CEITEC

Central European Institute of Technology

BRNO | CZECH REPUBLIC

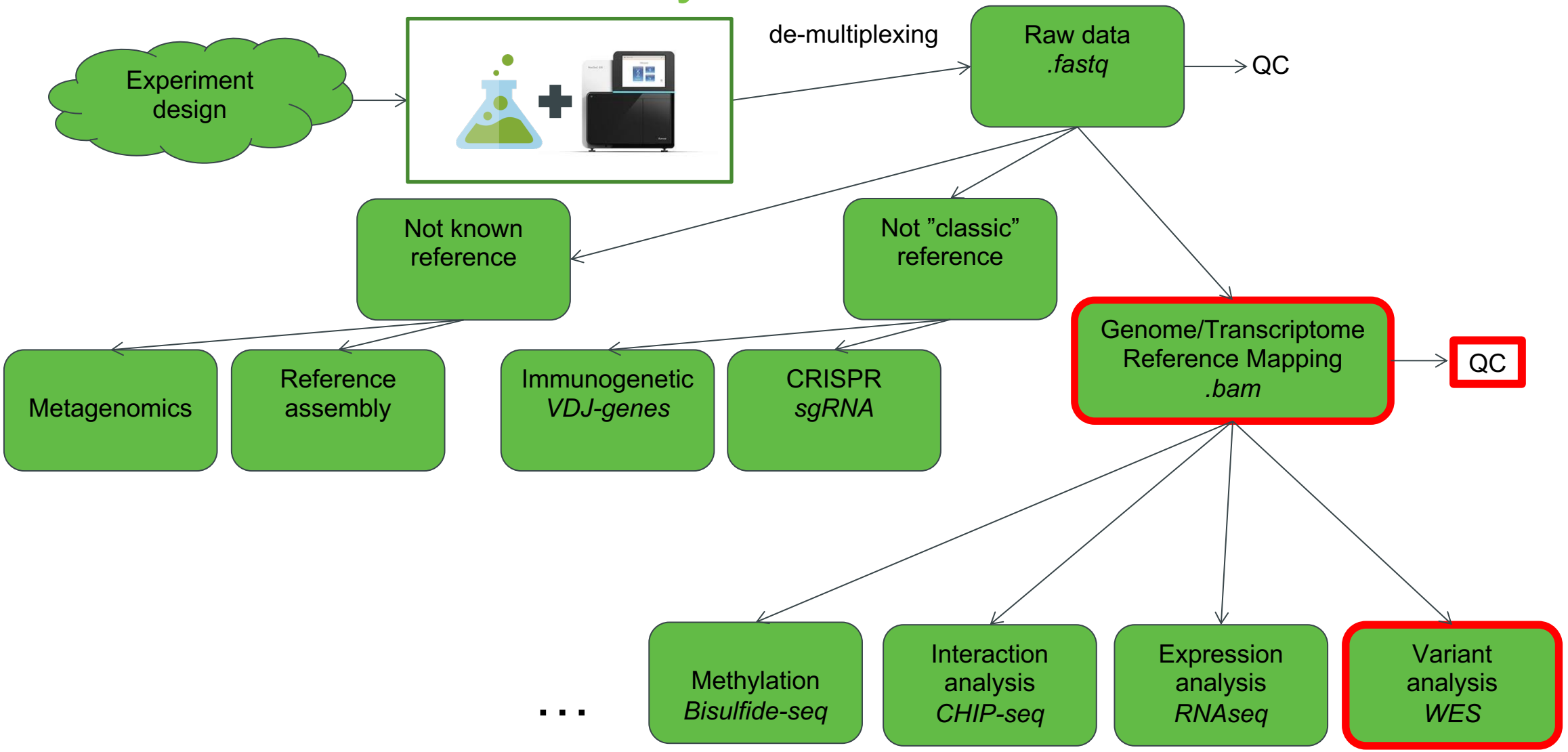
UNIVERSITAS MASARYKIANA BRUNENSIS

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

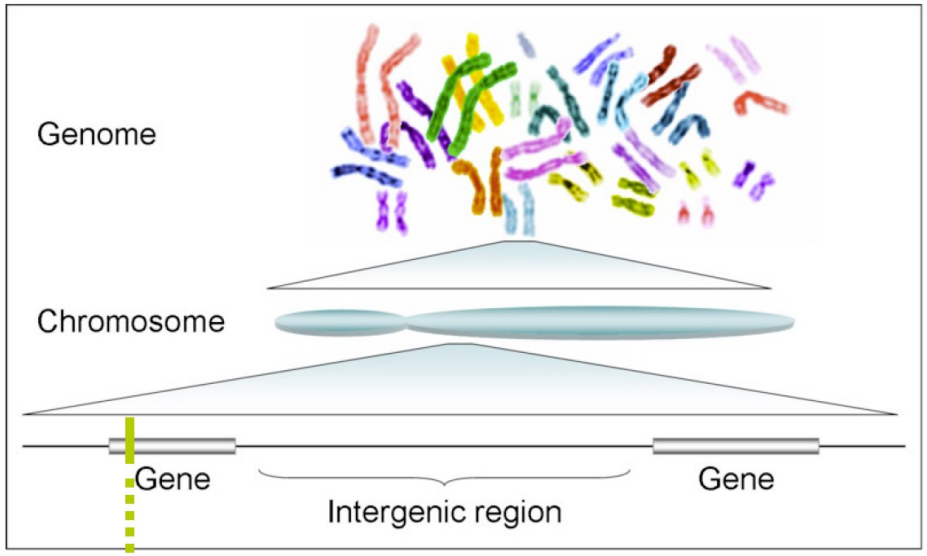


DNA re-sequencing

- Variant Calling
- Medical genomics
  - Cancer genomics

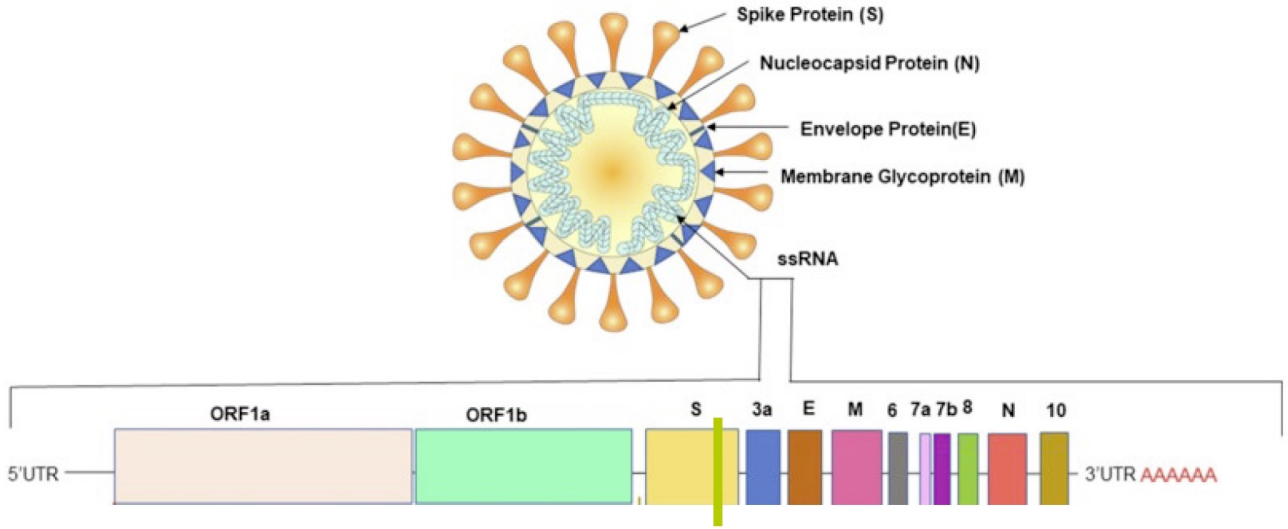
Genome Variation

Human Genome



TP53 mutation
predisposing to cancer

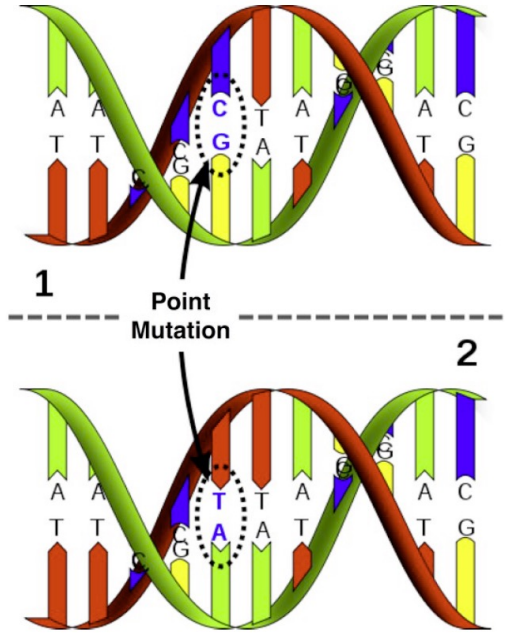
SARS-CoV-2 Genome

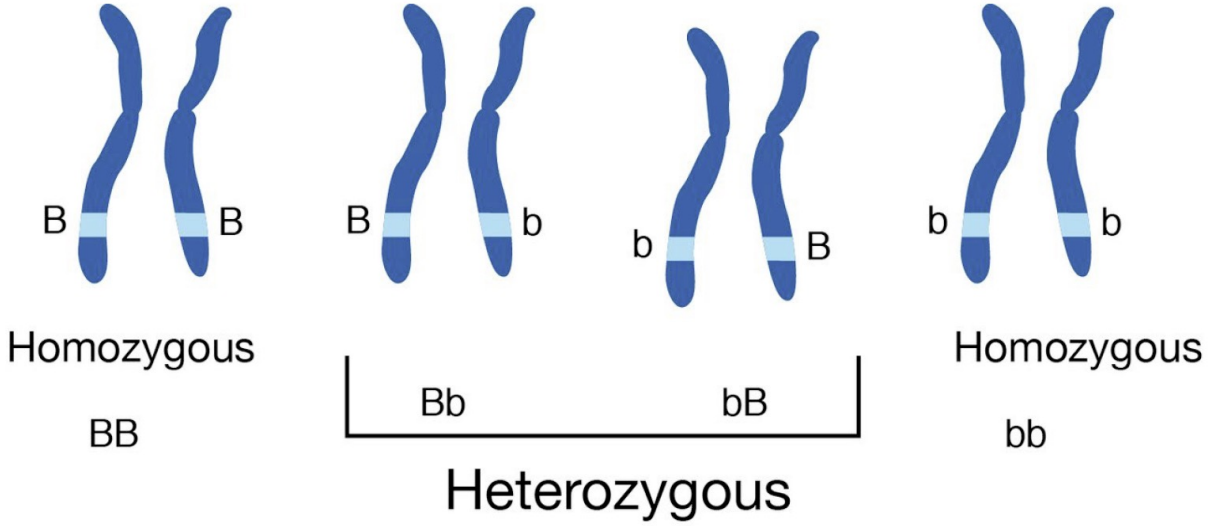


S gene mutation
increasing virus infectiousness

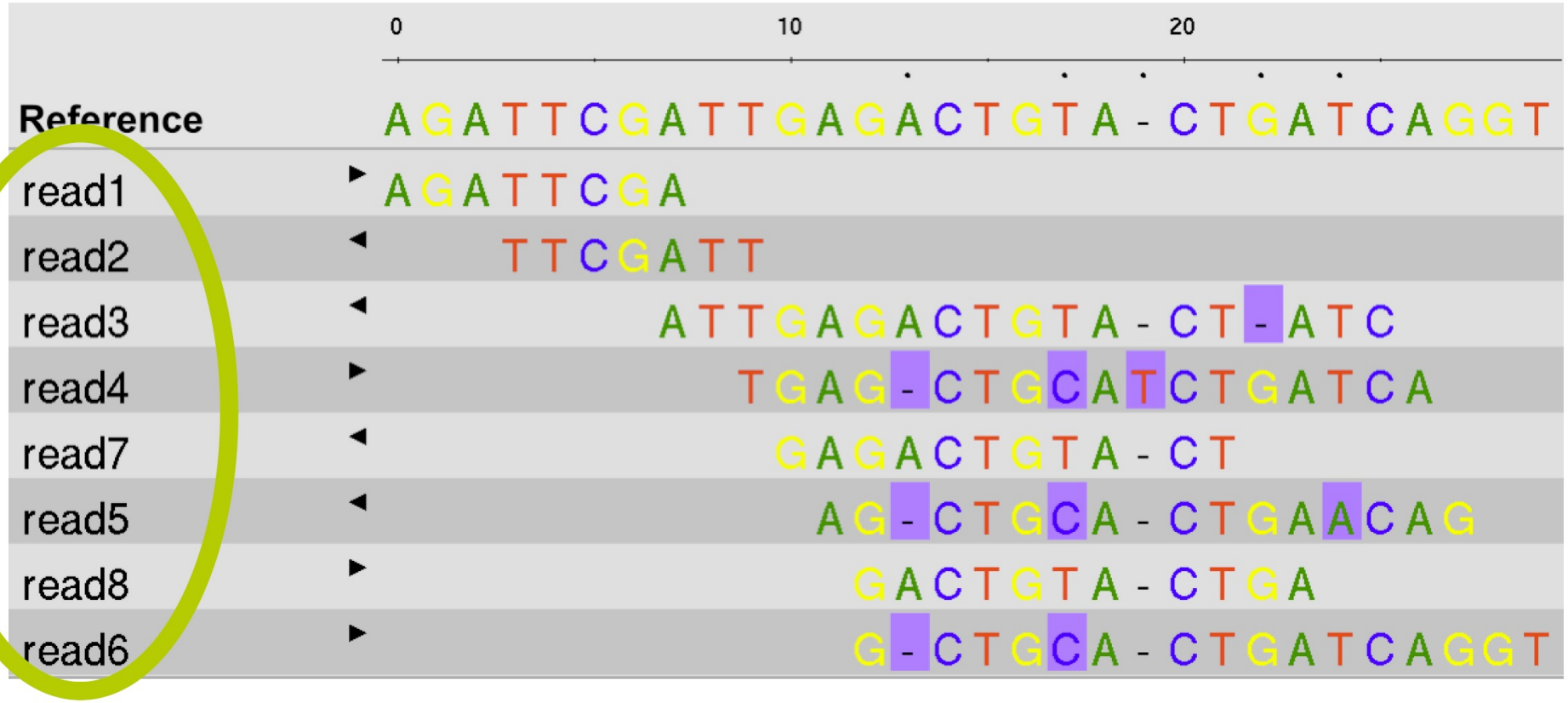
Human Genome Variation

- Humans genomes are >99% similar by sequence
- A typical human genome has ~5 million variants with 3-4 million single-nucleotide variants
- Humans are diploid





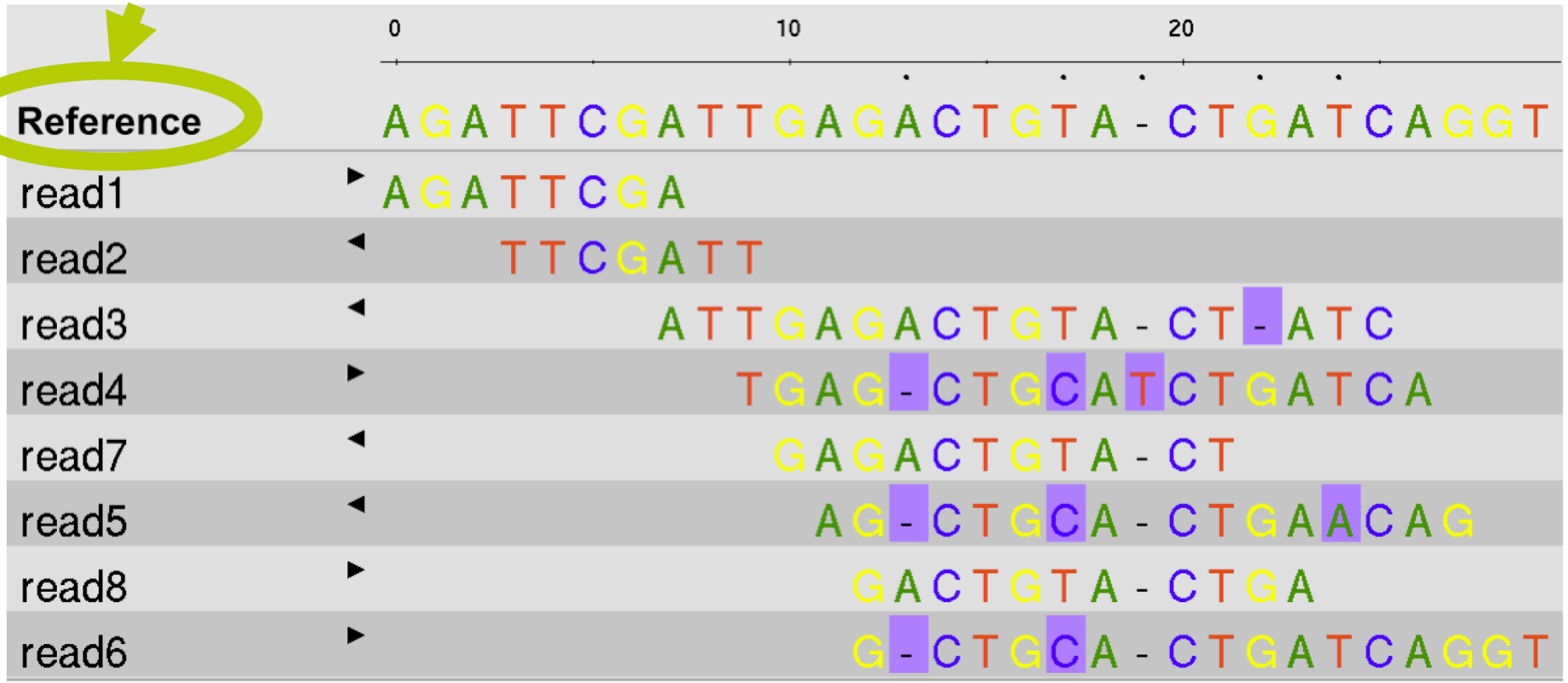
Read Alignment



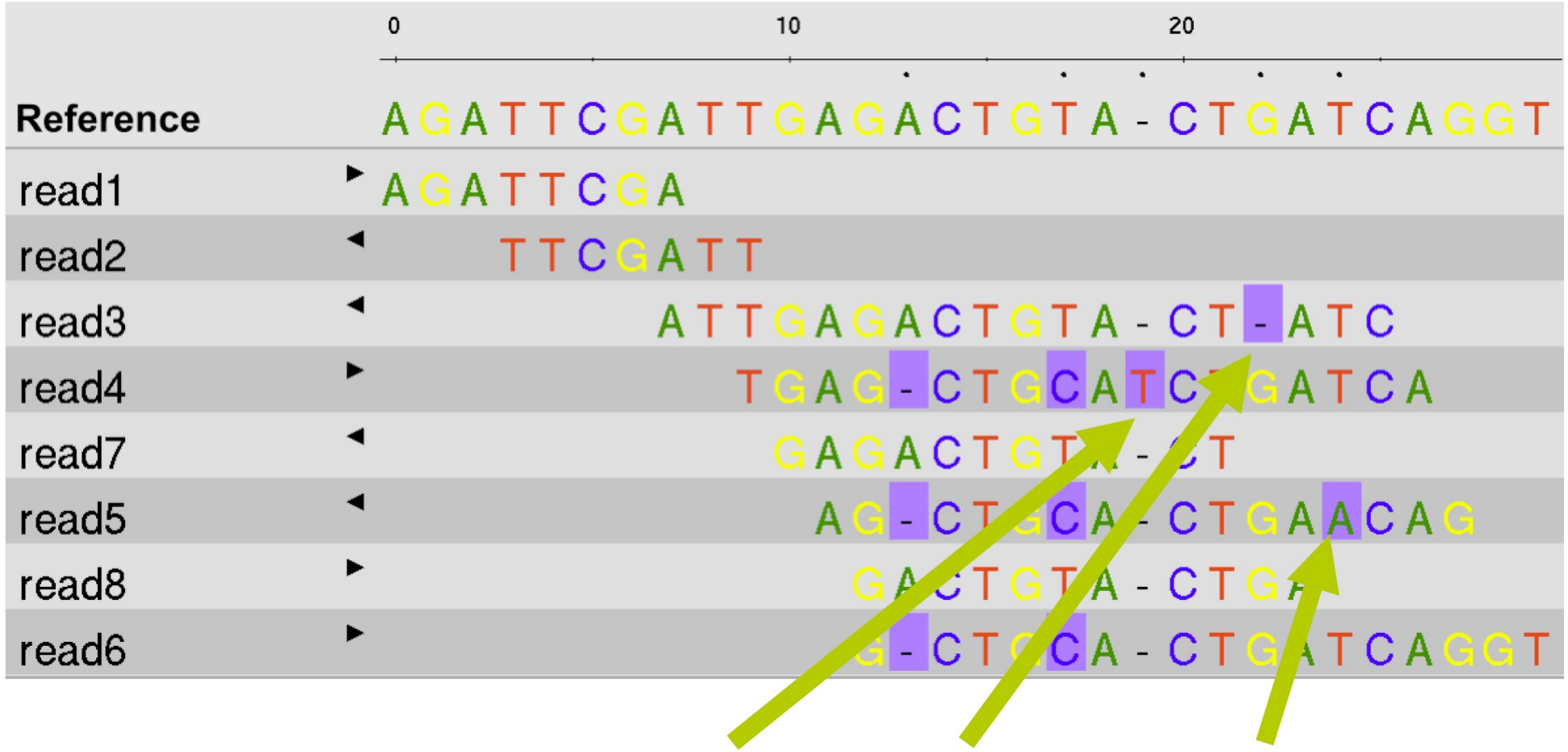
Set of reads

Read Alignment

Reference sequence

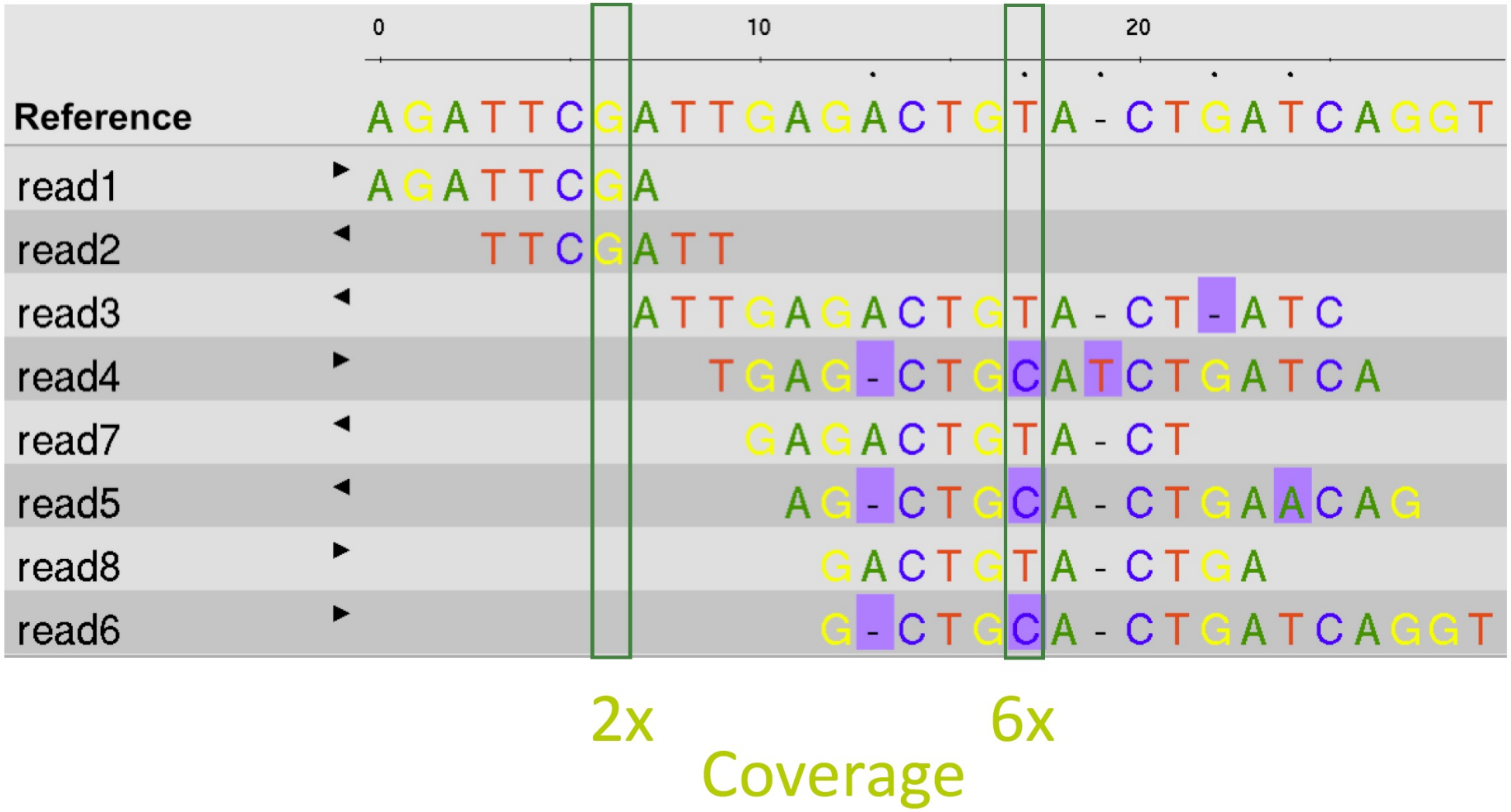


Sequencing Errors

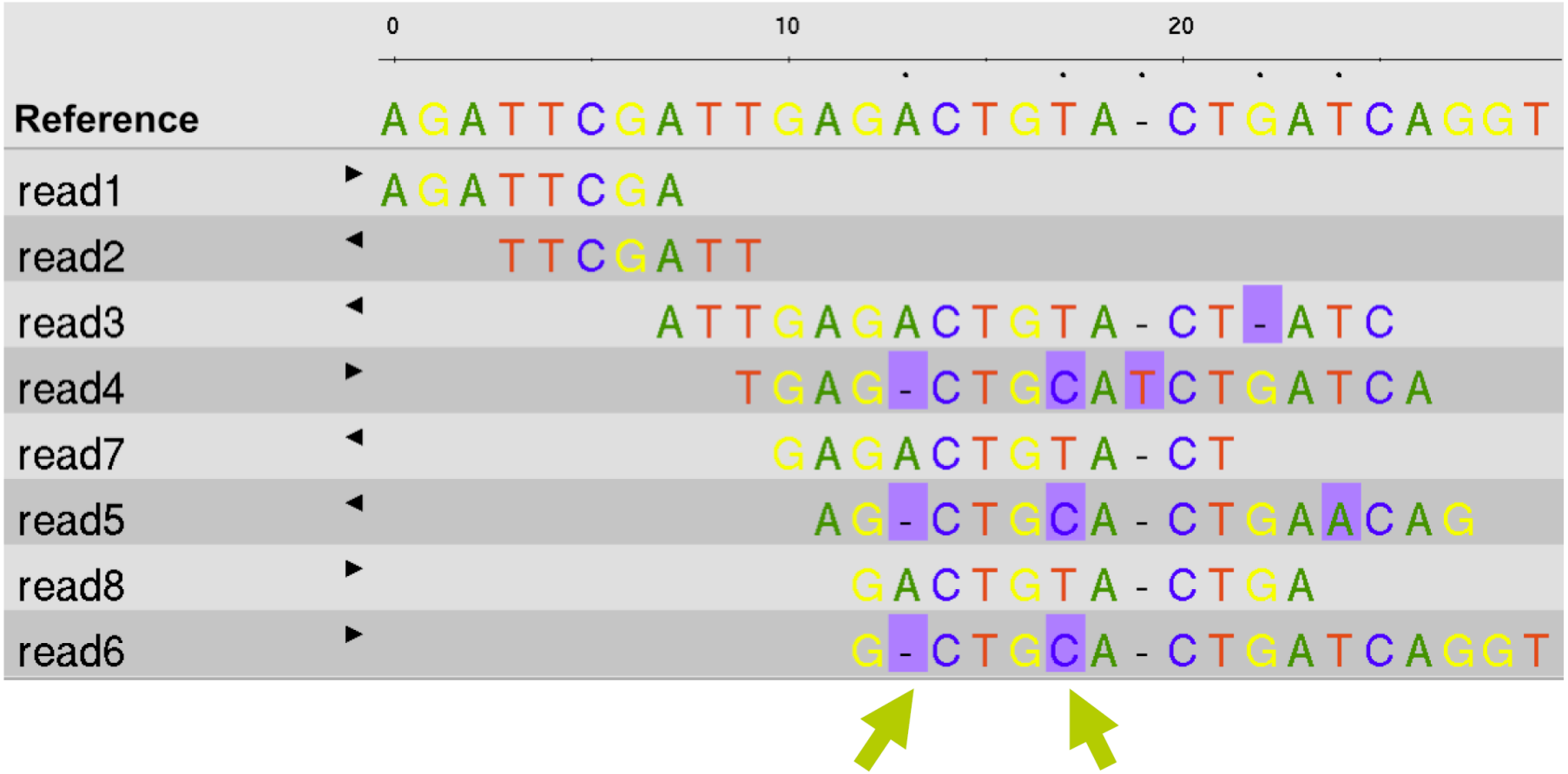


Sequencing errors: Insertions, deletions & basecalling errors

Sequencing Coverage



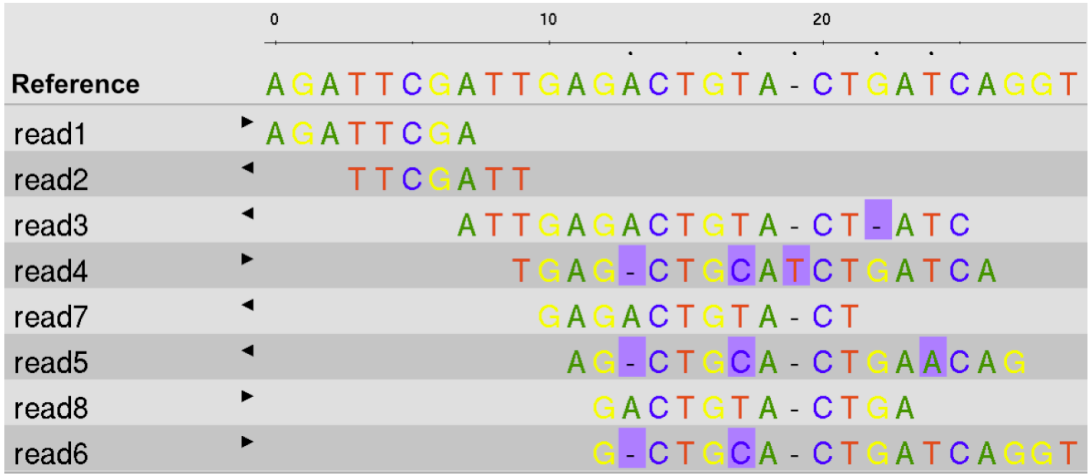
DNA Variant Detection



Variations: Deletion & Single-nucleotide variant

Variant Calling - Data Transformation

Alignment



SAM/BAM file

List of Variants

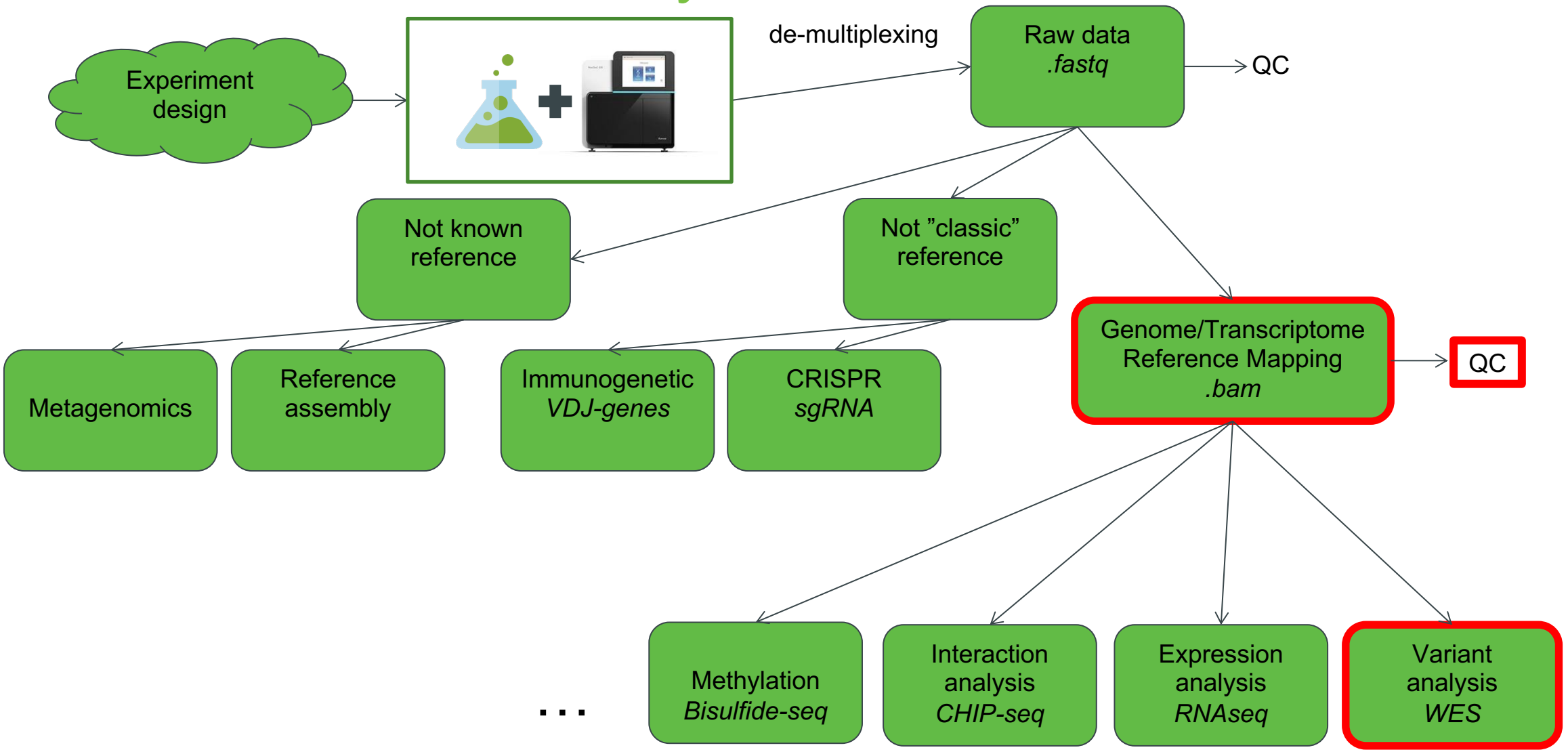
CHR	POS	ID	REF	ALT	GT
chr1	12	.	GA	G	0/1
chr1	17	rs123	T	C	0/1

Genotype (GT):

0/0: hom. reference
0/1: heterozygous
1/1: hom. alternative

VCF/BCF file

NGS data analysis



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

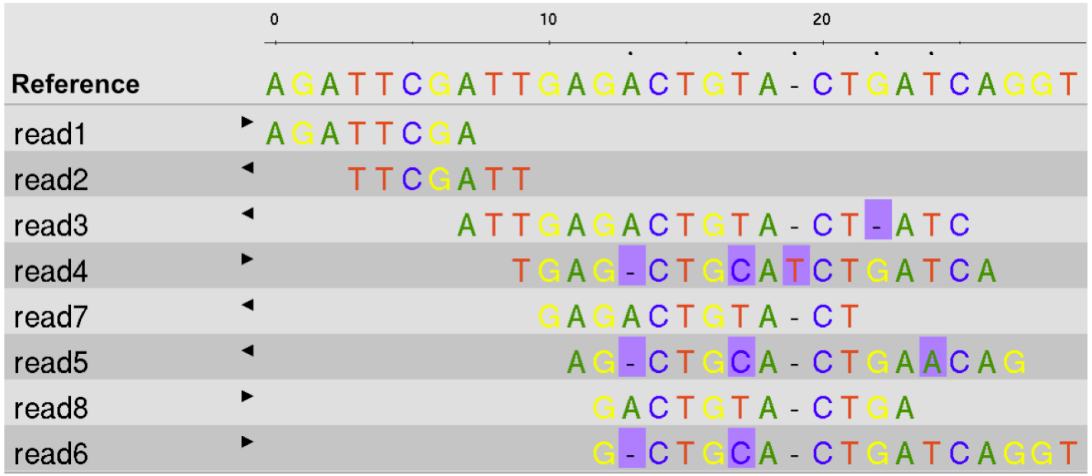
General Statistics

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

Variant Calling - Data Transformation

Alignment



SAM/BAM file

List of Variants

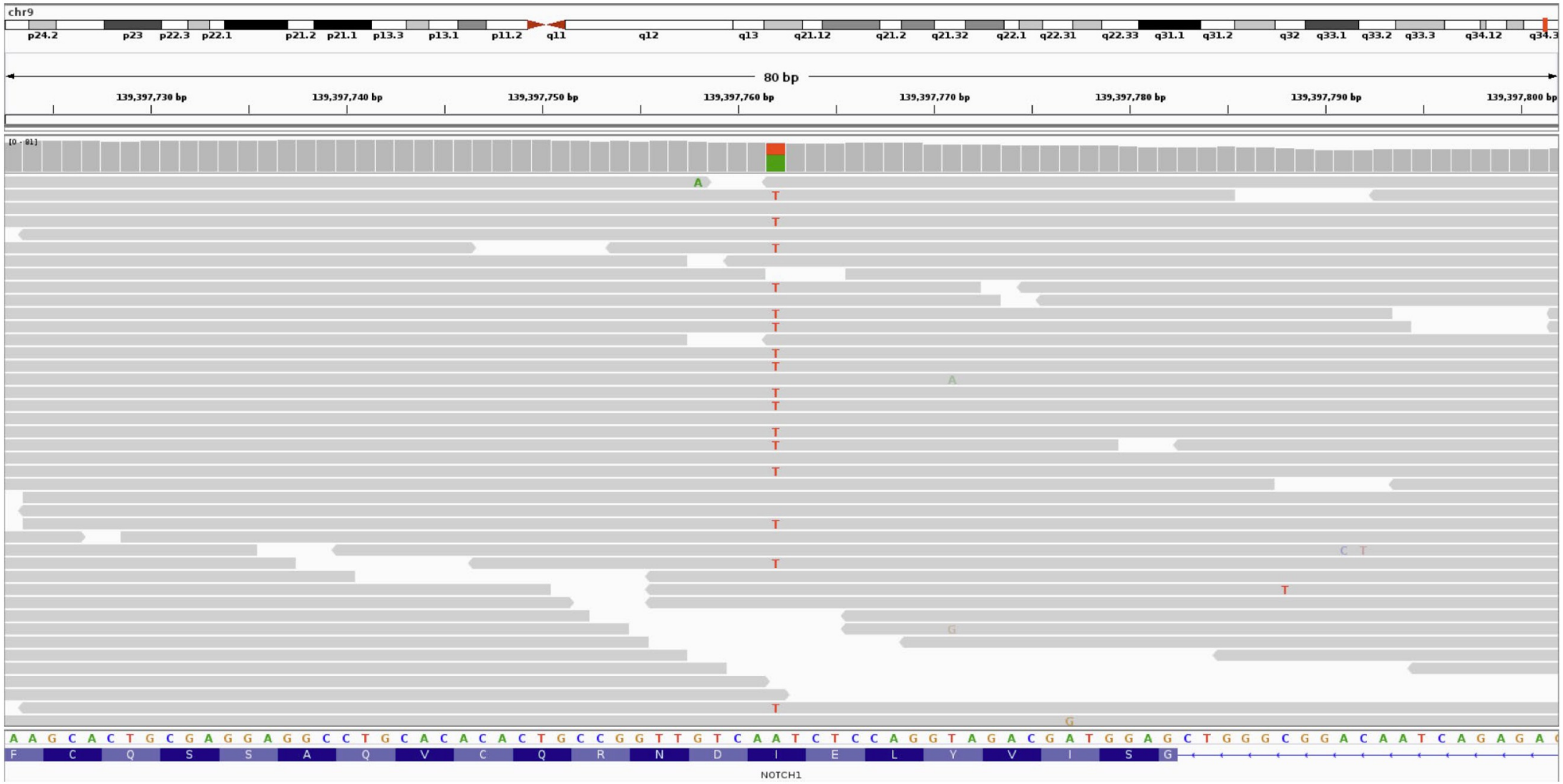
CHR	POS	ID	REF	ALT	GT
chr1	12	.	GA	G	0/1
chr1	17	rs123	T	C	0/1

Genotype (GT):

0/0: hom. reference
0/1: heterozygous
1/1: hom. alternative

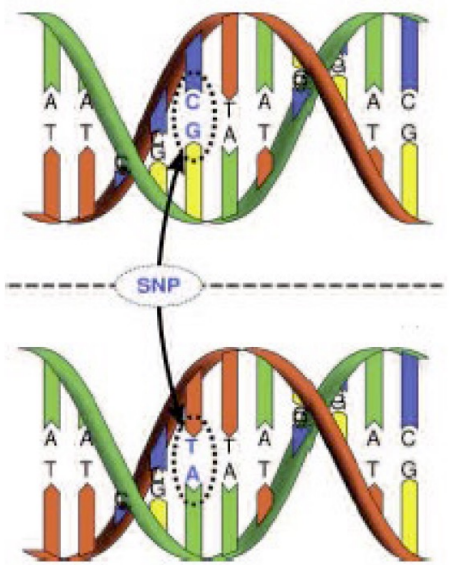
VCF/BCF file

Alignment and variant viewers

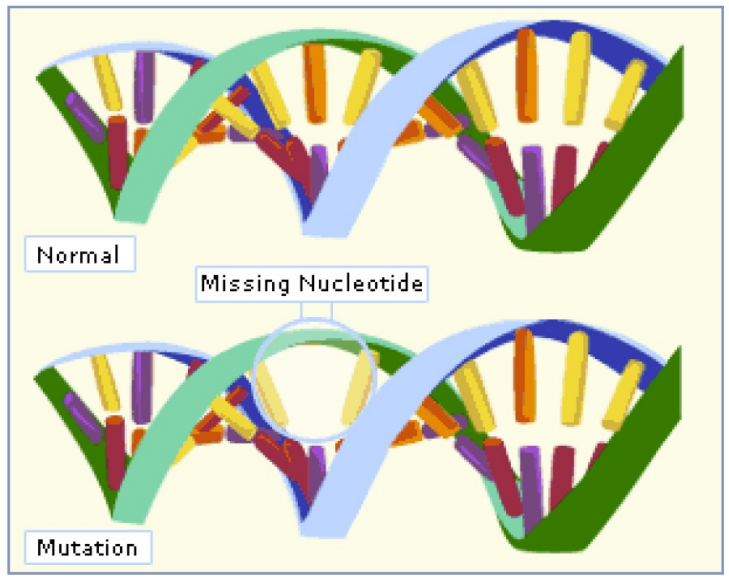


Types of Variants

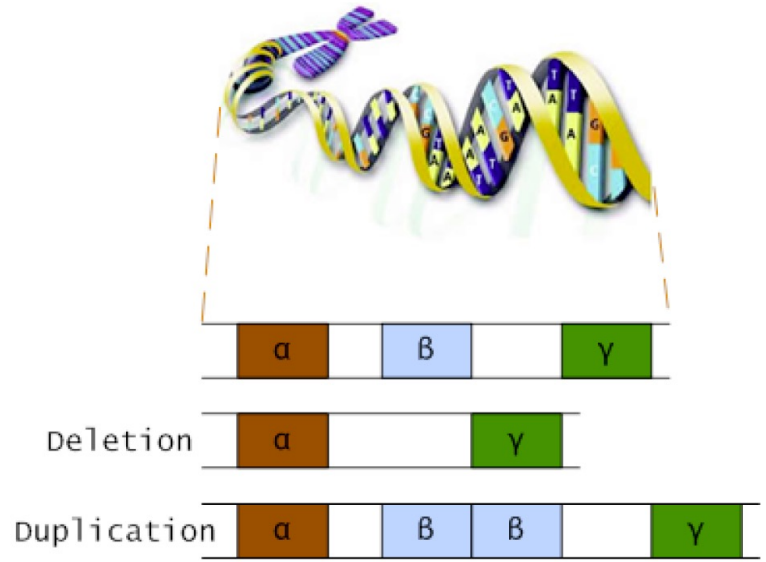
Single-nucleotide variants (SNVs)



Short insertions & deletions (InDels)

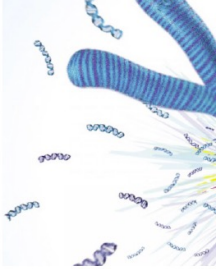


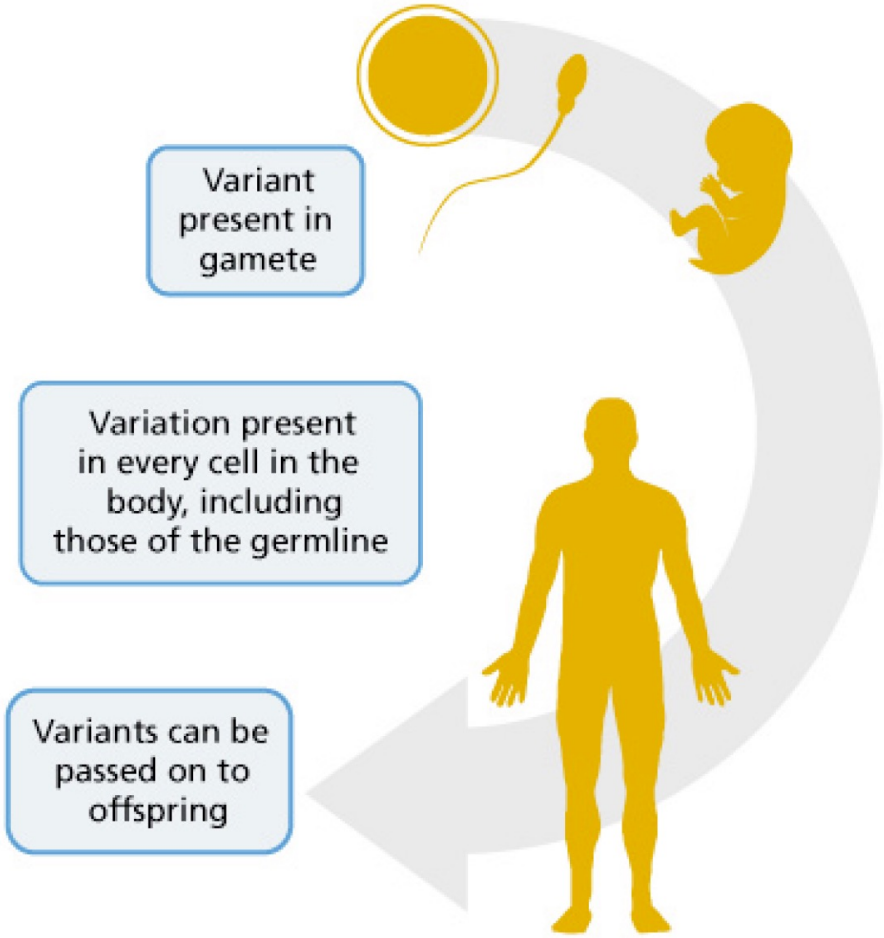
Copy-Number Variants (CNVs) & Structural Variants (SVs)



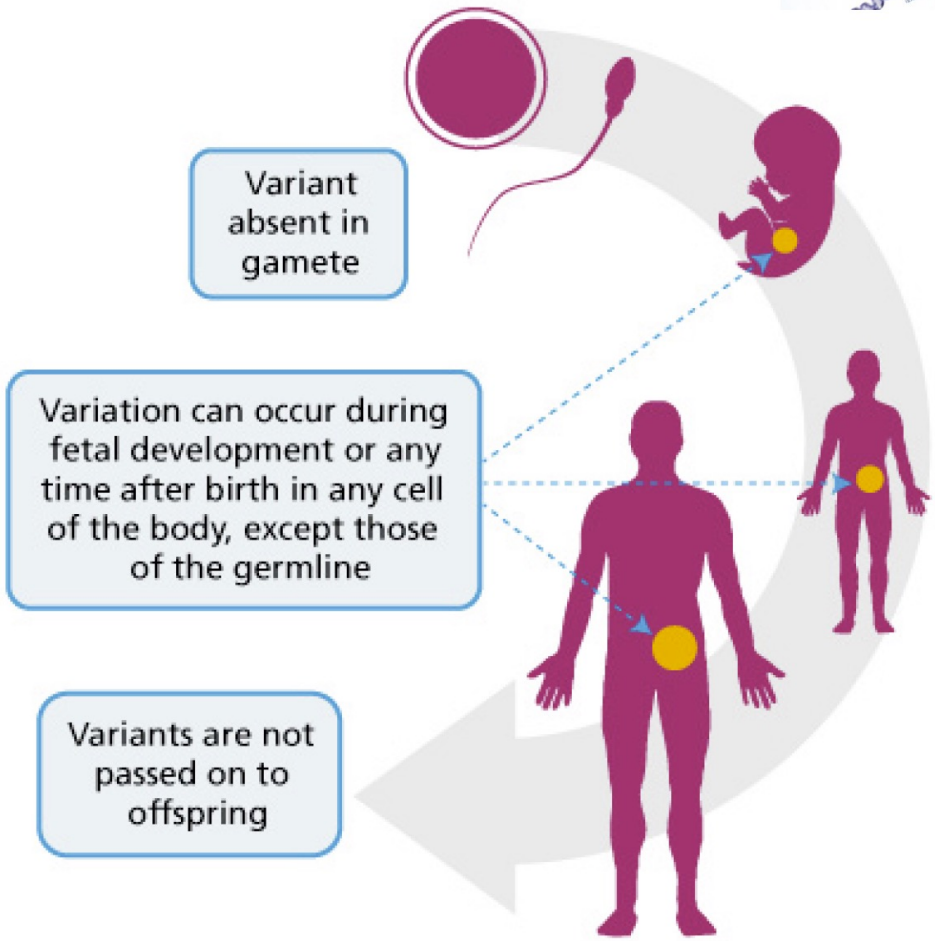
Different methods are used to discover and genotype SNVs, InDels, SVs and CNVs.

Germline and Somatic Variants
„Cancer is a disease of the genome“



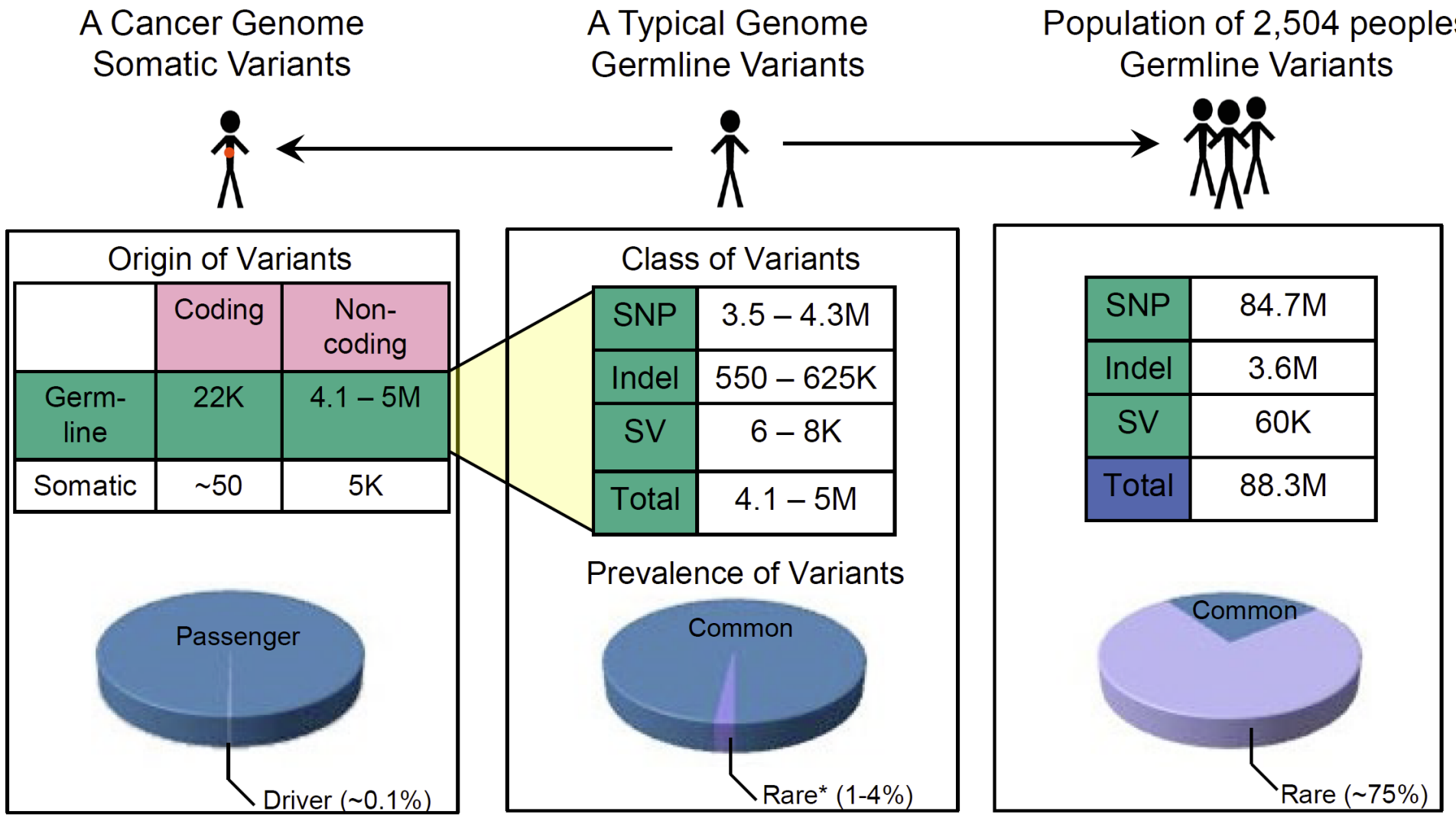


Germline variants



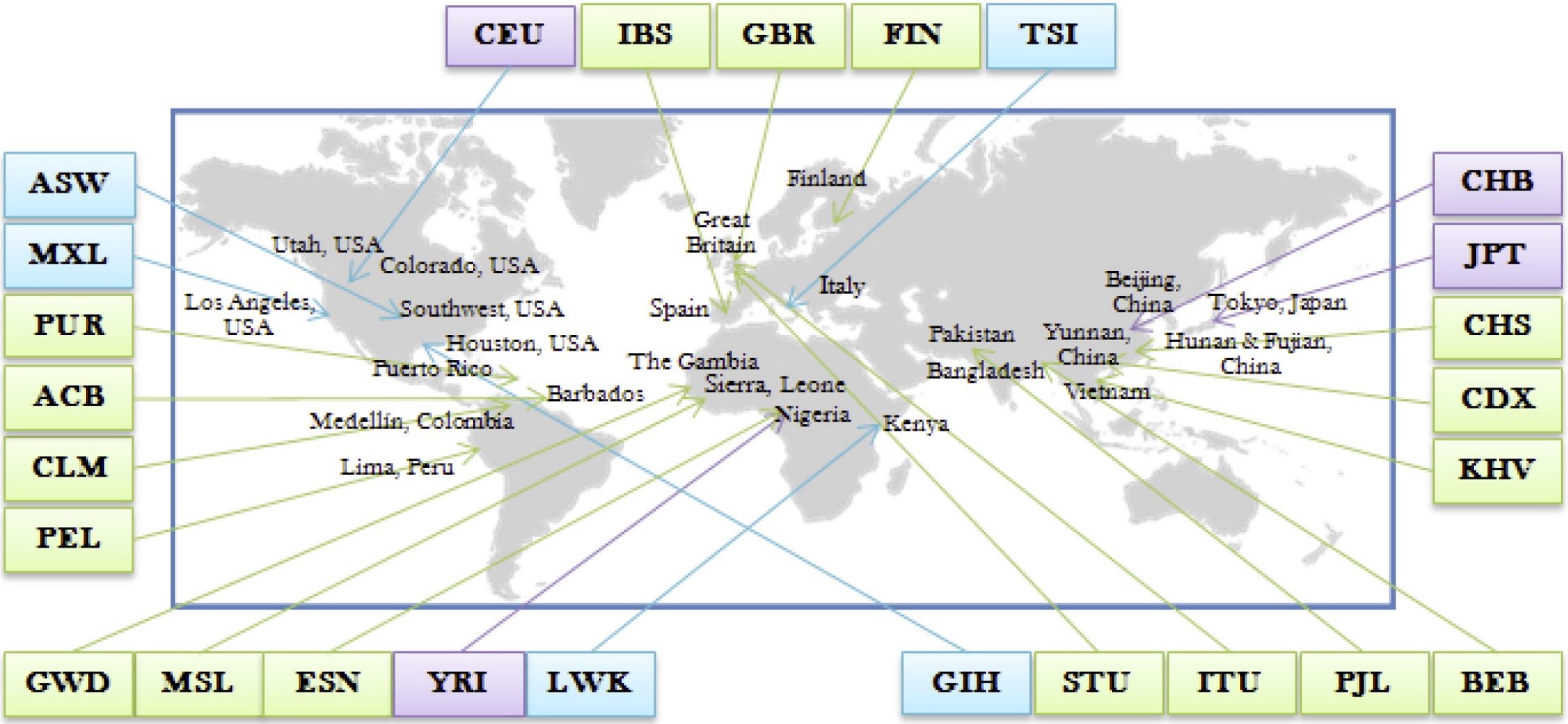
Somatic variants

Human Genome Variation

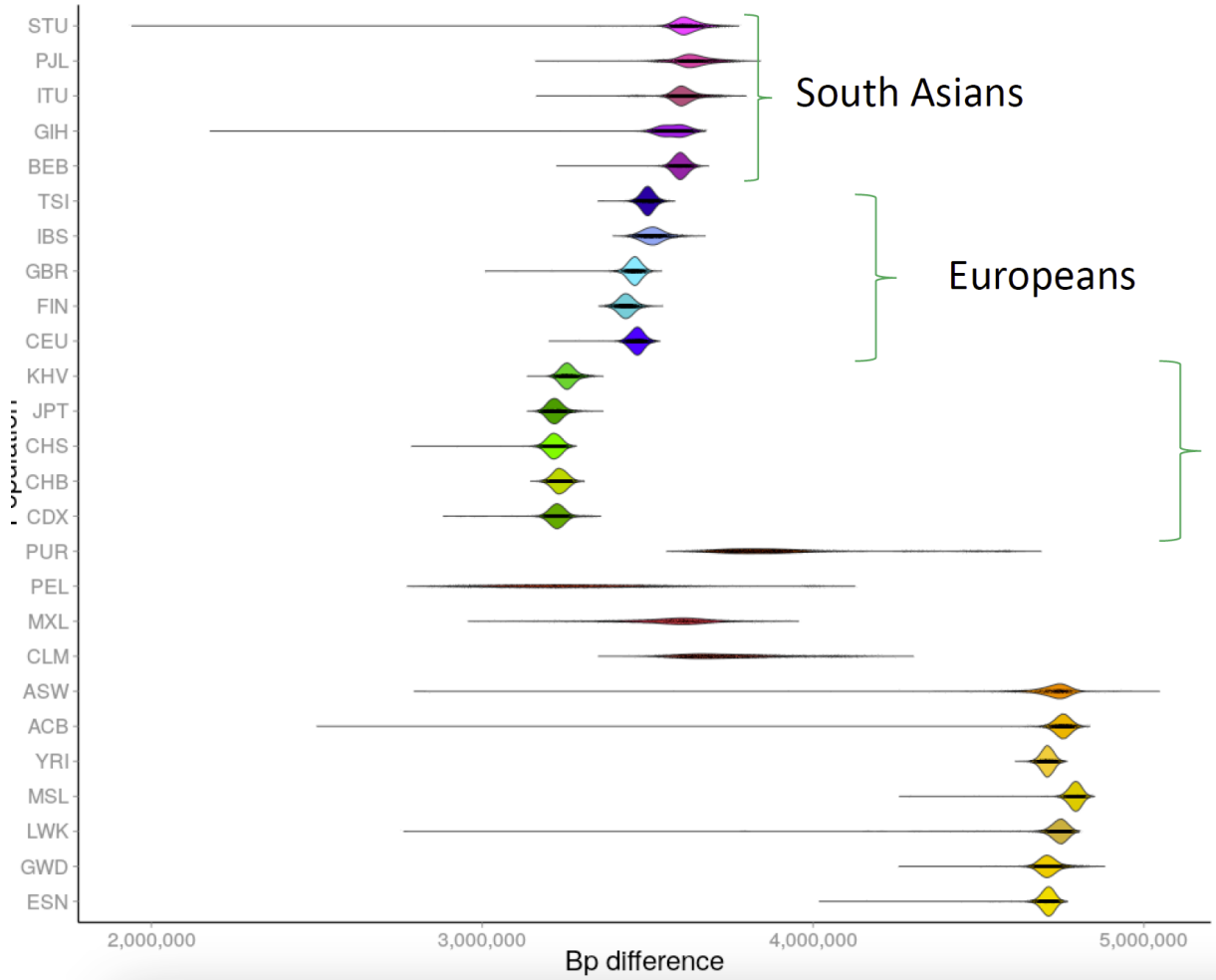


* Variants with allele frequency < 0.5% are considered as rare variants in 1000 genomes project.

1000 Genomes Project Populations



Basepair Difference between 2 Individuals (SNPs only)



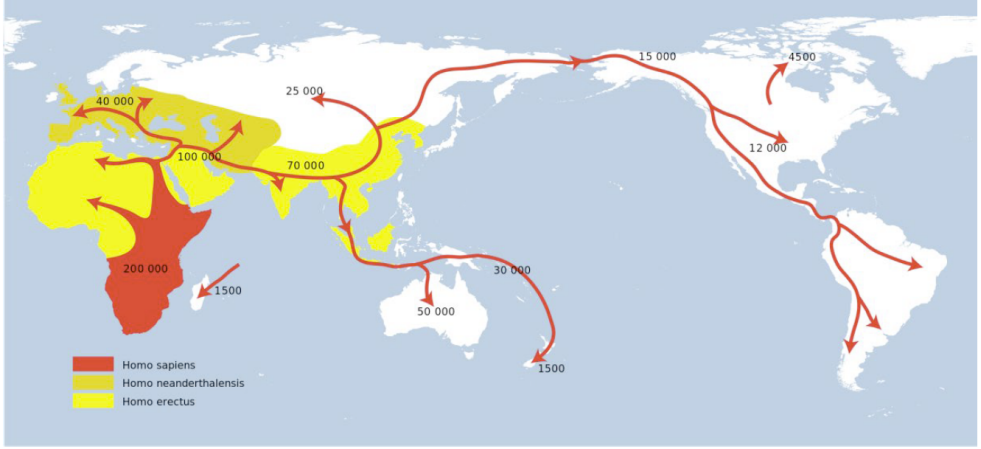
South Asians

Europeans

East Asiens

Admixed Populations

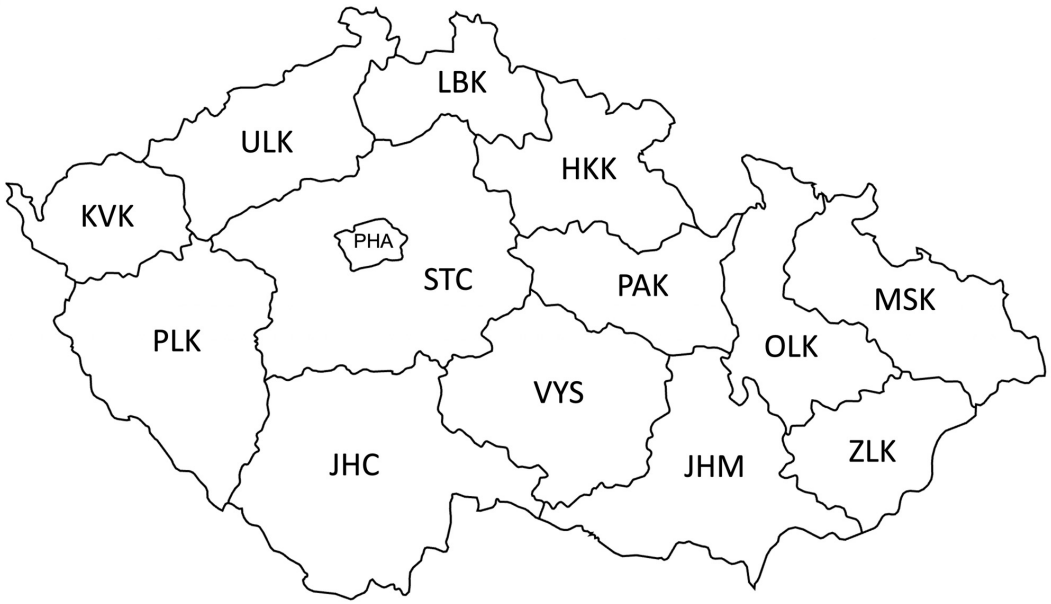
Africans

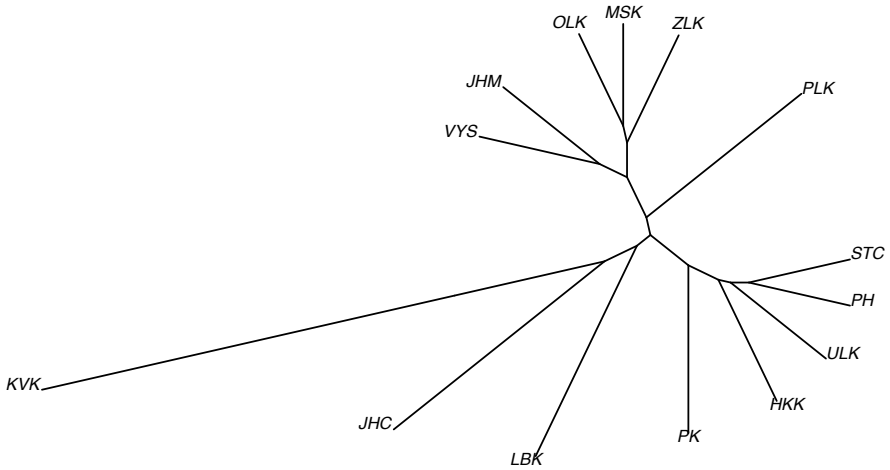


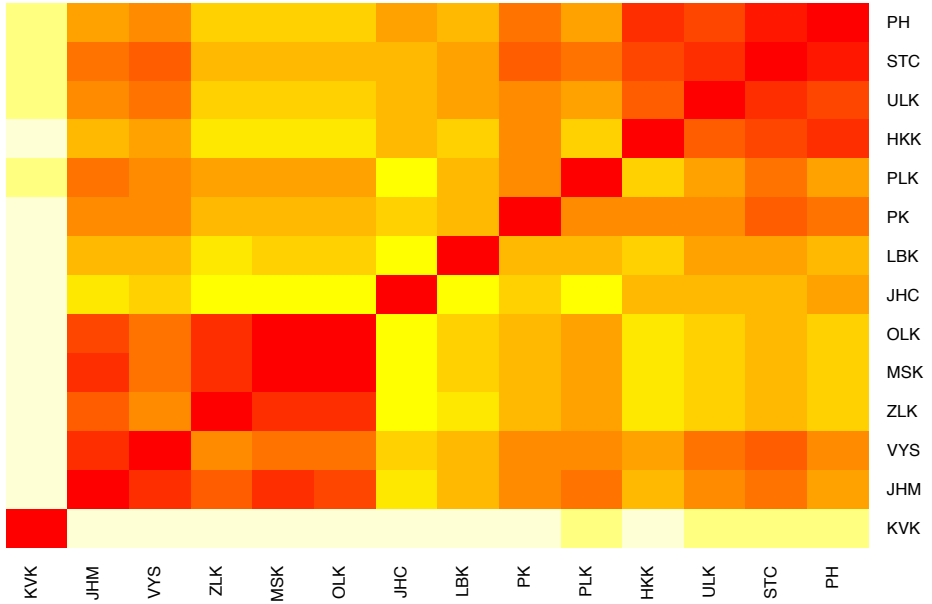
Out of Africa Migration

EMBL

Klastrování regionů podle AF

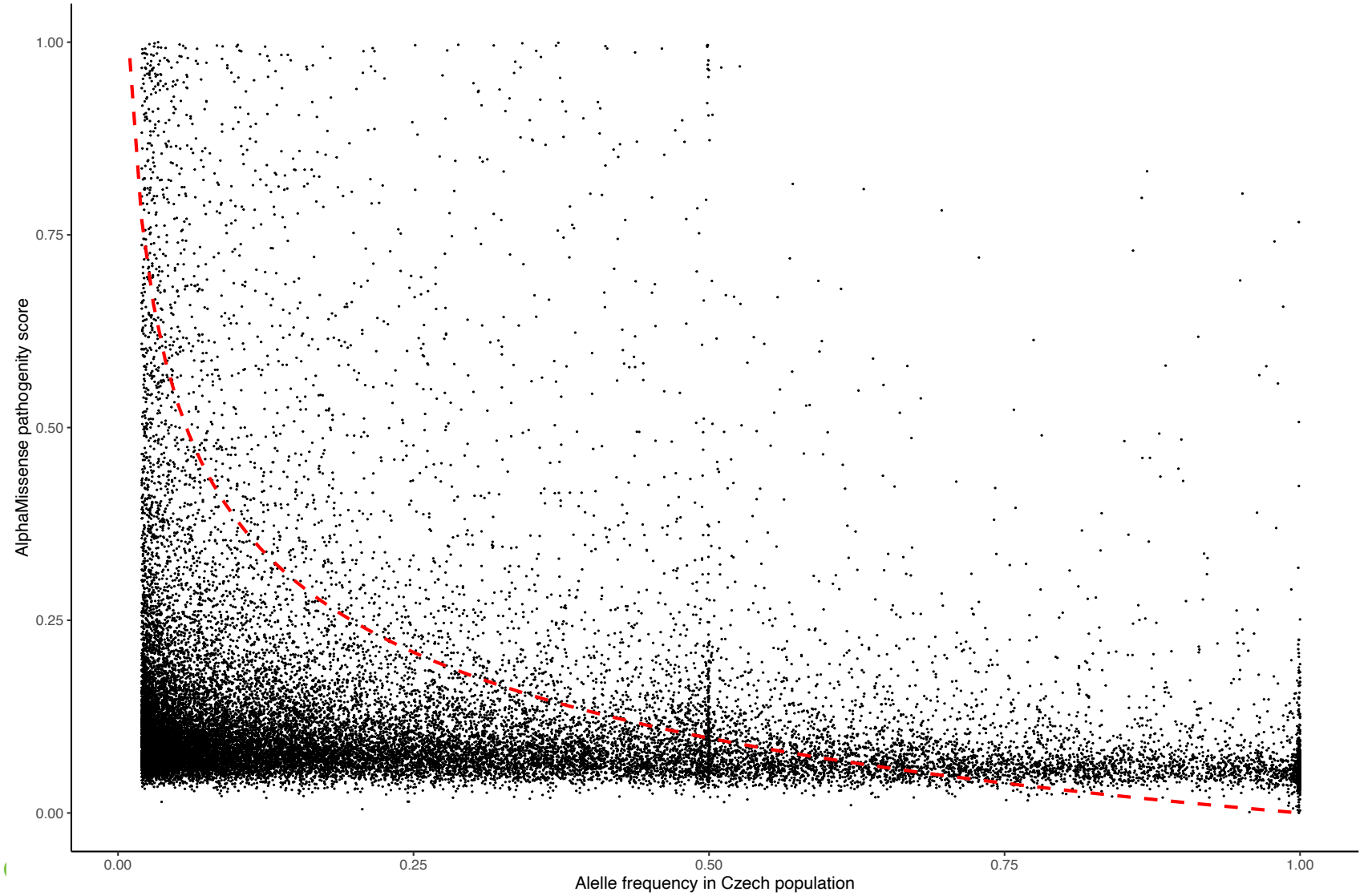






CEITEC

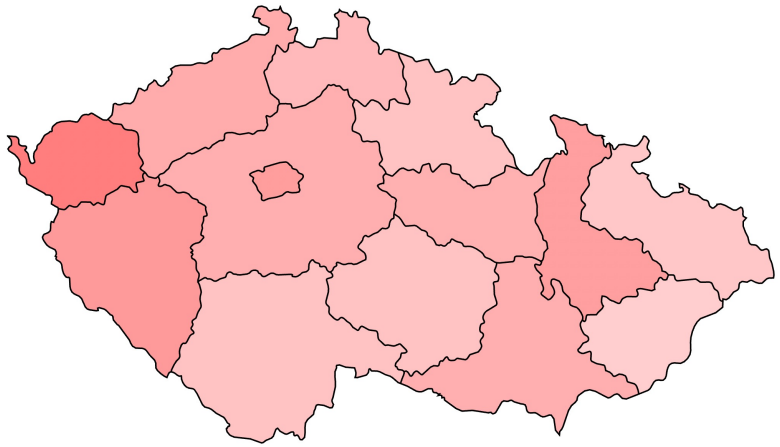
Korelace patogenity a AF

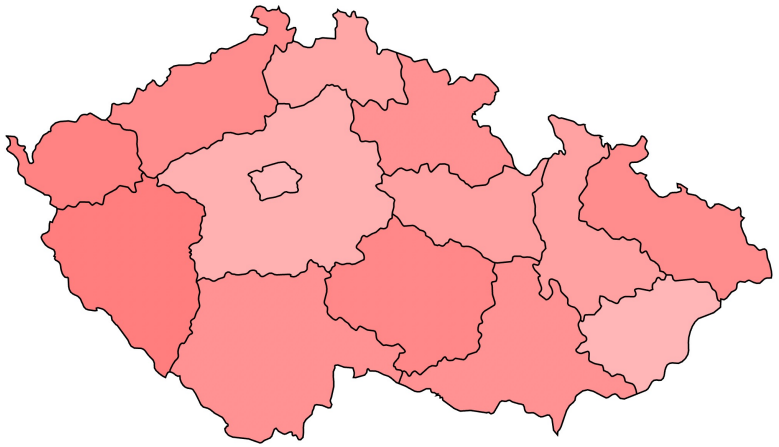


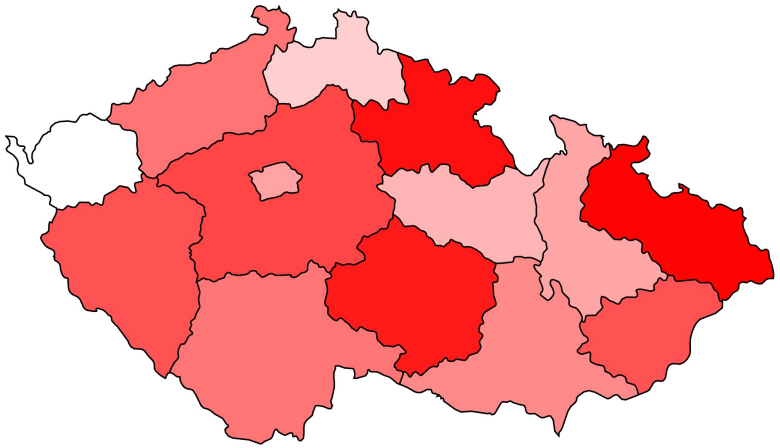
Zajímavé varianty - F5 - Trombofilie

Gene_name	chrom	pos	ref	alt	HGVSc	HGVSp
F5	chr1	169542640	T	G	c.2450A>C	p.N817T
F5	chr1	169514323	T	C	c.6665A>G	p.D2222G
F5	chr1	169549811	C	T	c.1601G>A	p.R534Q

AF_CZE	EUR_AF	EAS_AF	AFR_AF	AMR_AF	SAS_AF
8,4%	6,2%	3,3%	0,5%	9,2%	6,4%
8,2%	6,6%	3,2%	0,2%	9,1%	6,4%
3,8%	1,2%	0,0%	0,0%	1,0%	1,1%

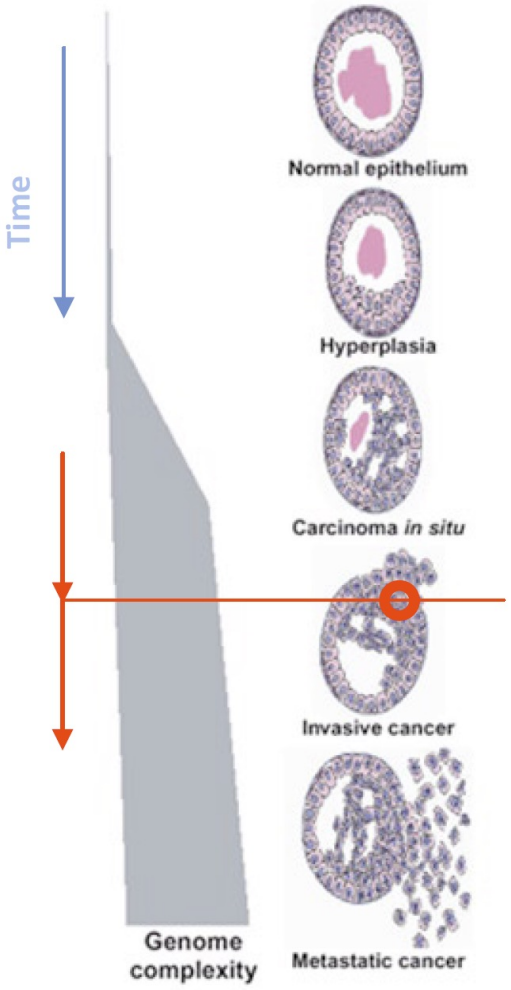






Somatic Variants – Cancer Genome Sequencing

Sequencing provides a Snapshot in Time & Space



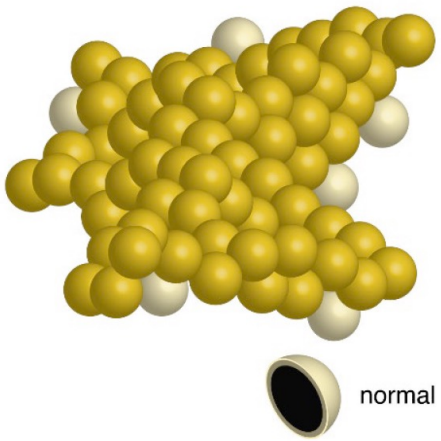
DNA-Sequencing

-Tumor cell content (tumor purity)

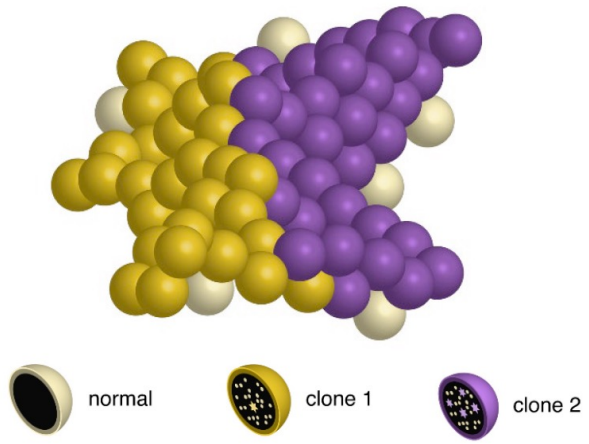
-Tumor heterogeneity (subclonality)

-Tumor ploidy

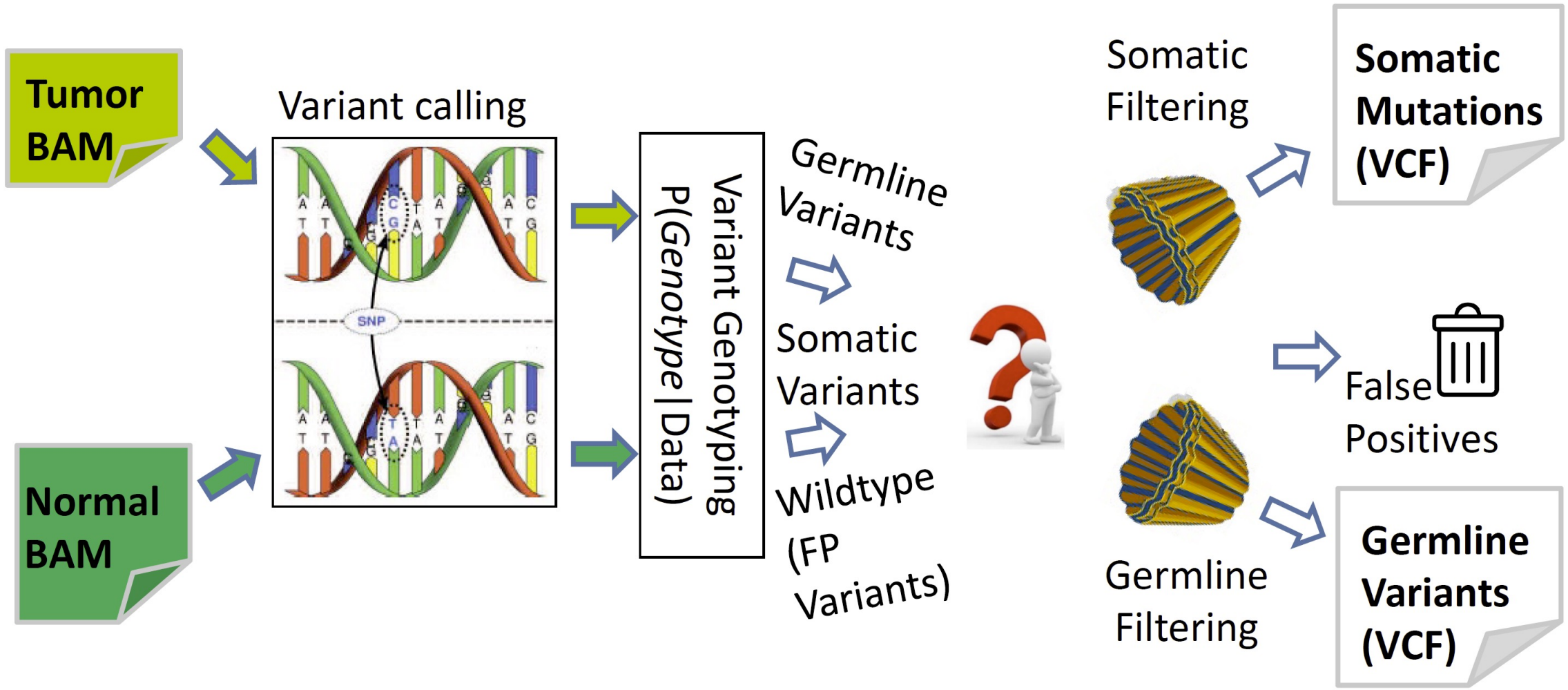
Schematic depiction of a mono-clonal tumor



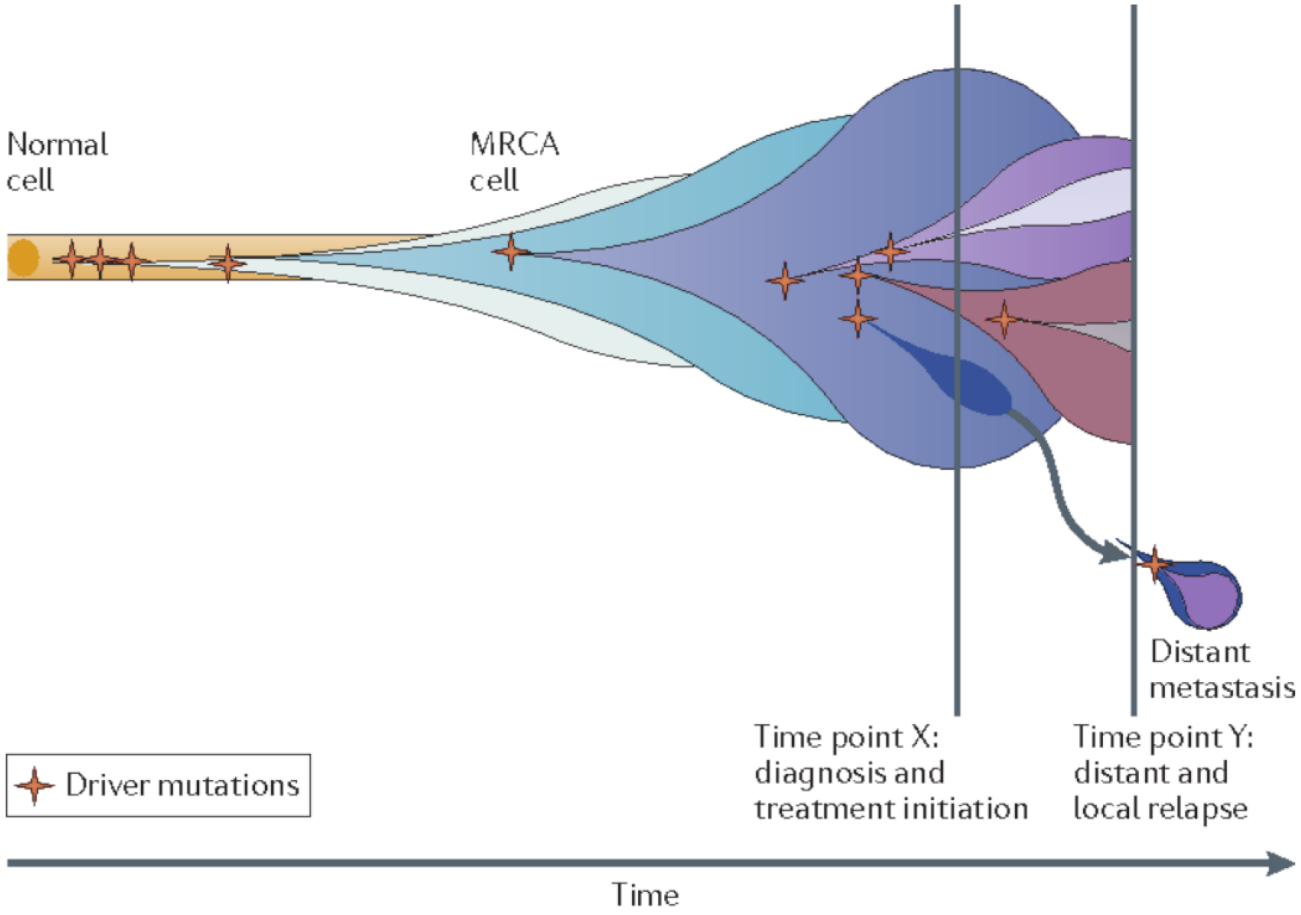
Schematic depiction of a bi-clonal tumor

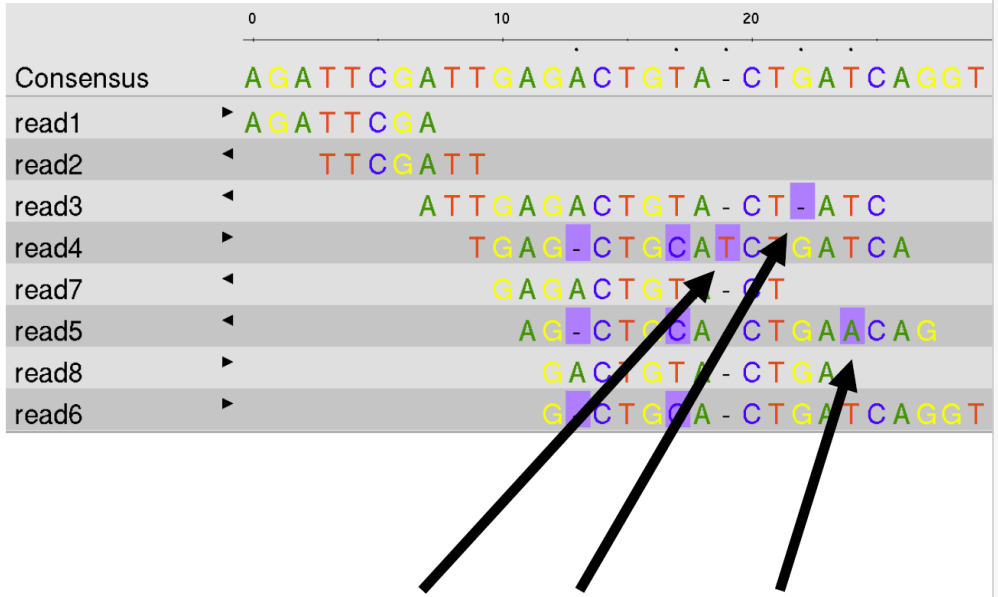


Somatic vs. Germline Variants



Tumor Heterogeneity



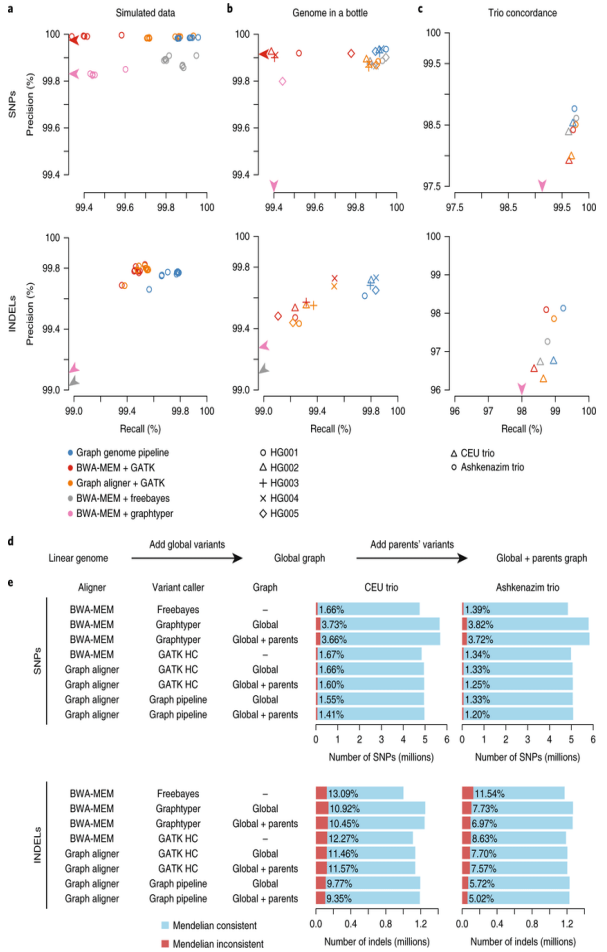


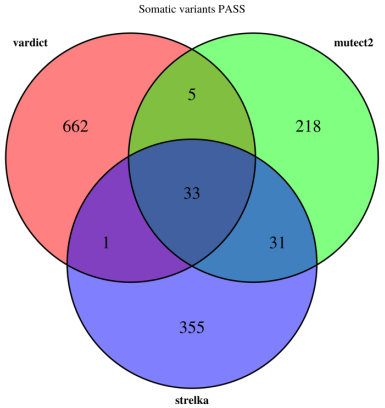
Sequencing errors or possibly subclonal variants?

→ Cancer genomes are often sequenced to >=60x because of tumor purity, tumor heterogeneity and many chromosomal aberrations

Variant Calling - Tools

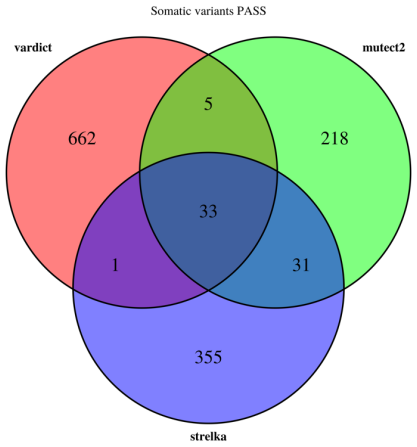
- Multiple tools:
  - strelka2, verdict, mutect2, somaticsniper, lofreq, muse, varscan
- Ensemble/meta callers usually outperformes individual
  - SomaticSeq
- Benchmarking
  - Genome in a Bottle
  - GIAB
  - son/father/mother trios of Ashkenazi Jewish

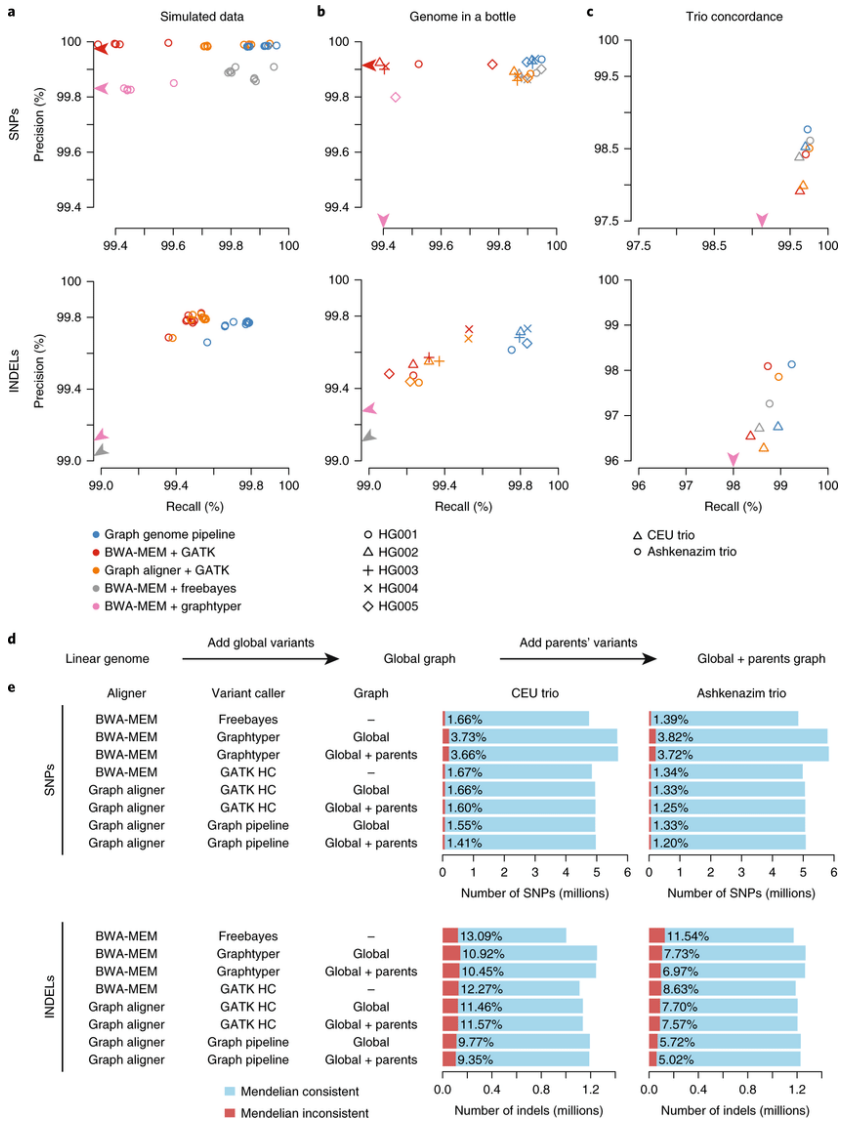




Variant Calling - Tools

- 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

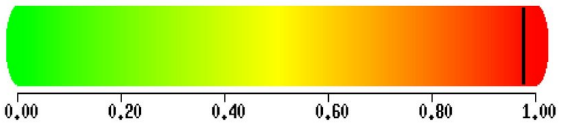
Small Variant annotation – functional prediction

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

POLYPHEN-2

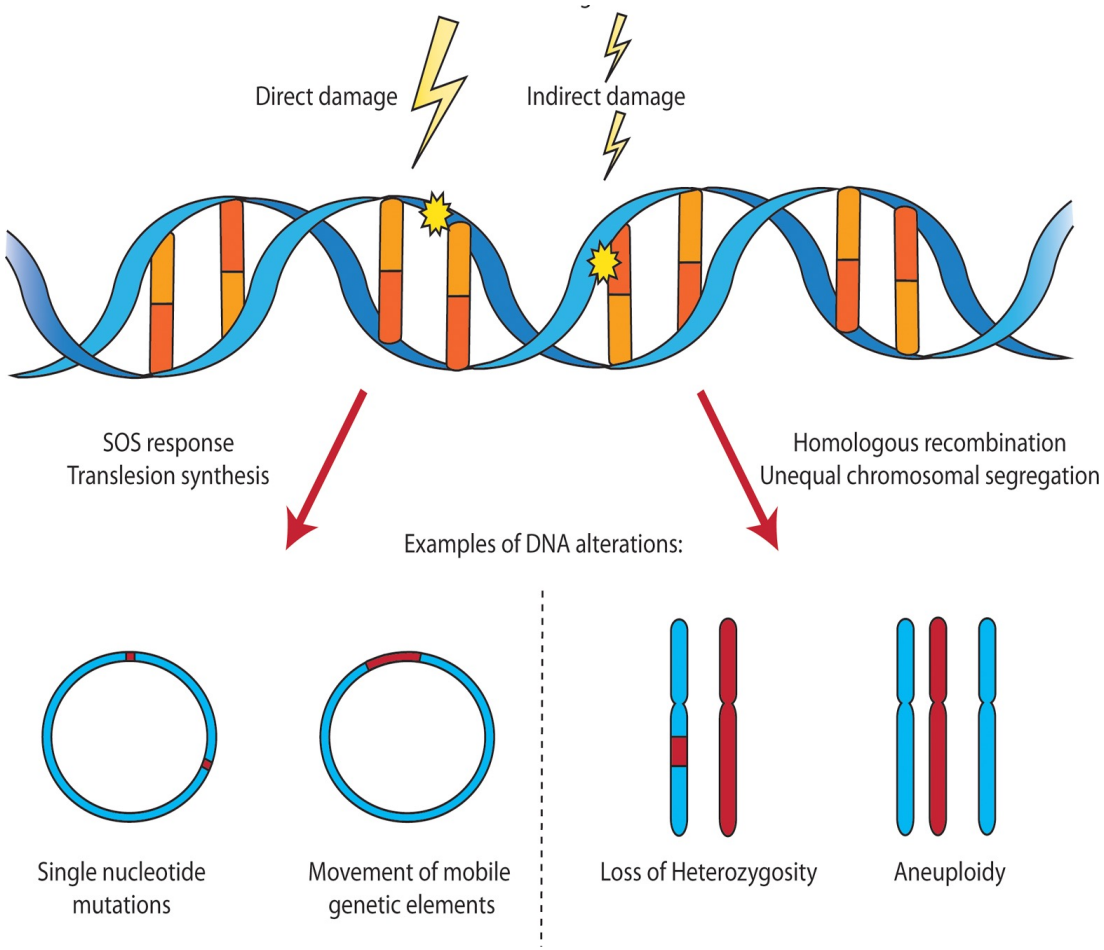
This mutation is predicted to be **PROBABLY DAMAGING** with a score of **0.976**

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



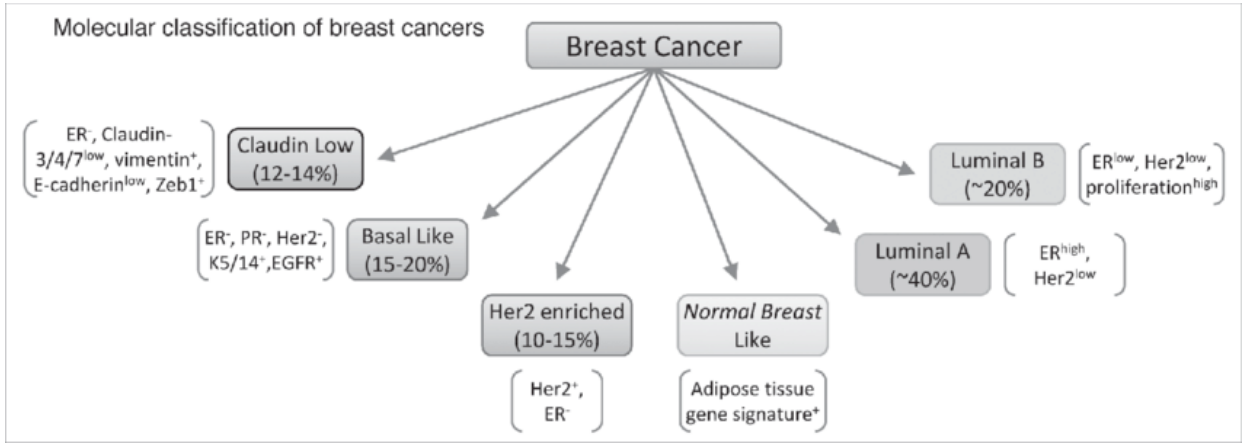
* SO term	SO description	SO accession	Display term	IMPACT
transcript_ablation	A feature ablation whereby the deleted region includes a transcript feature	SO:0001893	Transcript ablation	HIGH
splice_acceptor_variant	A splice variant that changes the 2 base region at the 3' end of an intron	SO:0001574	Splice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



Cancer genomics introduction

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

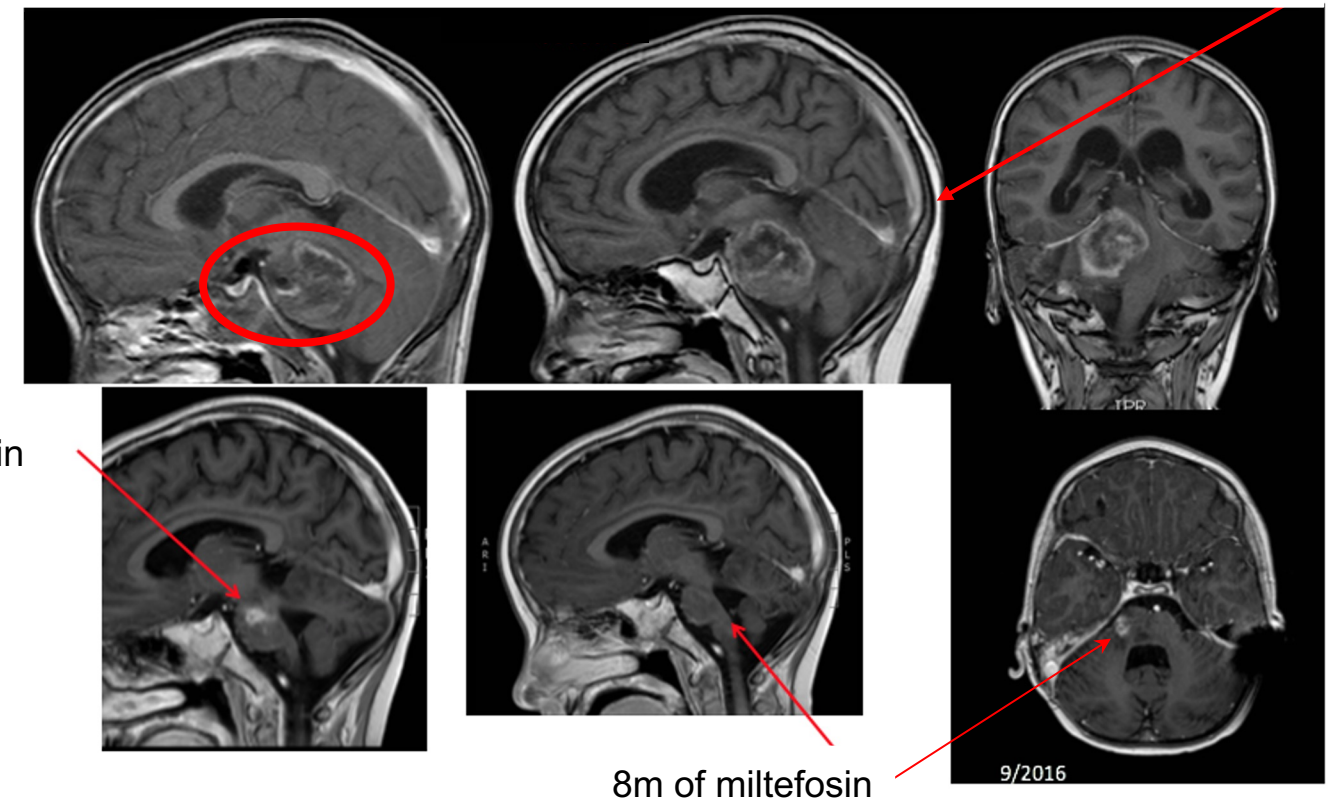


Cancer genomics introduction - Case report

- **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

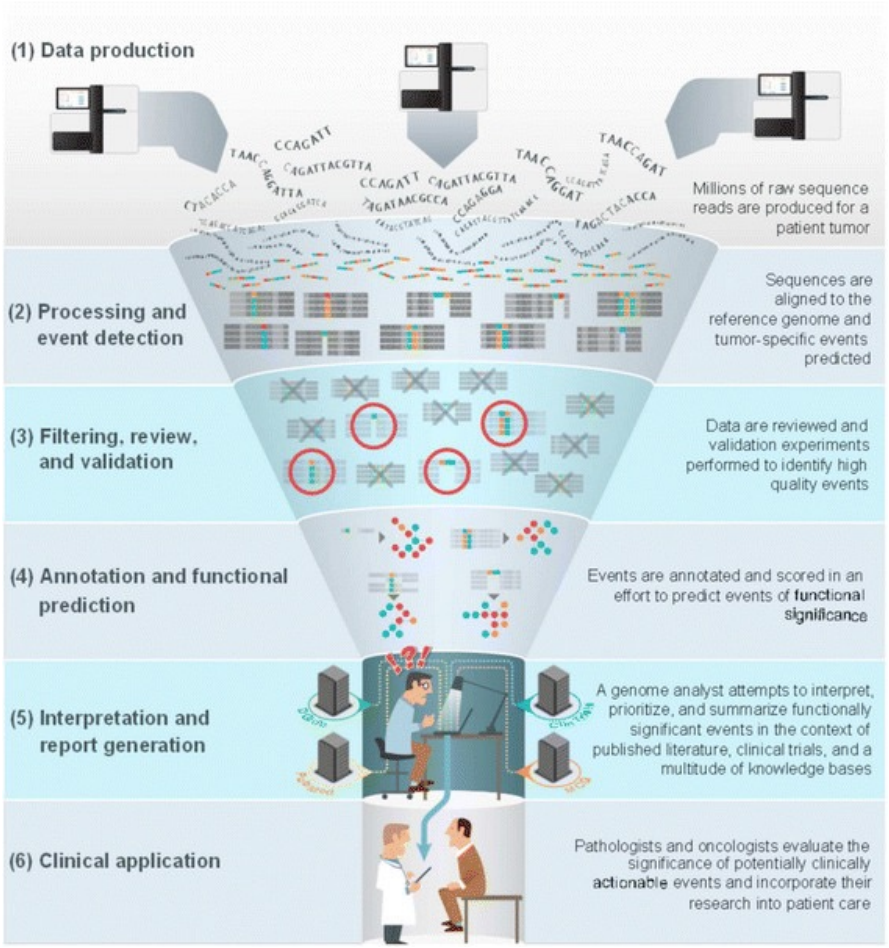


Miltefosin/impavido
(only approved Akt inhibitor)

Leishmaniasis

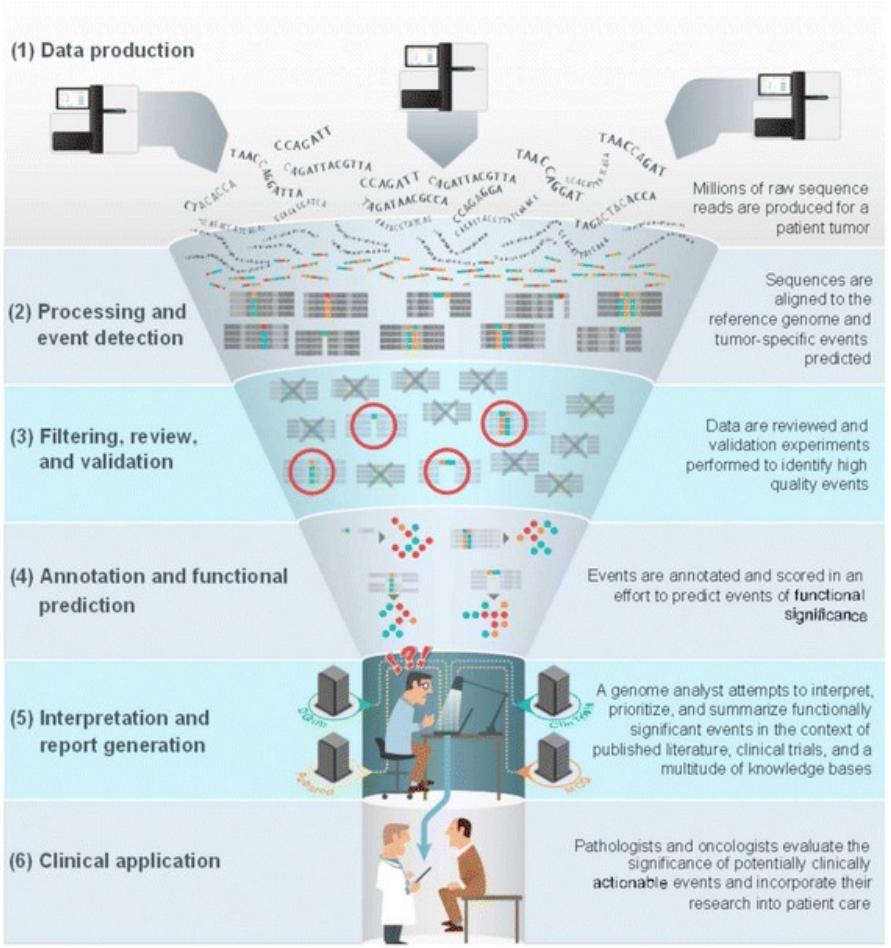
DRUG REPURPOSING

Somatic variant NGS data analysis



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

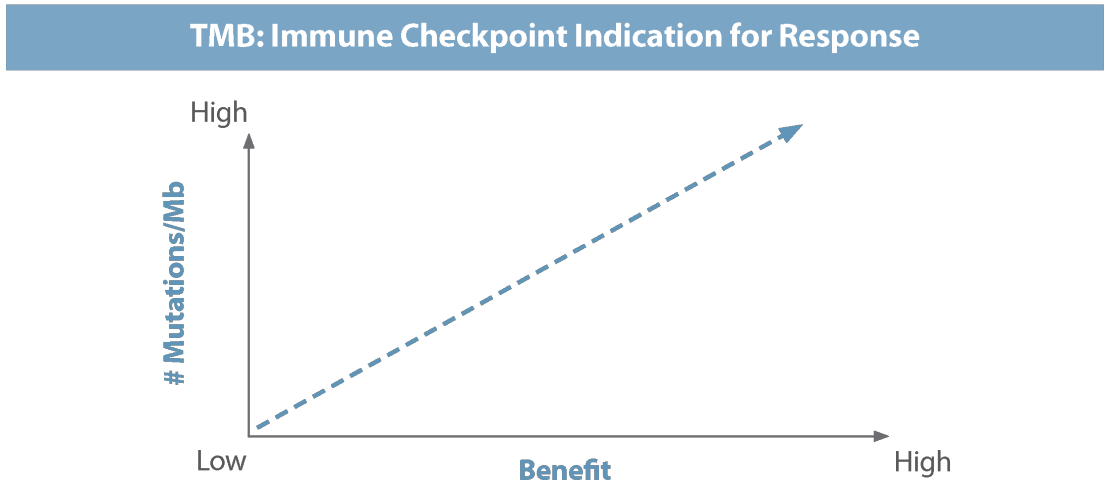
Somatic variant NGS data analysis



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

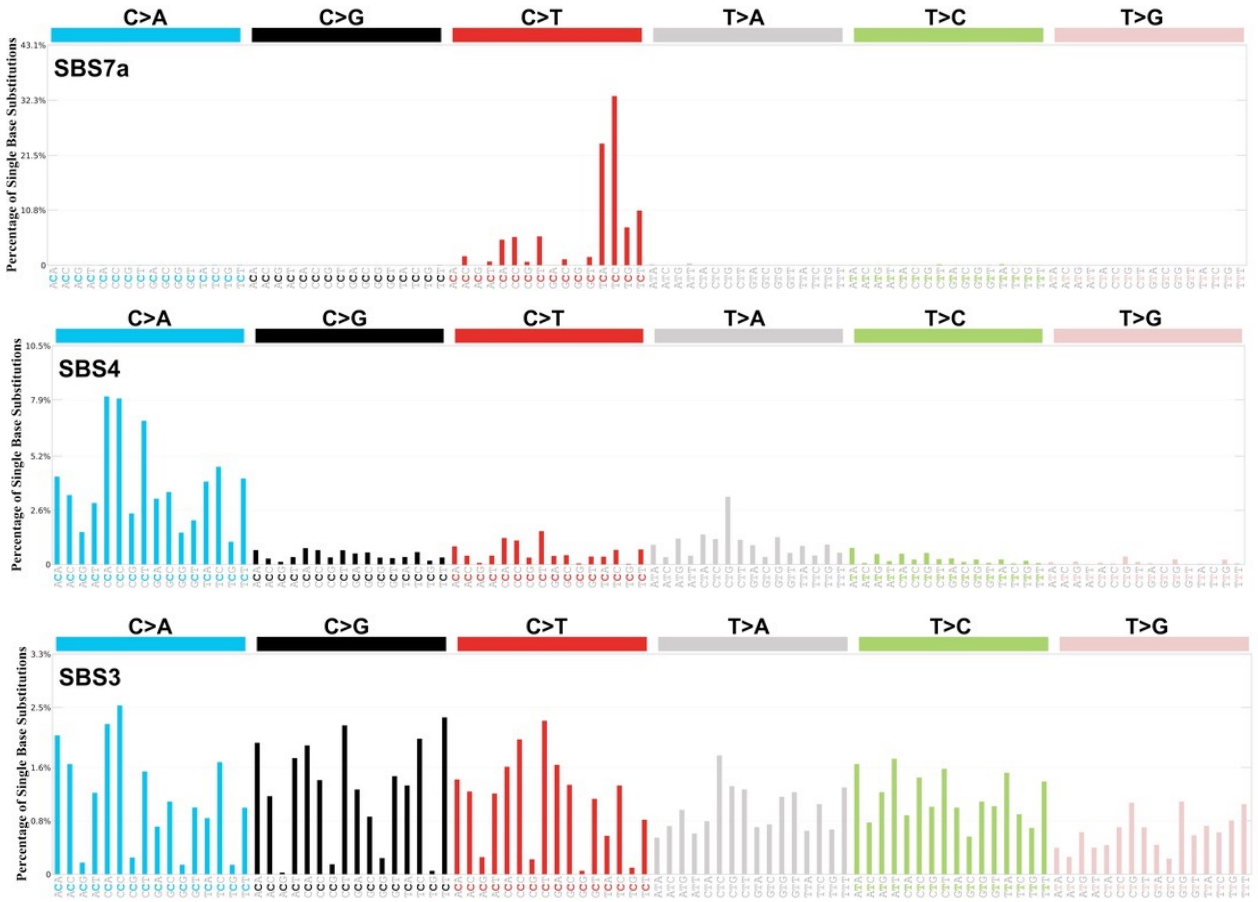
Variant interpretation – derived informations

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



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

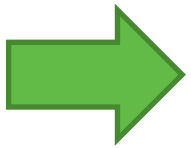


Genomic variant predictive modeling

- 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

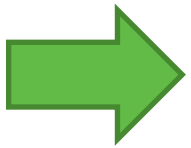
- Genomic variant data are very problematic for modeling
  - Enormous feature space
    - ~ 100 000 features
  - Limited number of data points
    - Only one predictive label per patient
- Feature selection/extraction
- Increase number of samples



Curse of dimensionality

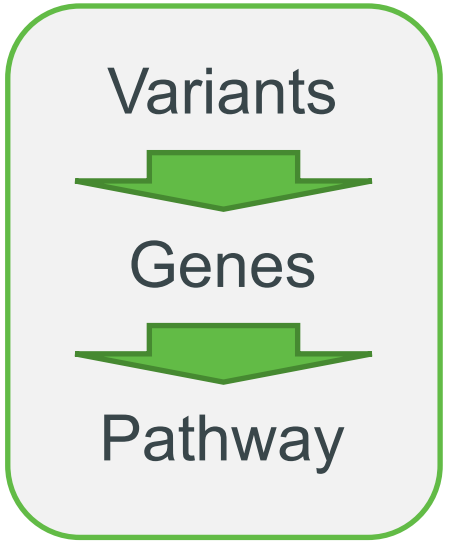
Genomic variant predictive modeling

- 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

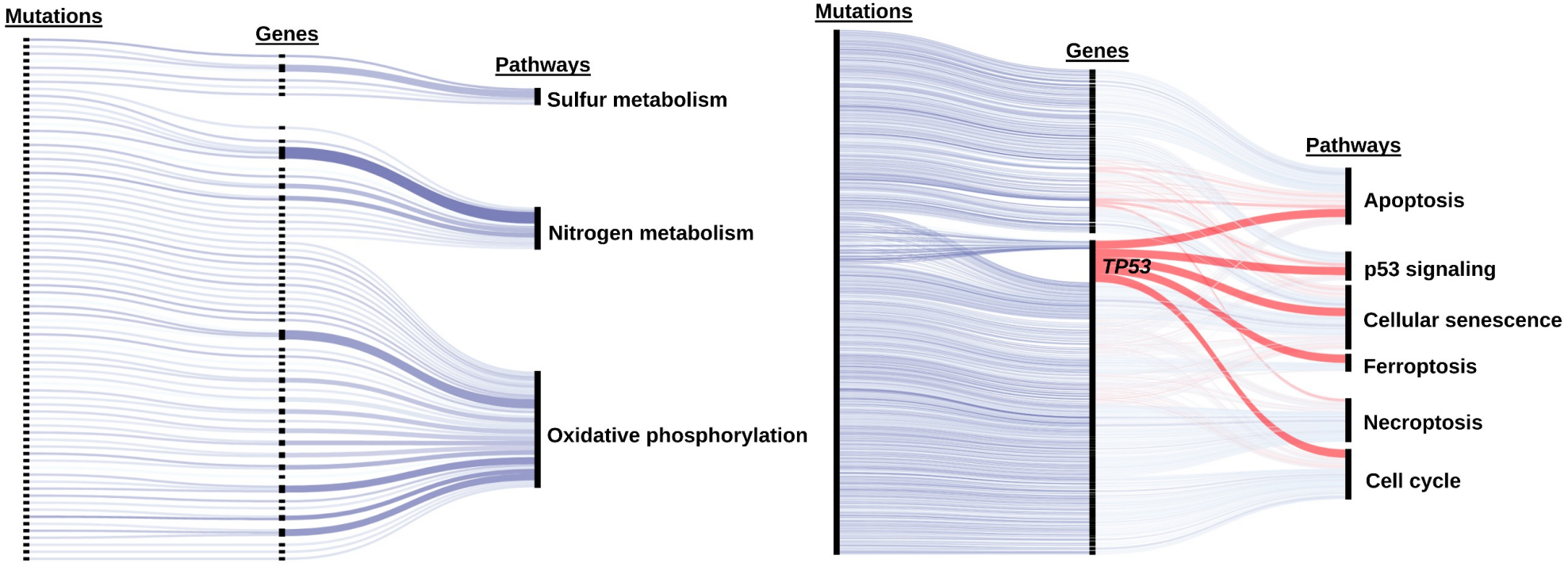
- Feature selection/extraction
- Increase number of samples



- 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