Somatic Mutation Calling

Human mutation

Mutations are:

- Changes in the DNA
- Caused by exogenous or endogenous processes
- Sometimes but not always heritable

Types of mutation:

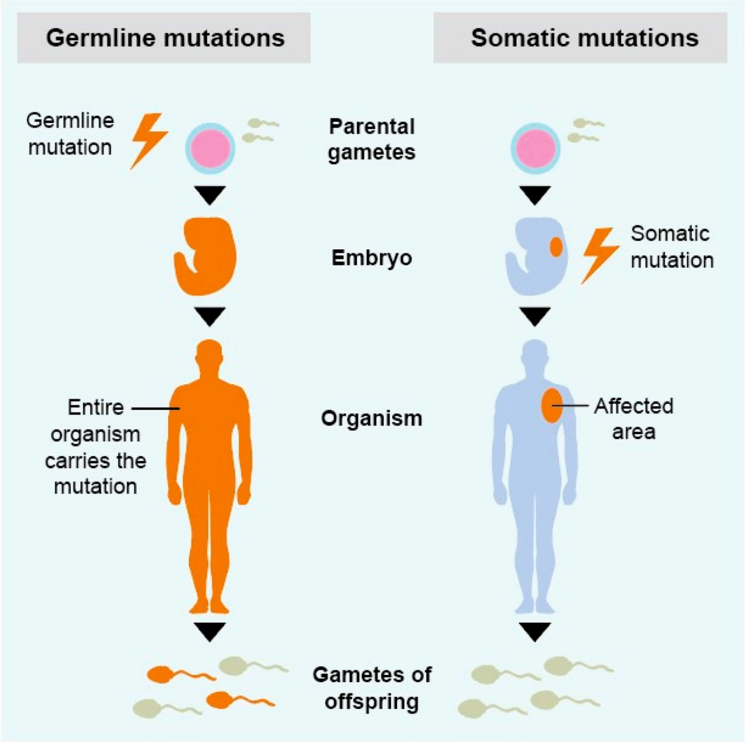
- Single nucleotide variants (SNVs)
  - Somatic simple variants (SSVs)
- Insertion-deletion variants (indels)
- Structural variants (SVs)

Somatic mutations

Somatic mutations occur in cells after conception which do not belong to the germline.

Somatic mutations may occur in almost any cell in the body.

Somatic mutations are never passed on to future generations.



https://www.mdpi.com/1422-0067/22/1/336/htm

Calling mutations in sequencing reads

Mutations are called relative to a reference genome

Every human differs every ~1000bp compared to the reference (“variation”)

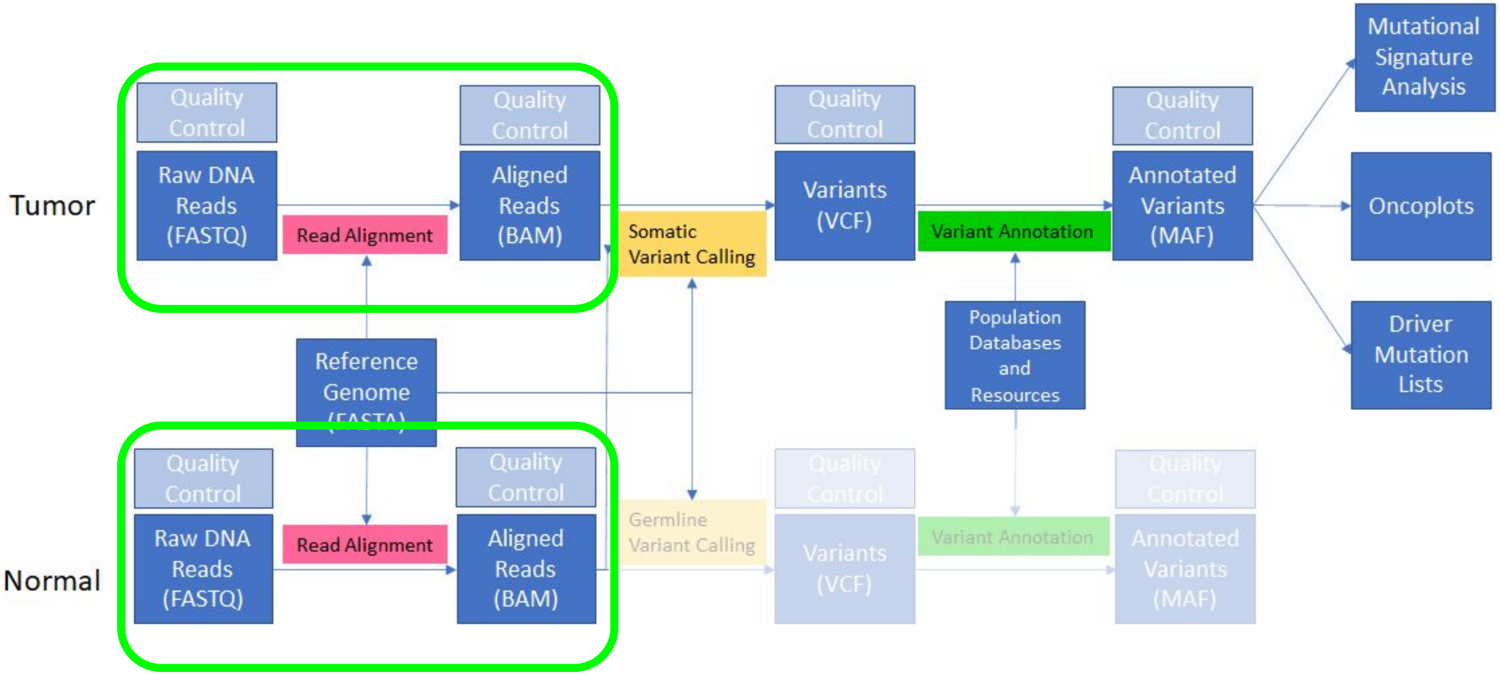
“Mutation” term is often reserved for variants with impact (e.g. driver mutations)

To call somatic mutations, a variant caller does the following:

1. Determine if a given site differs from the reference in the tumor and normal separately
2. Determine the allele and genotype of the site in the separate tumor / normal
3. Classify the site as “germline” or “somatic” by filtering any sites observed in both the tumor and normal
4. Sites that are observed in the tumor, but not the normal, are considered somatic variants.

Read alignment

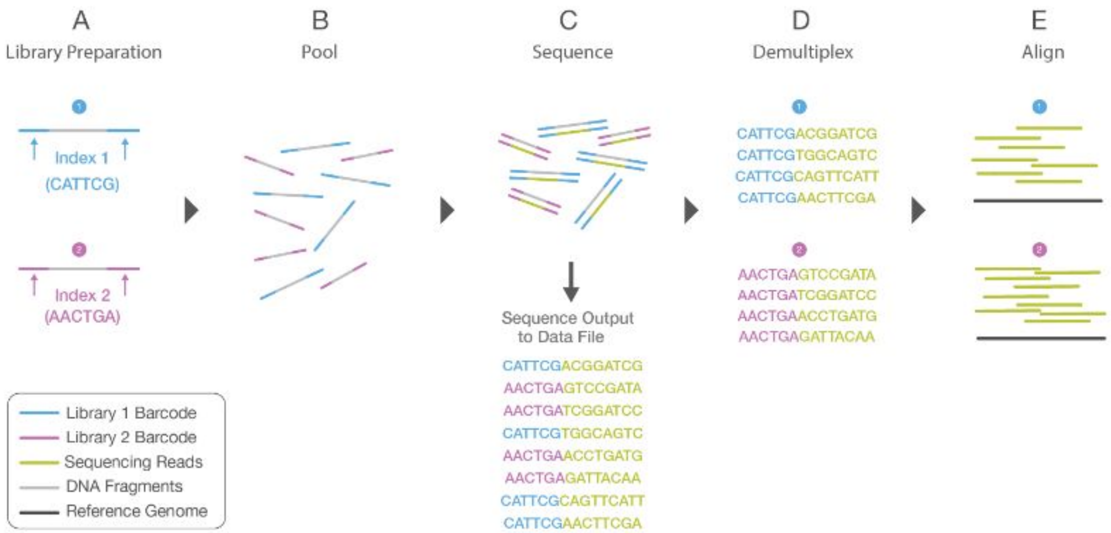
Review from yesterday:



Read alignment

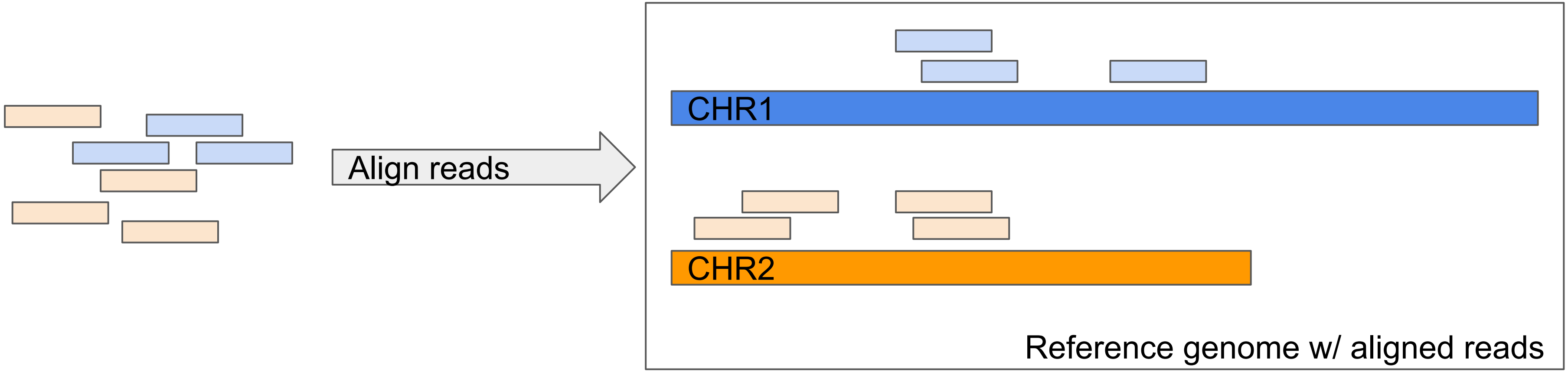
Remember - we have no idea where in the genome our FASTQ reads came from!

- We can only sequence small (150bp) portions of the genome
- We must “shatter” the DNA into sequenceable pieces (but these are random bits of DNA)



Read alignment

In *read alignment*, we probabilistically assign the random DNA fragments (“reads”) to a position within the reference genome.



For a full discussion of sequence alignment, see: https://github.com/WCSCourses/NGS_Bio_Africa/blob/main/Modules/Module_4_Alignment_to_Reference/read-alignment.lecture_slides.20210504.pdf

Read Alignment

We perform the following conceptual procedure to align a single read (using BWA mem):

1. Search the reference genome for the longest shared sequence(s) in both the read and the reference
   a. Record the positions for these maximal-exact matches (MEMs)
2. For each MEM above some length, perform Smith-Waterman alignment between the flanking sequences of the read / reference at the MEM
   a. Generates the *optimal alignment* between two sequences
   b. Provides us a score to judge the best alignment

Variant calling

Our aligned reads will have variation relative to the reference. When we call variants, we <u>probabilistically assign alleles and genotypes to these variations based on the evidence provided by our alignments</u>

```
               GAATTGGTCAAAAAT
                      CAAAAAT-CTTA
Aligned read(s):  TGGTCAAAAAT-C
Reference: ACTGGAATGGCCAAAAATGCTTAAGGCCTTATGGAAATGGAATCCACCA
                      ⇧       ⇧
```

Rules of thumb:

- humans have a germline mutation on average every 1000 bp (99.9% identical)
- SNPs are roughly 10x more frequent than indels (except in hypermutator genotypes like Microsatellite Instability)
- While each tumor has a unique number of mutations, within each cancer type the number of observed somatic mutations is often remarkably consistent (e.g. in thyroid cancer hundreds of mutations, in colon cancer 1000s, etc.)

What types of variants might we find?

Single-nucleotide variants

- Mutation substitutes a single basepair (e.g., C->T)

Multi-nucleotide variants

- Two or more adjacent nucleotides are substituted (e.g. CC->TT)

Insertion-deletion variants

- One or more nucleotides are inserted into or deleted from the reference (e.g. AATTGGCC -> AAT-GGCC and AATTGGCC -> AATT<u>TTT</u>GGCC)

Structural variants

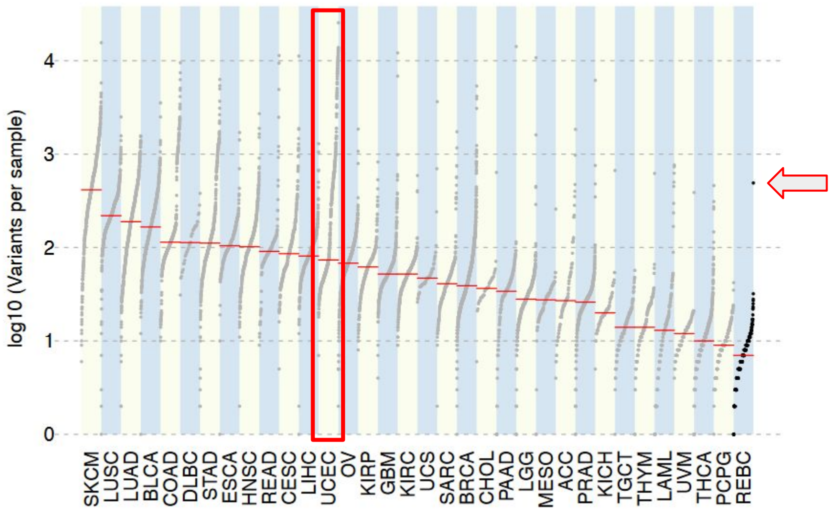
- Large (>50bp) changes to the genome that may alter copy number, orientation, chromosomal organization, etc.

Somatic SNVs and Indels

Somatic single-nucleotide variants (sSNVs) and indels are sometimes collectively referred to as SSVs (simple somatic variants)

Every tumor will have some, but total count ("mutation burden") can vary both within and across cancer types.

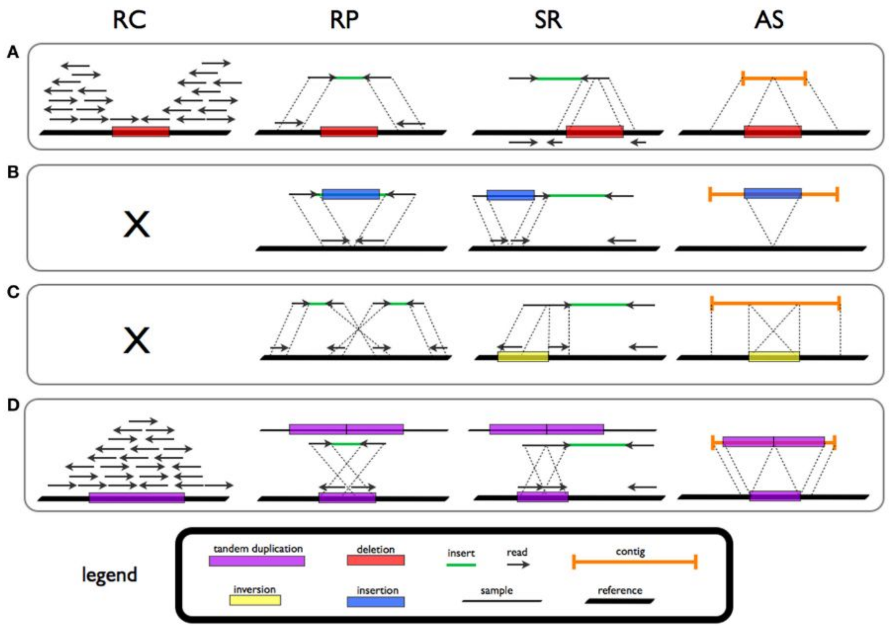
Generally, tens to tens-of-thousands of somatic mutations (vs. millions of germline variants)



Structural variants

Structural variants are large (>50bp) alterations within the genome

- Hard to detect in short reads
- Historically undersurveyed in short-read studies
- Occur more rarely than SSVs, but affect more basepairs of sequence
- Incredibly powerful impacts on genome
  - Gene Fusions
  - Promoter swapping
  - Chromothripsis, kataegis, etc.
- Short read callers: delly, lumpy, Manta, svaba
- Long read callers: SNIFFLES2, PBSV



Types of structural variants and various evidence types (from Tattini et al 2015.)

Variant calling software

Variant callers use various mathematical methods to generate variant calls, genotype calls, and quality metrics (e.g., Genotype Quality)

- Heuristic methods
- Bayesian probability
- Deep learning

There are probably hundreds of germline and somatic variant callers, but some of the most popular are:

Germline: GATK HaplotypeCaller, DeepVariant, bcftools, strelka2

Somatic: GATK mutect2, strelka2, MuSE, loFreq

Often, running multiple variant callers produces better results than a single caller (see the GDC pipeline for an overview of one such approach)

MuTect2

MuTect2 is a somatic variant caller that 1. Leverages the assembly approach of HaplotypeCaller 2. Applies a Bayesian classifier to detect low-allele fraction mutations 3. Uses a set of strict filters to increase specificity

Originally published in 2013

Pros:

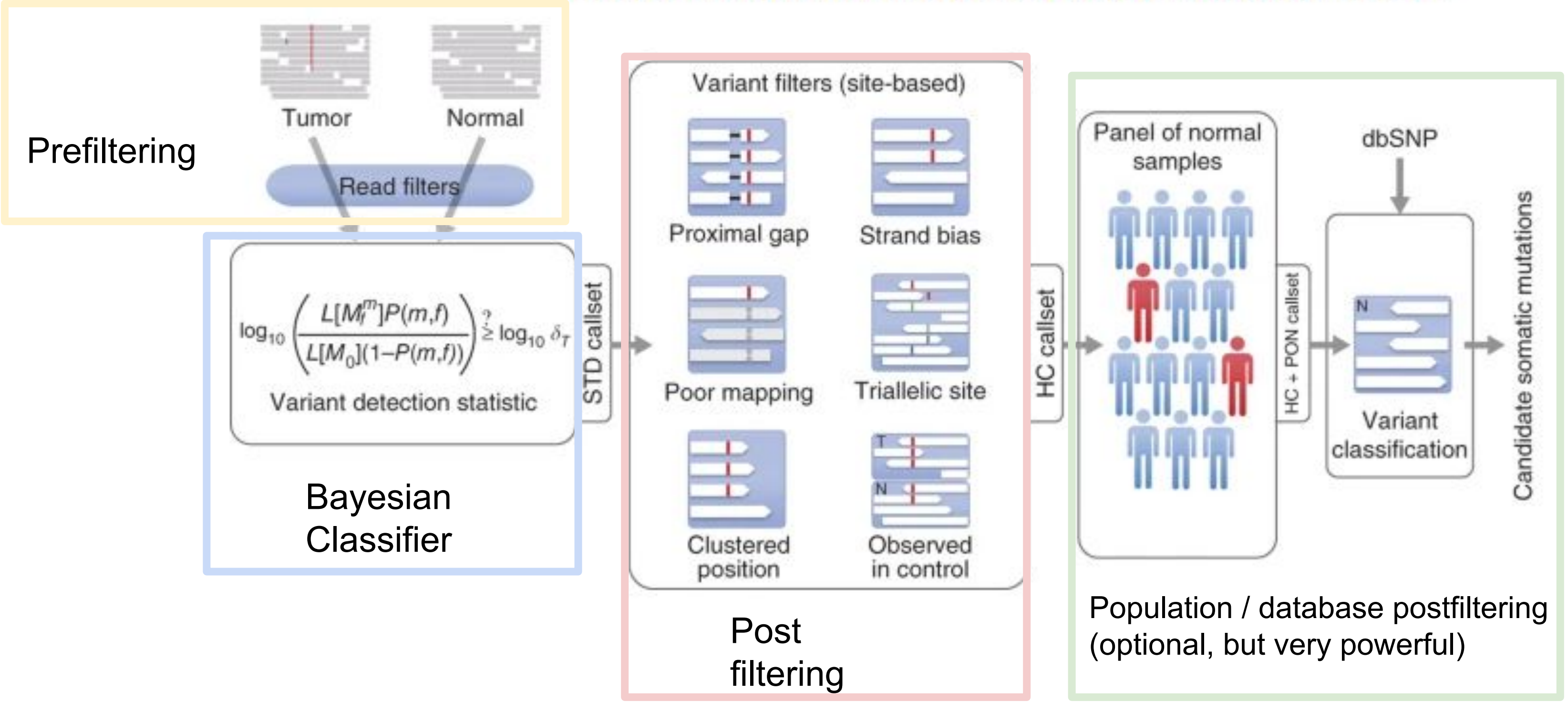
- Written in Java, so installation is easy (and it’s included as part of the GATK)
- Widely used
- Repeatedly validated in studies and competitions

Cons:

- Very slow (but accelerated alternatives available)
-

Mutect2

Figure 1: Overview of the detection of a somatic point mutation using MuTect.

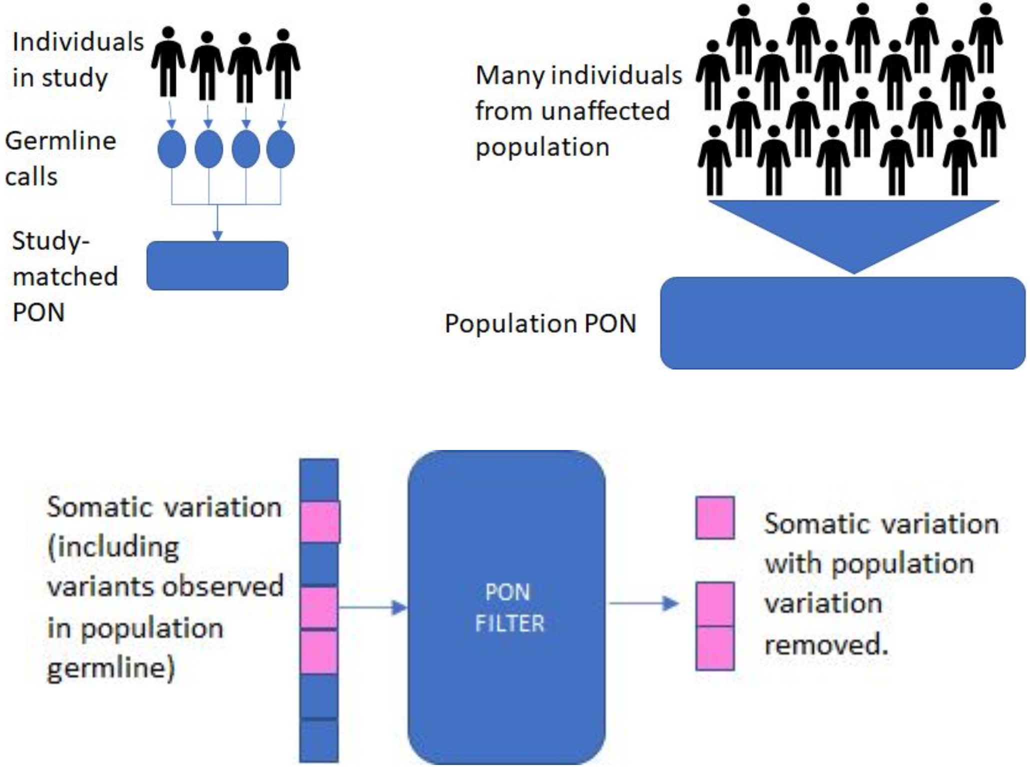


Mutect2

Basic usage:

```
gatk Mutect2 \
     -R reference.fa \
     -I tumor.bam \
     -I normal.bam \
     -normal normal_sample_name \
     -O somatic.vcf.gz
```

Mutect2 - Panel-of-Normals (PON)



Mutect2

Usage with germline resource + PON:

```
gatk Mutect2 \
     -R reference.fa \
     -I tumor.bam \
     -I normal.bam \
     -normal normal_sample_name \
     --germline-resource af-only-gnomad.vcf.gz \
     --panel-of-normals pon.vcf.gz \
     -O somatic.vcf.gz
```

Creating a PON:

1. Call normal samples in tumor-only mode
2. Create a PON using mutect

   or

Download a public PON

Mutect2

Tumor-only mode (used to create PON):

```
gatk Mutect2 \
  -R reference.fa \
  -I sample.bam \
  -O single_sample.vcf.gz
```

Tumor only mode with PON + germline resource (for when you don’t have a matched normal):

```
gatk Mutect2 \
  -R reference.fa \
  -I sample.bam \
  --germline-resource
af-only-gnomad.vcf.gz \
  --panel-of-normals pon.vcf.gz \
  -O single_sample.vcf.gz
```

Mutect2 outputs

A VCF file of somatic variant calls

Important fields: DP, GQ, GT, VAF

Useful options:

--annotations <annotation> : add fields like DP (depth) to output

--annotation-group <annotation group> : add groups of fields to output

--native-pair-hmm-threads : improve performance slightly

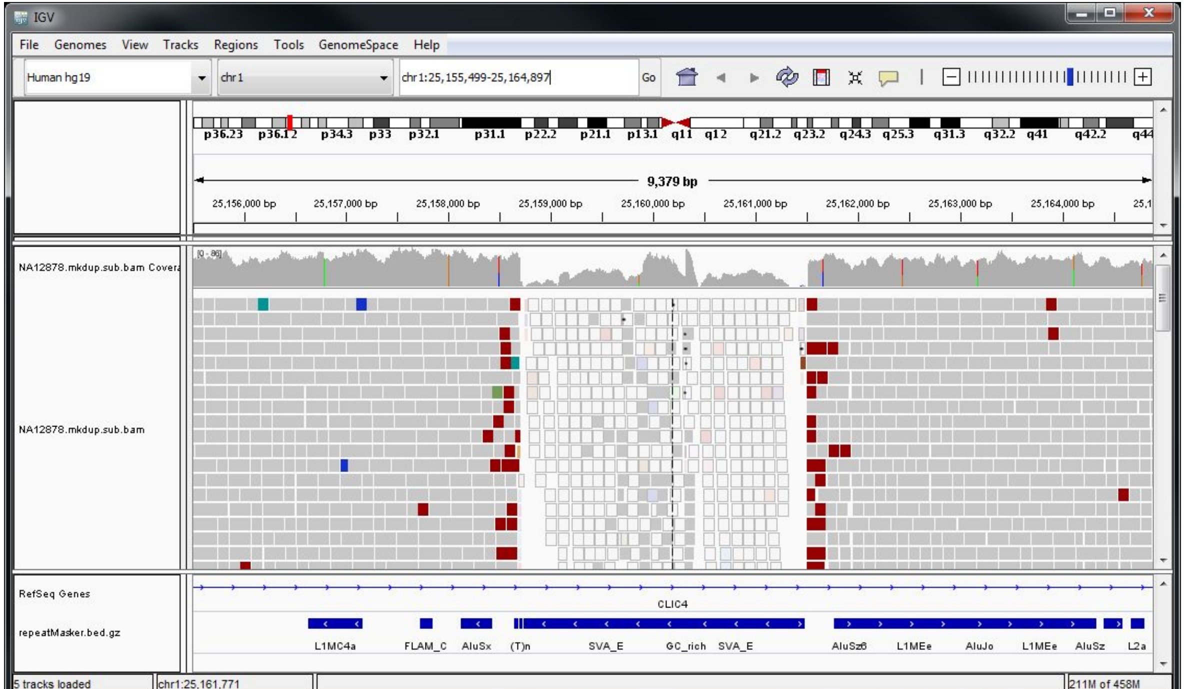
Mutect2

Mutect2 is very slow. Some ways to speed it up:

- Split analysis across genomic intervals (standard way)
- Add more threads (slight boost)
- Use an accelerated version (e.g., Nvidia Clara Parabricks, Sentieon, etc.)

Integrated Genomics Viewer (IGV)

IGV is a program for viewing alignments (https://software.broadinstitute.org/software/igv/)



IGV

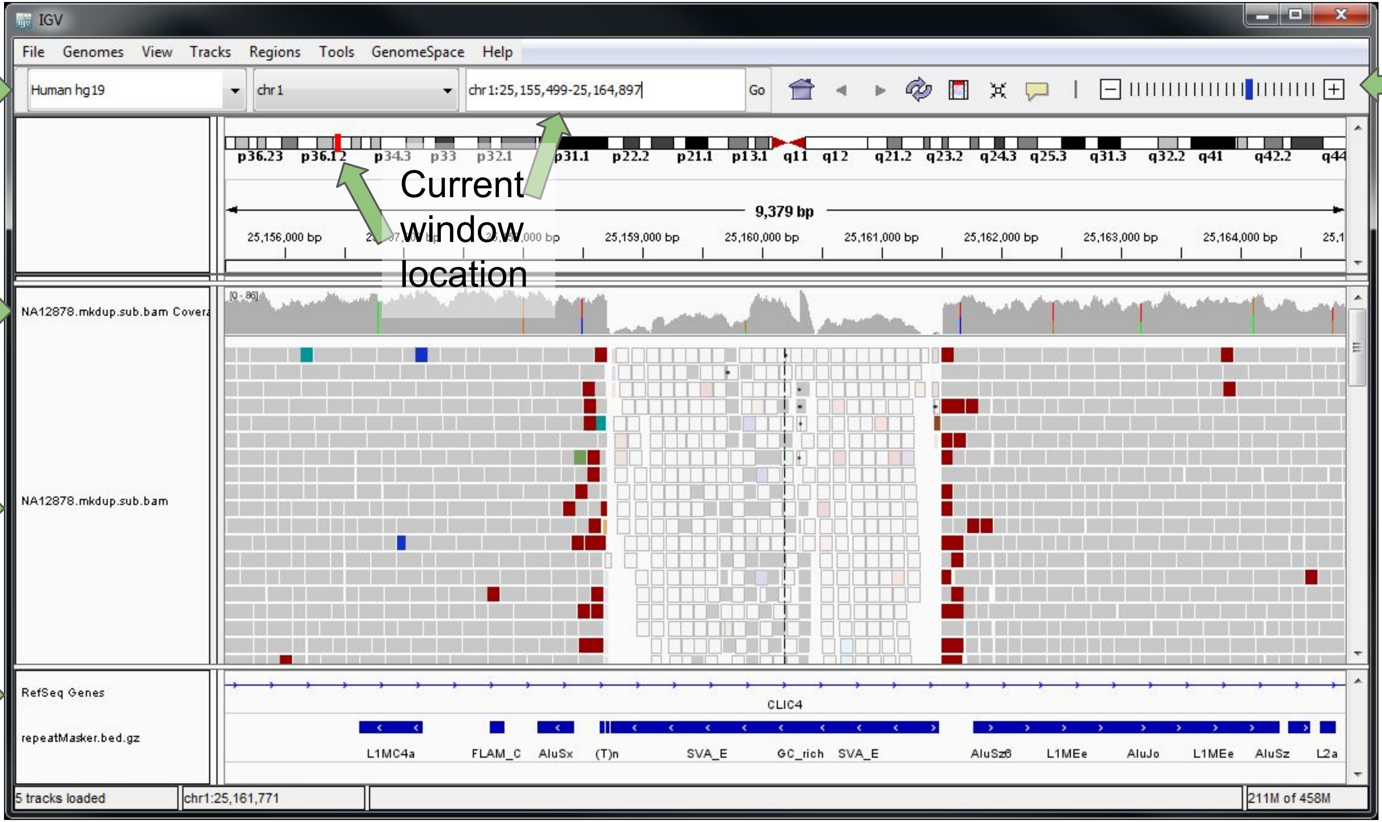
Select Genome

Coverage Track

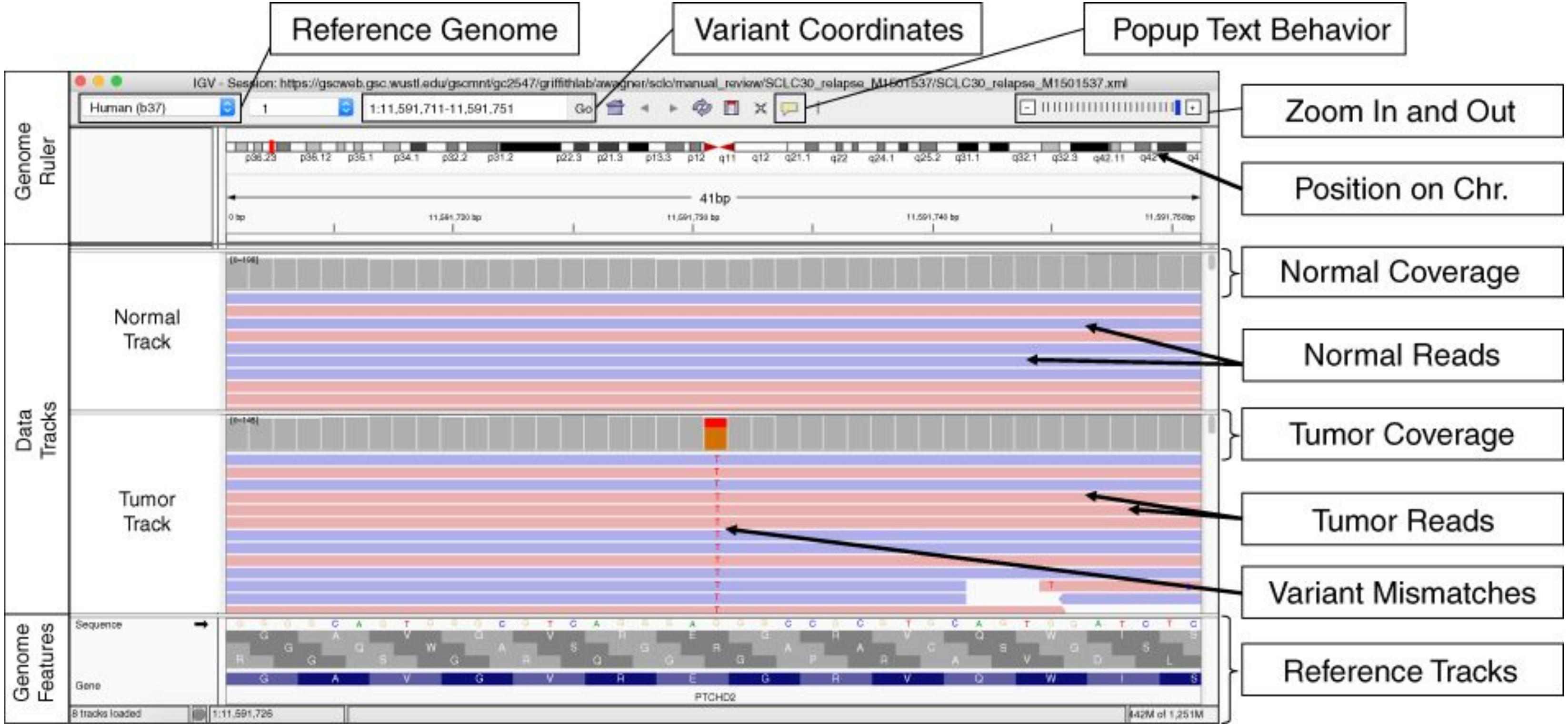
Read Track

Annotation Track

Zoom



IGV



Useful paper:https://www.nature.com/articles/s41436-018-0278-z

Variant annotation

Variant calls alone tell us only where a variant occurred in the genome and the allele / genotype.

To assess a variant in the context of cancer, we want to know:

- Its frequency in the unaffected population
- Its frequency in cancer cases (from tumors of same/different type)
- Its possible impact on gene expression
- Its possible impact when translated to protein
- Any association to disease

We annotate variants to link our calls with such information

Variant annotations

Where do annotations come from?

Gene and transcript databases (ENSEMBL, UCSC)

Population databases (1000Genomes, dbSNP, COSMIC, PCAWG)

Impact prediction software (CADD, REVEL, dbNSFP)

Manual annotation (ClinVar, BRCA Exchange)

How do we annotate variants?

We use annotation software, which adds fields to our data

Examples:

- Variant Effect Predictor (VEP)
- Funcotator
- VCF2MAF
- Vcfanno
- SNPSwift

Output: VCF or MAF (both tab-delimited files)

Annotation

#sequenced_samples: P-0000004-T01-IM3 P-0000015-T01-IM3 P-0000023-T01-IM3 P-0000024-T01-IM3 P-0000025-T02-IM5 P-0000025-T01-IM3 P-0000026-T01-IM3 P-0000027-T01-IM3 P-0000030-T01-IM3 P-0000034-T01-IM3 P-0000036-T01-IM3 P-000003

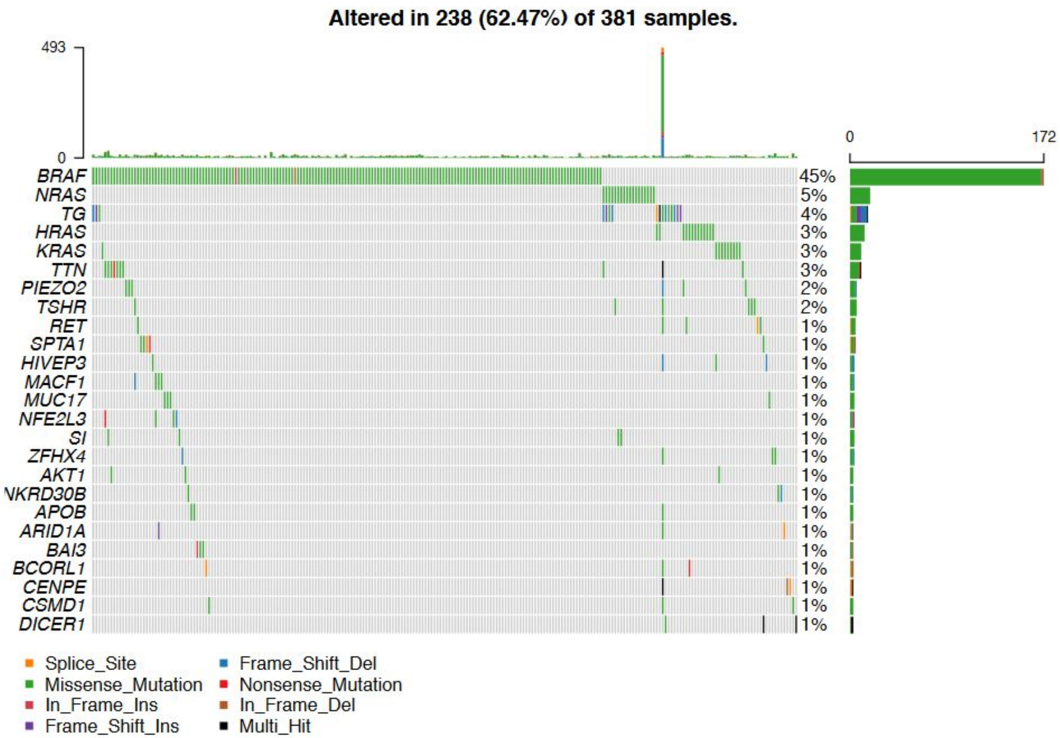
Hugo_Sym	Entrez_Ge	Center	NCBI_Buil	Chromoso	Start_Posi	End_Posit	Strand	Consequence	Variant_Classification	Variant_T	Reference	Tumor_Se	Tumor_Se	dbSNP_RS	dbSNP_Va	Tumor_Sa	Matched_	Match_Nc
SPEN			GRCh37	1	16265908	16265908	+	missense_variant	Missense_Mutation	SNP	A	A	T			P-0000004-T01-IM3		
ALK			GRCh37	2	29543736	29543736	+	missense_variant	Missense_Mutation	SNP	A	A	G			P-0000004-T01-IM3		
PDCD1			GRCh37	2	2.43E+08	2.43E+08	+	missense_variant	Missense_Mutation	SNP	G	G	A			P-0000004-T01-IM3		
MAP3K1			GRCh37	5	56177843	56177843	+	missense_variant	Missense_Mutation	SNP	C	C	G			P-0000004-T01-IM3		
FLT4			GRCh37	5	1.8E+08	1.8E+08	+	missense_variant	Missense_Mutation	SNP	C	C	A			P-0000004-T01-IM3		
FLT4			GRCh37	5	1.8E+08	1.8E+08	+	missense_variant	Missense_Mutation	SNP	T	T	C			P-0000004-T01-IM3		
NOTCH4			GRCh37	6	32178570	32178570	+	missense_variant	Missense_Mutation	SNP	C	C	T			P-0000004-T01-IM3		
NOTCH4			GRCh37	6	32188823	32188823	+	missense_variant	Missense_Mutation	SNP	G	G	A			P-0000004-T01-IM3		
MLL3			GRCh37	7	1.52E+08	1.52E+08	+	missense_variant	Missense_Mutation	SNP	C	C	T			P-0000004-T01-IM3		
MLL2			GRCh37	12	49433883	49433883	+	missense_variant	Missense_Mutation	SNP	G	G	A			P-0000004-T01-IM3		
TSHR			GRCh37	14	81422178	81422178	+	missense_variant	Missense_Mutation	SNP	C	C	A			P-0000004-T01-IM3		
AKT1			GRCh37	14	1.05E+08	1.05E+08	+	missense_variant,splice_region_variant	Missense_Mutation	SNP	C	C	T			P-0000004-T01-IM3		
TSC2			GRCh37	16	2110795	2110795	+	missense_variant	Missense_Mutation	SNP	G	G	A			P-0000004-T01-IM3		
RNF43			GRCh37	17	56440643	56440643	+	missense_variant	Missense_Mutation	SNP	G	G	A			P-0000004-T01-IM3		
NOTCH3			GRCh37	19	15303190	15303190	+	missense_variant,splice_region_variant	Missense_Mutation	SNP	C	C	T			P-0000004-T01-IM3		
TP53			GRCh37	17	7578503	7578518	+	frameshift_variant	Frame_Shift_Del	DEL	CAGGGCA	CAGGGCA	-			P-0000004-T01-IM3		
ALK			GRCh37	2	29450535	29450535	+	missense_variant	Missense_Mutation	SNP	C	C	T			P-0000015-T01-IM3		
PIK3CA			GRCh37	3	1.79E+08	1.79E+08	+	missense_variant	Missense_Mutation	SNP	G	G	A			P-0000015-T01-IM3		

Annotation

More useful MAF columns:

t_ref_count	t_alt_count	n_ref_count	n_alt_count	HGVSc	HGVSp	HGVSp_Short	Transcript	RefSeq	Protein_p	Codons	Hotspot	cDNA_change
400	73			ENST00000375759	p.Ile3661Phe	p.I3661F	ENST0000	NM_01500	3661	Att/Ttt	0	c.10981A>T
180	13			ENST00000389048	p.Val476Ala	p.V476A	ENST0000	NM_00430	476	gTg/gCg	0	c.1427T>C
225	15			ENST00000334409	p.Ala215Val	p.A215V	ENST0000	NM_00501	215	gCc/gTc	0	c.644C>T
370	12			ENST00000399503	p.Ser939Cys	p.S939C	ENST0000	NM_00592	939	tCt/tGt	0	c.2816C>G
360	25			ENST00000261937	p.Arg1324Leu	p.R1324L	ENST0000	NM_18292	1324	cGg/cTg	0	c.3971G>T
273	22			ENST00000261937	p.Thr494Ala	p.T494A	ENST0000	NM_18292	494	Acg/Gcg	0	c.1480A>G
279	17			ENST00000375023	p.Gly942Arg	p.G942R	ENST0000	NM_00455	942	Ggg/Agg	0	c.2824G>A
207	11			ENST00000375023	p.Ser244Leu	p.S244L	ENST0000	NM_00455	244	tCg/tTg	0	c.731C>T
84	11			ENST00000262189	p.Met812Ile	p.M812I	ENST0000	NM_17060	812	atG/atA	0	c.2436G>A
247	16			ENST00000301067	p.Pro2557Leu	p.P2557L	ENST0000	NM_00348	2557	cCg/cTg	0	c.7670C>T
195	13			ENST00000298171	p.Pro52Thr	p.P52T	ENST0000	NM_00036	52	Ccc/Acc	0	c.154C>A

Annotation -> oncoplots



From [Dawson 2020]

Annotation -> significantly mutated genes

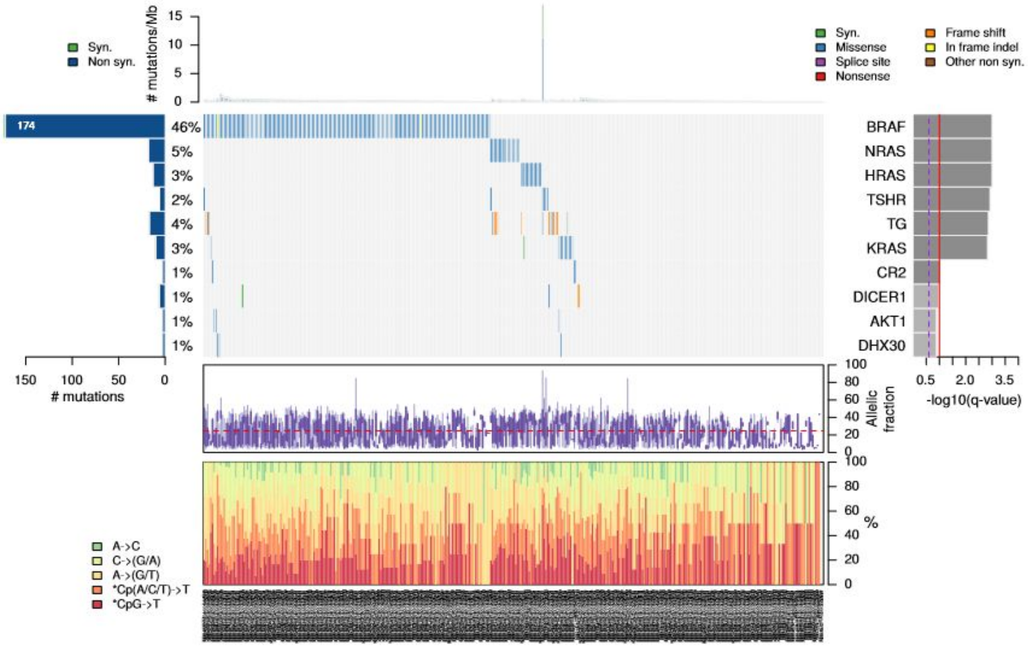
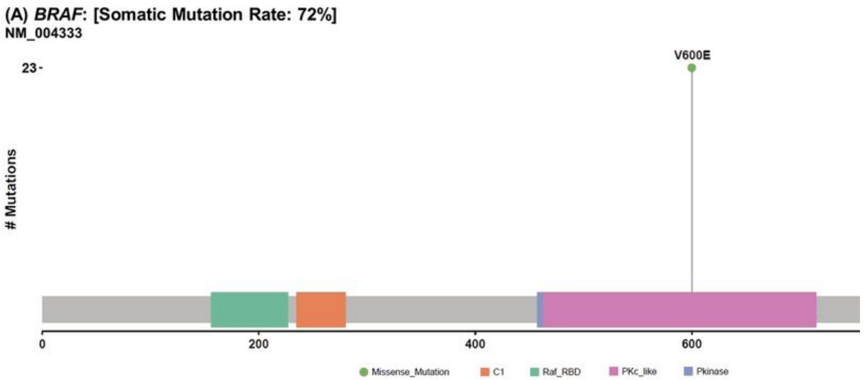


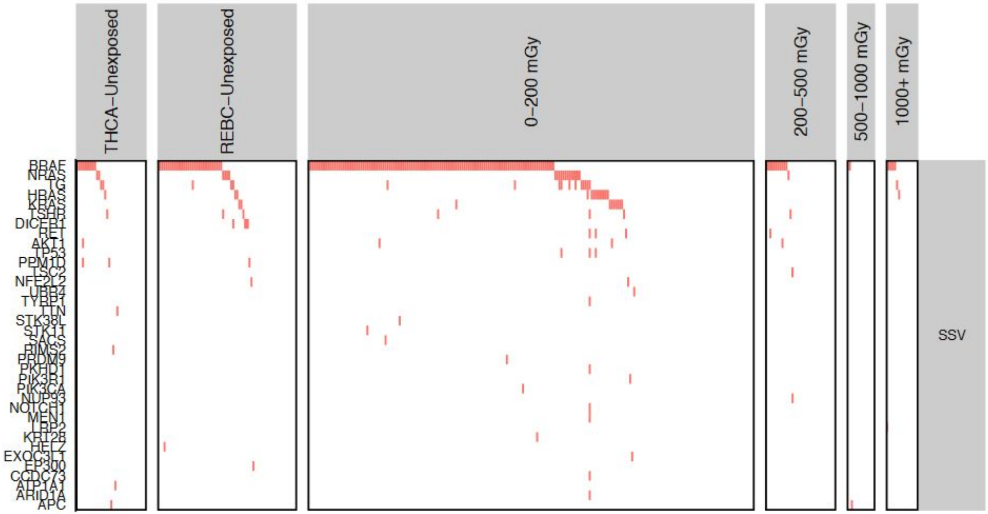
Fig. 2.10 Significantly mutated genes from MutSig2CV.

From Dawson 2020

Annotations -> driver mutations



Modified from: https://link.springer.com/article/10.1007/s12020-019-01842-y

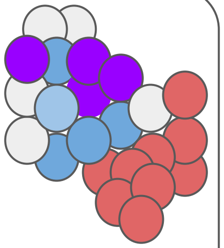


From [Dawson 2020]

VAF, CCF, Tumor Purity, Copy Number, and Contamination

Tumors are complex, bulk mixtures of cells

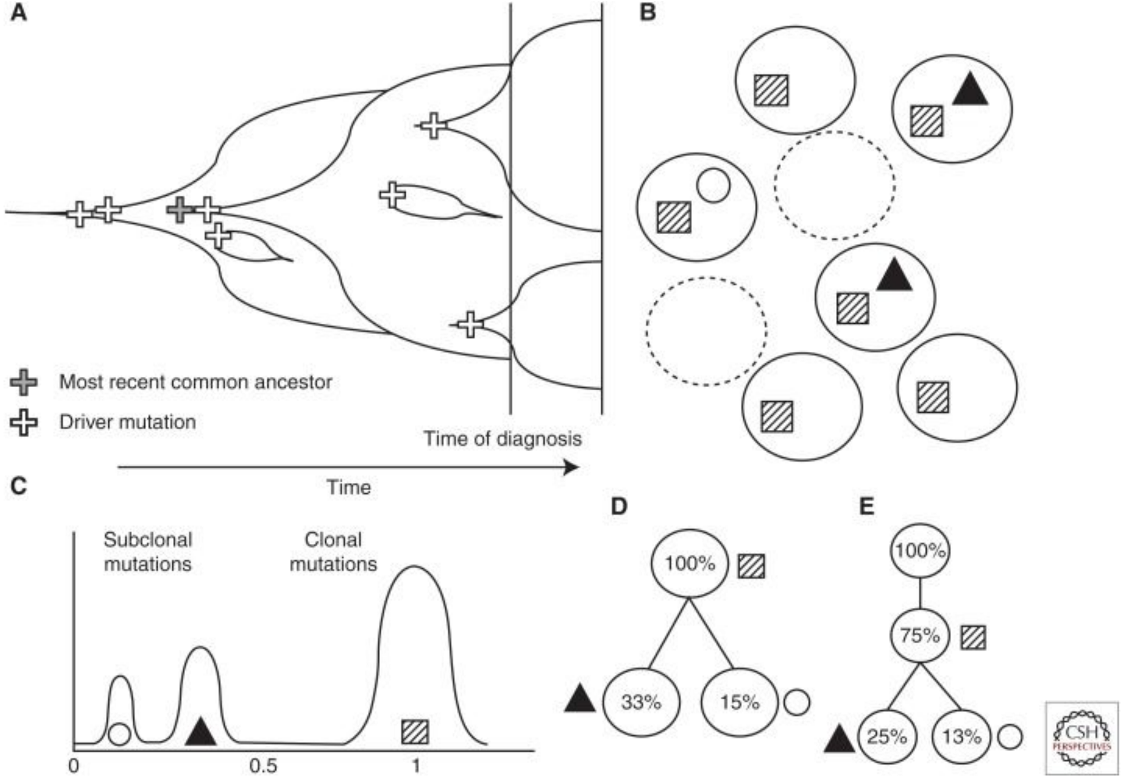
- Tumors continue to mutate while under constant evolutionary pressure
  - Significant benefits for developing faster growth, immune evasion, angiogenesis and other hallmarks of cancer
  - May be composed of multiple subclones with complex architecture



Resected tumor

A tumor is a cellular mixture of (sub)clones with various genetic mutations over a common background; normal cells; immune cells; etc.

Tumor evolution



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5538405/

VAF: Variant Allele Frequency

The proportion of reads supporting a given allele at a site:

$$\text{VAF} = \frac{\text{\# Reads supporting alt}}{(\text{\# reads supporting alt}) + (\text{\# reads supporting reference})}$$

Simple! Except:

- Not all tumor cells will contain a given mutation
- Tumors are contaminated with normal cells
- Copy number changes can change the ratio of alleles

Therefore, we can’t rely on VAF as an accurate descriptor of a tumor’s mutations (as we can for germline)

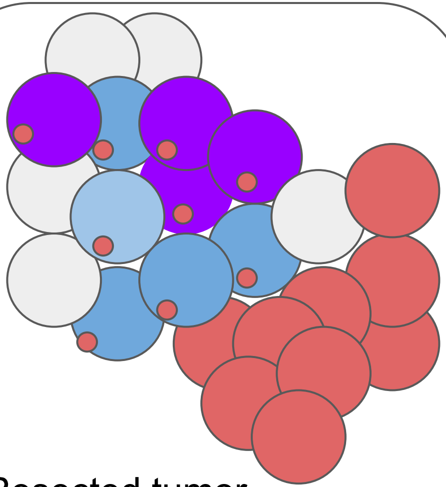
Tumor purity

Tumor purity: the percentage of cells in the “tumor” sample that are actually from the tumor.

- We can’t extract tumors cell-by-cell - there is **always** some level of normal cell infiltration

Tumor purity: 18 / 23 ~78% (pretty good!)

Typical purity range: 20%-80%

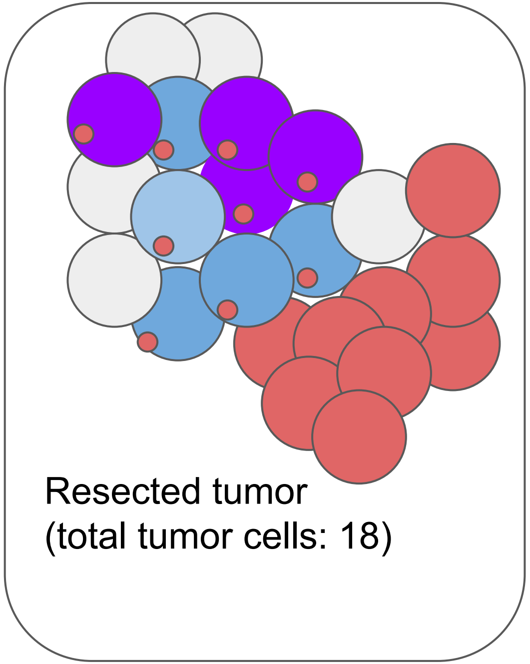


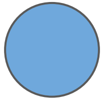
Resected tumor
(Total cells: 23
total tumor cells: 18)

Cancer cell fraction (CCF)

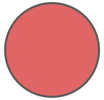
CCF: the proportion of cells within a tumor containing a given variant

- Think of this as an allele fraction weighted to cells within the tumor
- CCF = 1.0: all cells in the tumor have this mutation ("clonal")
- CCF = < 1.0: some proportion of cells in the tumor have this mutation ("subclonal")
  - Common subclonal VAFs: 0.5, 0.33 (think of dividing the tumor)

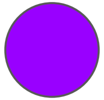




CCF = 5 / 18 =
~0.33 (subclonal)

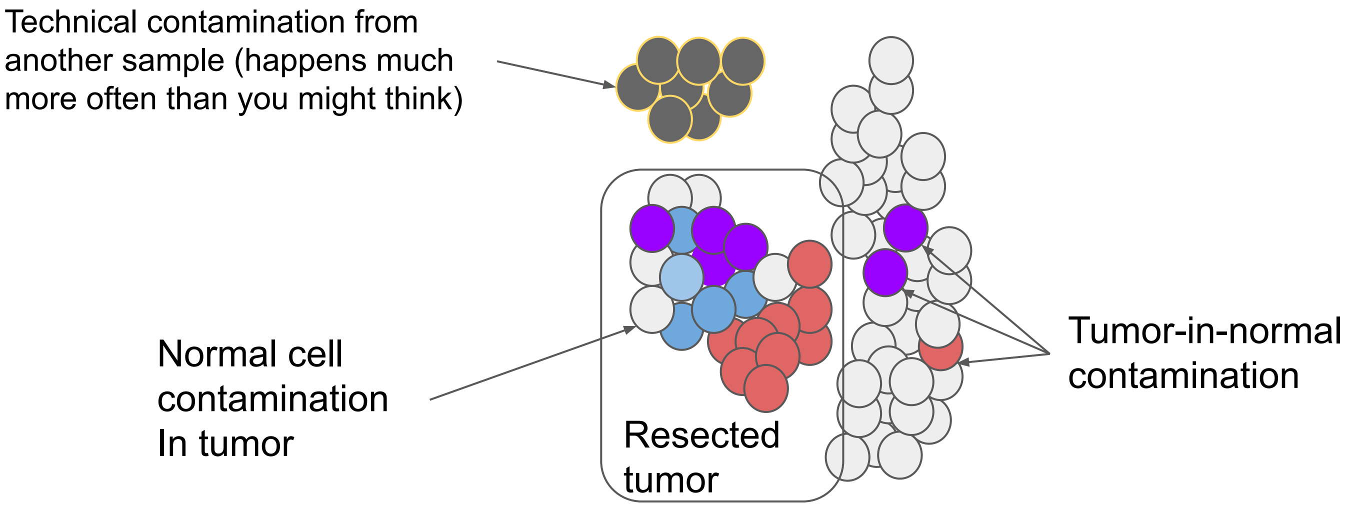


CCF = 18 / 18 = 1
(possible driver)



CCF = ?? = ??
(clonal or subclonal?)

Contamination



Copy number variation

Copy number = 1
(Ploidy = 2)

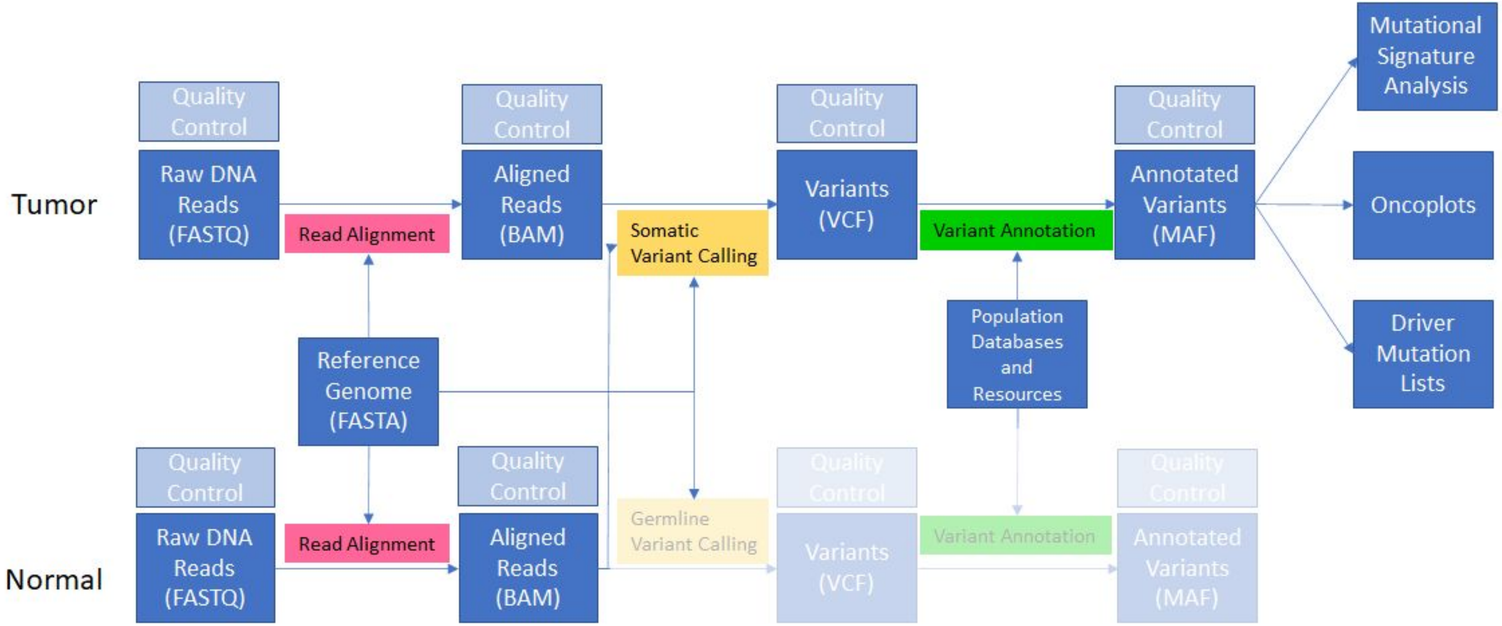
Copy number = 2
(Ploidy = 4)

Distorted VAF / CCF

- Original CCF: 0.5
- Original VAF: 0.5
- One haplotype duplicated -> new VAF ~0.66

If you observe CCFs not divisible by 2, and your tumor is relatively pure, consider looking at copy number (e.g. with ASCAT)

Conclusions



Conclusions

- If you just want to get SSV calls: BWA mem + MuTect2
- Variant calls are almost useless without annotation
- Filter your calls extensively

- Tumors are complex, and somatic calling is fraught with error
- Be careful, think statistically, and consider common sources of error
- Expectations from one study (e.g., SSV burden) may or may not apply in others