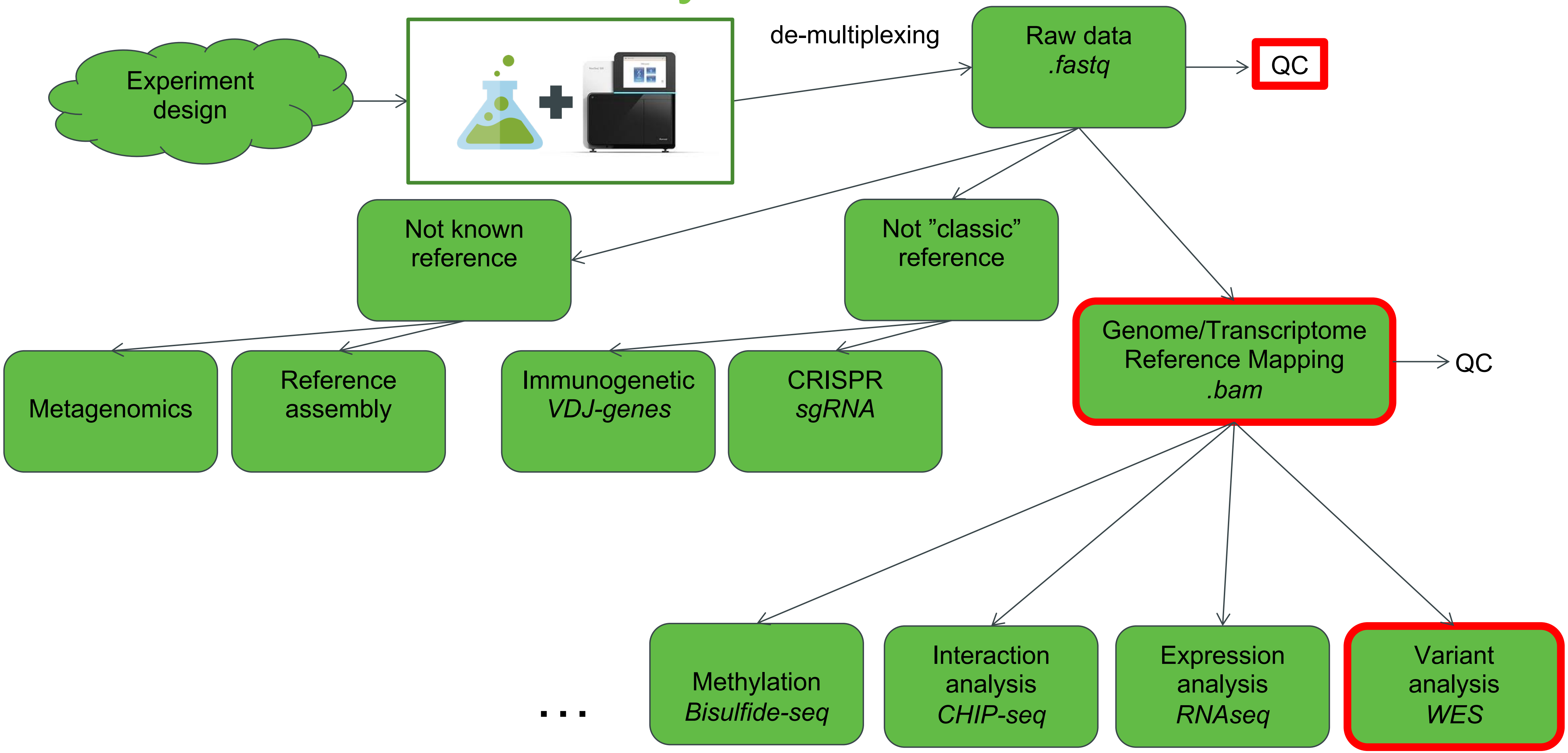
CEITEC
Central European Institute of Technology
BRNO | CZECH REPUBLIC

Modern methods for genome analysis
(PřF:Bi7420)

Lecture 3 : DNA re-sequencing + Small variant calling

Vojta Bystry
vojtech.bystry@ceitec.muni.cz

NGS data analysis



Experiment design

de-multiplexing

Raw data
.fastq

QC

Not known reference

Not ”classic” reference

Genome/Transcriptome Reference Mapping
.bam

QC

Metagenomics

Reference assembly

Immunogenetic
VDJ-genes

CRISPR
sgRNA

. . .

Methylation
Bisulfide-seq

Interaction analysis
CHIP-seq

Expression analysis
RNAseq

Variant analysis
WES

DNA re-sequencing

- Variant Calling
- Medical genomics
  - Cancer genomics

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

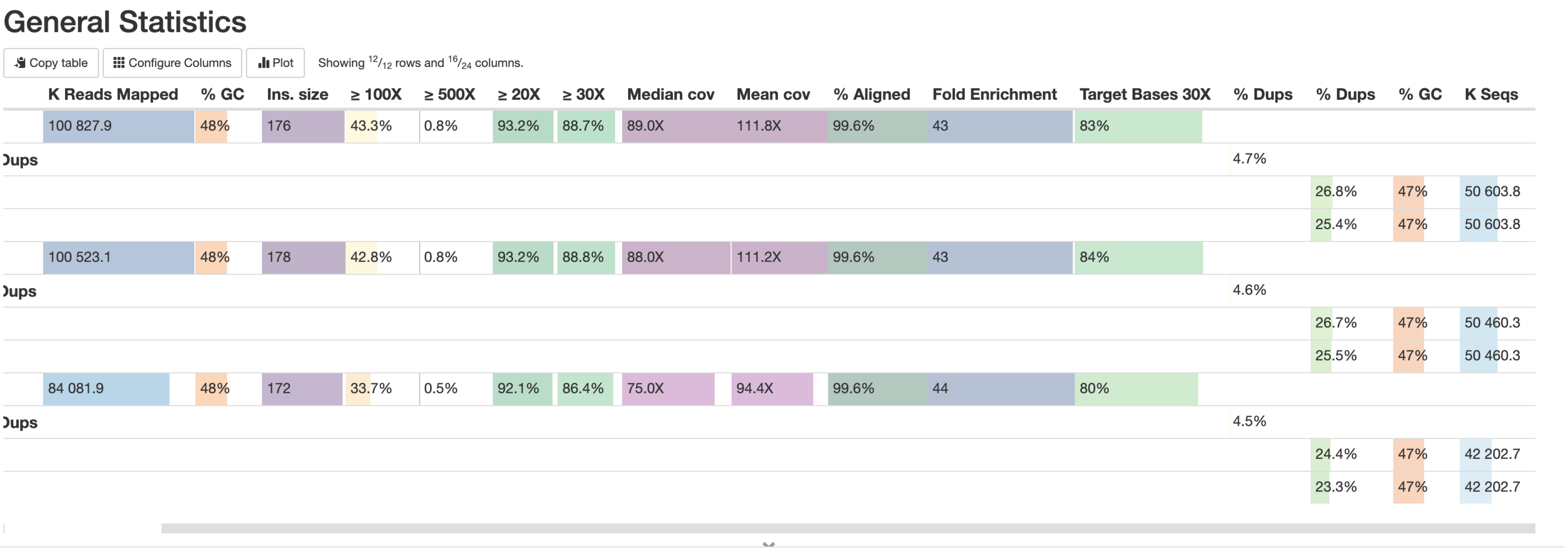
- Germline vs. Somatic

Mapping

- Computationally most demanding
- More or less standardized
- Output .bam
  - .bam = binary (ziped) .sam
  - .sam = Sequence Alignment Map   DNA re-sequencing

- Tools
  - BWA - DNA
  - STAR – RNA (eucaryotic)

Mapping QC

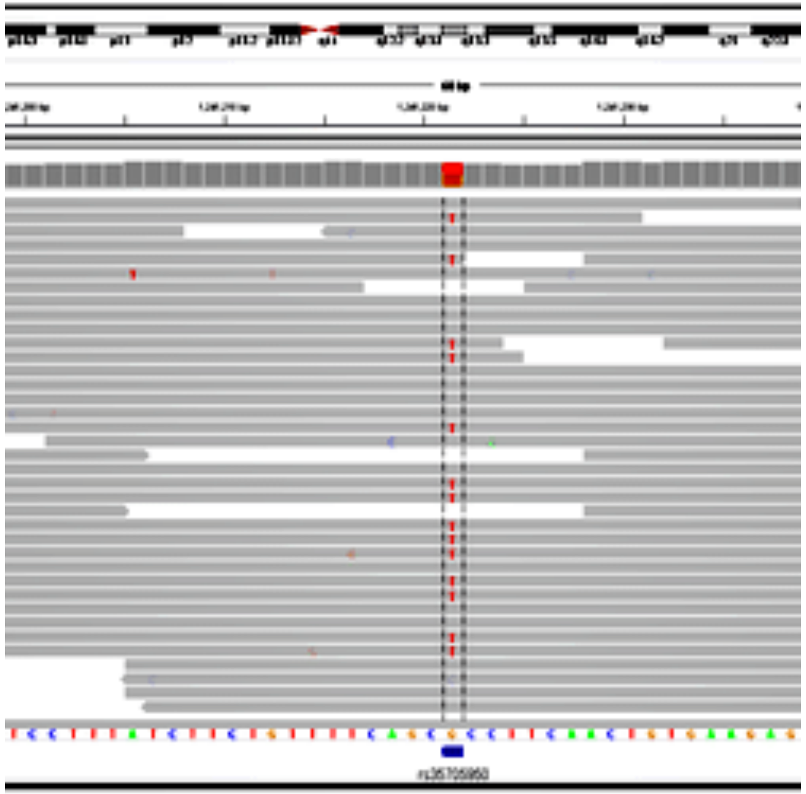


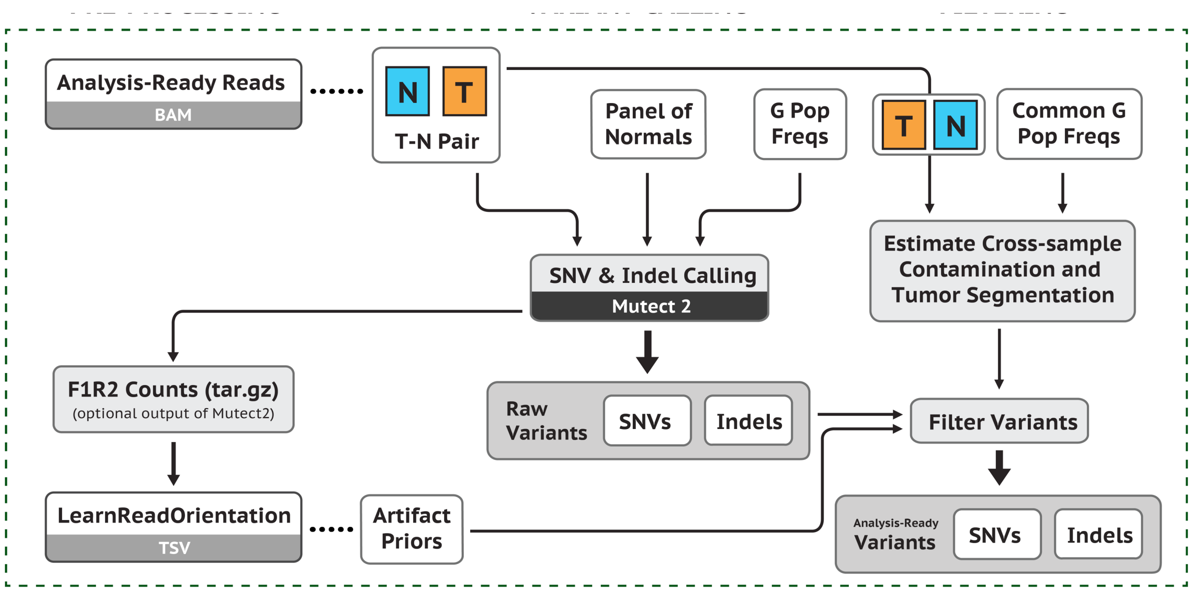
General Statistics

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%				
)ups													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%				
)ups													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%				
)ups													4.5%			
														24.4%	47%	42 202.7
														23.3%	47%	42 202.7

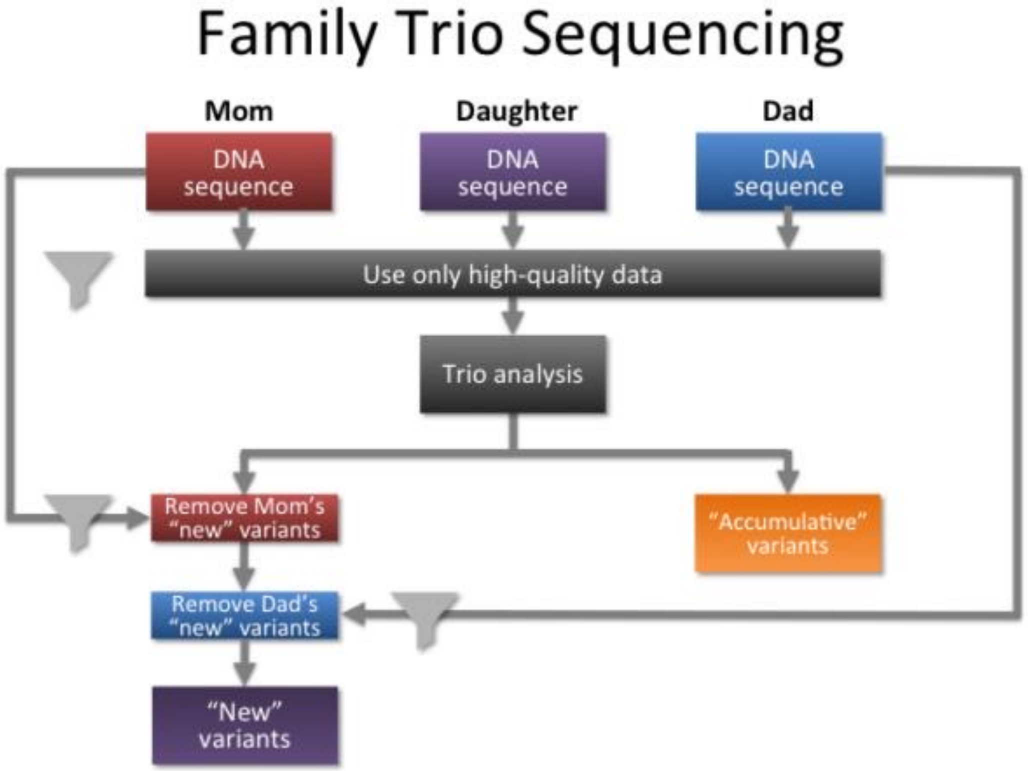
Small Variant calling





Variant Calling - Germline

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



Family Trio Sequencing

Mom

DNA sequence

Daughter

DNA sequence

Dad

DNA sequence

Use only high-quality data

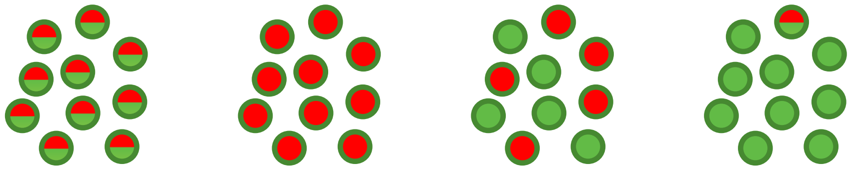
Trio analysis

Remove Mom's "new" variants

"Accumulative" variants

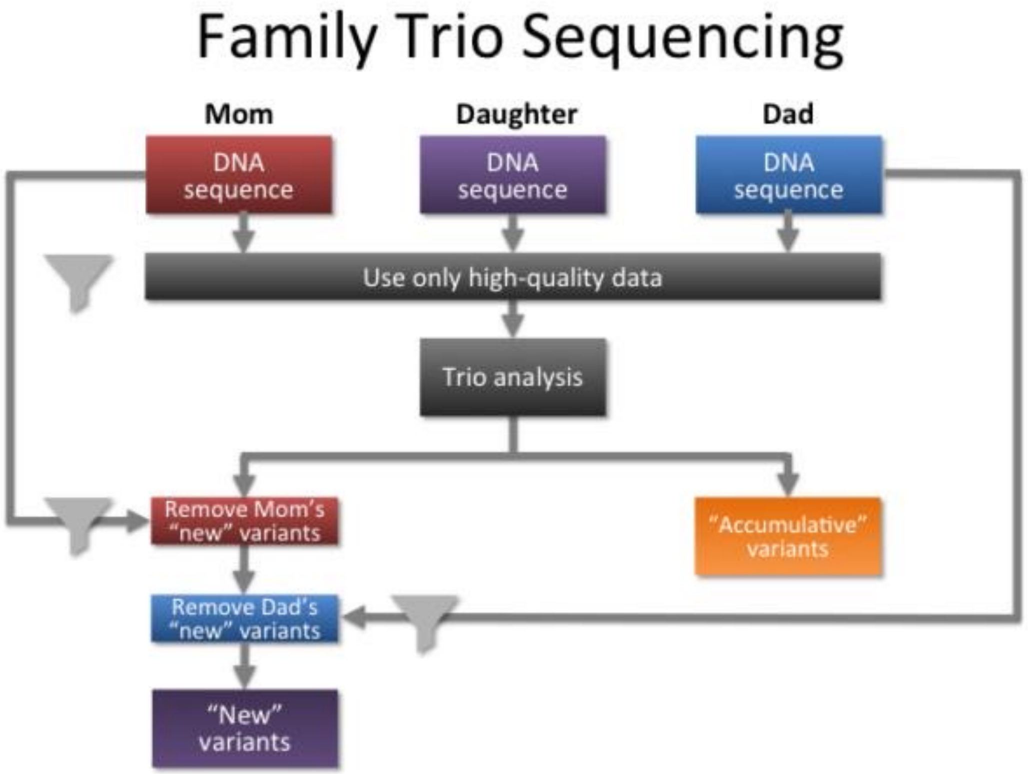
Remove Dad's "new" variants

"New" variants



Variant Calling - Germline

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



Family Trio Sequencing

Mom

Daughter

Dad

DNA sequence

DNA sequence

DNA sequence

Use only high-quality data

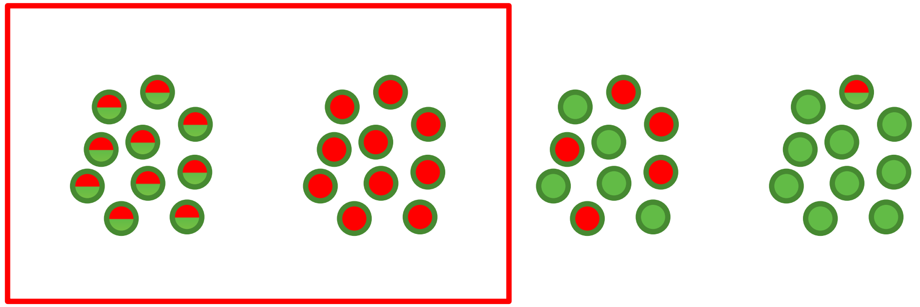
Trio analysis

Remove Mom's "new" variants

"Accumulative" variants

Remove Dad's "new" variants

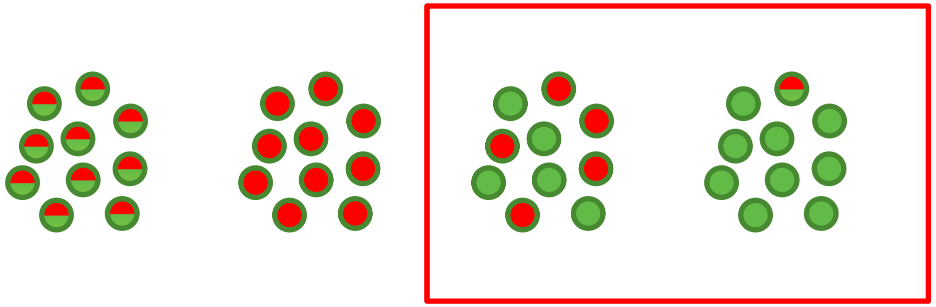
"New" variants



Variant Calling - Somatic

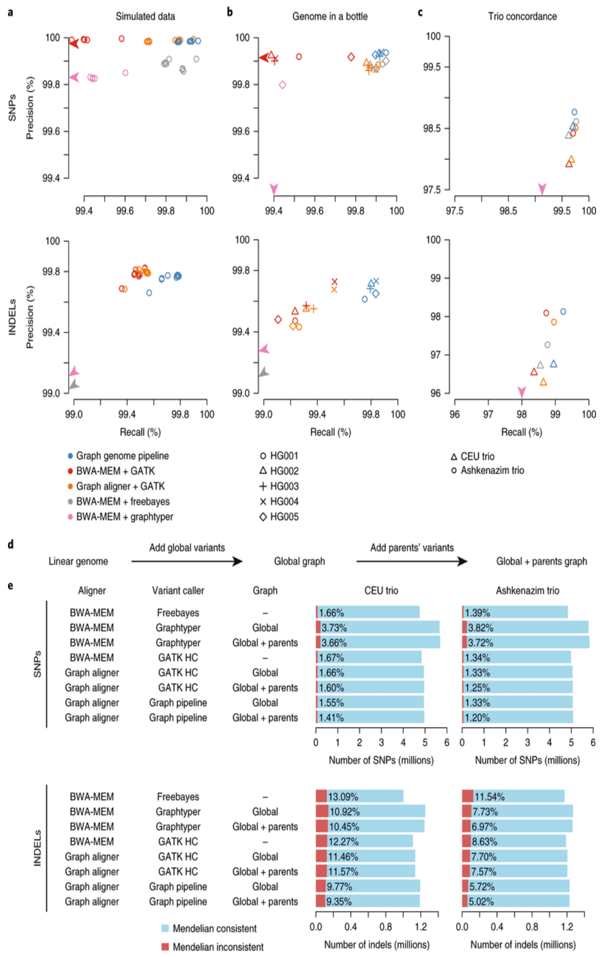
- Diagnostics / prognostic / therapy decision
- Tumor – normal paired
  - Somatic variant calling without normal needs high coverage (200x >)
    - not all germline variants will be filtered
- Expected variant heterogeneity
- Expected variant allelic frequency (VAF)
  - Histopathology prediction overestimate tumor load
  - Negative correlation to the necessary coverage

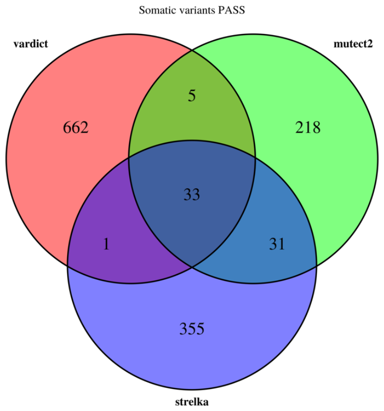
Tumor purity estiamtion
Tumor composition



Variant Calling - Tools

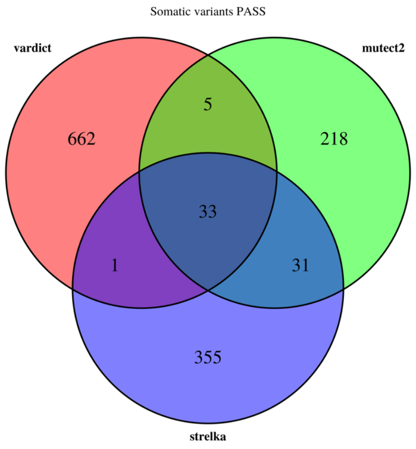
- Multiple tools:
  - strelka2, verdict, mutect2, somaticsniper, lofreq, muse, varscan
- Ensemble/meta callers usually outperformes individual
  - SomaticSeq
- Benchmarking
  - Genome in a Bottle
  - GIAB
  - son/father/mother trios of Ashkenazi Jewish

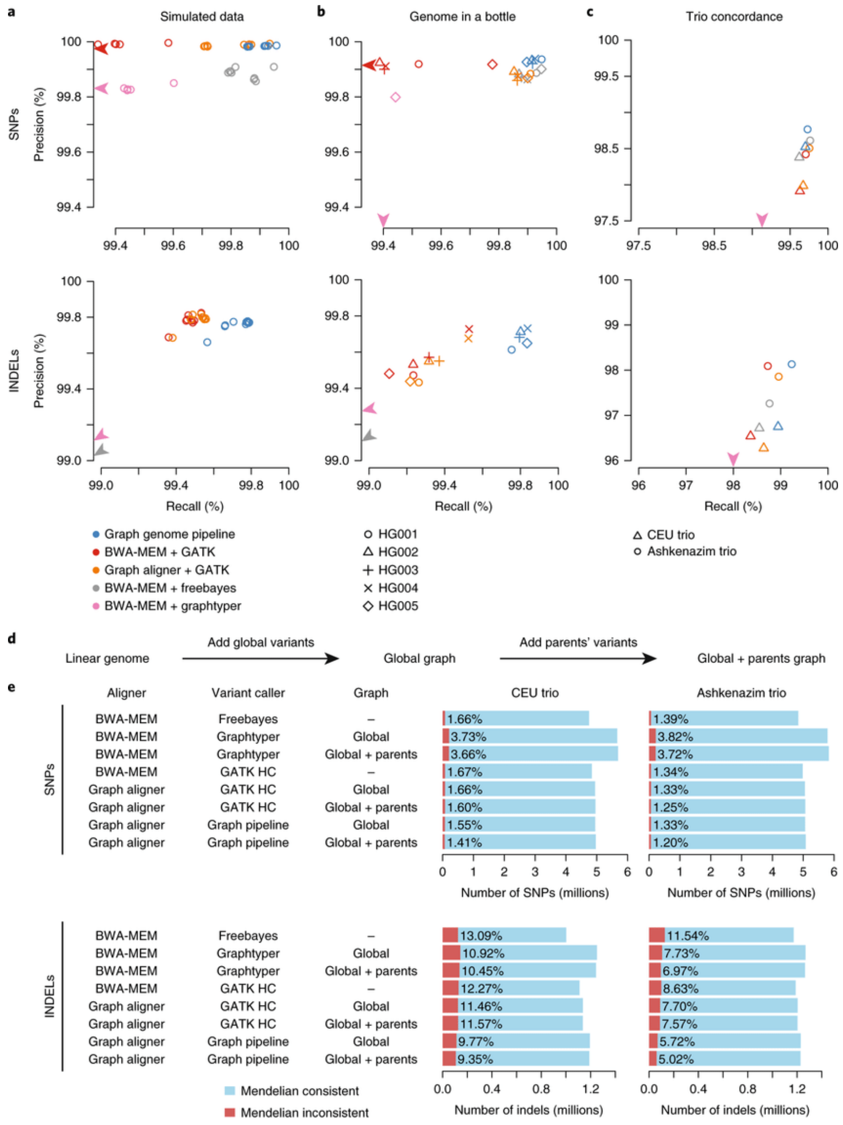




Variant Calling - Tools

- Problem is variant filtering
  - Complex regions
  - Pseudo-genes
- Sensitivity vs. specificity tradeoff
  - Preferred sensitivity
  - Preferred accuracy for automated processing



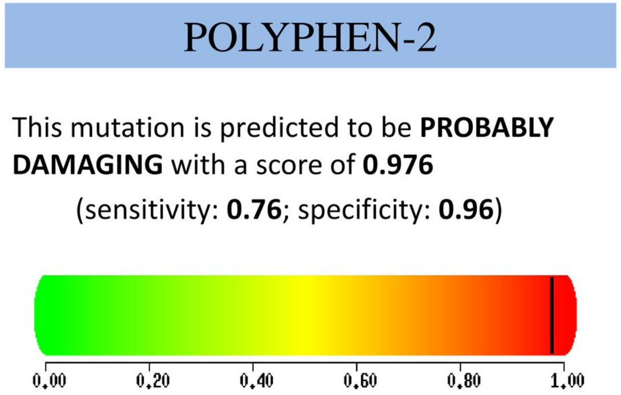


Small Variant annotation

- 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

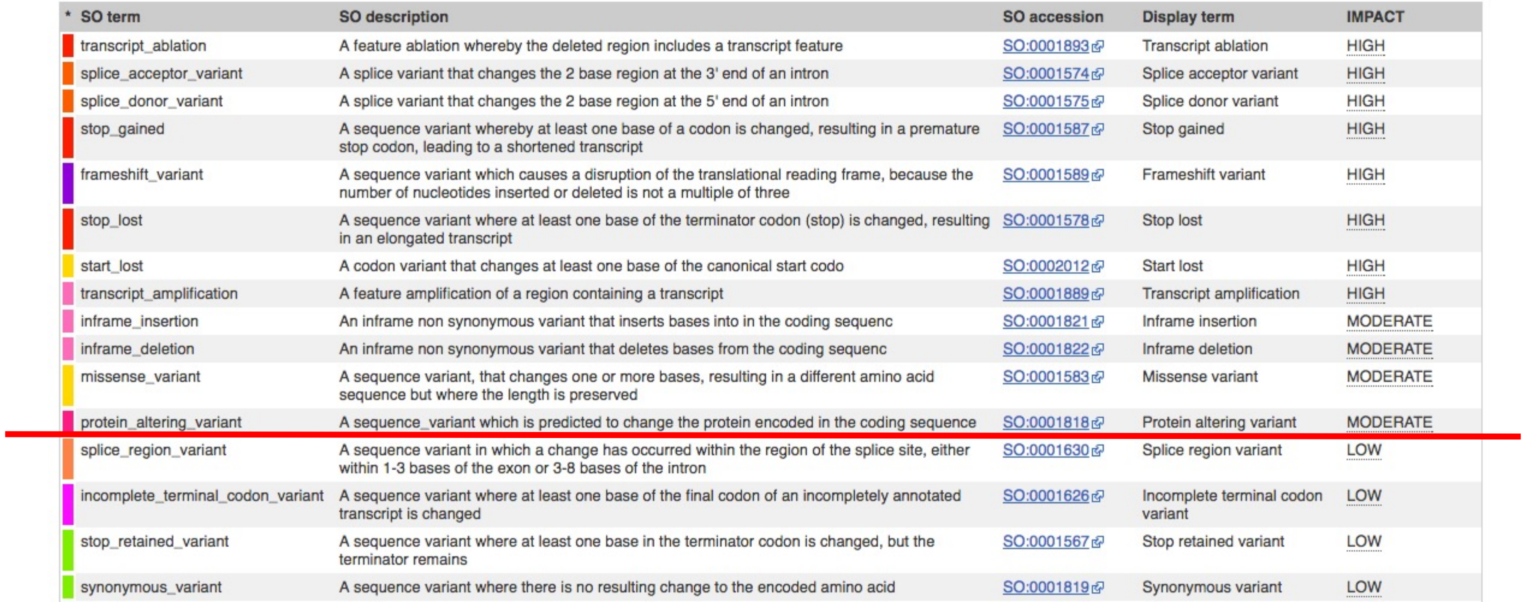
- General variant consequence
  - Based on the position
  - Impact
- Effect of the variant on protein structure
  - PolyPhen
  - SIFT



POLYPHEN-2

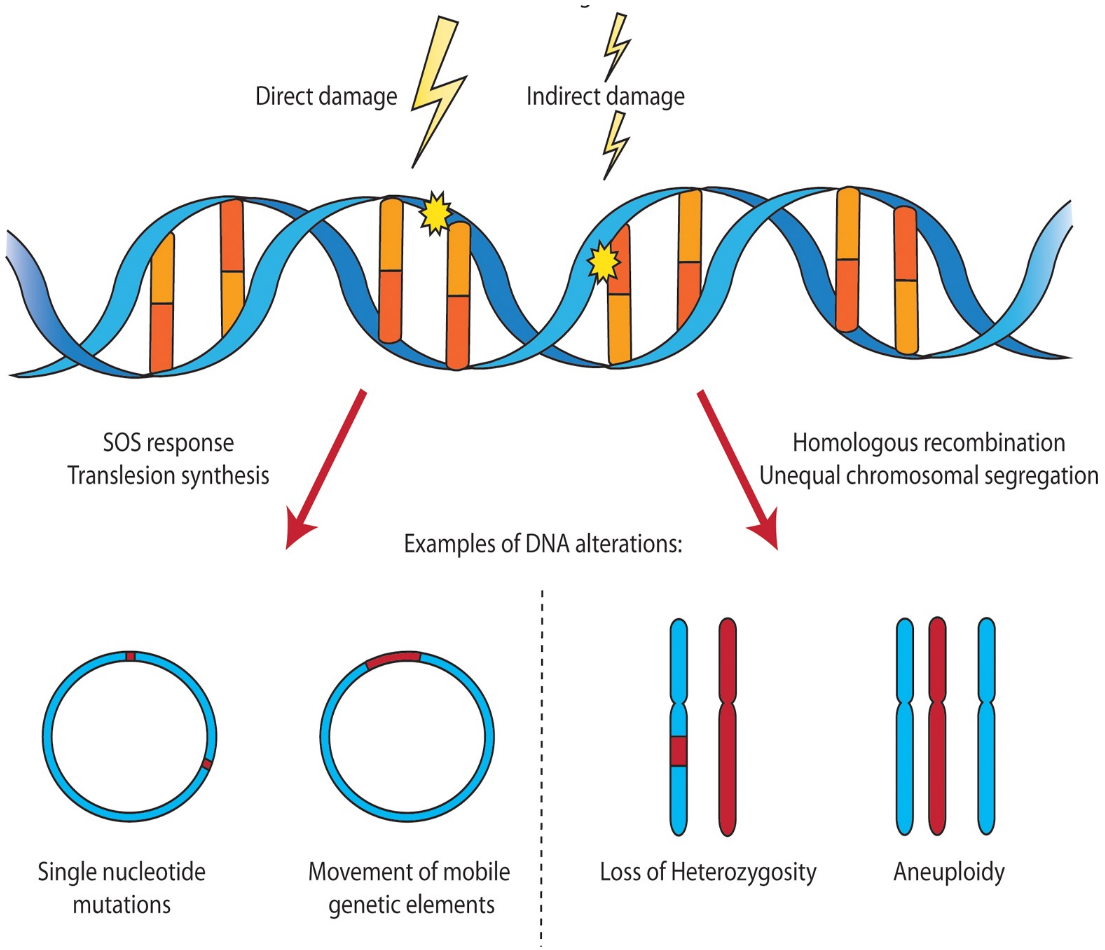
This mutation is predicted to be **PROBABLY DAMAGING** with a score of **0.976**

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



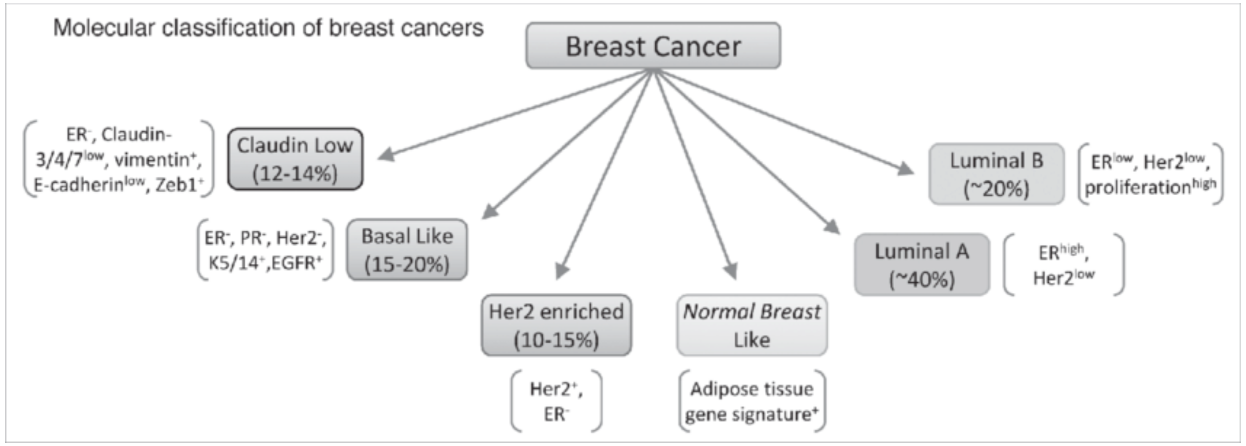
*	SO term	SO description	SO accession	Display term	IMPACT
■	transcript_ablation	A feature ablation whereby the deleted region includes a transcript feature	SO:0001893	Transcript ablation	HIGH
■	splice_acceptor_variant	A splice variant that changes the 2 base region at the 3' end of an intron	SO:0001574	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 codo	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
■	missense_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	Missense 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

Cancer genomics introduction


Direct damage
Indirect damage
SOS response
Translesion synthesis
Homologous recombination
Unequal chromosomal segregation
Examples of DNA alterations:
Single nucleotide mutations
Movement of mobile genetic elements
Loss of Heterozygosity
Aneuploidy

Cancer genomics introduction

- Based on molecular state
  - Classification
  - Prognostic
  - Treatment selection
    - Precission medicine

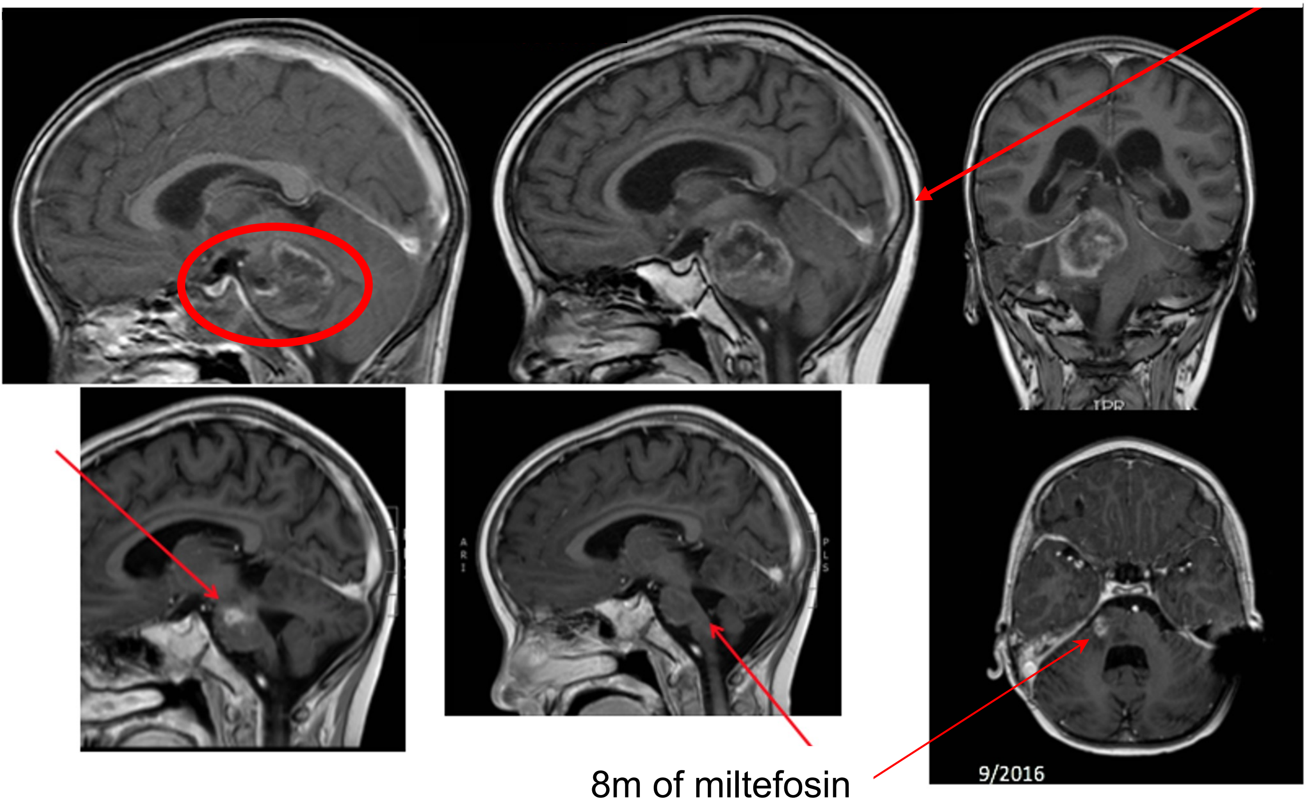

Molecular classification of breast cancers
Breast Cancer
ER^-, Claudin-3/4/7^{low}, vimentin$^+$, E-cadherin^{low}, Zeb1$^+$
Claudin Low (12-14%)
ER^-, PR^-, Her2$^-$, K5/14$^+$,EGFR$^+$
Basal Like (15-20%)
Her2 enriched (10-15%)
Her2$^+$, ER^-
Normal Breast Like
Adipose tissue gene signature$^+$
Luminal A (~40%)
ER^{high}, Her2^{low}
Luminal B (~20%)
ER^{low}, Her2^{low}, proliferation^{high}

Cancer genomics introduction - Case report

- **5 years old boy with diffuse intrinsic pontine glioma (DIPG),** 6 months of standard chemo/radiotherapy -> tumor progression, only 6 months to live
- WES identified activation mutation in PI3K kinase -> Akt oncogenic signalling pathway

At the beggining

6m treatment



4m of miltefosin

8m of miltefosin

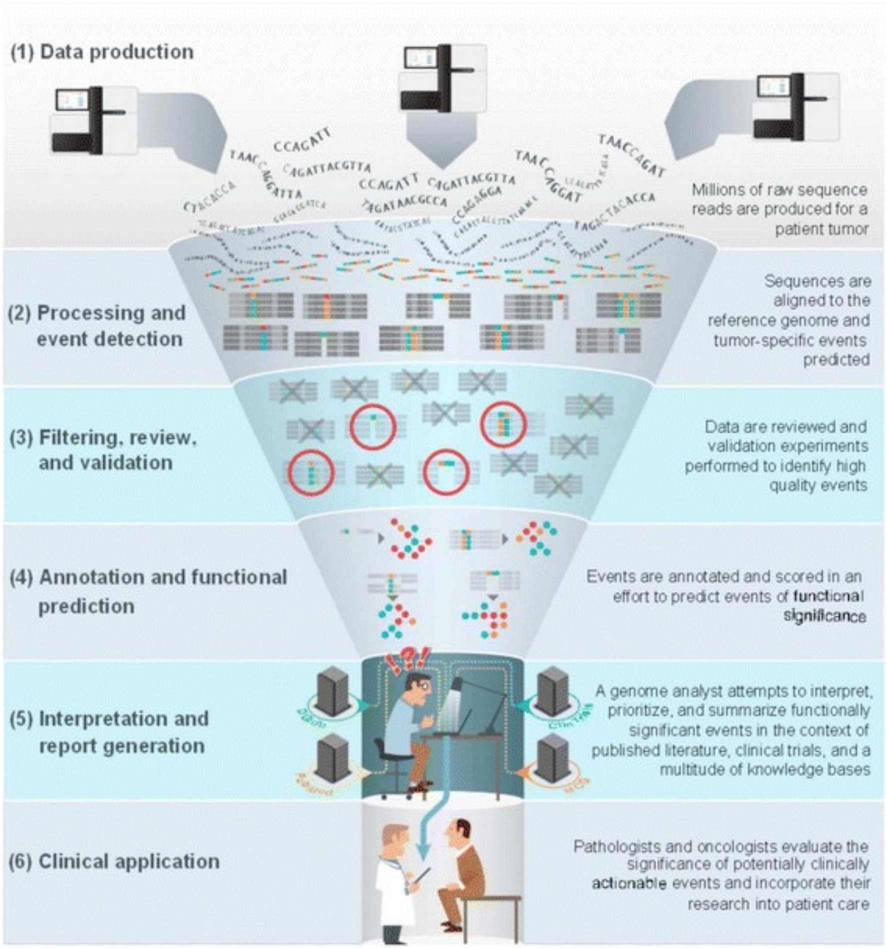
9/2016

Miltefosin/impavido
(only approved Akt inhibitor)

Leishmaniasis

DRUG REPURPOSING

Somatic variant NGS data analysis


(1) Data production
(2) Processing and event detection
(3) Filtering, review, and validation
(4) Annotation and functional prediction
(5) Interpretation and report generation
(6) Clinical application

Millions of raw sequence reads are produced for a patient tumor

Sequences are aligned to the reference genome and tumor-specific events predicted

Data are reviewed and validation experiments performed to identify high quality events

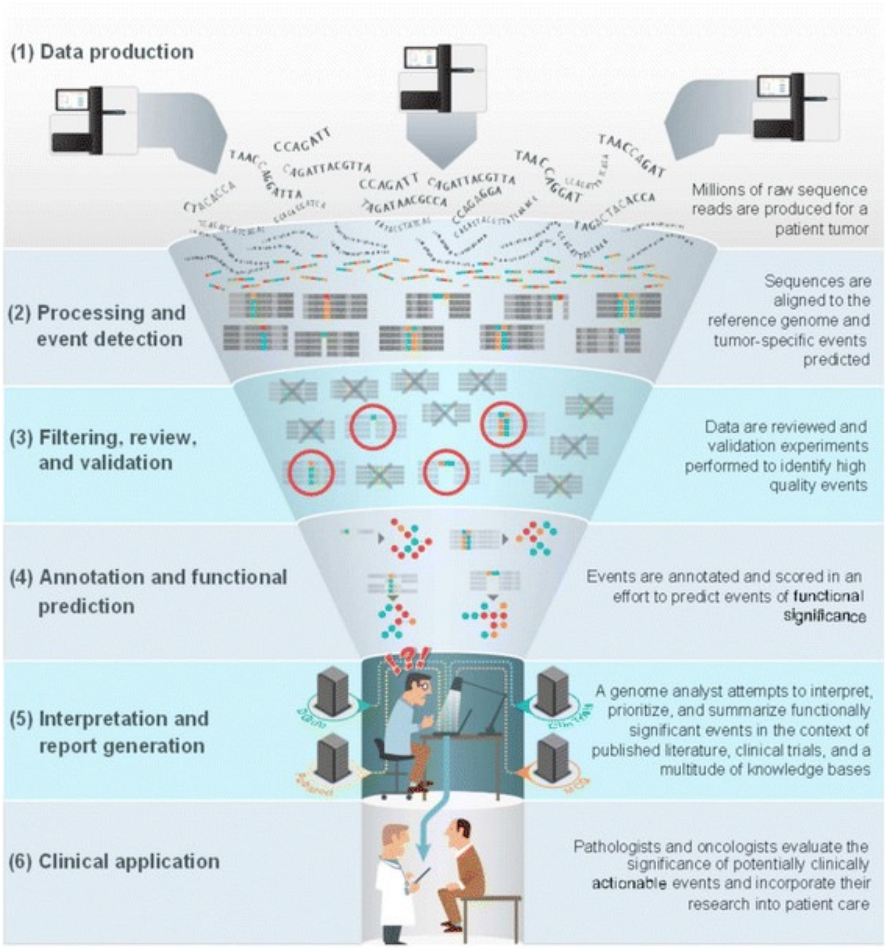
Events are annotated and scored in an effort to predict events of functional significance

A genome analyst attempts to interpret, prioritize, and summarize functionally significant events in the context of published literature, clinical trials, and a multitude of knowledge bases

Pathologists and oncologists evaluate the significance of potentially clinically actionable events and incorporate their research into patient care

- Primary analysis and QC
- Variant calling
- Variant annotation
- Variant interpretation
- Clinical application

Somatic variant NGS data analysis



(1) Data production

Millions of raw sequence reads are produced for a patient tumor

(2) Processing and event detection

Sequences are aligned to the reference genome and tumor-specific events predicted

(3) Filtering, review, and validation

Data are reviewed and validation experiments performed to identify high quality events

(4) Annotation and functional prediction

Events are annotated and scored in an effort to predict events of functional significance

(5) Interpretation and report generation

A genome analyst attempts to interpret, prioritize, and summarize functionally significant events in the context of published literature, clinical trials, and a multitude of knowledge bases

(6) Clinical application

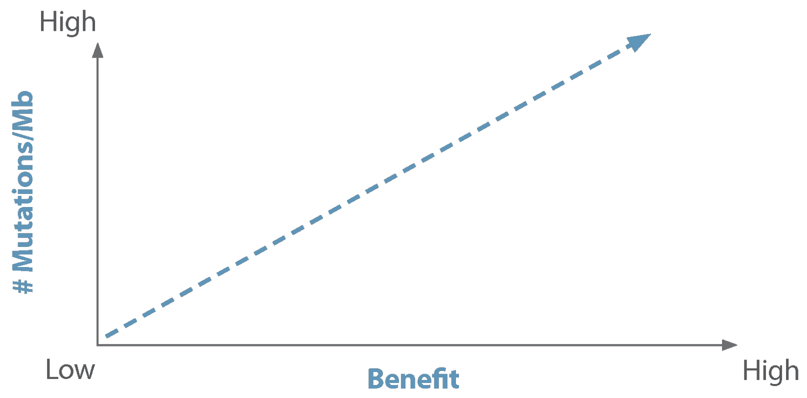
Pathologists and oncologists evaluate the significance of potentially clinically actionable events and incorporate their research into patient care

- Primary analysis and QC
- Variant calling
- Variant annotation
- ~~Variant interpretation~~
- Aggregated feature extraction
- Predictive modeling
- …
- Clinical application

Variant interpretation – derived informations

- Tumor mutational burden
  - Several definitions
  - Mutations per million bases
  - Good indicator for immunotherapy to work
- Microsatellite Instability
  - Specific variants occurence
- HPV status

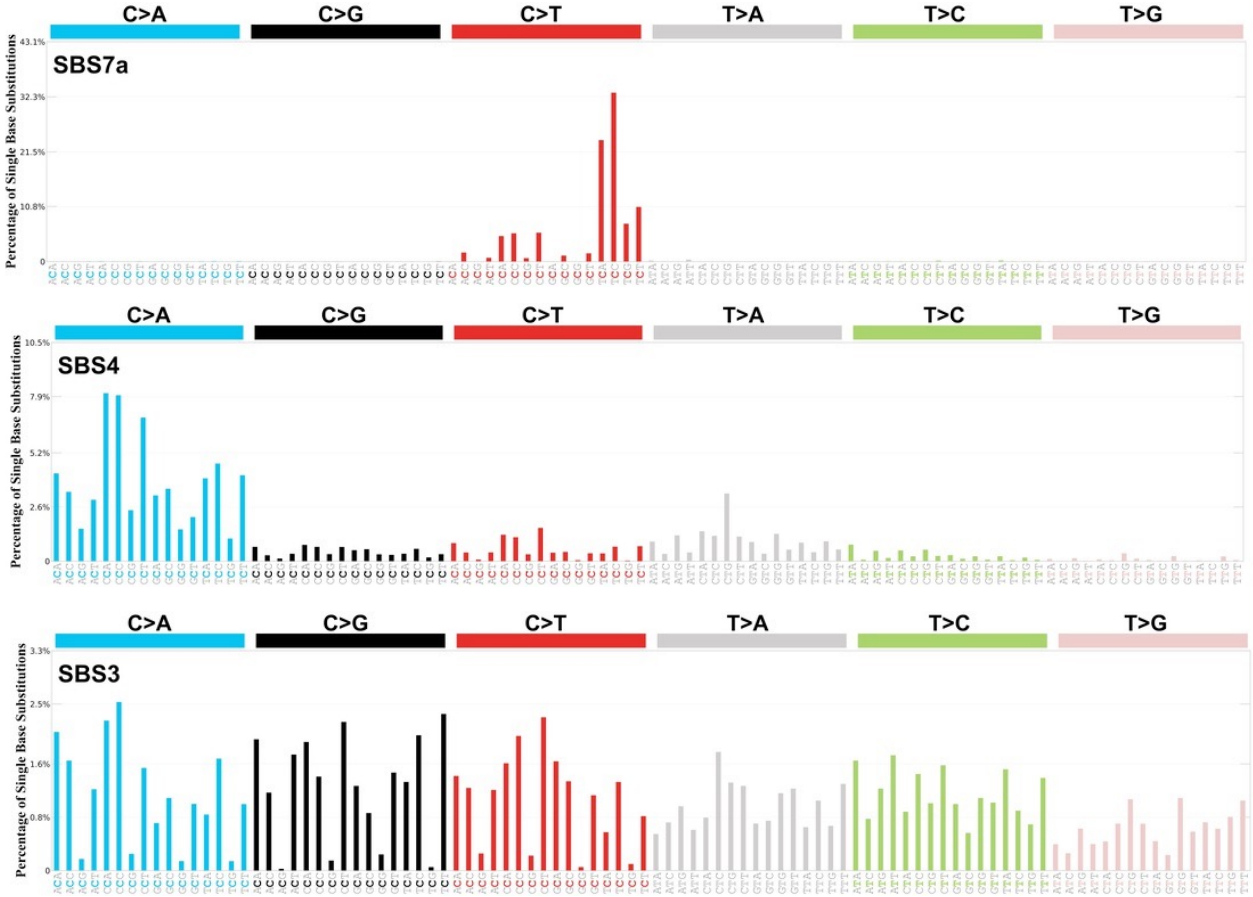
TMB: Immune Checkpoint Indication for Response



Tumors with significant numbers of mutations resulting in altered proteins (neo-antigens) may respond more effectively to immunotherapies.[1,2]

Variant interpretation – derived informations

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



Genomic variant predictive modeling

- Genomic variant data are very problematic for modeling
  - Enormous feature space
    - ~ 100 000 features
  - Limited number of data points
    - Only one predictive label per patient
- Feature selection/extraction
- Increase number of samples

Genomic variant predictive modeling

- Genomic variant data are very problematic for modeling
  - Enormous feature space
    - ~ 100 000 features
  - Limited number of data points
    - Only one predictive label per patient

→ Curse of dimensionality

- Feature selection/extraction
- Increase number of samples

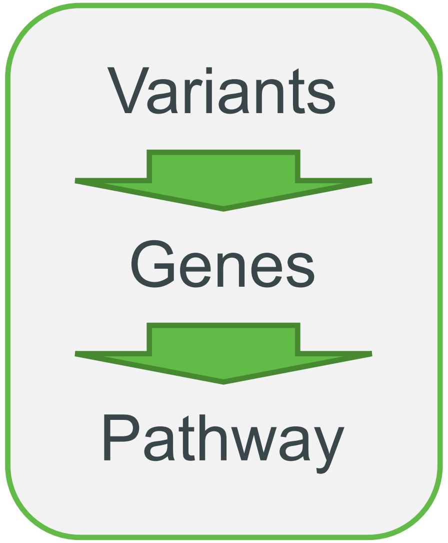
Genomic variant predictive modeling

- Genomic variant data are very problematic for modeling
  - Enormous feature space
    - ~ 100 000 features
  - Limited number of data points
    - Only one predictive label per patient

Curse of dimensionality

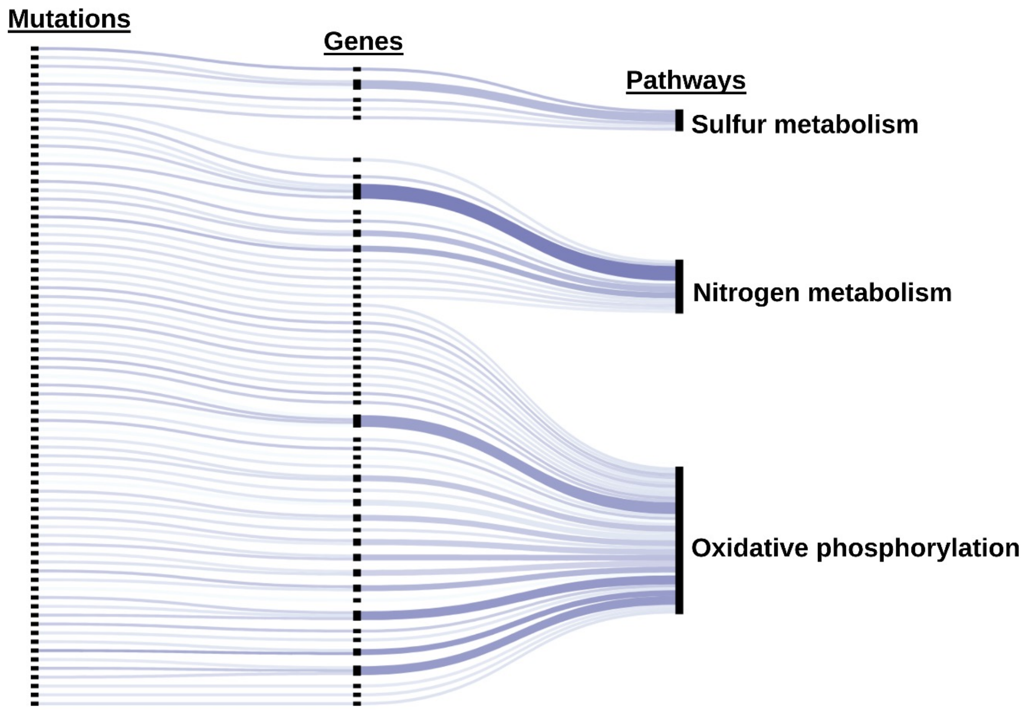
- Feature selection/extraction
- Increase number of samples

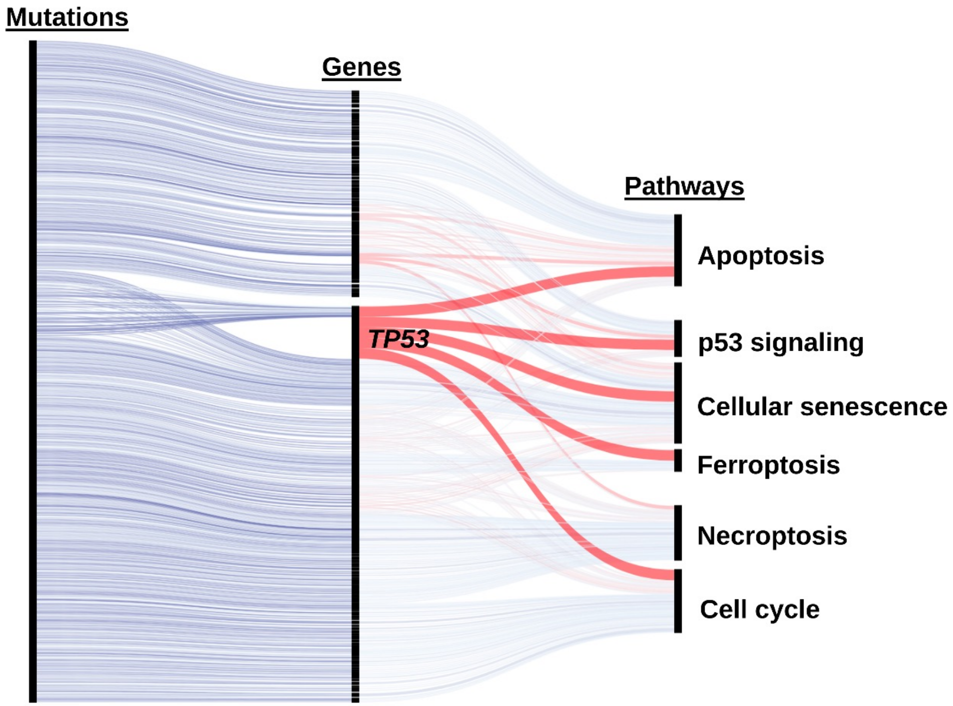
- Biologically meaningful data extraction
- Usage of publicly available data



Genomic variant predictive modeling

- Pathway level “disruption” score from gene- and mutation-level scores
  - KEGG pathways
  - Mutation effect combination of CADD, EVE, Polyphen2 scores





CEITEC

@CEITEC_Brno

Thank you for your attention!

Vojta Bystry
vojtech.bystry@ceitec.muni.cz