs. specificity
 - Preferred sensitivity
 - Preferred accuracy for derived information



Small Variant annotation

- VEP – variant effect predictor
- Transcript "selection"
 - Refseq vs. ensemble
- Population frequency
 - 1000 genome project
 - Gnomad
- Many clinical variant DBs
 - Gene based vs. variant based
 - snpDB
 - COSMIC
 - clinvar
 - CGC

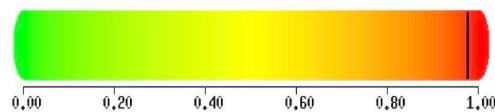
Small Variant annotation – functional prediction

- General variant consequence
 - Based on the position
 - Impact
- Effect of the variant on protein structure
 - PolyPhen
 - SIFT

POLYPHEN-2

This mutation is predicted to be **PROBABLY DAMAGING** with a score of **0.976**

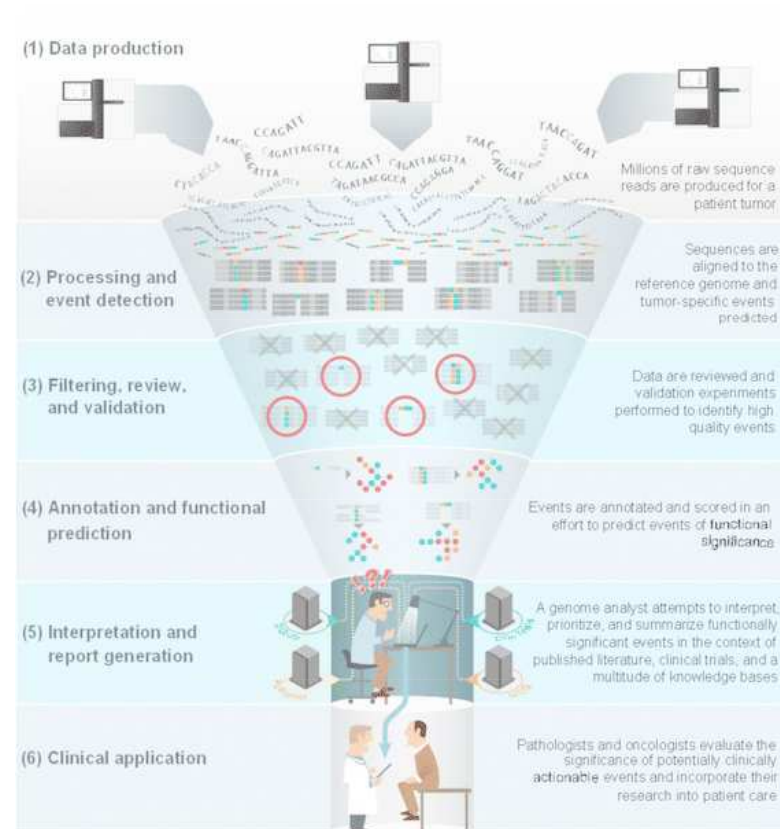
(sensitivity: **0.76**; specificity: **0.96**)



SO term	SO description	SO accession	Display term	IMPACT
transcript_ablation	A feature ablation whereby the deleted region includes a transcript feature	SO:0001893	Transcript ablation	HIGH
splice_acceptor_variant	A splice variant that changes the 2 base region at the 3' end of an intron	SO:0001574	Splice acceptor variant	HIGH
splice_donor_variant	A splice variant that changes the 2 base region at the 5' end of an intron	SO:0001575	Splice donor variant	HIGH
stop_gained	A sequence variant whereby at least one base of a codon is changed, resulting in a premature stop codon, leading to a shortened transcript	SO:0001587	Stop gained	HIGH
frameshift_variant	A sequence variant which causes a disruption of the translational reading frame, because the number of nucleotides inserted or deleted is not a multiple of three	SO:0001589	Frameshift variant	HIGH
stop_lost	A sequence variant where at least one base of the terminator codon (stop) is changed, resulting in an elongated transcript	SO:0001578	Stop lost	HIGH
start_lost	A codon variant that changes at least one base of the canonical start codon	SO:0002012	Start lost	HIGH
transcript_amplification	A feature amplification of a region containing a transcript	SO:0001889	Transcript amplification	HIGH
inframe_insertion	An inframe non synonymous variant that inserts bases into in the coding sequenc	SO:0001821	Inframe insertion	MODERATE
inframe_deletion	An inframe non synonymous variant that deletes bases from the coding sequenc	SO:0001822	Inframe deletion	MODERATE
misceneo_variant	A sequence variant, that changes one or more bases, resulting in a different amino acid sequence but where the length is preserved	SO:0001583	Misceneo variant	MODERATE
protein_altering_variant	A sequence_variant which is predicted to change the protein encoded in the coding sequence	SO:0001818	Protein altering variant	MODERATE
splice_region_variant	A sequence variant in which a change has occurred within the region of the splice site, either within 1-3 bases of the exon or 3-8 bases of the intron	SO:0001630	Splice region variant	LOW
incomplete_terminal_codon_variant	A sequence variant where at least one base of the final codon of an incompletely annotated transcript is changed	SO:0001626	Incomplete terminal codon variant	LOW
stop_retained_variant	A sequence variant where at least one base in the terminator codon is changed, but the terminator remains	SO:0001567	Stop retained variant	LOW
synonymous_variant	A sequence variant where there is no resulting change to the encoded amino acid	SO:0001819	Synonymous variant	LOW

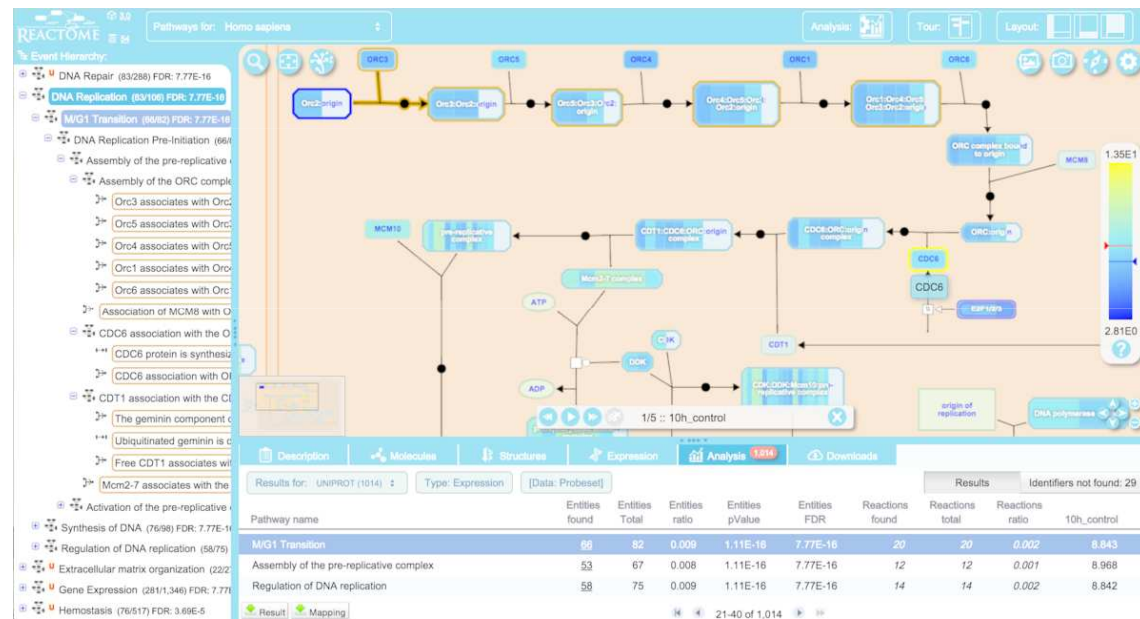
Small Variant interpretation

- Hardest part
- Usually manual work
 - Clinical genetics
 - Select 5 probable causal from ~1000
- Bioinformatics can help



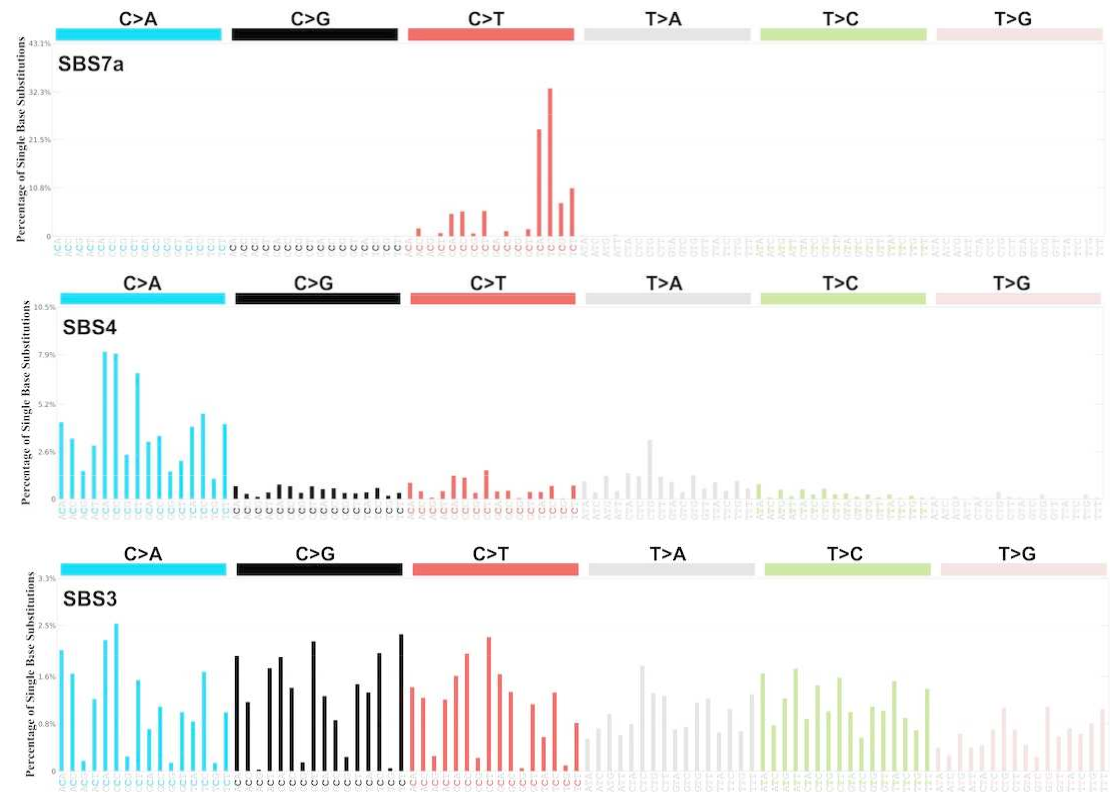
Variant interpretation – gene networks

- Gene ontology
- Biological pathway DB
 - KEGG
 - Reactome
 - WikiPathways



Variant interpretation – derived informations

- Tumor mutational burden
 - Several definitions
 - Mutations per million bases
- Mutational Signatures
 - COSMIC
 - exposure to ultraviolet light
 - Tobacco smoking
 - Defective DNA damage repair





f CEITEC

🐦 @CEITEC_Brno

Thank you for your attention!