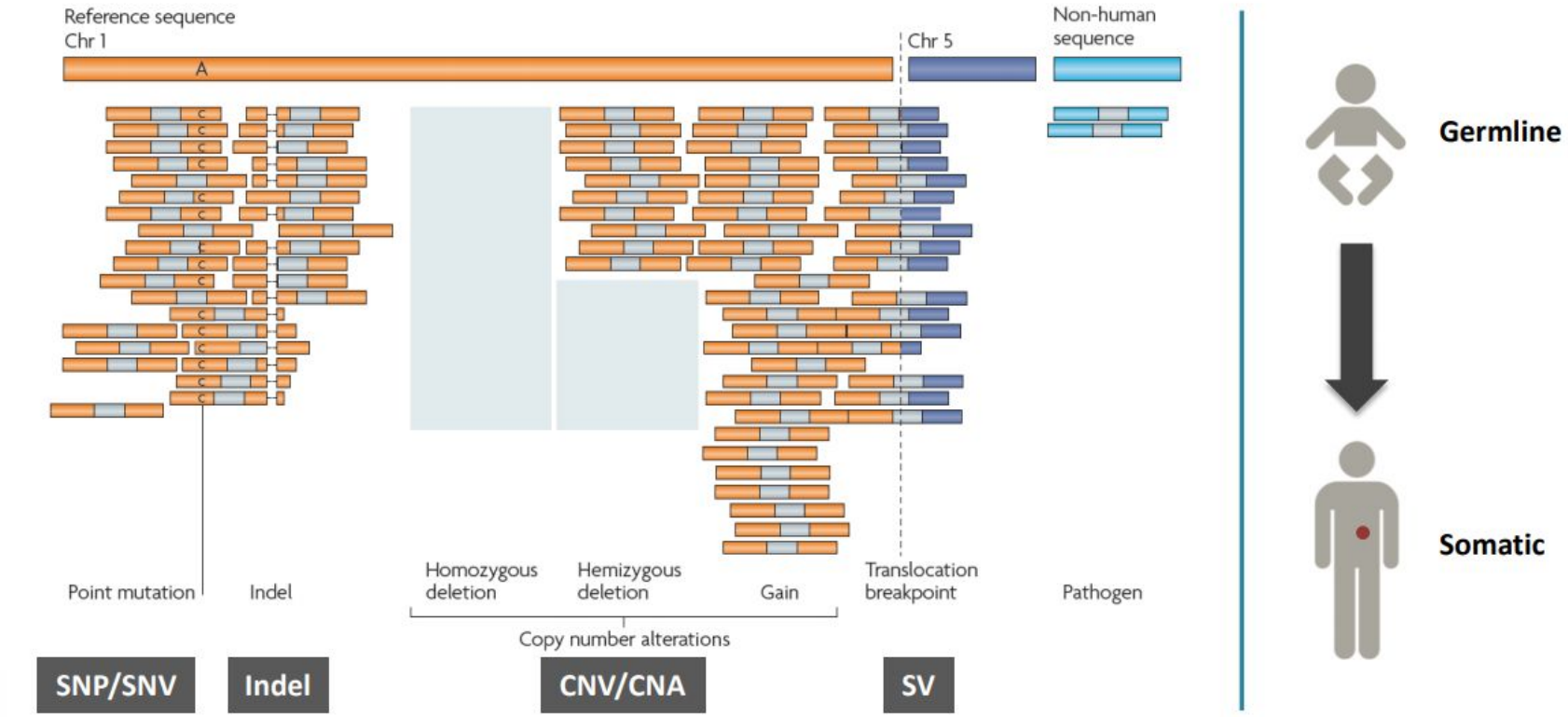
Somatic Variant Calling

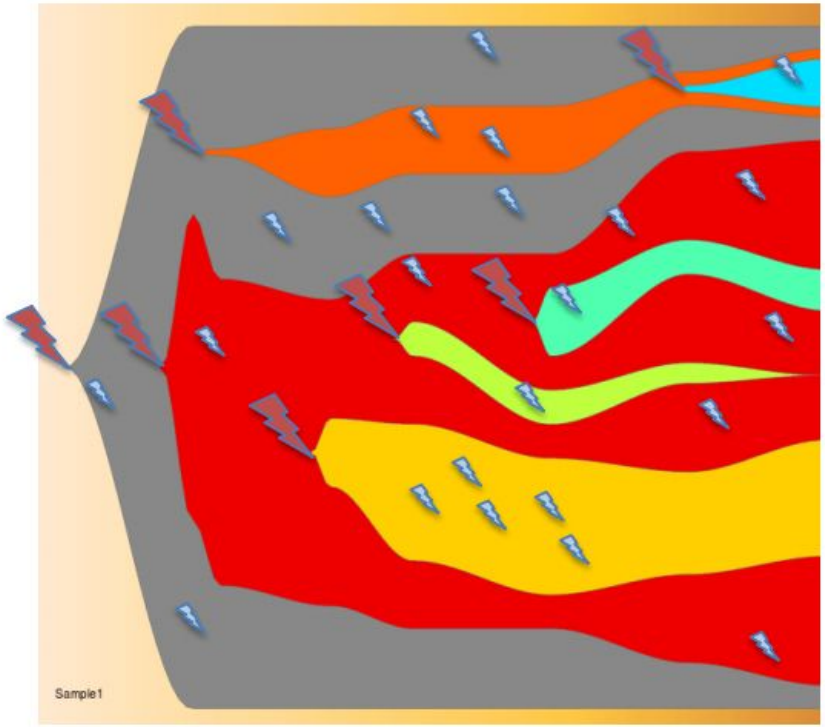
Different types of variants



from Introduction to Somatic Variant Discovery, GATK Best Practices for Variant Discovery

Role of somatic mutation events in tumor progression

Increasing tumor heterogeneity as genomic instability increases



Driver mutations

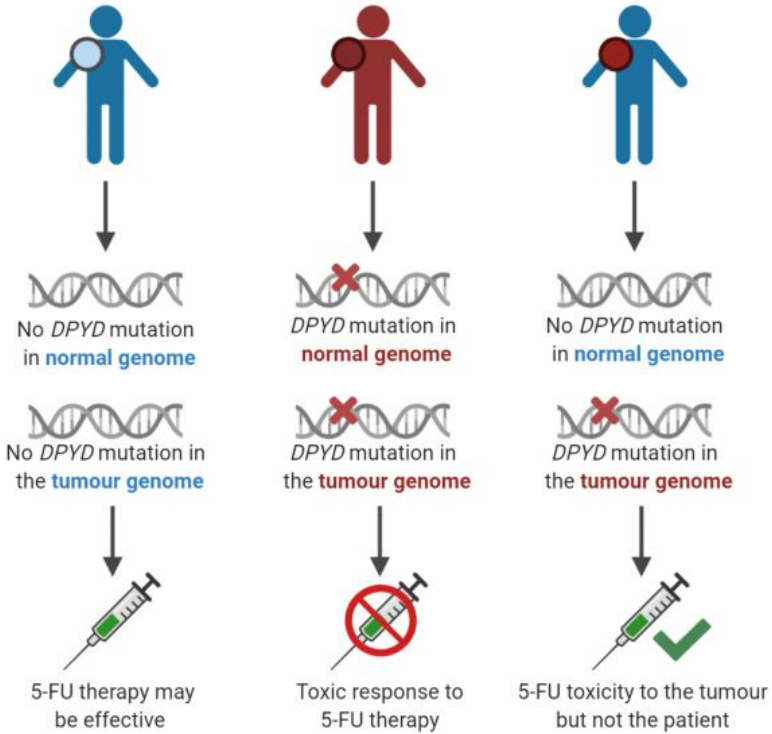
Passenger mutations

Dittmar & Zanker 2015
http://www.mdpi.com/1422-0067/16/12/26240/htm

Illustration made with
https://github.com/chrisamiller/fishplot

from Introduction to Somatic Variant Discovery, GATK Best Practices for Variant Discovery

Cancer sequencing helps prioritize cancer treatment options



https://www.bcgsc.ca/news/genome-sequencing-helps-prioritize-cancer-treatment-options

Why is SNV calling in cancer samples more complicate ?

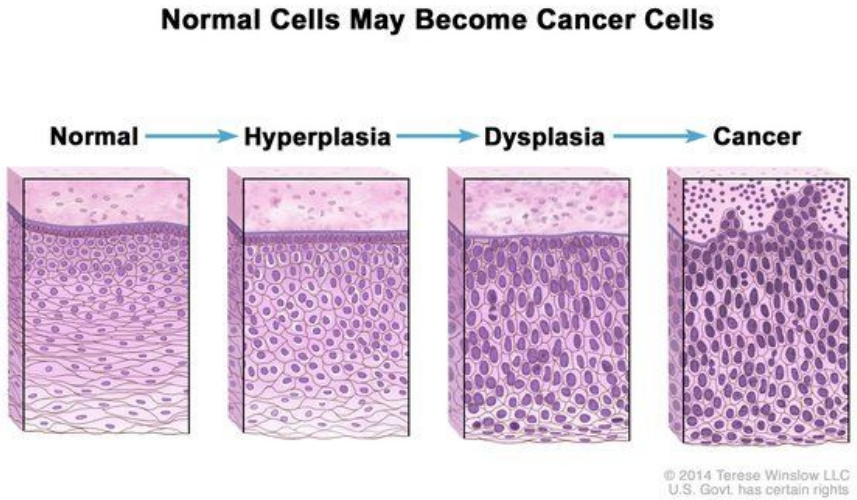
MANY REASONS

Why is SNV calling in cancer samples more complicate ?

Low cellularity (tumour DNA content)

Tumor samples may have lower DNA content

→ need sensitivity in variant calling +++



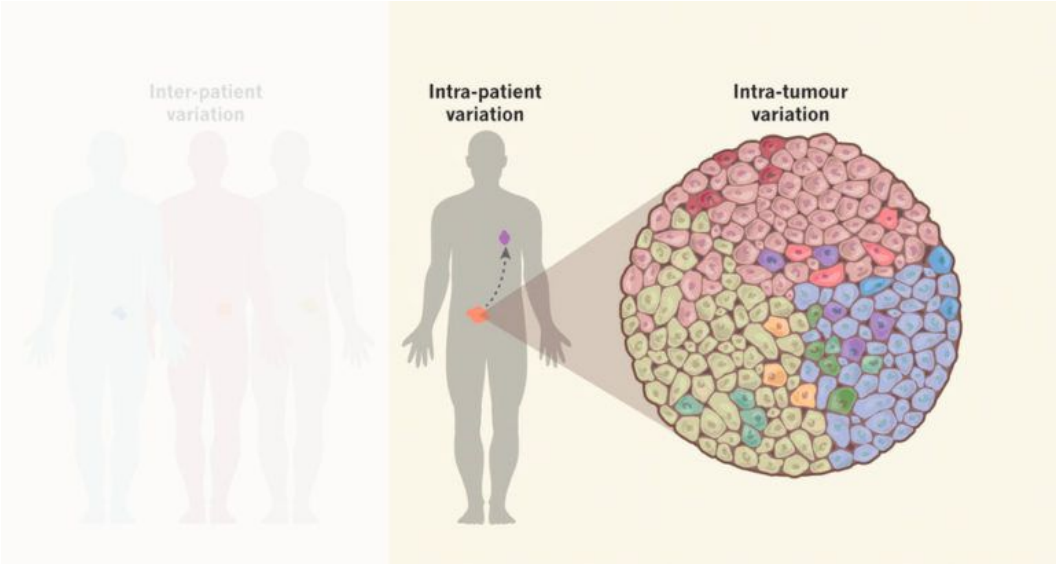
https://visualsonline.cancer.gov/details.cfm?imageid=9752

Why is SNV calling in cancer samples more complicate ?

Intra-tumour heterogeneity in which multiple tumour cell populations (subclones) exist

Multiple subclonal populations that are constantly evolving.

→ variants can be present in only one subclonal population



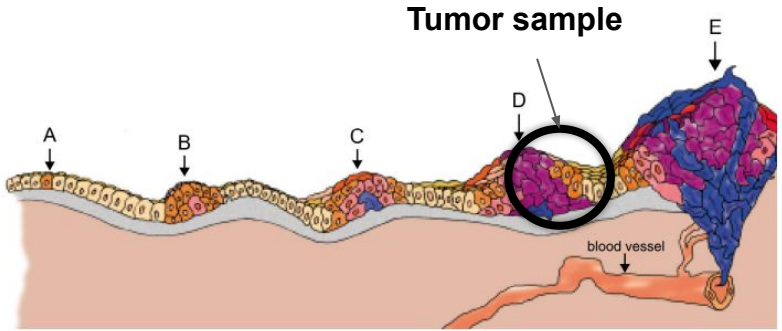
https://hugosnippert.nl/research-genetic-heterogeneity/

Why is SNV calling in cancer samples more complicate ?

Normal contamination

Normal cells can “contaminate” the tumor biopsy.

$$Tumor\ purity = \frac{tumor\ cells}{(normal + tumor\ cells)}$$

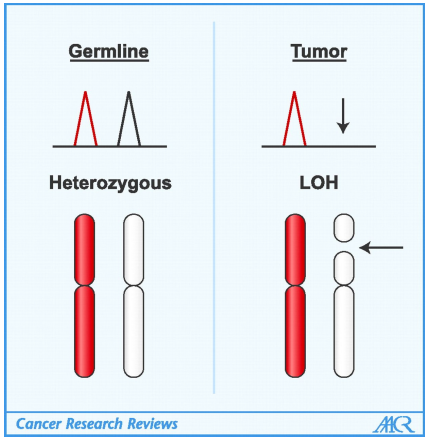


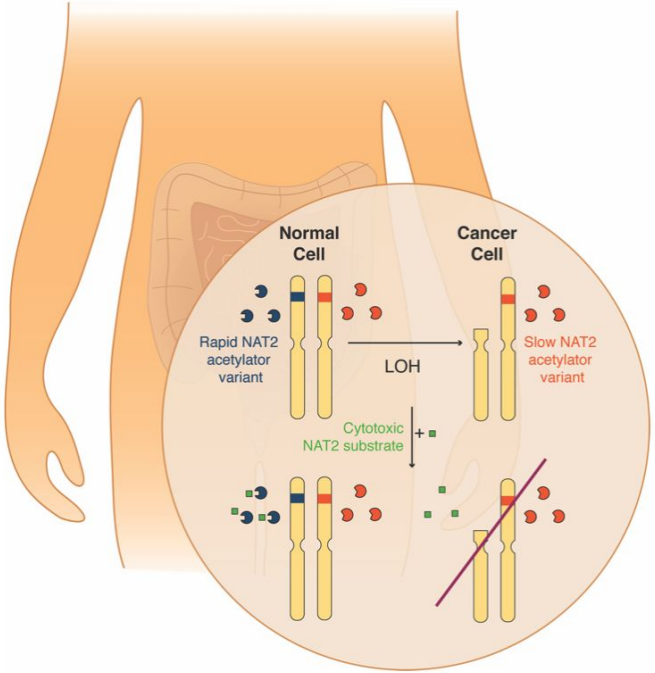
Adapted from Introduction to Somatic Variant Discovery, GATK
https://science.education.nih.gov/supplements/nih_cancer.pdf

Why is SNV calling in cancer samples more complicate ?

Unbalanced structural variations (deletions, duplications, etc.)

→ need to detect LOH events





https://cancerres.aacrjournals.org/content/69/17/6759

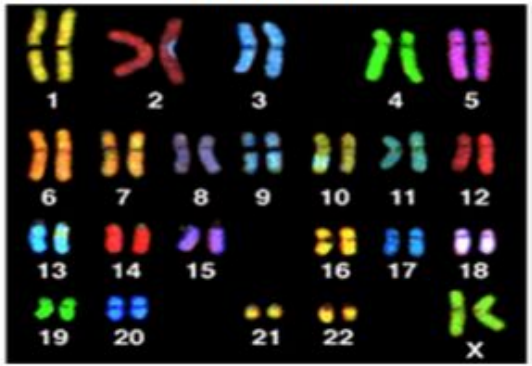
https://cancercommunity.nature.com/posts/61527-a-tumor-s-loss-of-heterozygosity-is-our-gain

Why is SNV calling in cancer samples more complicate ?

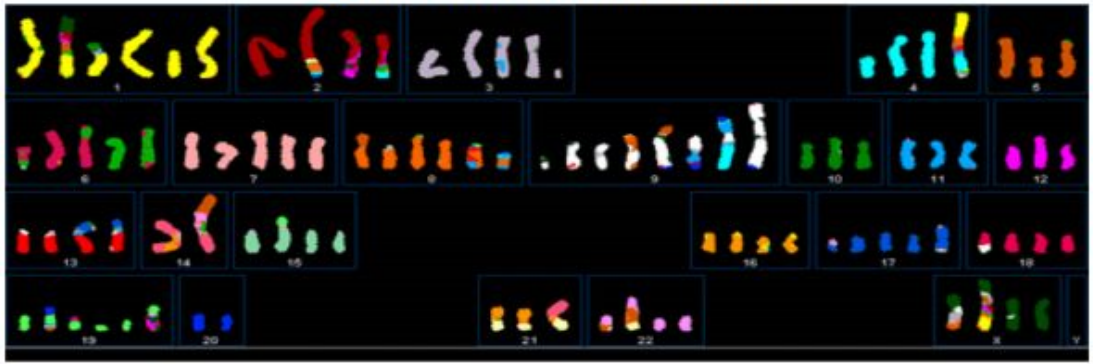
Aneuploidy → need for variant calling algorithm with no assumption on ploidy

Somatic alterations can be dramatic

Normal Cell



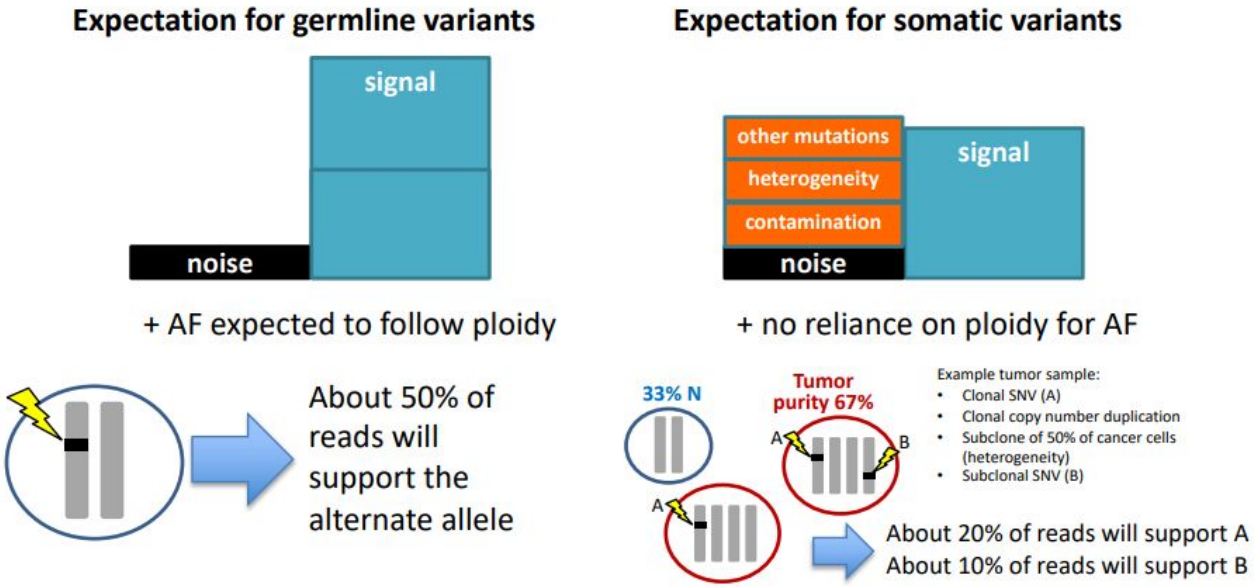
Cancer Cell Line HCC1954



from Introduction to Somatic Variant Discovery, GATK Best Practices for Variant Discovery

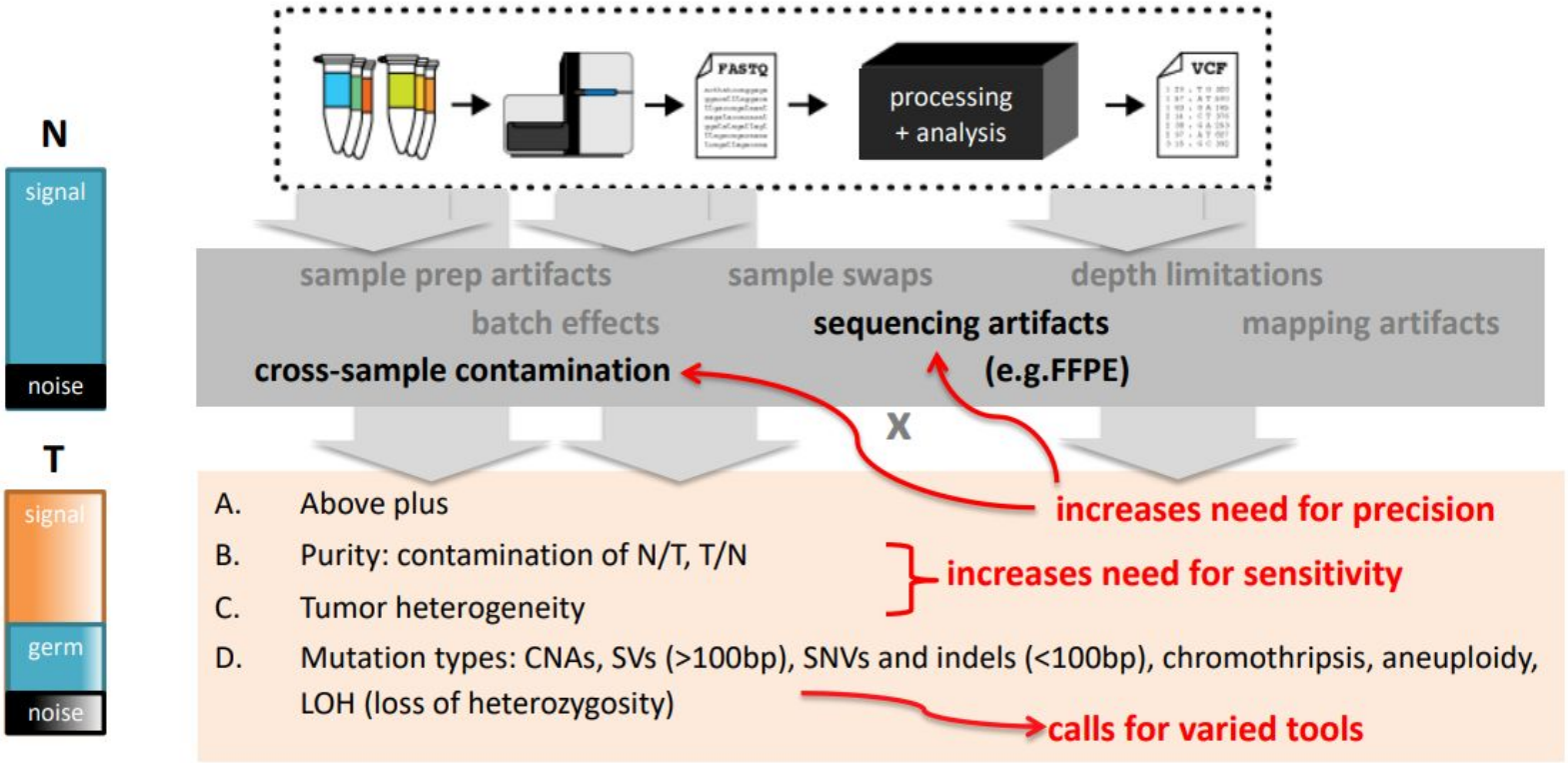
Why is SNV calling in cancer samples more complicate ?

Amount of signal may be comparable to noise



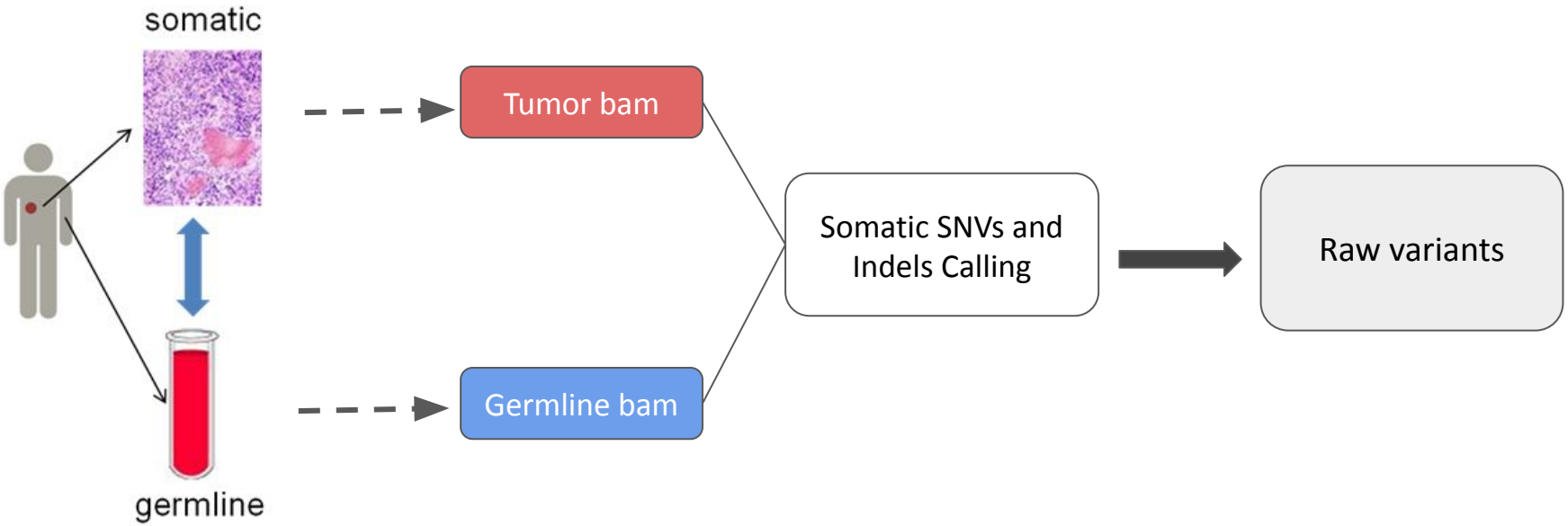
from Introduction to Somatic Variant Discovery, GATK Best Practices for Variant Discovery

Cancer-specific challenges confound analyses



from Introduction to Somatic Variant Discovery, GATK Best Practices for Variant Discovery

Tumor - Normal pair analysis pipeline

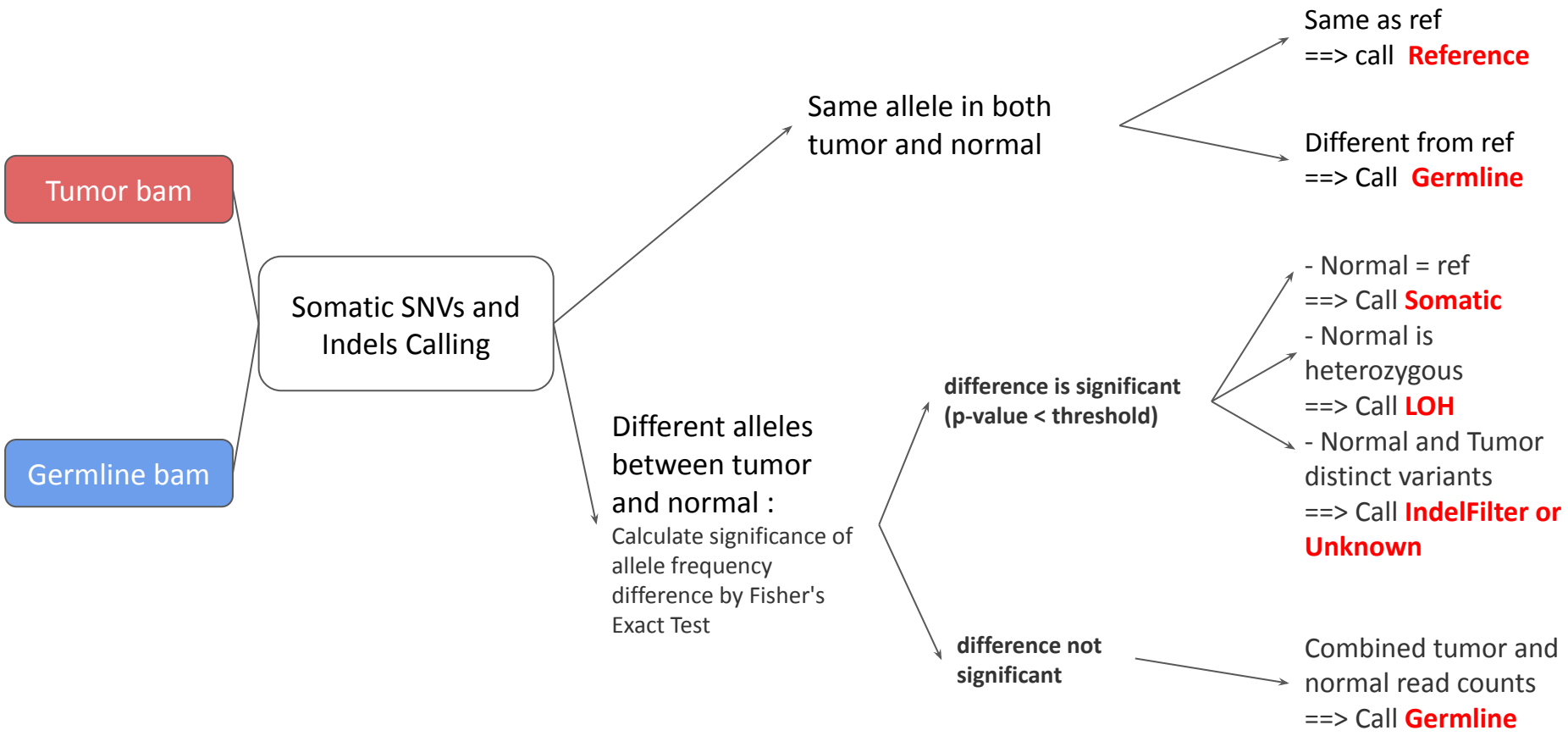


Logic of the Tumor-Normal workflow

Comparison to matched normal -> subtraction of germline background

from Introduction to Somatic Variant Discovery, GATK Best Practices for Variant Discovery

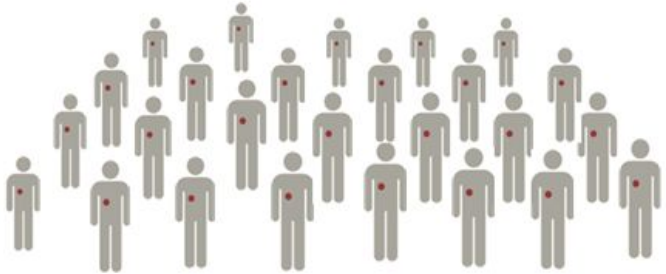
Tumor - Normal pair analysis by Varscan



Tumor-only analysis

If matched normal sample NOT available → Tumor only analysis

Pool of normals (PON) : used to eliminate common germline variation



Somatic variant filtering

Somatic variant callers output specific informations on variant :

➔ Somatic likelihood like score as phred-scale somatic p-value

➔ Status : - Germline,
- Somatic,
- LOH

⇒ first metrics to filter variants on

Somatic variant annotation

Somatic Annotation Databases

Databases of variant-disease and gene-disease associations

- Cancer HotSpots
  - Single residue and in-frame indel mutation hotspots identified in 24,592 tumor samples.
- COSMIC
  - COSMIC (Catalogue of Somatic Mutations in Cancer) is a data resource that is designed to store and display somatic mutation information and related details and contains information relating to human cancers.
  - Data in COSMIC is curated from known Cancer Genes Literature and Systematic Screens.
- CIViC
  - CIViC (Clinical Interpretation of Variant in Cancer) is a an open access, open source, community-driven web resource for Clinical Interpretation of Variants in Cancer. The goal is to enable precision medicine by providing an educational forum for dissemination of knowledge and active discussion of the clinical significance of cancer genome alterations.

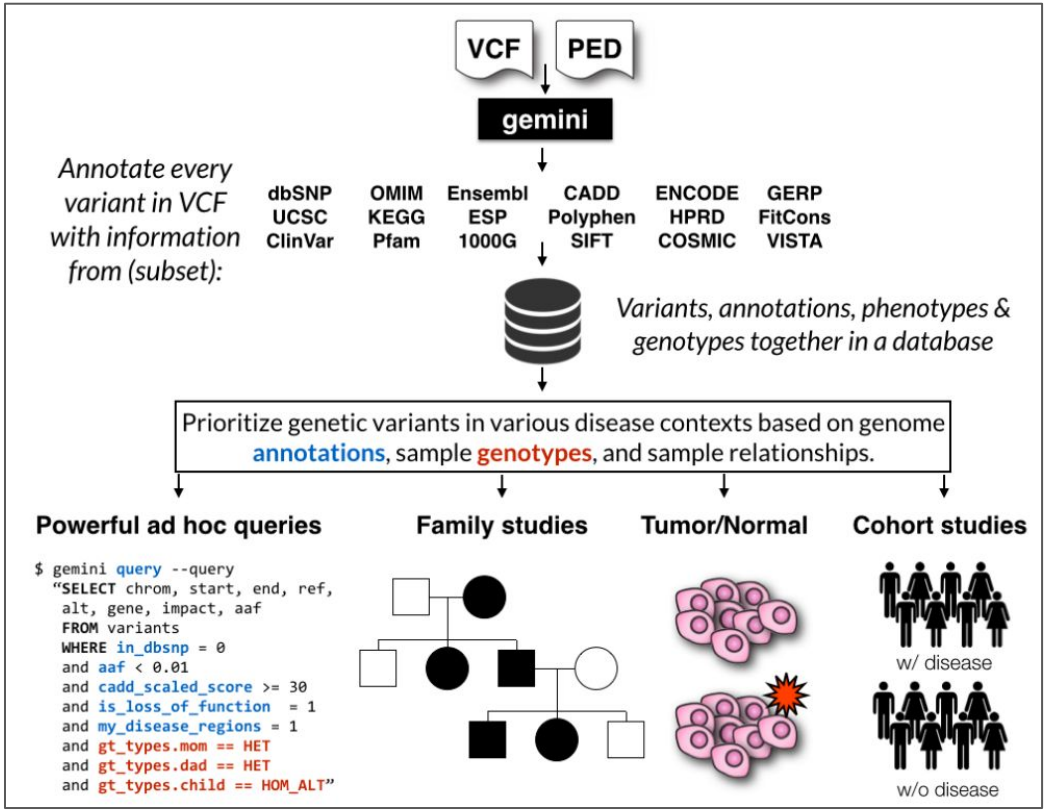
Somatic Annotation Databases

Databases of variant-disease and gene-disease associations

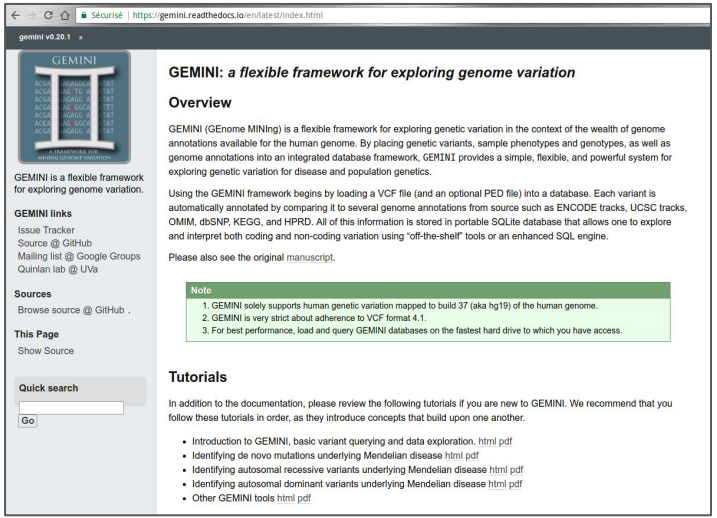
- Cancer Genome Interpreter (CGI)
  - Cancer Genome Interpreter (CGI) is designed to support the identification of tumor alterations that drive the disease and detect those that may be therapeutically actionable. CGI relies on existing knowledge collected from several resources and on computational methods that annotate the alterations in a tumor according to distinct levels of evidence.
    It contains : a Cancer Biomarkers database, a Catalog of Validated Oncogenic Mutations and a Catalog of Validated Oncogenic Mutations

- The Cancer Genome Atlas (TCGA)
  - The Cancer Genome Atlas (TCGA), a landmark cancer genomics program, molecularly characterized over 20,000 primary cancer and matched normal samples spanning 33 cancer types.TCGA generated over 2.5 petabytes of genomic, epigenomic, transcriptomic, and proteomic data.

GEMINI presentation



Documentation :
http://gemini.readthedocs.io
Tutorials :
https://speakerdeck.com/arq5x/



GEMINI database overview

The variants table

Gene information

gene
transcript
is_exonic
is_coding
is_lof
is_splicing
exon
codon_change
aa_change
aa_length
biotype
impact
impact_so
impact_severity
polyphen_pred
polyphen_score
sift_pred
sift_score
pfam_domain

Genotype information

gts	
gt_types	
gt_phases	
gt_depths	
gt_ref_depths	
gt_alt_depths	
gt_alt_freqs	
gt_quals	BLOB A compressed binary vector of the genc - Extracted from the VCF GQ genotype t

Variant and PopGen info

type		
sub_type		
call_rate		
num_hom_ref		
num_het		
num_hom_alt		
num_unknown		
aaf		
hwe		
inbreeding_coeff	FLOAT	The inb
pi	FLOAT	The co

Core VCF fields

column_name
chrom
start
end
vcf_id
variant_id
anno_id
ref
alt
qual
filter

Population information

in_dbsnp
rs_ids
in_hm2
in_hm3
in_esp
in_1kg
aaf_esp_ea
aaf_esp_aa
aaf_esp_all
aaf_1kg_amr
aaf_1kg_eas
aaf_1kg_sas
aaf_1kg_afr
aaf_1kg_eur
aaf_1kg_all
in_exac
aaf_exac_all
aaf_adj_exac_all
aaf_adj_exac_afr
aaf_adj_exac_amr
aaf_adj_exac_eas

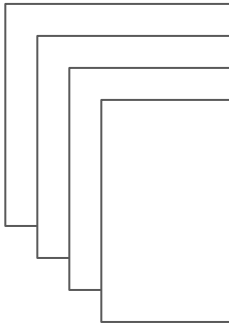
Disease phenotype info (from ClinVar).

in_omim	BOOL	0 : Absence of the variant in OMIM databas 1 : Presence of the variant in OMIM databa
clinvar_causal_allele	STRING	The allele(s) that are associated or causal f
clinvar_sig	STRING	The clinical significance scores for each of the variant according to ClinVar: *unknown*, *untested*, *non-pathogenic* *probable-non-pathogenic*, *probable-pathog* *pathogenic*, *drug-response*, *histocompatibil* *other*
clinvar_disease_name	STRING	The name of the disease to which the varia
clinvar_dbsource	STRING	Variant Clinical Channel IDs
clinvar_dbsource_id	STRING	The record id in the above database
clinvar_origin	STRING	The type of variant. Any of: *unknown*, *germline*, *somatic*, *inherited*, *paternal*, *maternal*, *de-novo*, *biparental*, *uniparental*, *not-tested*, *tested-inconclusive*, *other*
clinvar_dsdb	STRING	Variant disease database name
clinvar_dsdbid	STRING	Variant disease database ID
clinvar_disease_acc	STRING	Variant Accession and Versions
clinvar_in_locus_spec_db	BOOL	Submitted from a locus-specific database?
clinvar_on_diag_assay	BOOL	Variation is interrogated in a clinical diagnostic assay?
clinvar_gene_phenotype	STRING	'|' delimited list of phenotypes associated with this gene (includes any variant in the same gene in clinvar not just the current variant).

The variant_impacts table

variant_id	INTEGER	PRIMARY_KEY (Foreign key to *variants* table)
anno_id	INTEGER	PRIMARY_KEY (Based on variant transcripts)
gene	STRING	The gene affected by the variant.
transcript	STRING	The transcript affected by the variant.
is_exonic	BOOL	Does the variant affect an exon for this transcript?
is_coding	BOOL	Does the variant fall in a coding region (excludes 3' & 5' UTR's of exons)?
is_lof	BOOL	Based on the value of the impact col, is the variant LOF?
exon	STRING	Exon information for the variants that are exonic
codon_change	STRING	What is the codon change?
aa_change	STRING	What is the amino acid change?
aa_length	STRING	The length of CDS in terms of number of amino acids (`SnpEff only`)
biotype	STRING	The type of transcript (e.g., protein-coding, pseudogene, rRNA etc.) (`SnpEff only`)
impact	STRING	Impacts due to variation (ref.impact category)
impact_so	STRING	The sequence ontology term for the impact
impact_severity	STRING	Severity of the impact based on the impact column value (ref.impact category)
polyphen_pred	STRING	Impact of the SNP as given by PolyPhen (`VEP only`) benign, possibly_damaging, probably_damaging, unknown
polyphen_scores	FLOAT	Polyphen score reflecting severity (higher the impact, *higher* the score) (`VEP only`)
sift_pred	STRING	Impact of the SNP as given by SIFT (`VEP only`) neutral, deleterious
sift_scores	FLOAT	SIFT prob. scores reflecting severity (Higher the impact, *lower* the score) (`VEP only`)

Etc.



Tables/fields descriptions :
http://gemini.readthedocs.io/en/latest/content/database_schema.html

GEMINI usages

```
SELECT column-names
FROM table-name
WHERE condition
ORDER BY sort-order
```

ad hoc data exploration

- ➔ Use SQL language to create queries and report data matching your requirements
- ➔ Can personalize your query to answer complex questions

`gemini query -q “SELECT gene, chrom, clinvar_gene_phenotype FROM variants”`

column_name	type
chrom	VARCHAR(20)
start	INTEGER
ref	TEXT
alt	TEXT
qual	FLOAT
filter	TEXT
in_omim	BOOLEAN
clinvar_sig	TEXT
clinvar_gene_phenotype	TEXT
gene	VARCHAR(60)

Table variants in Gemini database

https://gemini.readthedocs.io/en/latest/content/querying.html