Session 4: Mutation calling and analysis

Emerging Approaches For Tumor Analyses
in Epidemiological Studies

January 11, 2023
9:30 AM- 12:00 PM

Session Overview

- **Introduction to somatic variants**
- **Framework for short variant discovery (GATK)**
- **Variant callers**
- **Variant filtering and annotation**
- **IGV for visualization and validation**
- **Common somatic mutational analyses**

Workshop website: https://nci-iteb.github.io/tumor_epidemiology_approaches/

Session 4: Mutation calling and analysis

Emerging Approaches For Tumor Analyses
in Epidemiological Studies

January 11, 2023
9:30 AM- 12:00 PM

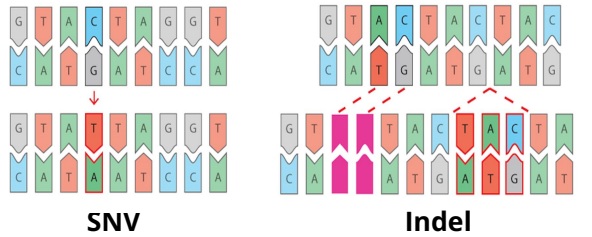
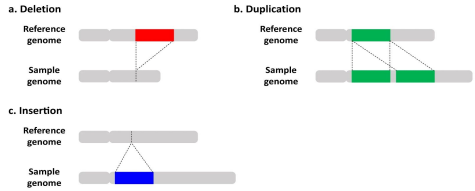
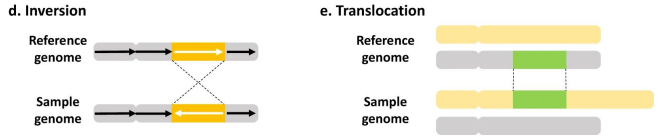
Introduction to somatic variants

Why call somatic variants?

The Cancer Genome Atlas (TCGA) Program Overview:

"There are at least 200 forms of cancer, and many more subtypes. Each of these is caused by errors in DNA that cause cells to grow uncontrolled. Identifying the changes in each cancer's complete set of DNA – its genome – and understanding how such changes interact to drive the disease will lay the foundation ***for improving cancer prevention, early detection and treatment.****"*

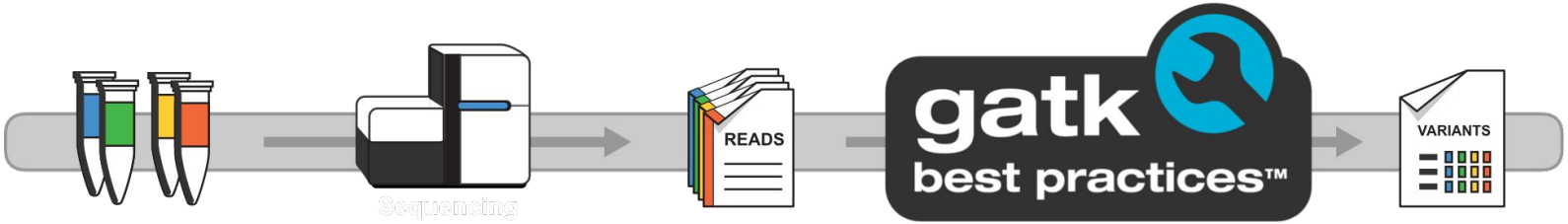
Types of somatic genetic variation

Type of variation	Description	Visualization
Short variants	Single Nucleotide Variants (**SNV**s) Short Insertions or Deletions (**Indel**s) (2bp- 1Kb)	
Unbalanced SVs	Somatic Copy Number Alterations (**SCNA**s) - copy number gain/loss (Session 7)	
Balanced SVs	Structural Variants (**SV**s): large insertion/deletion, inversion, translocation, tandem/dispersed duplication (Session 6)	

Short variant (SNVs and Indels) discovery and analyses will be the main focus of today's session

Framework for short variant discovery (GATK)

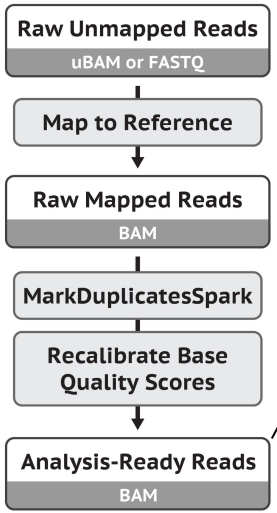
Framework for short variant discovery (GATK)



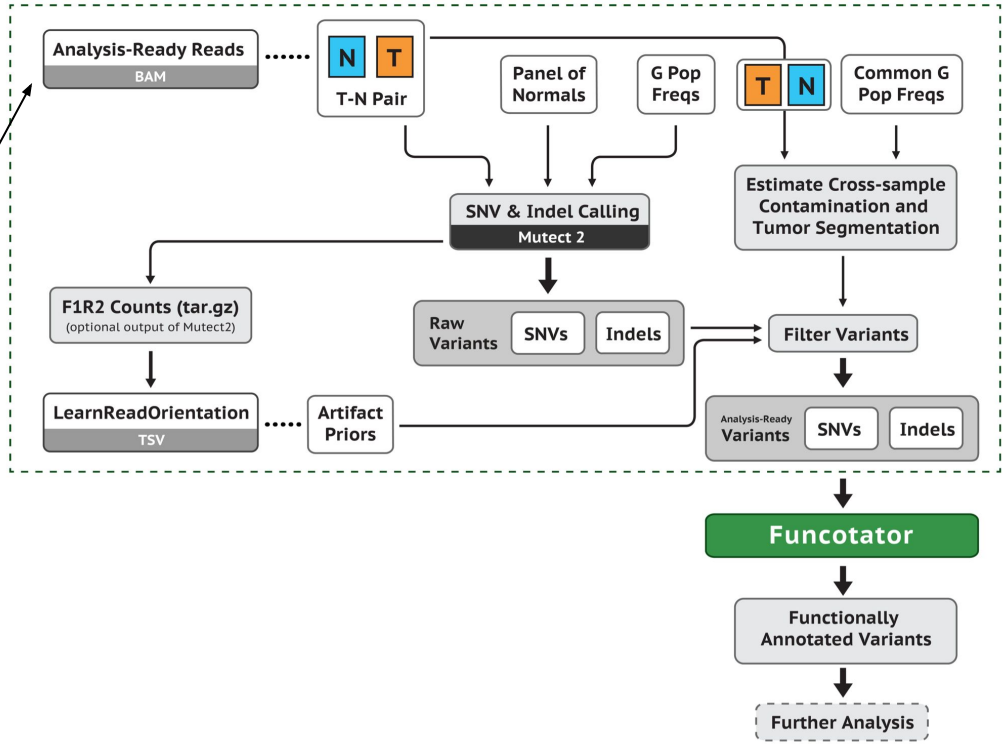
- Developed by the Broad Institute: https://gatk.broadinstitute.org/.
- Industry standard framework/pipeline for identifying somatic and germline short-variants.
- GATK workflow is expanding to include SCNAs and SVs.
- Also includes various tools to perform processing and QC
    - Designed mostly for WES and WGS Illumina data, but could be adapted to other technologies.
- GATK Best Practices
    - Step-by-step recommendations for processing and analysis from raw reads to variant discovery in high-throughput sequencing.

GATK somatic short variant discovery pipeline

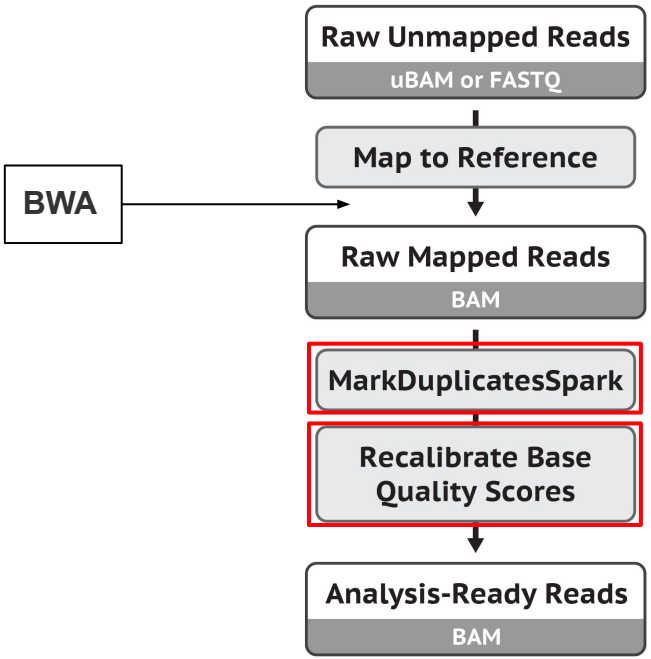
Data preprocessing



Variant discovery workflow



GATK - Data preprocessing for variant discovery



MarkDuplicatesSpark:

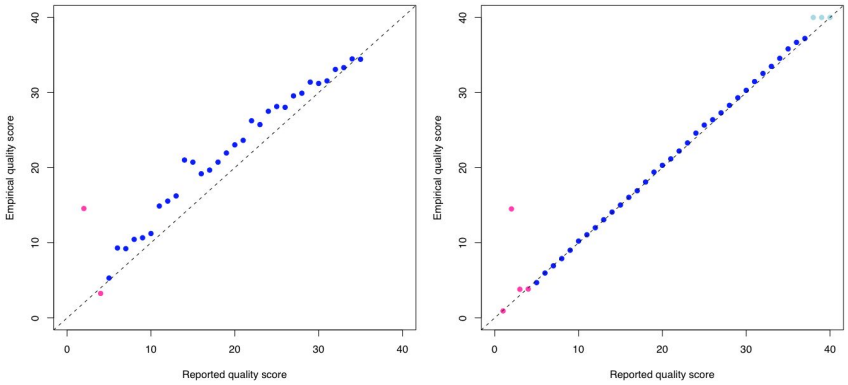
- Identifies and marks duplicates leaving only one read unmarked (lecture 3)
    - Sorts reads into coordinate-order
- Marked read pairs are ignored during variant discovery

Recalibrate Base Quality Scores:

- Base quality scores are confidence scores emitted by the sequencer for each base
    - Aid in weighing the evidence for or against possible variants
- Systematic quality score bias can occur during library prep and sequencing
- Recalibrated with machine learning to correct patterns of systematic bias

Effects of base quality score recalibration

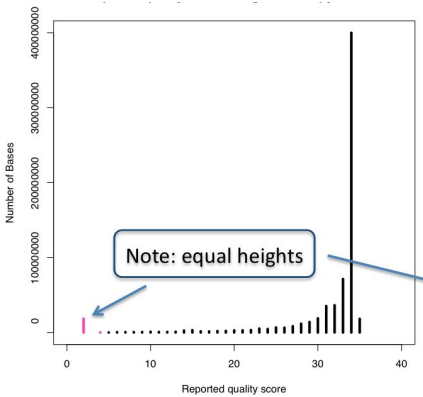
Empirical Quality Score vs. Reported Quality Score



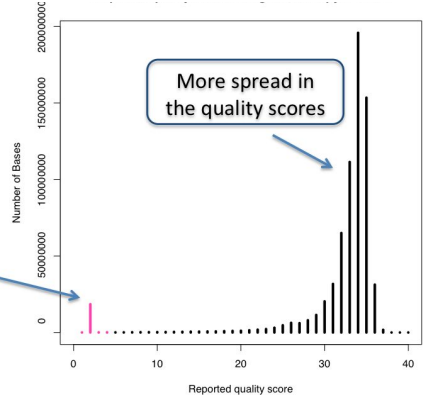
Original Data

After GATK Recalibration

Distribution of quality score



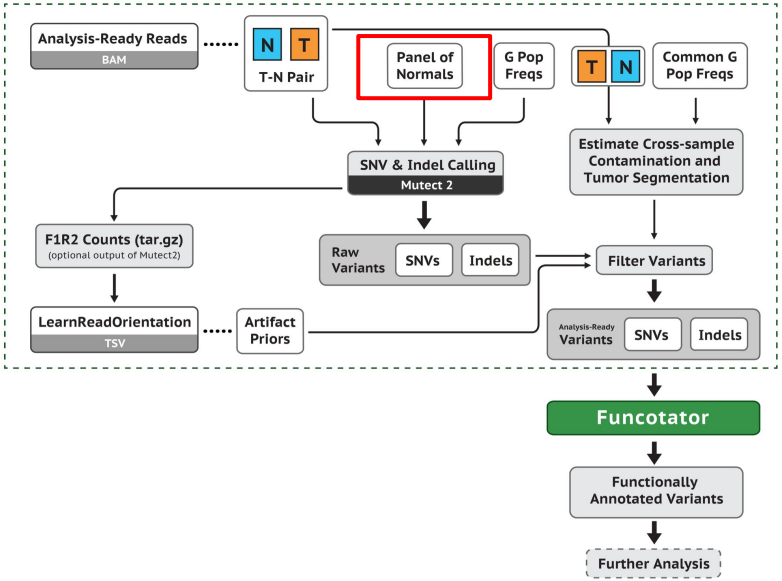
Original Data



After GATK Recalibration

From GATK

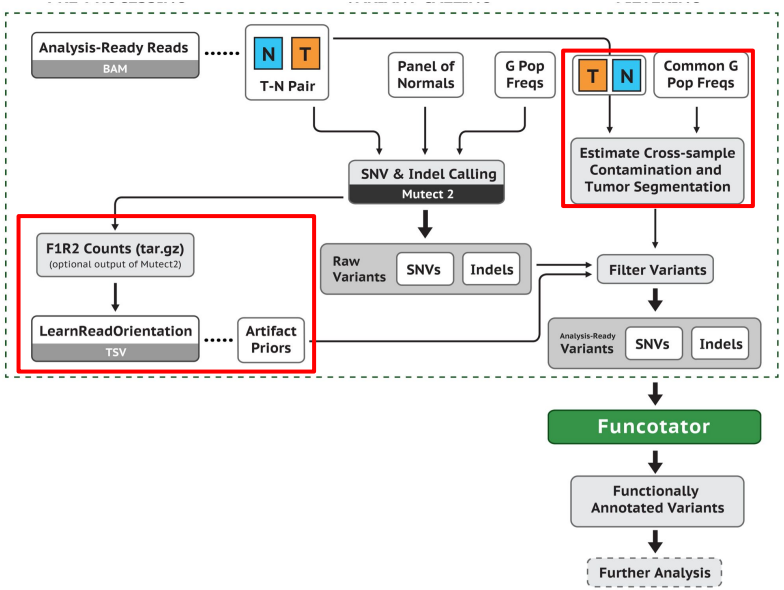
Panel of Normals (PoN)



- **Used to capture recurrent technical artifacts**
- **Common criteria for PoN:**
  - Derived from normal samples
  - As technically similar as possible to the tumors (e.g. same sample preparation methods, sequencing technology, etc.)
  - Recommended min. 40 samples
  - Usually from blood
    - Ideally young, healthy individuals with low chance of having undiagnosed cancers

- Public GATK PoN:
  - *Hg38:* gs://gatk-best-practices/somatic-hg38/1000g_pon.hg38.vcf.gz
  - *Hg19/b37:* gs://gatk-best-practices/somatic-b37/Mutect2-exome-panel.vcf

Learn Orientation Bias Artifacts & Cross-sample Contamination



LearnReadOrientation:

- Particularly important for FFPE and Illumina Novaseq samples
- Filters oxidation and deamination artifacts with higher frequency in one read pair orientation.
    - Read more about sequencing artifacts here

CalculateContamination:

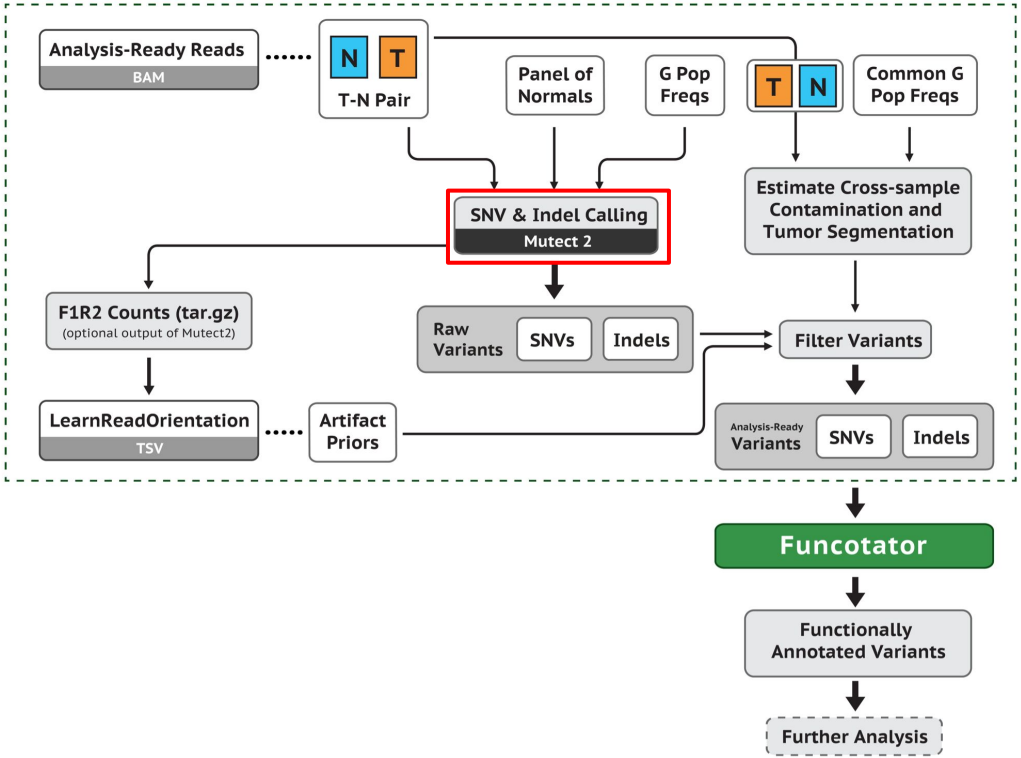
- Calculate fraction of reads coming from cross-sample contamination

Variant callers

Somatic variant calling methods - SNVs/indels

Algorithm Class	Details	Example Callers
Heuristic approaches	• Apply heuristics to filter sequencing errors • Use simple statistical tests to filter somatic variants from germline (e.g. fisher exact test of tumor vs normal VAF)	*VarScan2, qSNP, Shimmer, RADIA, SOAPsnv, VarDict*
Bayesian approaches	• Evaluate the Bayesian probability of a true variant given estimated sequencing error rates, somatic and germline mutation rates, etc.	*MuTect, Strelka, EBCall, deepSNV, SomaticSniper, FaSD-somatic, SAMtools, JointSNVMix2, Virmid, SNVSniffer, Seurat, CaVEMan, LoFreq, MuSE*
Haplotyping approaches	• *de novo* genome assembly around mutation sites • Allows for detection of larger indels + sensitive calling in areas with high mutation density	*MuTect2, Strelka2, Platypus, HapMuC, LocHap, FreeBayes, TNscope*
Machine learning approaches	• Use machine learning to distinguish true somatic variants from germline or artifacts • Needs a training set of true somatic variants or ensemble voting from other callers	*MutationSeq, SomaticSeq, SNooPer, BAYSIC*

GATK - Somatic variant caller Mutect2



Mutect2 currently can call somatic SNVs and indels in 3 modes:

- 1) Tumor with matched normal (joint calling of multiple tumors/normals from the same individual supported)
- 2) Tumor-only (additional filtering by functional significance by Functotator is recommended)
- 3) Mitochondrial

Sentieon genomics tools (commercial)

Sentieon DNASeq Variant Calling Workflow Demonstrates Strong Computational Performance and Accuracy

Kendig et al. 2019, Front. Genet

https://www.sentieon.com/

Sentieon® Analysis Pipelines & Tools

Alignment	Sentieon® BWA, STAR, Minimap2: identical match open source result with >2X speedup.
Germline SNV/INDEL Variant Calling	**DNAseq®**: PrecisionFDA award-winning software. Matches GATK 3.3-4.1, and without downsampling. Results up to 10x faster and 100% consistent every time. **DNAscope**: Improved accuracy and genome characterization. Machine learning enhanced filtering producing top variant calling accuracy. Supports both short reads and PacBio HiFi long reads.
Somatic SNV/INDEL Variant Calling	**TNseq®**: Matches MuTect, MuTect2 v3.8 - 4.1 without downsampling for higher accuracy and improved detections of low allelic fraction variants. **TNscope®**: Winner of ICGC-TCGA DREAM challenge. Improved accuracy, machine learning enhanced filtering. Supports molecular barcodes and unique molecular identifies.
Structural Variant Calling	Germline and somatic SV calling, including translocations, inversions, duplications and large INDELs.
Joint Calling	Supports large-cohort joint calling of over 200,000 WGS samples directly from gVCF and without intermediate steps.
BCL-FASTQ Tool	Sentieon®'s external library accelerates BCL to FASTQ conversion by 1.5 - 2x.
RNA Variant Calling	Matches GATK RNAseq variant calling Best Practices.

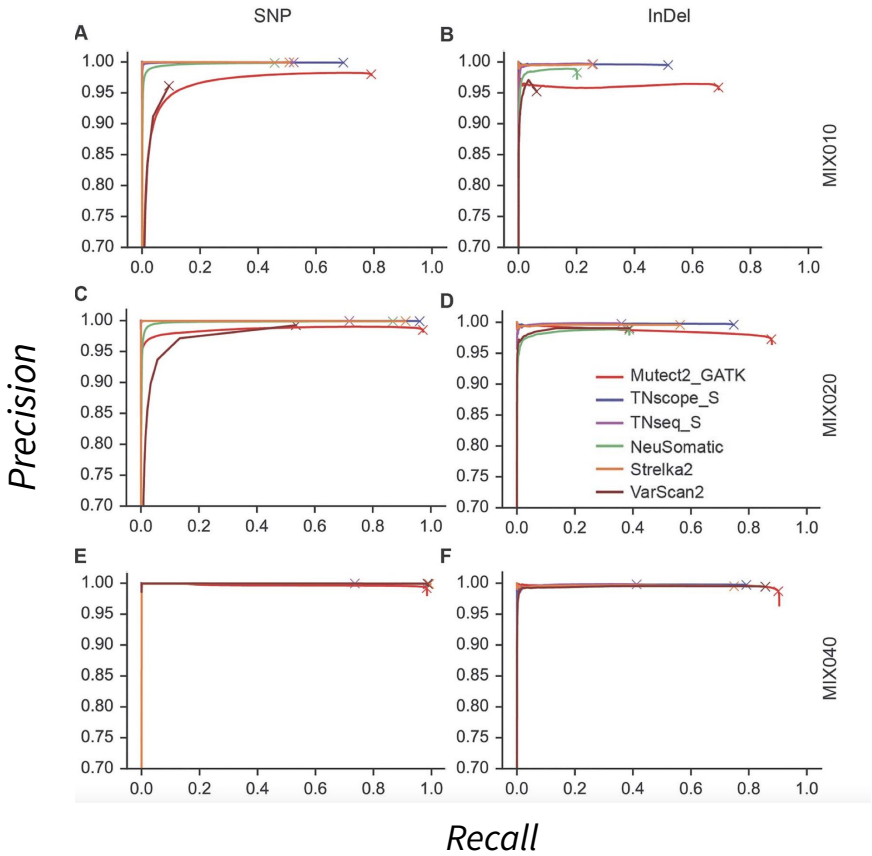
Popular somatic SNV/Indel callers

Name	Types of detection	Paired Normal Requirement	Panel of Normals (PoN)
MuTect2[1]	SNVs, Indels	Somatic tumor-only possible using PoN	Supported
VarScan2[2]	SNVs, Indels	Required for somatic filtering	No
Strelka[3]	SNVs, Indels	Required	No
CaVEman[4]	SNVs	Required	Supported
LoFreq[5]	SNVs, Indels	Required for somatic filtering	No
VarDict[6]	SNVs, Indels	Required for somatic filtering	No
MuSE[7]	SNVs	Required	No
TnScope/TnSeq (Sentieon, commercial)[8]	SNVs, Indels	Somatic tumor-only possible using PoN	Supported
NeuSomatic[9]	SNVs, Indels	Required	No

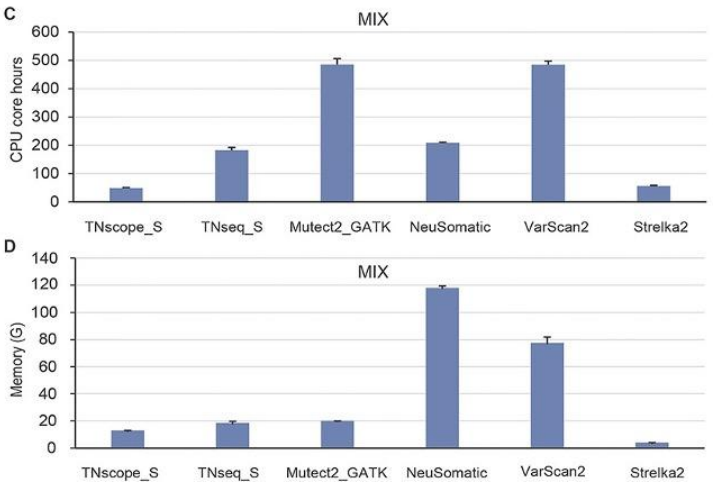
- Variant callers with tumor-only modes could potentially be useful for study without matched normal (additional downstream filtering is often recommended)
- These algorithms may work with PacBio, though somatic benchmarking is lacking

Benchmarking somatic variant callers

Precision-recall curves with different mixed tumor ratios



Computational cost of different callers



- Tumor purity affects variant calling accuracy and sensitivity
- Sentieon TNscope and GATK Mutect2 have best recall, particularly at low purity
  - TNscope also has low computing resource usage

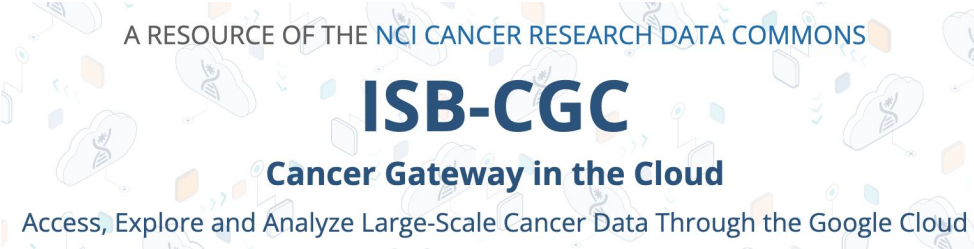
Pei et al. 2020, *Brief Bioinform*

An ENSEMBLE approach:
Example of somatic variant calling pipeline by CGR

- Substantial discrepancies exist among the calls from different callers
  - Multiple variant callers needed in pipeline
- Somaticseq2
  - https://github.com/NCI-CGR/somatic-conda
- Somatic-combiner
  - https://github.com/mingyi-wang/somatic-combiner
  - A consensus ensemble approach which can combine somatic variants generated from seven popular callers:
    - **LoFreq, MuSE, MuTect2, MuTect, Strelka, VarDict and VarScan**.

Wang et al. 2020, Sci Rep

NCI Cloud resources for somatic variant calling



- NCI Cloud Resources are components of the NCI Cancer Research Data Commons (see lecture 2).
- Eliminate the need to download and store extremely large datasets.
- Hosts large datasets such as TCGA, TARGET, Human Cell Atlas.

Selected somatic short variant workflows from NCI Cloud resources:

- Strelka
- VarDict
- VarScan
- GATK
- Other variant calling (germline, SVs, SCNAs), multi-omics and imaging workflow also available.

Mutation Calling with Unique Molecular Identifiers (UMIs)

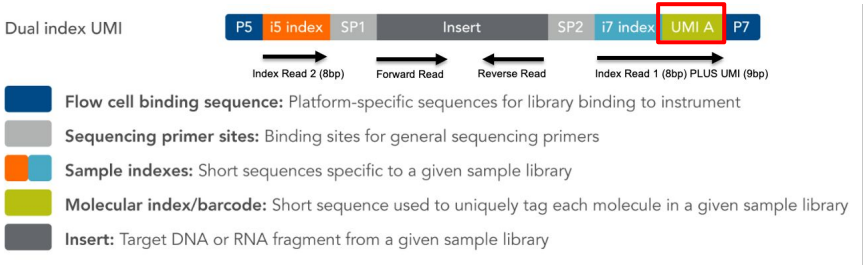


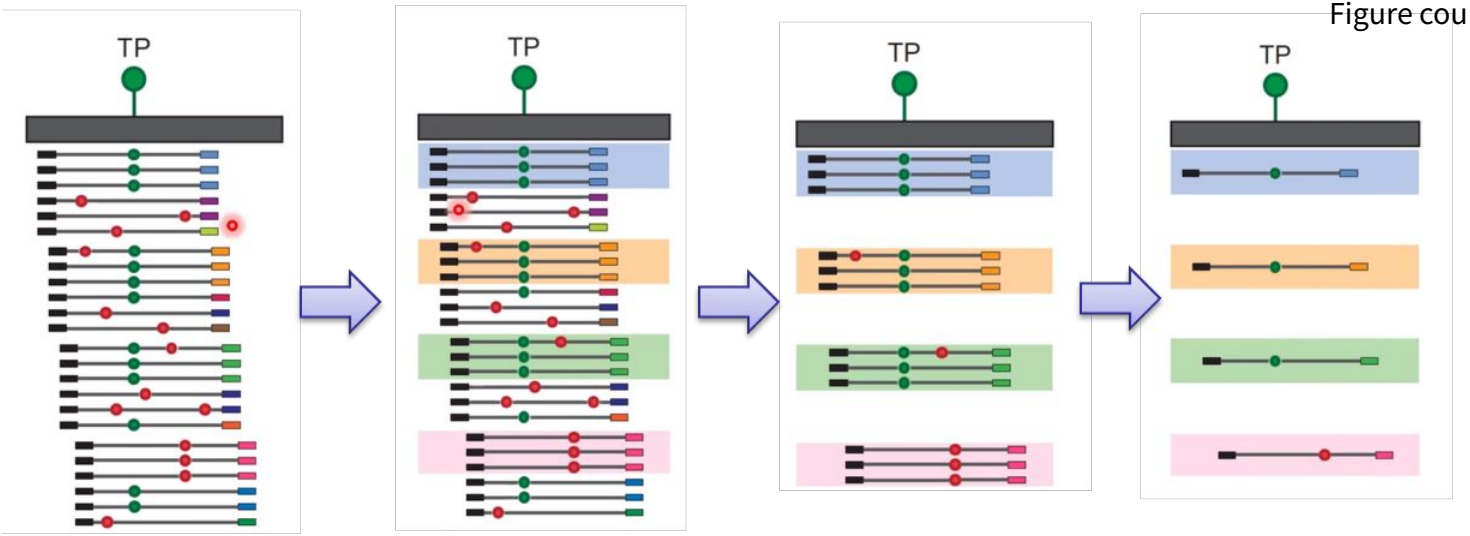
Figure courtesy of CGR

UMIs

- 9-base sequences embedded in the adapter sequence of all CGR library prep. UMI is unique per molecule prior to DNA amplification
- Intended for error-correction
    - Reads from the same molecule are collapsed into a single consensus read
- An analytical choice of whether to use UMIs, requires very deep coverage (>500 - 1000x)
    - Due to cost, mostly used for targeted sequencing
- Can improve sensitivity of low frequency calls (<5%)

UMIs for error correction

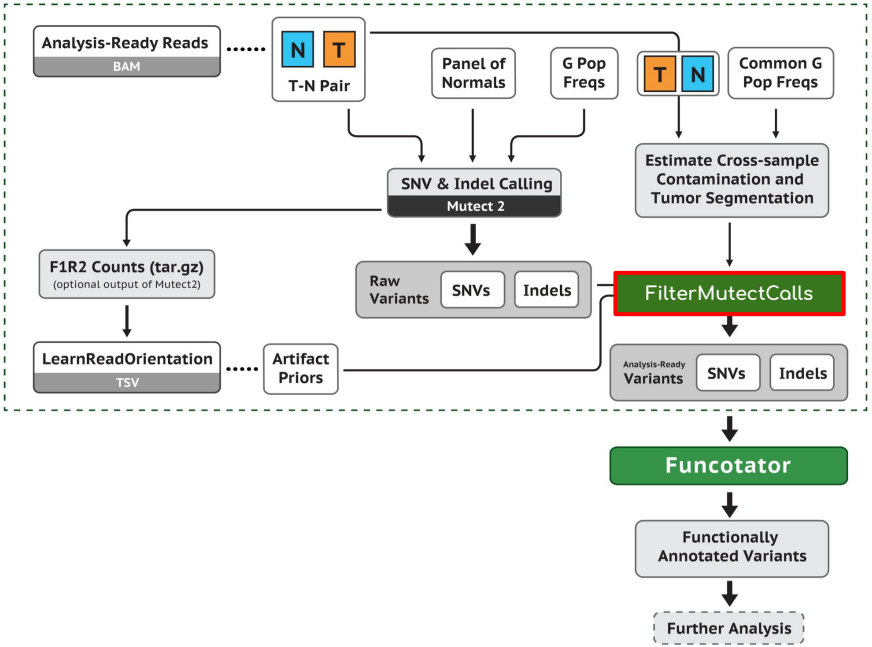
Figure courtesy of CGR



- If the start/stop sites of multiple read pairs are the same ***and*** they have the same UMI tag, they are identified as duplicates.
- Only molecules that have multiple read pairs are kept
    - Singleton read pairs are discarded, reducing final coverage for variant calling.
- Many late-stage PCR errors and sequencing errors can be eliminated, reducing noise.
    - Because of the reduction in errors, true low frequency variants are more readily called.

Variant filtering & annotation

Somatic variant filtering



- To remove potential artifact signals caused by sequencing platform, alignment software, and incomplete or complex nature of reference genomes.

  [**GATK - FilterMutectCalls**]

- To remove potential contaminated signals from germline variants.

  [**Germline Population Variant Database (PVDs)**]

Friendly Reminder: *Raw reads preprocessing & Somatic variant calling & Filtering & Annotation will be shown in the later practical session. Welcome!*

Commonly used tags for somatic variants filtering

Filtering Tags	Description
alt_allele_in_normal	filter out variants with enough evidence as being present in the paired normal sample
germline_risk	filter out variants that show sufficient evidence of being germline variants based on Population Variant Databases
homologous_mapping_event	filter out variants with more than three events present nearby, as they are usually indicative of false-positive calls
clustered_events	filter out variants linked to multiple events present on the same haplotype as they are usually indicative of false-positive calls
str_contraction	filter out variants which are contraction of short tandem repeat
Strand Bias Events	filter out variants which locates only in one strand (Forward/Reverse)
low_t_alt_frac	filter out variants with low alternate allele fraction in tumor sample

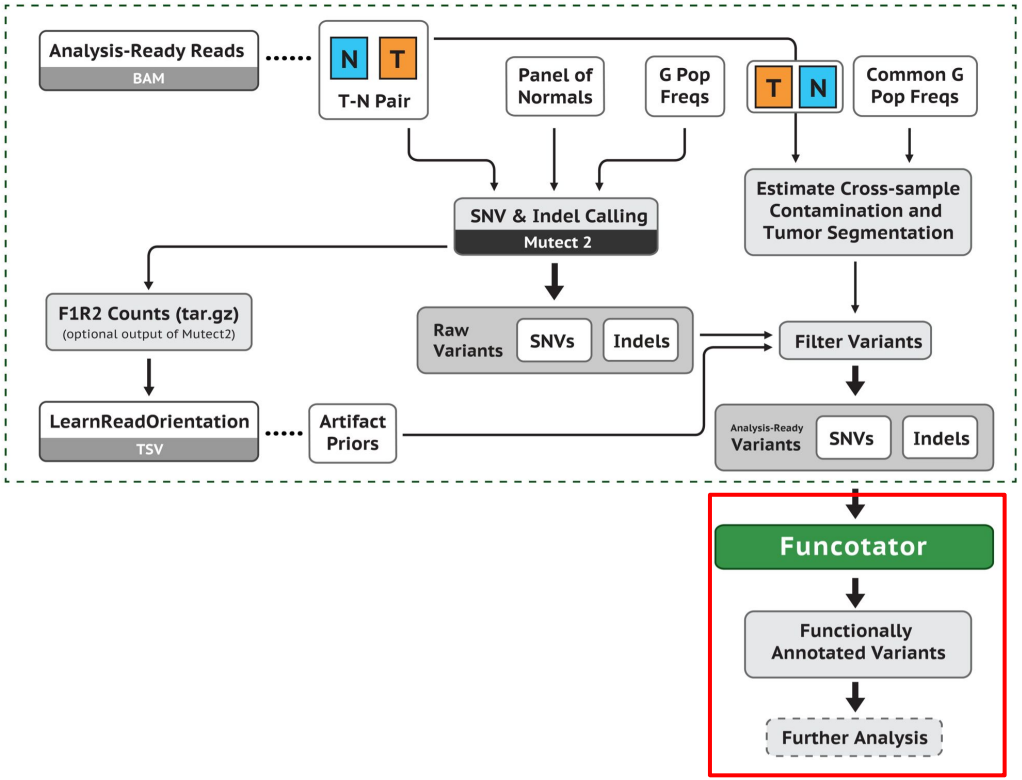
(URL: https://support.sentieon.com/appnotes/out_fields/)

Friendly Reminder: *in the next IGV visualization and validation section, detailed examples of the common filtered events*

Common population databases for germline-based filtering

Databases	Description	Filtering criteria suggestions	Links
The 1000 Genomes Project	Germline SNPs and SVs from >3200 individuals covering 25 populations	population specific MAF (Minor Allele frequency) <0.1%.	https://www.internationalgenome.org/
ESP	Germline mutations from the NHLBI Exome Sequencing Project.	population specific MAF <0.1%.	https://evs.gs.washington.edu/EVS/ **Biowulf: /fdb/annovar/current/hg38/hg38_esp6500siv2_all.txt**
gnomAD v2	Germline mutations from 125,748 whole exomes and 15,708 whole genomes	population specific MAF <0.1%.	https://gnomad.broadinstitute.org/ **Biowulf: /fdb/gnomad/release-2.0/**
gnomAD v3	Germline mutations from 76,156 whole genomes	population specific MAF <0.1%.	https://gnomad.broadinstitute.org/ **Biowulf: /fdb/gnomad/release-3.0/**

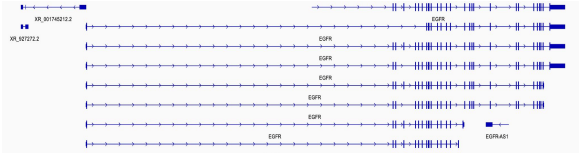
Variant annotation



- Variant annotation is the process of collection of detailed information from knowledge bases and publications to interpret a given genomic alteration.
- **Funcotator** is a useful module to provide functional annotation of somatic alterations.
- Example annotations: *gene name, position, frequency, classification (Intron, UTR, Splicing Site), etc.*
- Users can add their own annotation sources based on a set of public database resources.

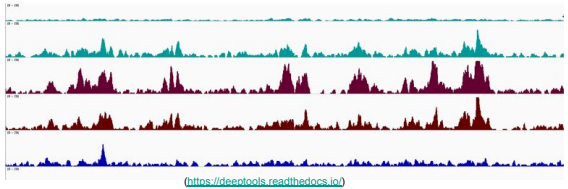
Different strategies of annotations

- **Gene-based annotation:**



Whether a genomic mutation happened **in known protein coding / non-coding regions**. Users can easily customize their analysis using different public databases with detailed gene annotation information.

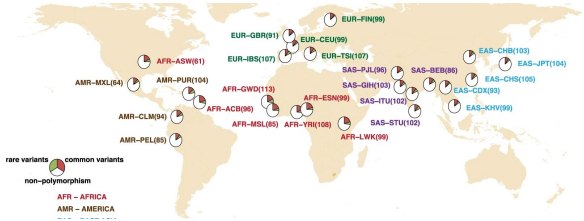
- **Region-based annotation:**



(https://deeptools.readthedocs.io/)

Whether a genomic mutation happened **in specific genomic elements**. For example, Open chromatin (ATAC-seq peaks), Histone mark enrichment (ChIP-Seq peaks), or many other annotations on genomic intervals.

- **Filter-based annotation:**



(PMID: 33547344)

Whether a genomic mutation is reported **in specific cancer databases** like cosmic. What is the frequency of this mutation among global populations?

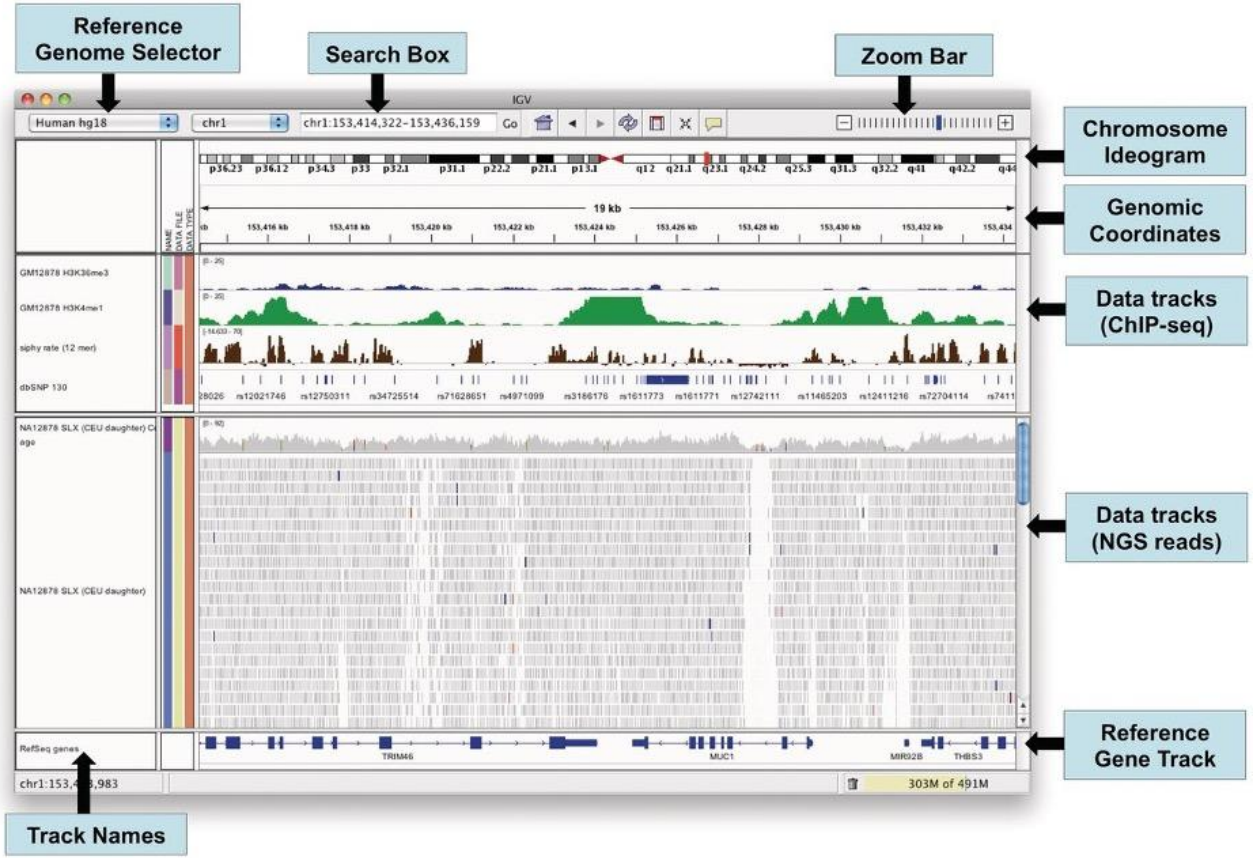
Useful resources for variants annotation

Category	Resource	Descriptions	Links
Gene-based annotation	UCSC	Comprehensive annotations at both genes and transcripts levels	https://genome.ucsc.edu/ **Biowulf: /fdb/annovar/current/hg38/hg38_refGene.txt**
Region-based annotation	ENCODE	Encyclopedia of DNA Elements, including transcription factor binding sites, ChIP-Seq peaks, and ATAC-seq peaks	https://sites.google.com/site/jpopgen/dbNSFP **Biowulf: /fdb/annovar/current/hg38/hg38_wgEncodeGencodeBasicV40.txt**
Filter-based annotation	COSMIC	Catalogue of somatic mutations in cancer	https://cancer.sanger.ac.uk/cosmic **Biowulf: /fdb/annovar/current/hg38/hg38_cosmic92_coding.txt**
	ICGC	Somatic mutations identified by International Cancer Genome Consortium	https://dcc.icgc.org/ **Biowulf: /fdb/annovar/current/hg38/hg38_icgc28.txt**
	Clinvar	Catalogue of human variants and phenotypes with supporting evidence	https://www.ncbi.nlm.nih.gov/clinvar/ **Biowulf: /fdb/annovar/current/hg38/hg38_clinvar_20220320.txt**

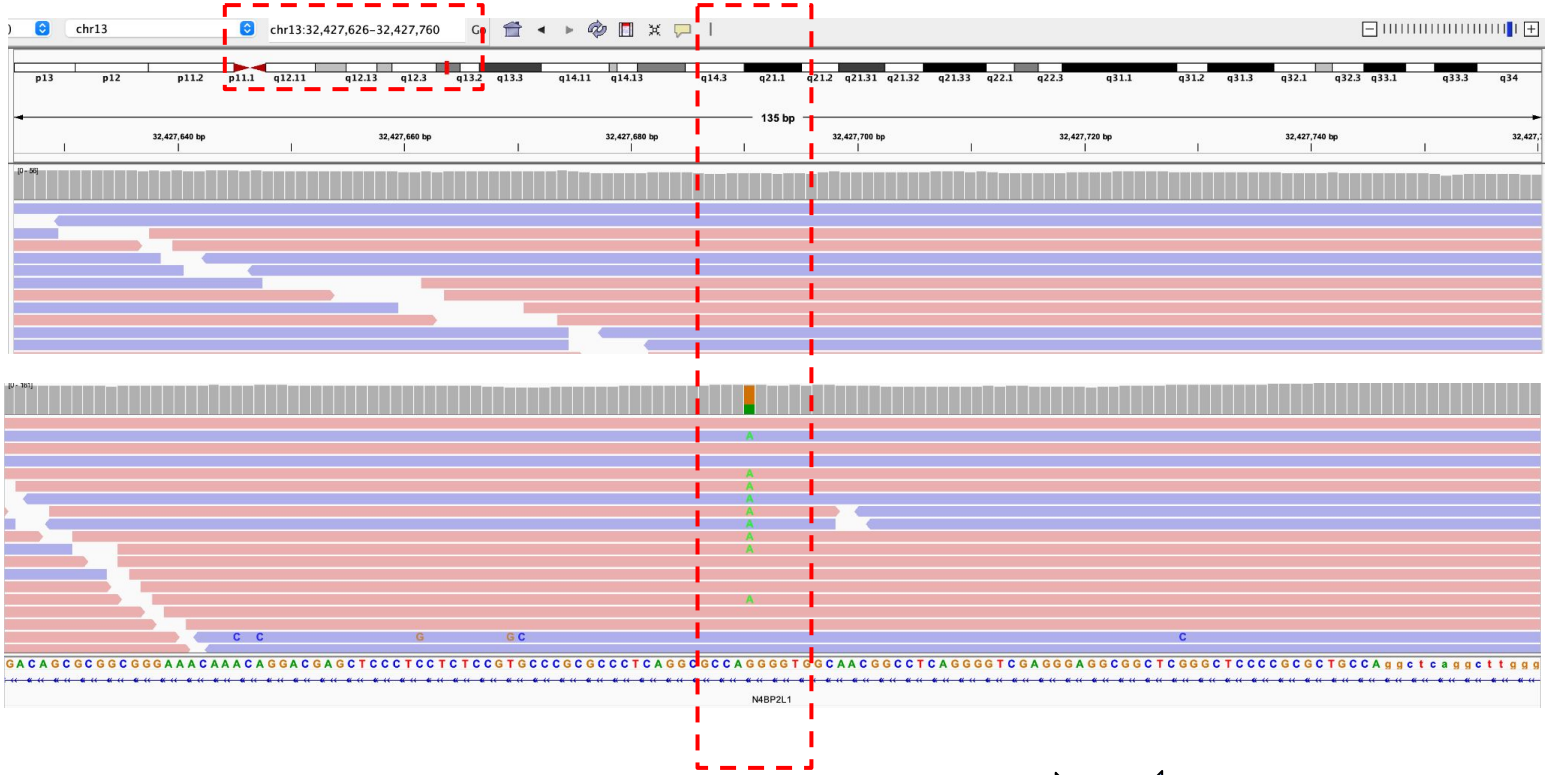
(URL: https://annovar.openbioinformatics.org/en/latest/user-guide/download)

IGV for visualization and validation

IGV - Integrative Genomics Viewer



1. Typical Good Examples of SNVs in IGV

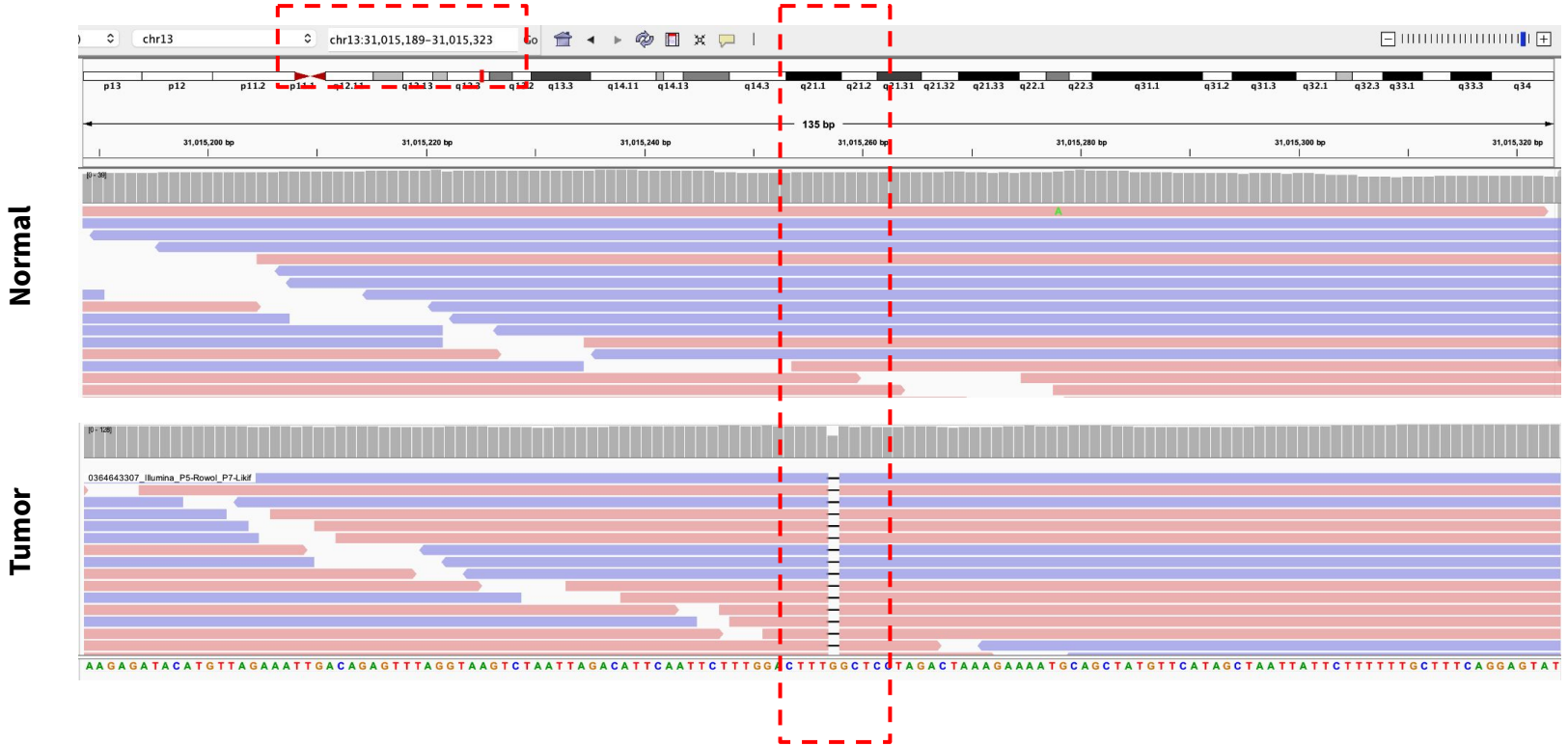


A: Green T: Red G: Brown C: Blue

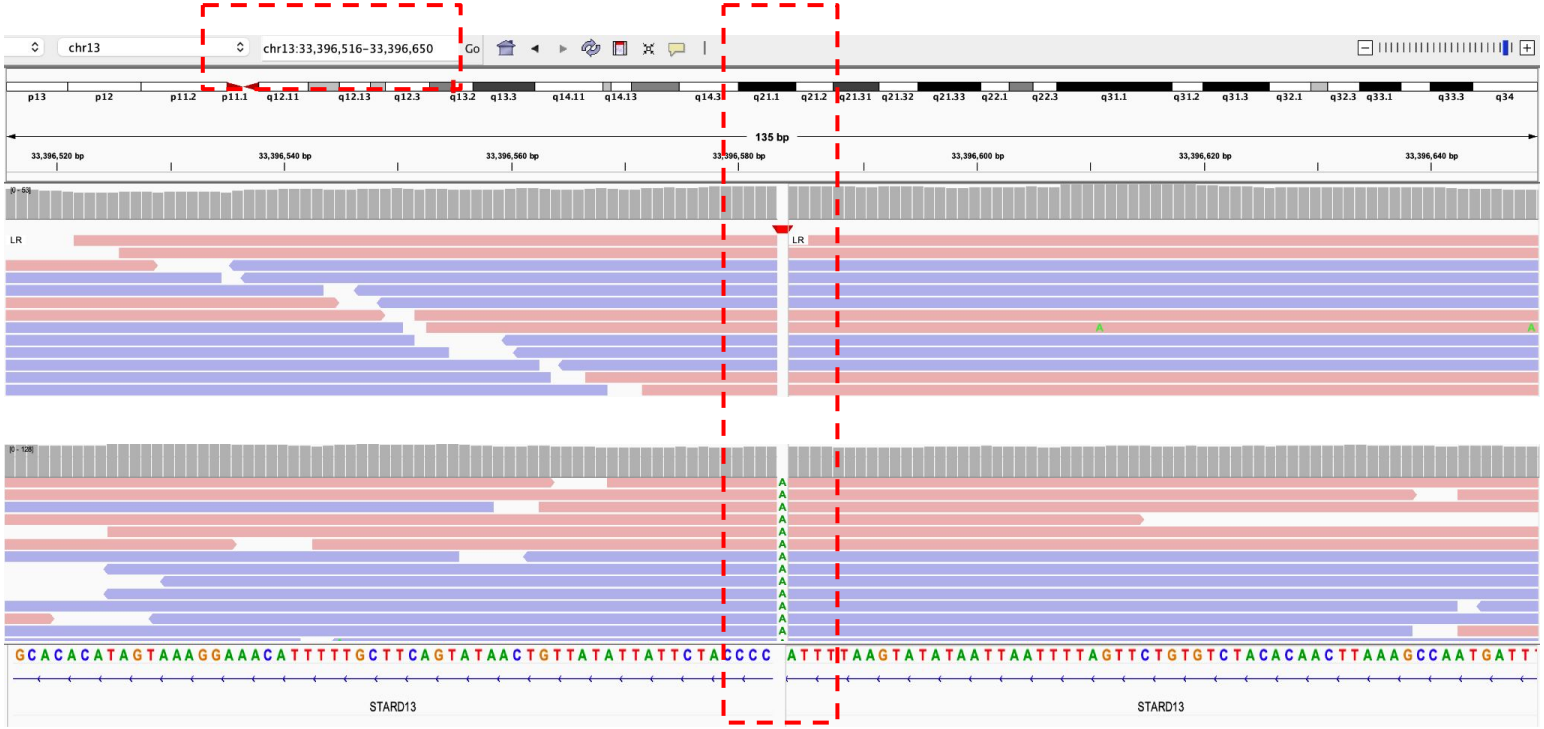
Forward Reads

Reverse Reads

2. Typical Good Examples of Deletion in IGV



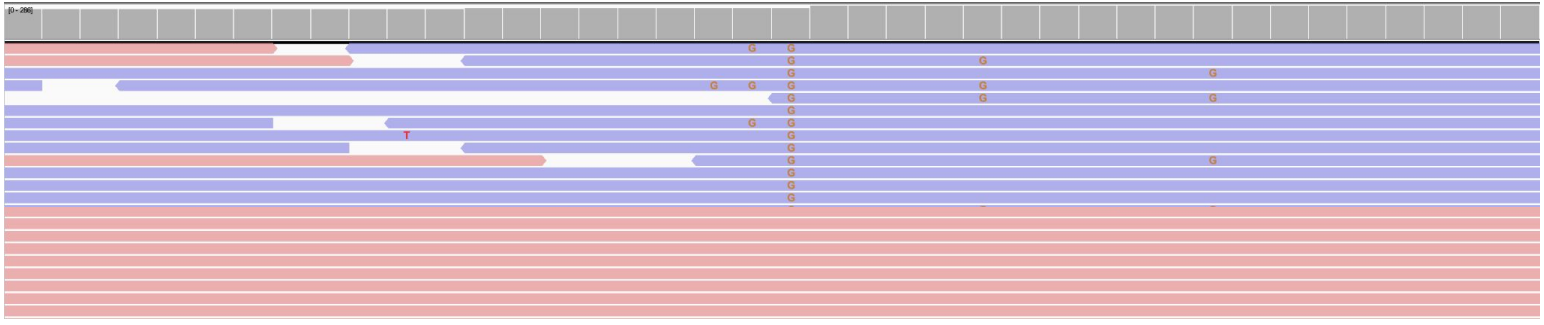
3. Typical Good Examples of Insertion in IGV



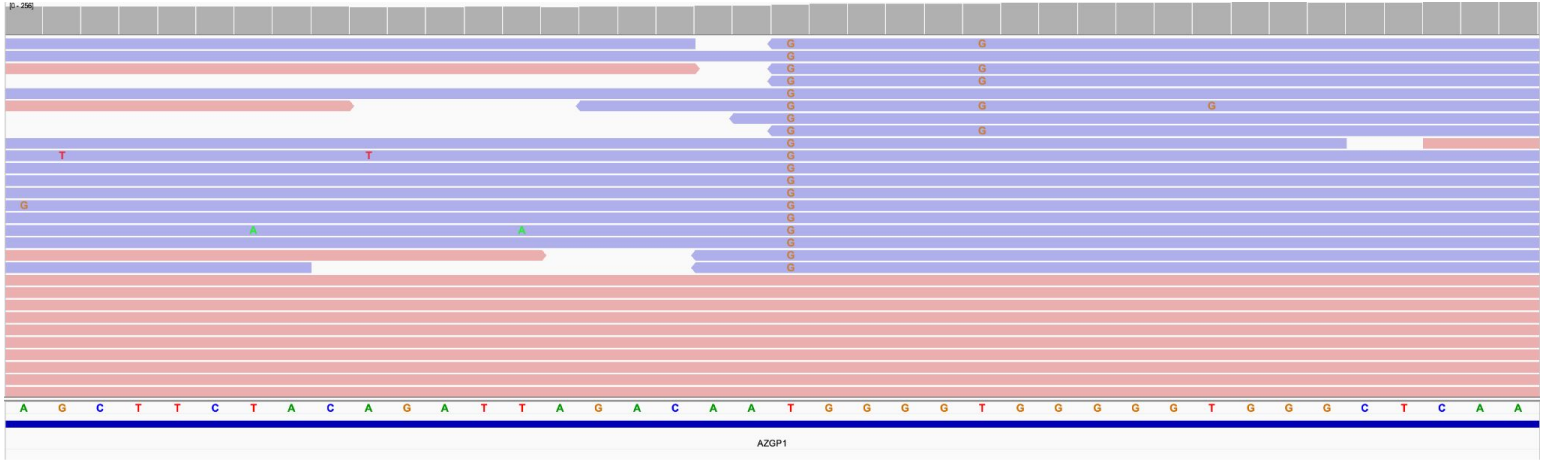
4. Suspicious Mutations - Germline Events

5. Suspicious Mutations - Strand Bias Events

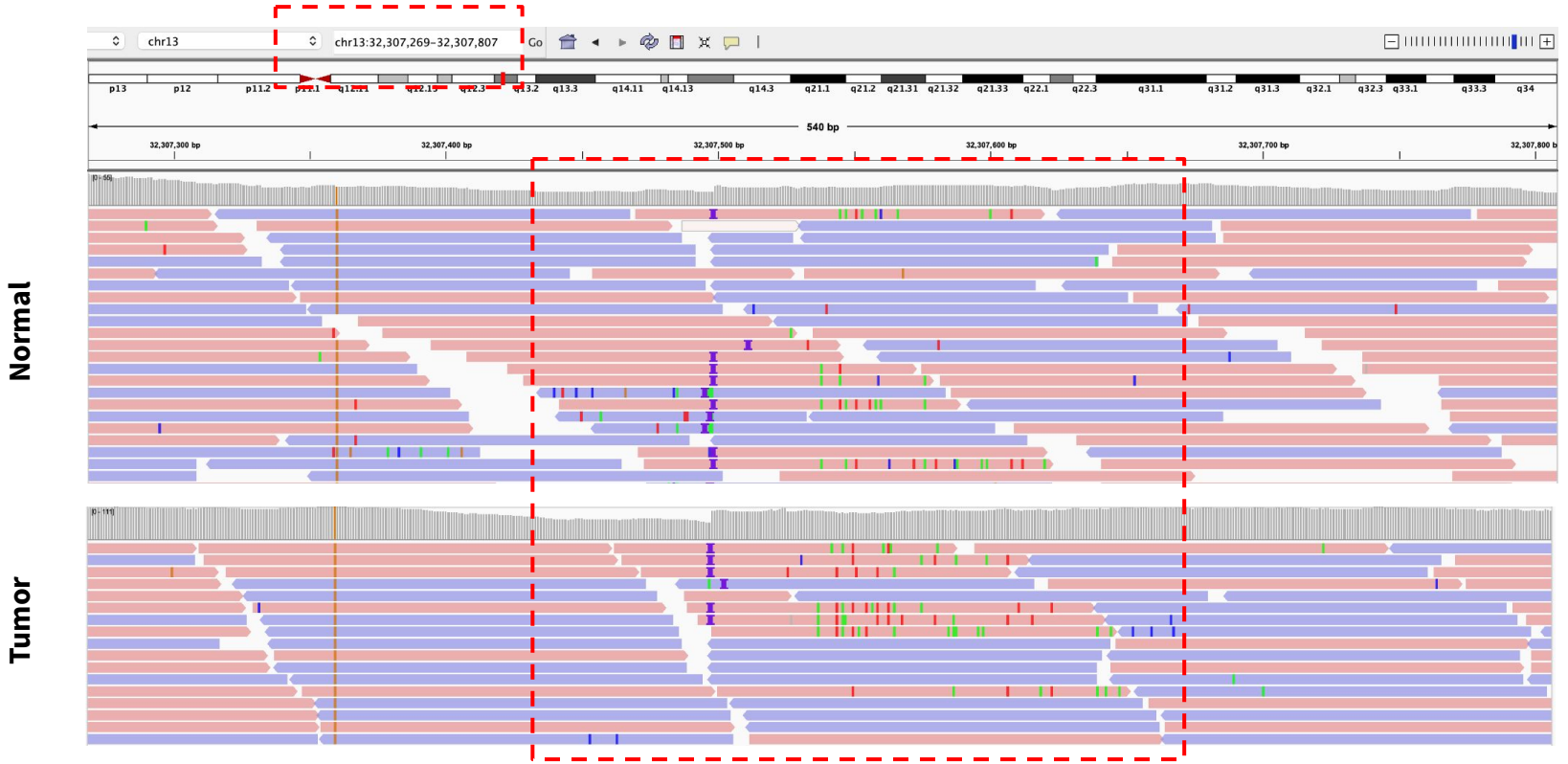
Normal



Tumor



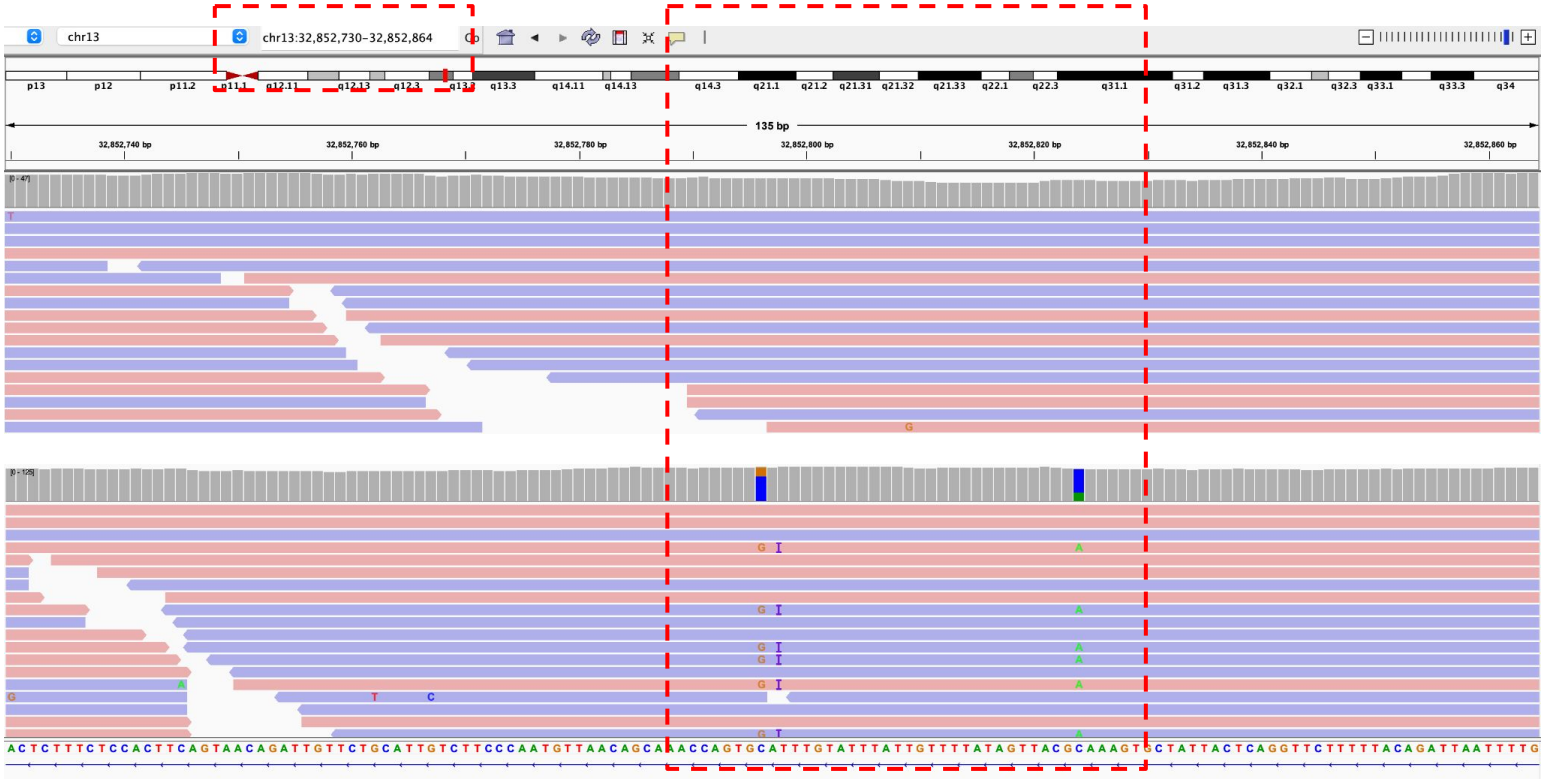
6. Suspicious SNV - Homologous Mapping Event



7. Suspicious SNV/Indel - Clustered Events

Normal

Tumor



Summary - Challenges for accurate somatic variant calling

Not as simple as identifying tumor-specific variants

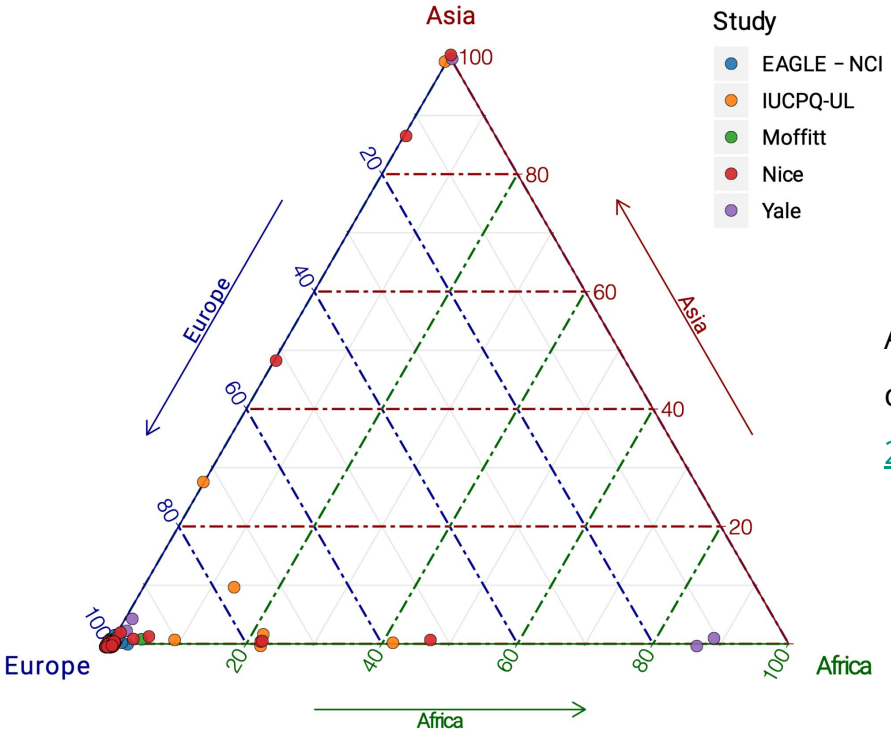
Challenges	Potential solutions
Artifacts from PCR amplification	Mark or even remove duplications before calling. Increase DNA sequencing library complexity. Utilize UMIs
Machine sequencing errors	Use strict variant hard filtering.
Incorrect local alignment of reads	Perform indel realignment before calling.
Tumor heterogeneity/Low tumor purity	Increase sequencing depth or manually call variants, which allows detection of low VAF variants.
Sequencing depth and read length, DNA library complexity and Insert size	Refer to Session 3 material for solutions.
Tumor-normal cross-contamination	Perform strict BAM-level data QC. Particularly check for any copy number alterations in normal tissue.
Artifacts related to the source materials (FFPEs or alcohol fixed tissue)	Use strict variant filtering and carefully check the VAF, mutational patterns, or known mutations after variant calling. Mutational signature analysis can be used to remove artifacts (e.g. FFPESig).
Complicated genome regions (long repeat region, GC/AT-rich regions, centromere etc.)	Define a blacklist of these regions and try to exclude any variants in this blacklist for downstream analysis.

Common somatic mutational analyses

Common somatic analyses (SNVs/Indels)

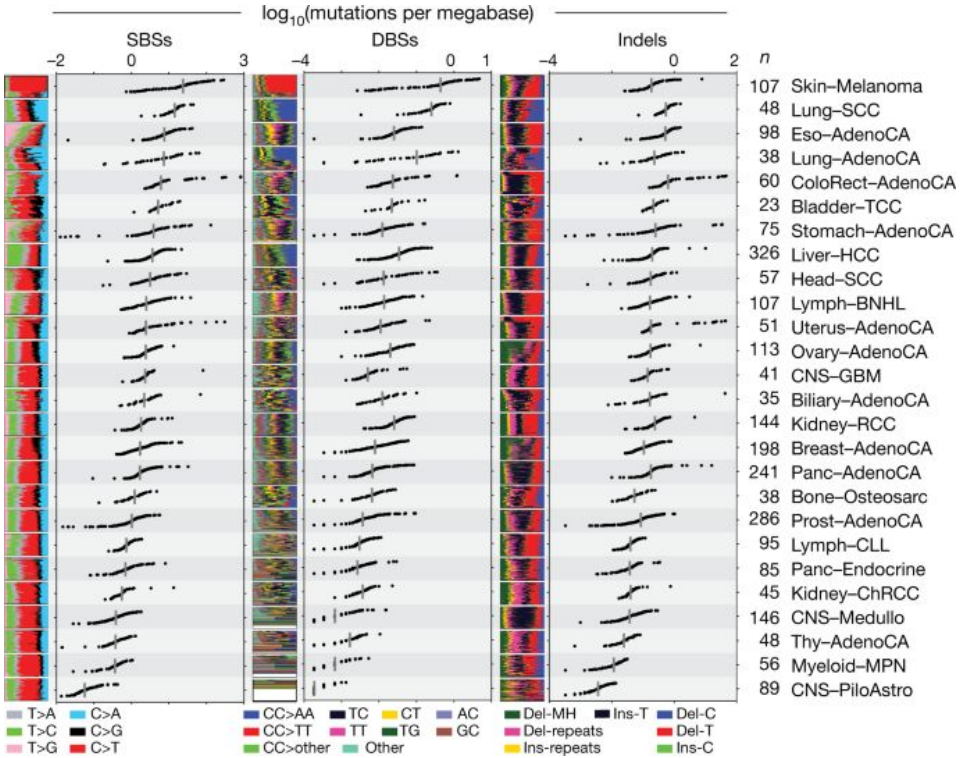
- Ancestry inferred from WGS (not somatic analysis, but very helpful)
- Tumor mutational burden (TMB) analysis
- Mutational exclusivity analysis
- Hotspot mutation analysis
- Pathway analysis
- Mutational signatures (session 5)
- Cancer driver genes (session 8)
- Mutation clonality analysis (session 9)

Ancestry inferred from WGS



ADMIXTURE analysis of 232 WGS lung cancer in never smokers (Zhang et al. 2020, *Nature Gen*)

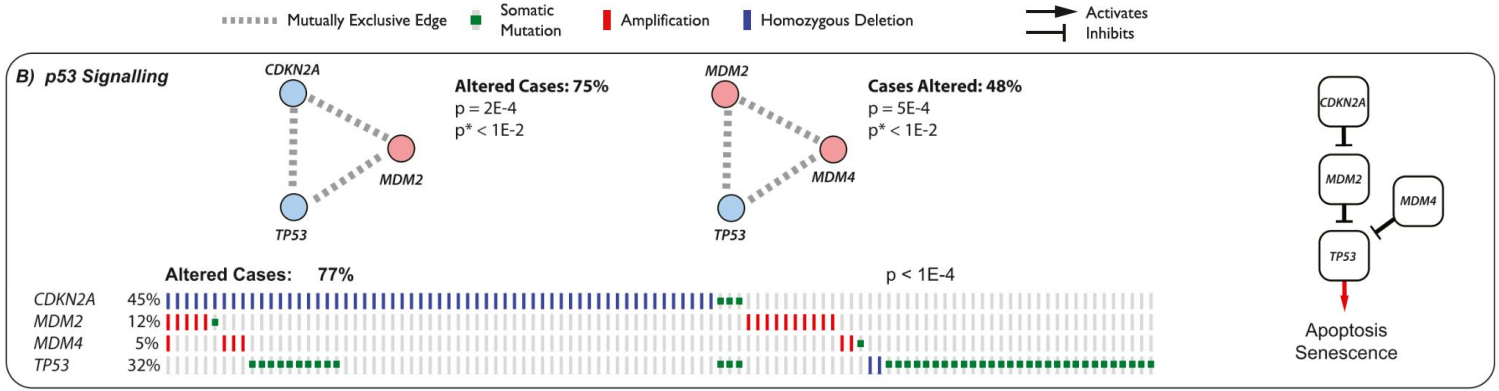
Tumor mutational burden (TMB) analysis



- Defined as total number of mutations (usually for short variants) in a cancer genome
- TMB is often considered a predictor of prognosis
    - High TMB is also considered as predictor of immunotherapy response
- Figure shows mutational burden across PCAWG tumor types (Alexandrov et al. 2020, *Nature*)

Mutational exclusivity

(Ciriello et al. 2011, *Genome Research*)

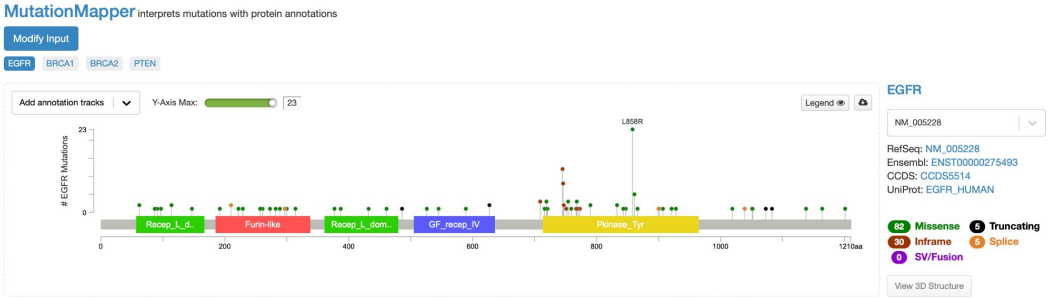


- Genetic alterations of a group of genes in the same pathway does not tend to occur in the same sample.

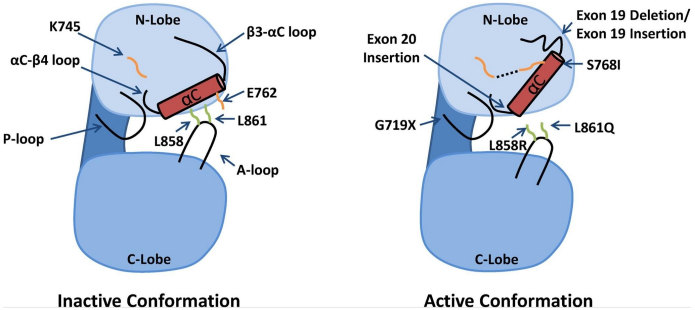
E.g. Mutations in *p53* signaling pathways often show mutual exclusivity in glioblastoma.

- *CDKN2A, MDM2, MDM4, and TP53* are in the same pathway regulating responses to DNA damage/apoptosis, etc.
- *Functional redundancy hypothesis:* The pathway is compromised and apoptosis evaded once one gene is altered. Additional alterations to the pathway do not change the effect on the apoptosis process and are not selected for.
- *Co-occurrence hypothesis:* Alteration to a second gene within the same pathway might lead to a disadvantage for the cell e.g. cell death.
- Application: synthetic lethality for therapeutic design. E.g. Targeting both *KRAS* and *EGFR* in LUAD

Hotspot mutation analysis



[MutationMapper]



[Harrison et al. 2020]

- Mutation hotspots can help reveal mutagenic mechanisms, or can reveal information about the functional domains of a target protein.
    - Could indicate preferential expansions of mutants with high fitness
- L858R is a common activating EGFR mutation in non-small cell lung cancer (right figure)
- Hotspot mutations occur more frequently in oncogenes than tumor suppressor genes

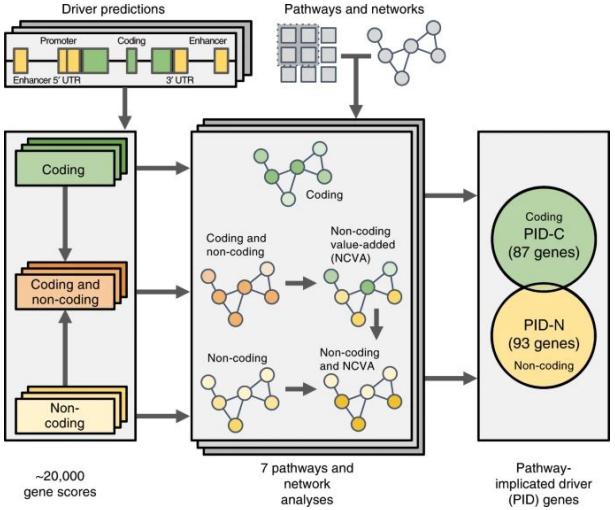
Pathway analysis

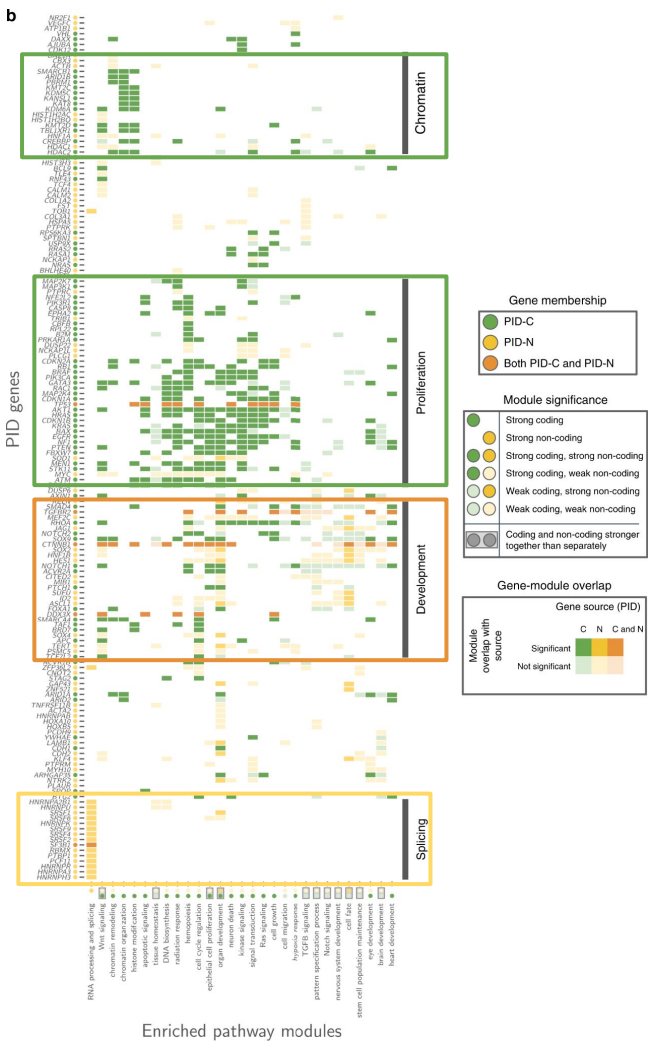
- To identify pathways disrupted by somatic mutations
  - Potential therapeutic treatments targeting the pathways could be developed
- Pathway and network analysis of more than 2500 whole cancer genomes (Reyna et al. 2020, *Nature Comm*)

Coding: chromatin remodelling and proliferation

Non-coding: RNA splicing

Coding and non-coding: Developmental pathways (Notch and Wnt signalling)





Mutational signatures

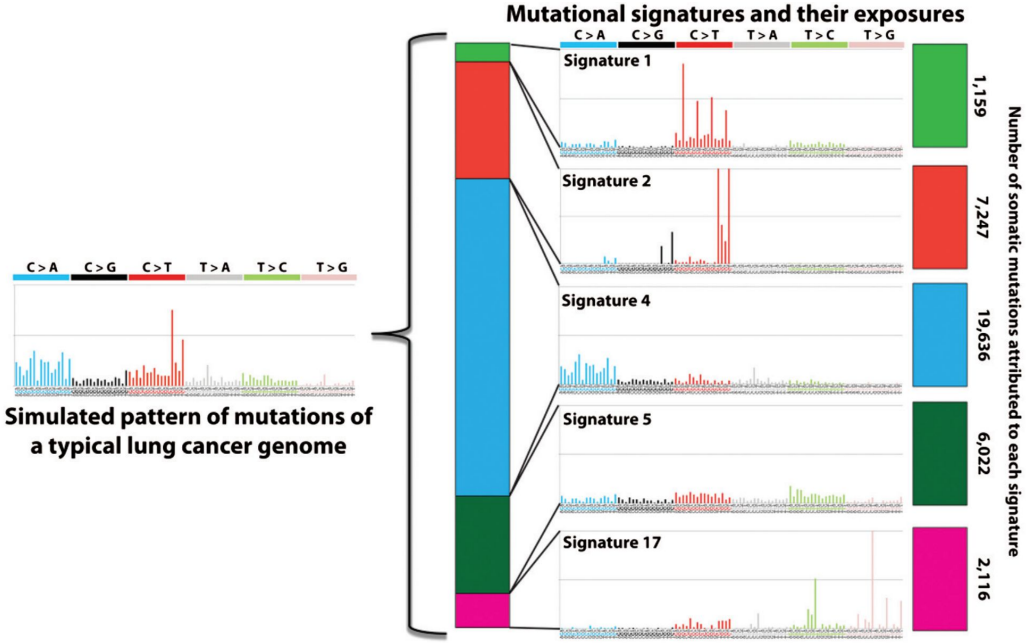


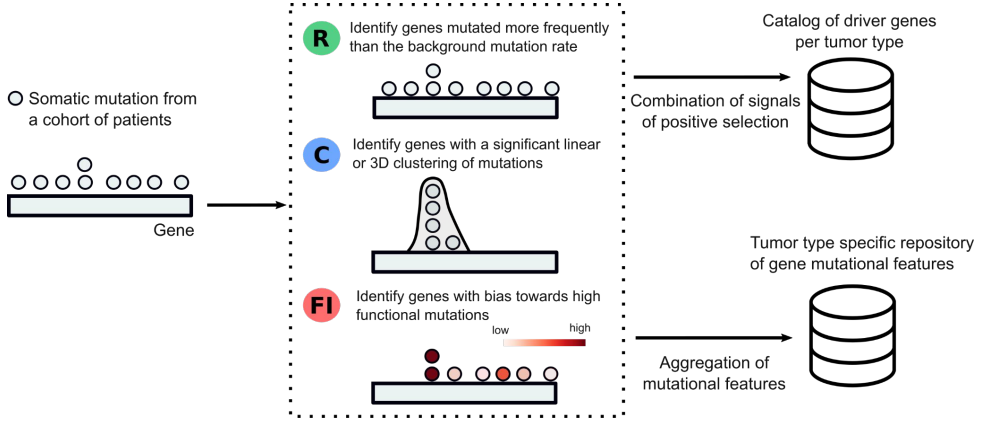
Figure from Petljak and Alexandrov, *Carcinogenesis*, 2016

- ***Different mutational processes generate unique combinations of mutation types, termed "Mutational Signatures"***
- Currently, four different variant classes are considered: **SBS** (Single Base Substitution), **DBS** (Double Base Substitution), **ID** (Small Insertion and Deletion), and **CN** (Copy Number Variation) Signatures
- COSMIC signatures: https://cancer.sanger.ac.uk/signatures/
- For additional databases for mutational signatures (see session 2) and details for mutational signature analyses (see session 5, January 18, 2023)

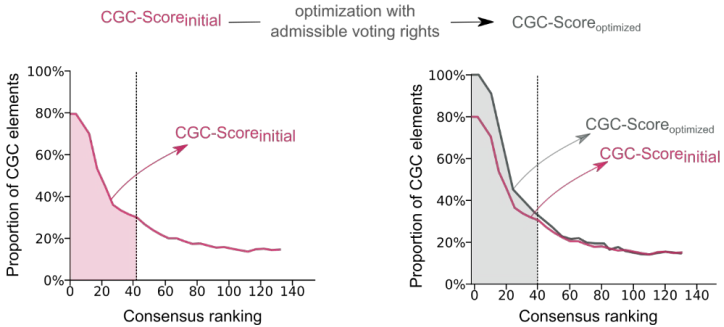
Cancer driver genes

IntOGen pipeline (Session 2) runs 7 different driver gene identification methods and combines their output.

More on this in session 8 (March 13, 2023)

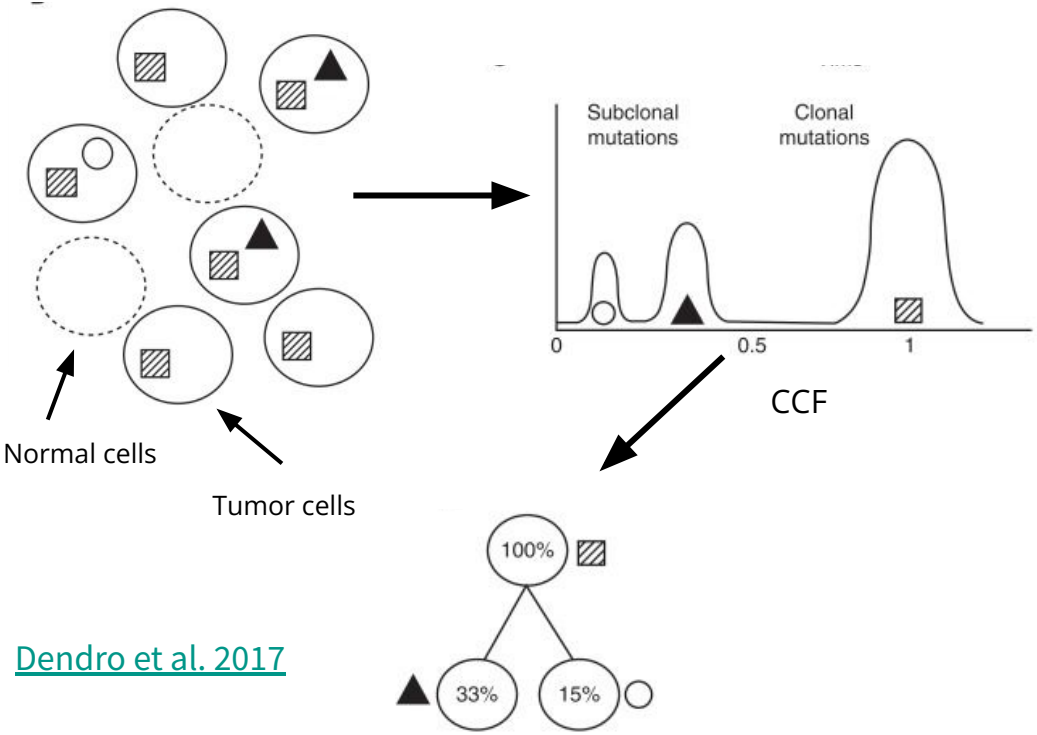


Method	Rationale
dNdScv	Negative binomial dN/dS estimation using regional covariates.
cBaSe	Bayesian inference of nonsynonymous counts given the observed synonymous counts
OncodriveCLUSTL	Positional recurrence in the DNA sequence.
HotMAPS3D	Positional recurrence in the 3D protein conformation
smregions	Enrichment of mutations in functional domains.
OncodriveFML	Functional impact score bias.
Mutpanning	Recurrence of nonsynonymous mutations combined with deviance from neutral characteristic mutational contexts.



Gene	Ranking	P-value	Q-value
PIK3CA	1st	1E-16	4E-15
TP53	2nd	1E-14	2E-13
PTEN	3rd	1E-12	1E-11
CDH1	4th	1E-8	6E-8
MAP2K4	5th	1E-4	7E-4
KMT2C	6th	1E-3	8E-3
TBX3	7th	5E-3	2E-2
DUSP16	8th	6E-1	8E-1

Mutational clonality analysis



Dendro et al. 2017

- Over time, clonal expansions can occur, resulting in the increase of subpopulations of cells harboring distinct sets of mutations.
- Using allele frequencies of mutations from sequencing data, accounting for copy number aberrations and tumor purity, an estimate of cancer cell fraction (CCF) for each mutation can be obtained.
- Clustering algorithms can be applied to mutations CCF to estimate number and CCF of each subclone.
- Relationship of each subclone can be then reconstructed and visualized.
- More information in session 9 (March 20, 2023)

Thank you for your attention!
Questions?

Next: Practical Session 4 (10:45am)