Module 9: Variant Calling and Annotation (Part 2)

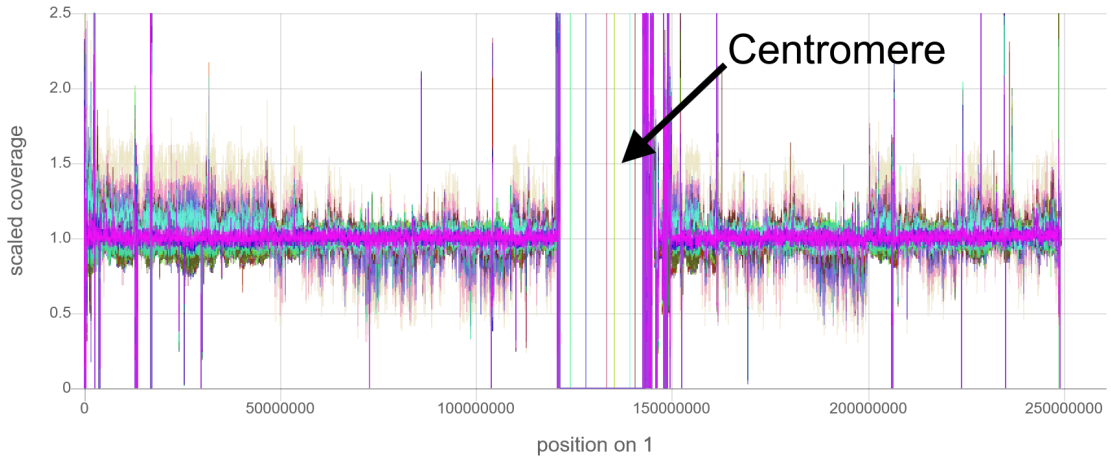
Key Concepts

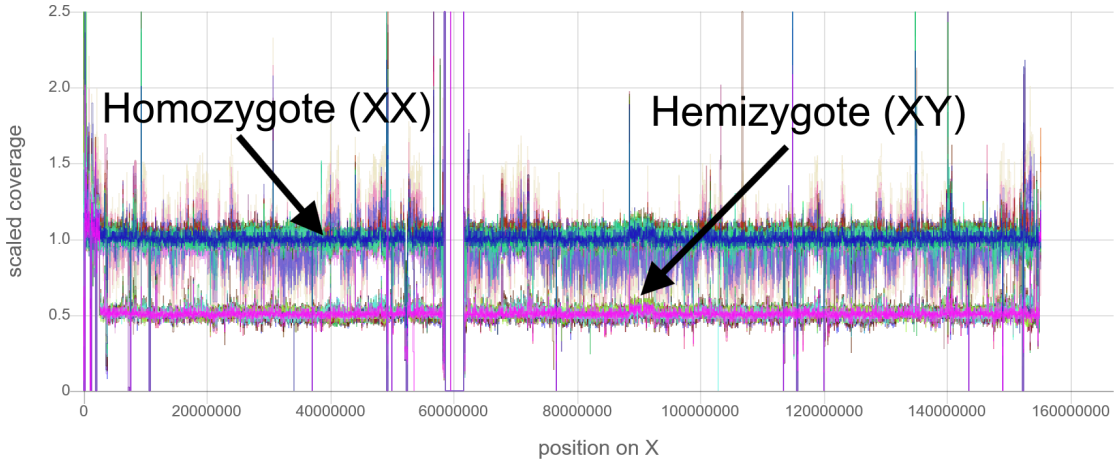
- Variant Calling
- Variant Annotation
- Types of input and output files
- Databases for clinical variants

Variant Calling: Structural Variants

Tool: Indexcov [1]

- Quickly estimate coverage from a **WGS** bam or cram index.
- A long stretch with values of 1.5 would be a heterozygous duplication.





Variant Calling: Structural Variants

Tool: Indexcov [1]

	Description
Purpose	Quickly estimates genome-wide coverage using BAM file index data.
Usage	Identifies regions with abnormal coverage levels, flagging potential large deletions, duplications, or aneuploidies.
Key Feature	Provides a fast QC overview of structural changes without full-scale variant calling.
Assumptions/Limitations	Assumes that index data accurately reflects read depth; cannot detect balanced rearrangements or provide detailed breakpoint information.

Variant Calling: Structural Variants

Tool: Manta[2]

	Description
Purpose	Detects a variety of SVs, including deletions, insertions, inversions, and translocations.
Usage	- Joint analysis of small sets of diploid individuals (where 'small' means family-scale <10 samples) - Subtractive analysis of a matched tumor/normal sample pair - Analysis of an individual tumor sample
Key Feature	High sensitivity and specificity; suitable for clinical diagnostics and research.

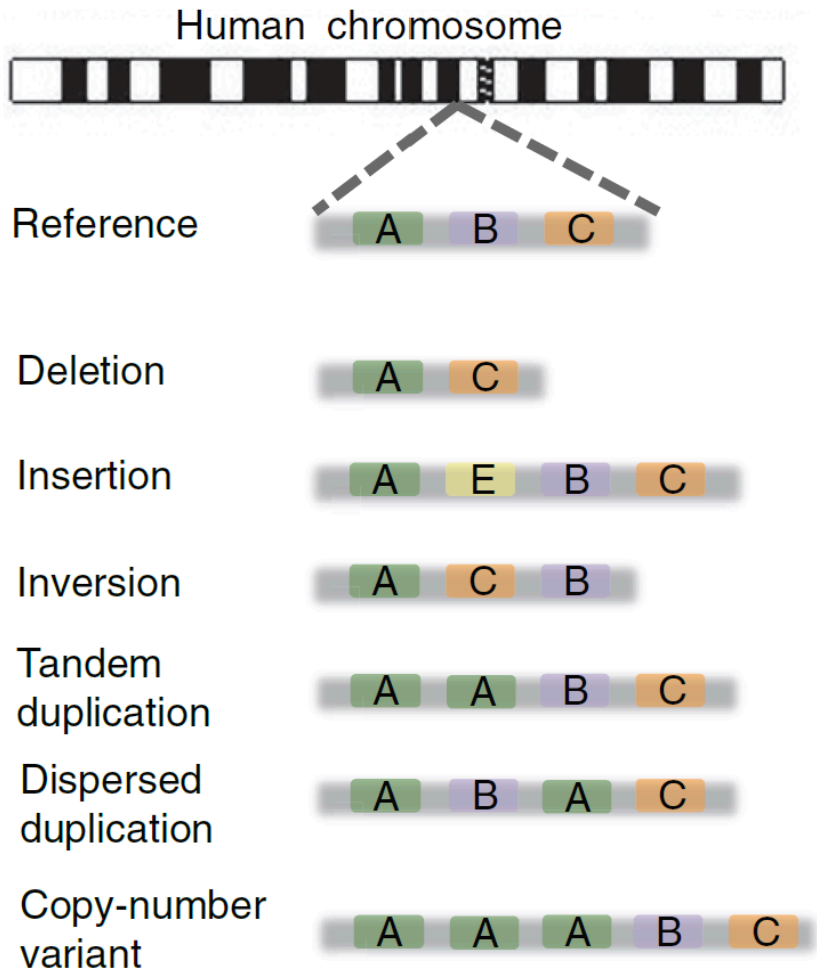
Variant Calling: Structural Variants

Tool: Manta[2]

Assumptions/Limitations: Cannot detect

- small inversions (<200bp),
- fully-assembled large insertions > 2 x read-pair fragment size,
- dispersed duplications

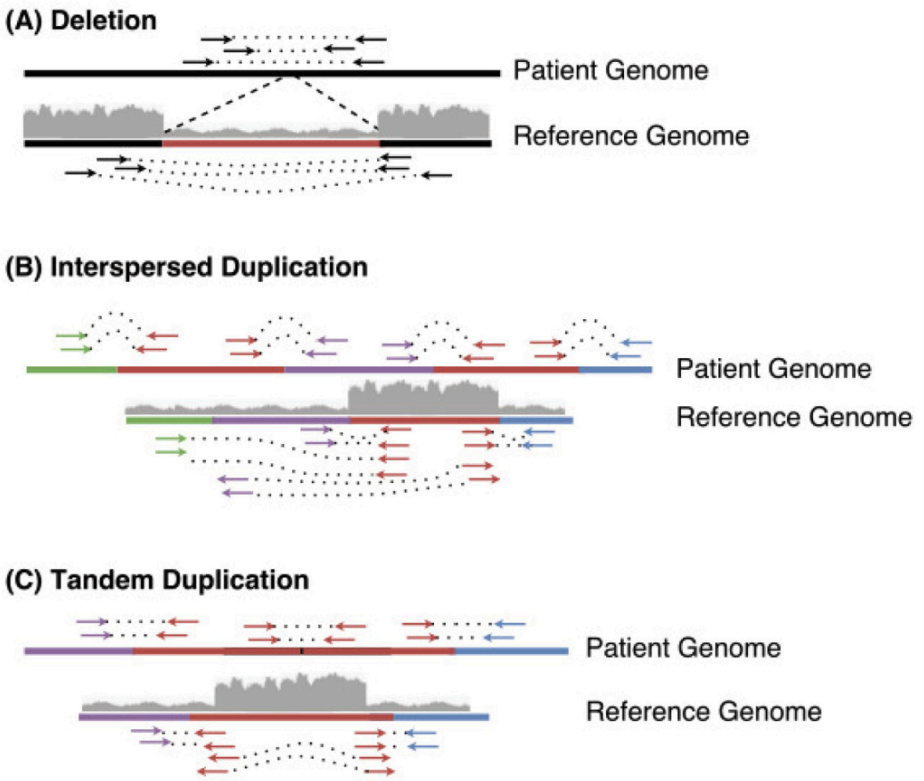
Does not support mate-pair libraries from public data sets

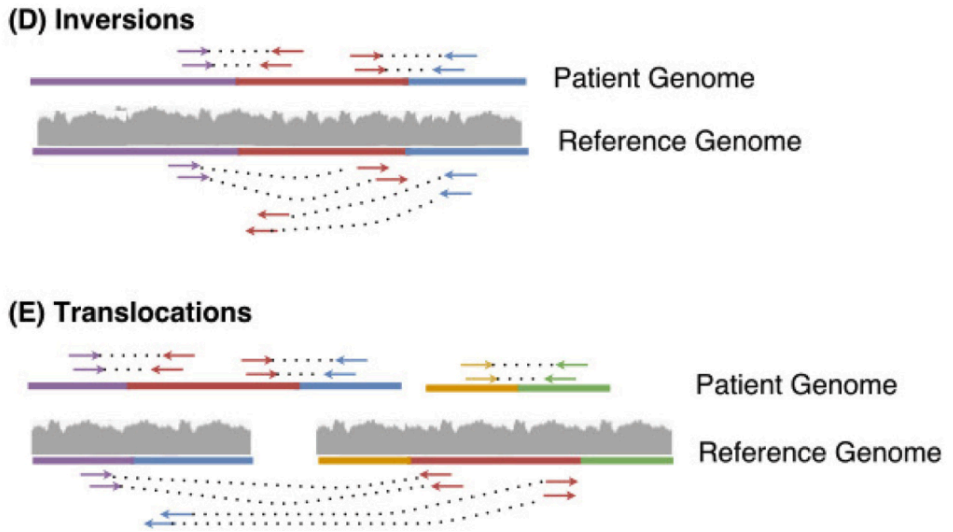


Variant Calling: Structural Variants

Tool: TIDDIT[3]

- Detects many structural variants





Variant Calling: Structural Variants

Tool: TIDDIT[3]

	Description
Usage	- Uses discordant pairs and split reads to detect the genomic location of structural variants - Uses the read depth information for classification and quality assessment of the variants
Key Feature	- Distributed with a database functionality SVDB (Structural Variant DataBase) to create structural variant frequency databases - Uses SVDB to call and evaluate rare disease causing structural variants
Assumptions/Limitations	Does **not** perform well on **small variants** but performs really well on **large variants**, especially **balanced variants** (translocations)

Variant Calling: Structural Variants

Tool: TIDDIT[3]

Table 2. Sensitivity and precision of the structural variant callers on a simulated dataset consisting of 6000 variants of each variant type. The variants were simulated using SVsim and Simseq.

SV detection on simulated data				
Caller	**Sensitivity**	**Precision**	**Sensitivity**	**Precision**
	Deletions		**Duplications**	
TIDDIT	0.96	0.99	0.96	0.99
CNVnator	0.9	0.92	0.86	0.91
Delly	0.94	1	0.95	1
Fermikit	0.41	1	0.33	1
Lumpy	0.95	0.97	0.95	1
Manta	0.95	1	0.95	1
	Inversions		**Translocations**	
TIDDIT	0.97	0.99	0.92	0.93
Delly	0.94	1	0.87	0.95
Fermikit	0.35	1	0.26	0.99
Lumpy	0.5	1	0.87	0.9
Manta	0.95	1	0.88	0.95

Table 4. CPU hour consumption of the structural variant callers. Each caller except Fermikit was run on a single core of a Intel Xeon E5-2660 CPU. Fermikit was run on 16 CPU cores. The CPU hour consumption of the Simseq data is reported as the median time consumption across the four Simseq samples.

CPU hour consumption on SV calling			
Caller	**NA12878**	**HG002**	**Simseq**
TIDDIT	2	1	1
CNVnator	2	1	1
Delly	30	15	7
FermiKit	640	120	15
Lumpy	45	2	7
Manta	3	NA	1

Variant Calling: Copy Number Variation (CNV)

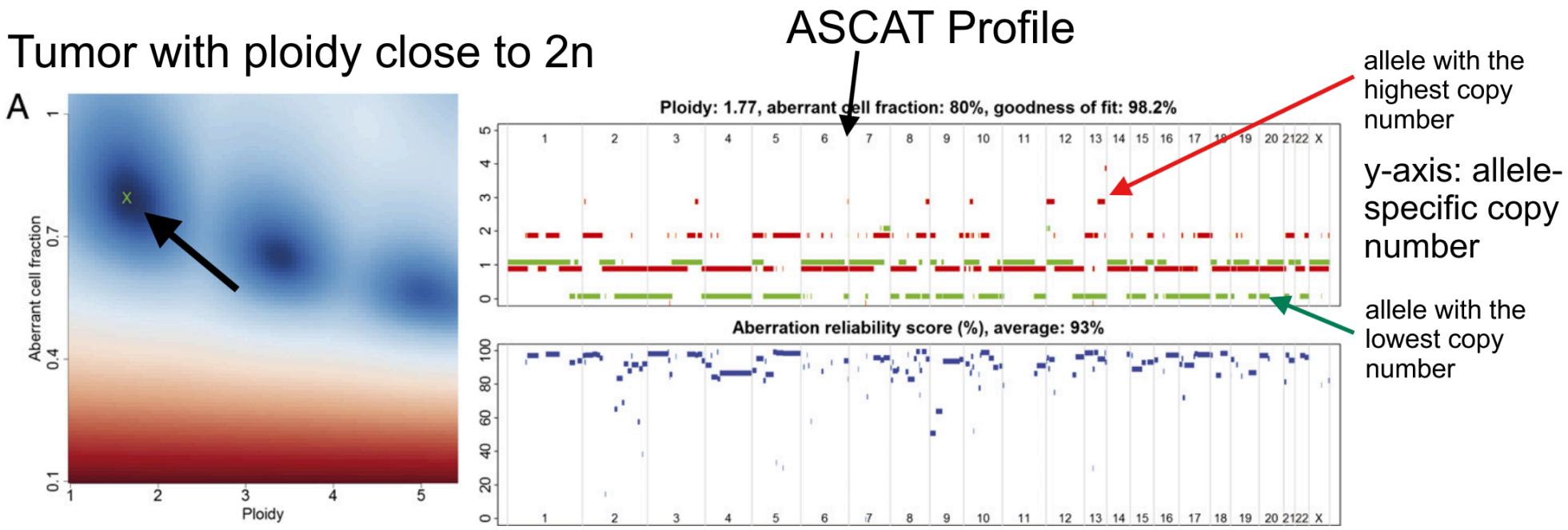
Tool: ASCAT[4]

	Description
Purpose	Allele-specific copy number analysis of tumors from WGS, WES, and TGS
How	Uses allele-specific frequencies to estimate copy number changes while accounting for tumor purity and ploidy in complex tumor genomes.
Key Feature	Profiles tumor ploidy, non-aberrant cell admixture, and frequency of gains, losses, LOH, and copy number-neutral events in heteregeneous tumor samples.
Assumptions/Limitations	- Assumes availability of high-quality SNP data and matched normal samples - May be less reliable with low tumor purity or highly rearranged genomes.

Variant Calling: Copy Number Variation (CNV)

Tool: ASCAT[4]

- The position of the green cross is selected based on the basis of the goodness of fit--for the estimation of ploidy of the tumor cells.



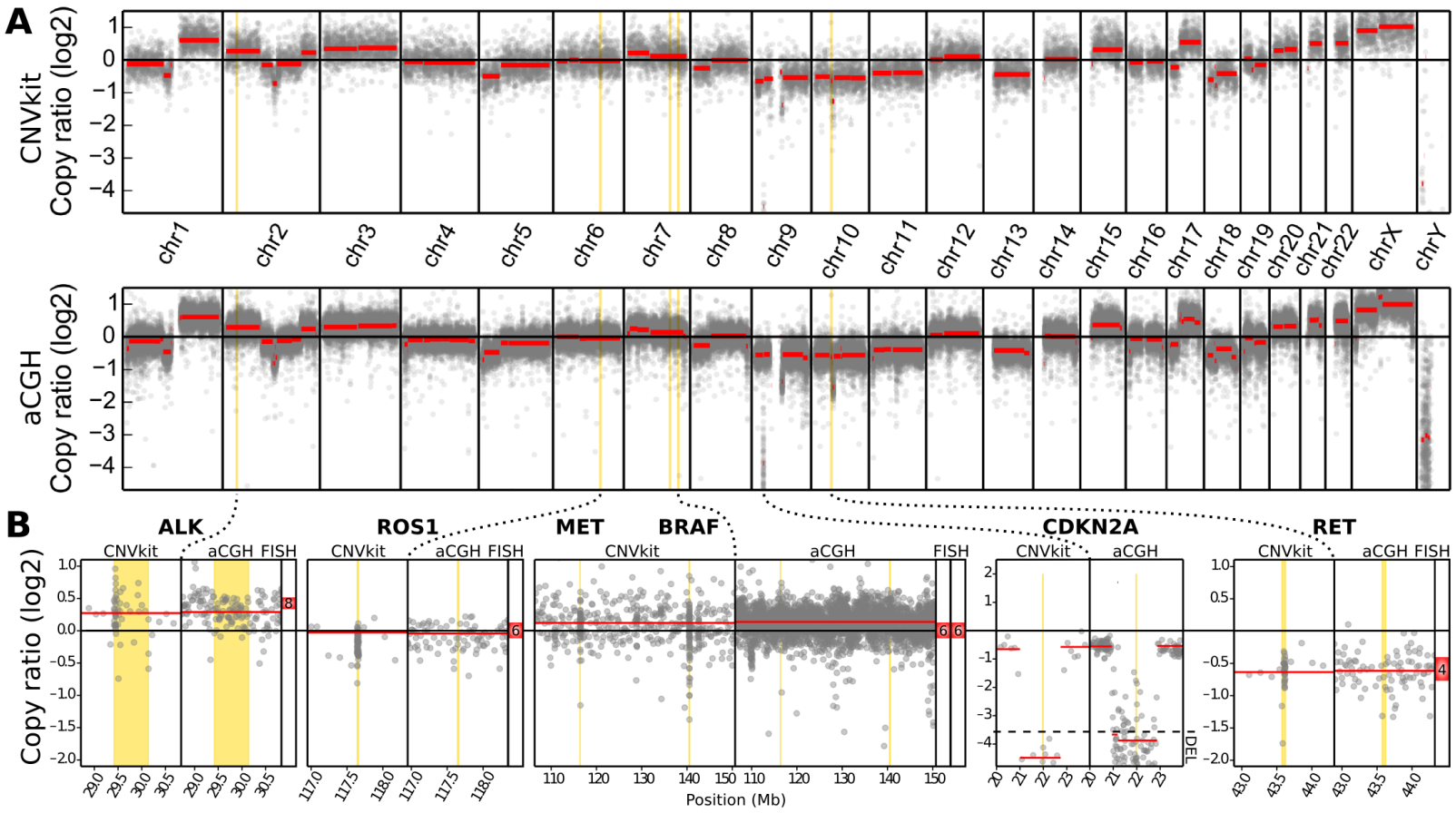
Variant Calling: Copy Number Variation (CNV)

Tool: CNVKit

	Description
Purpose	Detects copy number variations from targeted sequencing data (exomes, gene panels).
How	- Combines on-target and off-target read data to create CN profiles - Correct for biases in sequencing read depth: GC content, target footprint size and spacing, and repetitive sequences.
Key Feature	Optimized for germline copy number variants (CNVs) and somatic copy number alterations (SCNAs) in WES and TGS
Assumptions/Limitations	- Assumes that off-target reads provide sufficient coverage - May be sensitive to capture biases and not as robust when applied to whole-genome data without modifications.

Variant Calling: Copy Number Variation (CNV)

Tool: CNVKit



Variant Calling: Copy Number Variation (CNV)

Tool: Control-FREEC

	Description
Purpose	Detects CNVs and allelic imbalances from whole-genome or exome sequencing data.
How	Normalizes read depth (accounting for GC-content) and segments the genome.
Key Feature	You can use this with or without a matched normal sample
Assumptions/Limitations	- Assumes that read depth variations correlate with copy number changes - May struggle in regions with extreme GC content or highly repetitive sequences.

Variant Calling: Microsatellite Instability

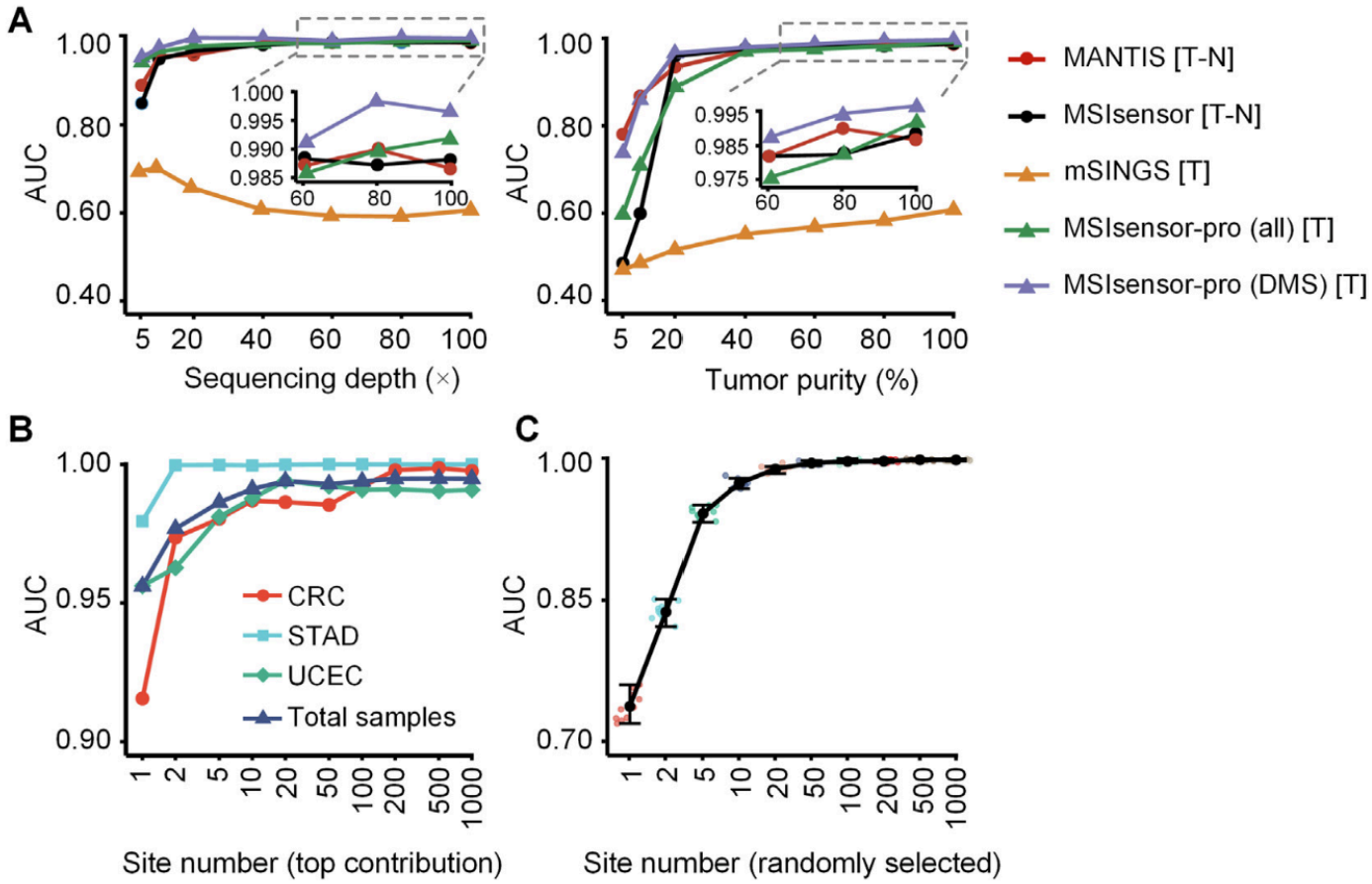
Tool: MSIsensorPro

- MSI is a molecular marker for DNA mismatch repair (MMR) defects, which can lead to cancer

	Description
Purpose	Detects microsatellite instability (MSI) by comparing microsatellite regions between tumor and normal samples.
How	Analyzes repeat length distributions at defined microsatellite loci to compute an MSI score that reflects the level of instability.
Key Feature	Provides a sensitive and specific measure of MSI status
Assumptions/Limitations	- Assumes sufficient coverage at microsatellite loci and availability of matched normal data (or a reliable reference) - Sensitivity may be impacted by sequencing quality.

Variant Calling: Microsatellite Instability

Tool: MSIsensorPro



Variant Annotation Tools

	snpEff	VEP (Variant Effect Predictor)	bcftools
Purpose	Predicts variant effects (e.g., missense, nonsense, frameshift) on genes.	Provides comprehensive variant annotations using Ensembl data.	Primarily a variant processing tool that also offers basic annotation capabilities.
Features	Uses pre-built databases for many organisms; highly configurable for custom annotations.	Annotates variants with gene, transcript, and regulatory information; integrates population frequency data and supports custom plugins.	Can add custom INFO tags and combine external annotations to VCF files.
Usage	Ideal for quick, high-throughput annotation of SNPs and small indels.	Suitable when extensive, detailed annotation is required and integration with Ensembl resources is desired.	Useful for post-calling processing and integrating annotations from other tools into your workflow.

Variant Calling: File Types

PON (Panel of Normals)

- Aggregated file (often in VCF format) created from multiple normal samples
- Used in somatic variant calling to filter out false positives by comparing tumor data against a baseline.
- Improves specificity by removing systematic errors.
- **Limitations:** Quality depends on the number and quality of normal samples; may not capture all artifacts.

Variant Calling and Annotation: File Types

Property	VCF (Variant Call Format)	MAF (Mutation Annotation Format)
Description	Text-based format for storing variant calls.	Tab-delimited format, commonly used in cancer genomics.
Key Fields / Features	CHROM, POS, ID, REF, ALT, QUAL, FILTER, INFO, FORMAT.	Contains curated annotations (e.g., gene name, variant classification, sample-specific details).
Usage	Widely used for both germline and somatic variant calls.	Ideal for generating human-readable, standardized reports of somatic mutations.
Strengths	Extensible with custom annotations; standard format in many pipelines.	Focused on detailed mutation annotation; facilitates downstream interpretation.
Limitations	Can become large with extensive annotations; requires proper indexing for efficient querying.	Less common for germline variants; may need conversion from VCF.

File Types in Variant Calling and Annotation

- **Alignment Files (BAM/CRAM):** For mapping reads and ensuring quality alignments.
    - BAM is standard but large; CRAM offers better compression.
- **Variant Files (VCF/MAF):** For storing and interpreting variant calls.
    - VCF and MAF provide variant-level annotations with MAF offering more cancer-specific details.
- **PON:** For filtering out recurrent artifacts, especially in somatic pipelines.

Variant Databases

	gnomAD
Description	Aggregates large-scale exome and genome sequencing data from diverse populations.
Purpose	Provides allele frequency data to distinguish common variants from rare variants.
Usage	Used for filtering variants based on population frequency; supports studies in population genetics and variant prioritization.
Key Points	- Extensive dataset with rigorous quality filters. - Continuously updated with diverse sample representation.
Limitations	- Lacks detailed clinical annotation. - May not capture all sub-population specific variants.

Variant Databases

	ClinVar
Description	A public archive aggregating information about the relationships between human variants and phenotypes.
Purpose	Provides clinical significance and supporting evidence for genetic variants.
Usage	Essential for clinical interpretation; used to assess variant pathogenicity and inform diagnostics.
Key Points	- Curated submissions from clinical laboratories and research groups. - Integrates multiple clinical assertions and evidence.
Limitations	- May contain conflicting interpretations. - Focused primarily on clinically relevant variants.

Variant Databases

	dbGaP (Database of Genotypes and Phenotypes)
Description	A controlled-access repository for genotype and phenotype data from a wide array of studies.
Purpose	Facilitates research by providing access to raw and processed genomic and phenotypic data.
Usage	Used for large-scale genomic studies, variant-disease association research, and validation of genetic findings.
Key Points	- Contains data from diverse study cohorts. - Access is governed by ethical and legal restrictions.
Limitations	- Data access is restricted and may require approval. - May require additional processing and harmonization.