Memorial Sloan Kettering Cancer Center

Cancer Genomics Analysis

Jian Carrot-Zhang, Ph.D.

Assistant Attending,
Computational Oncology

GSK Core Course Spring 2025

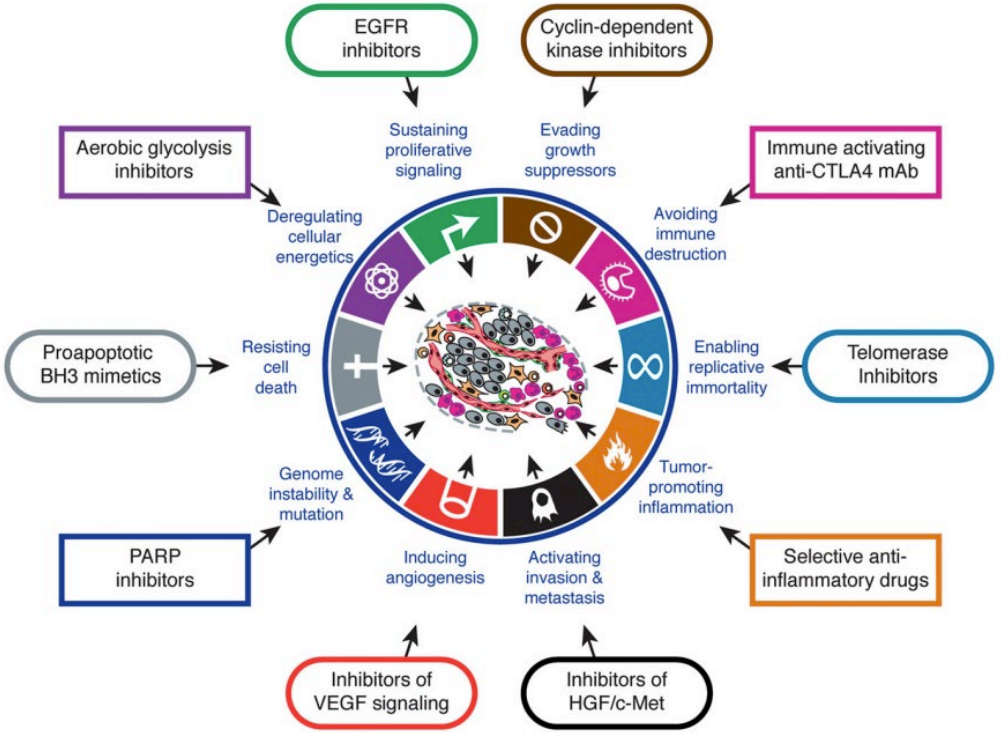
Outlines

- Overview of Cancer Genome Alterations

- Overview of Genomic Analysis
  - Somatic alteration detection
  - Cancer gene discovery
  - Clinical genomic sequencing

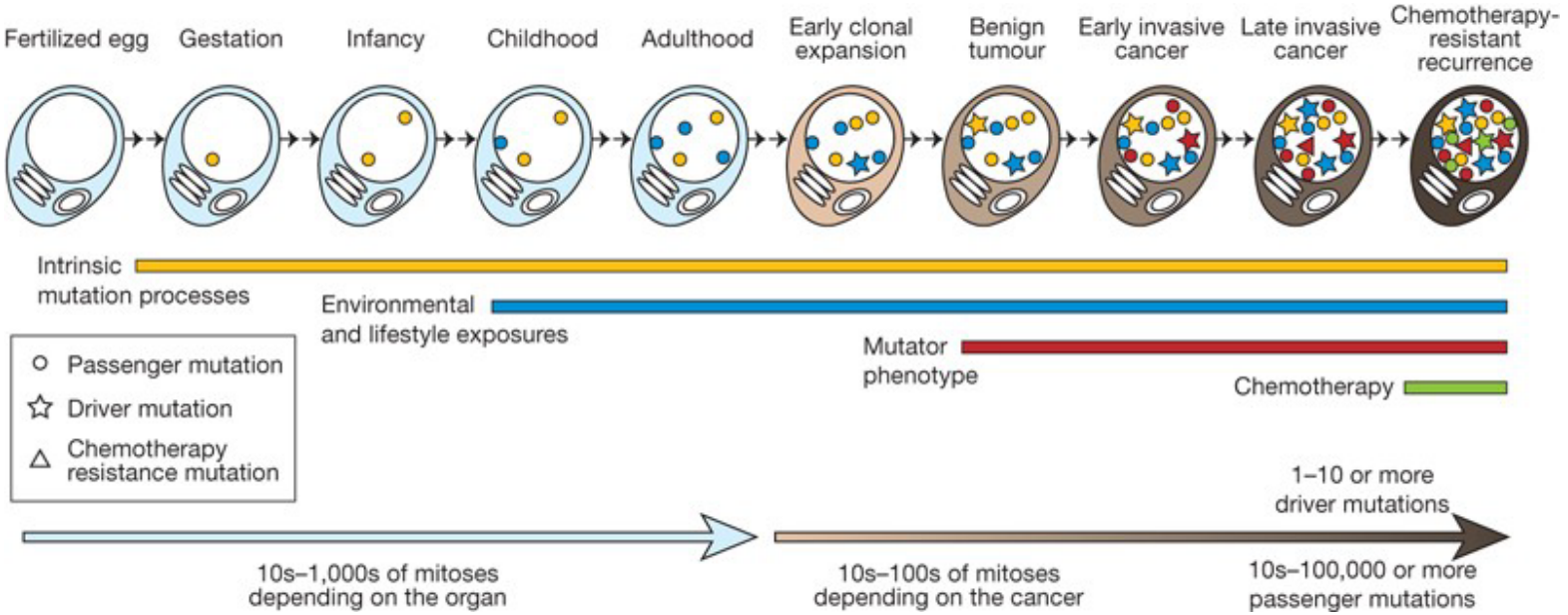
The hallmarks of cancer

- All cancers are driven by somatic mutations that lead to tumor growth and hallmarks of cancer



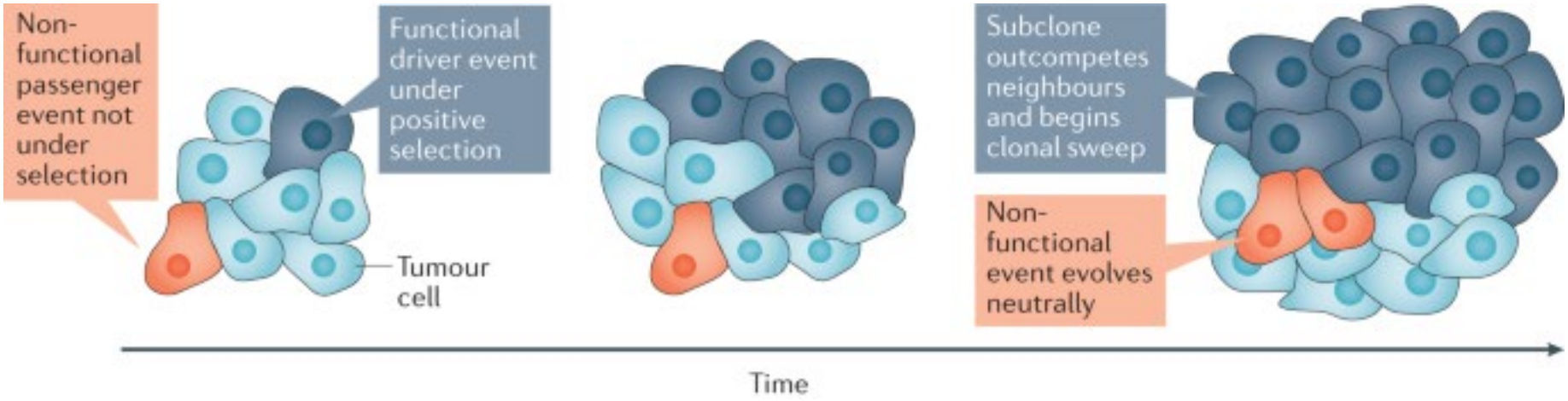
Hannahan & Weinberg *Cell* (2011)

Mutational processes



Stratton, Campbell & Futreal *Nature* (2009)

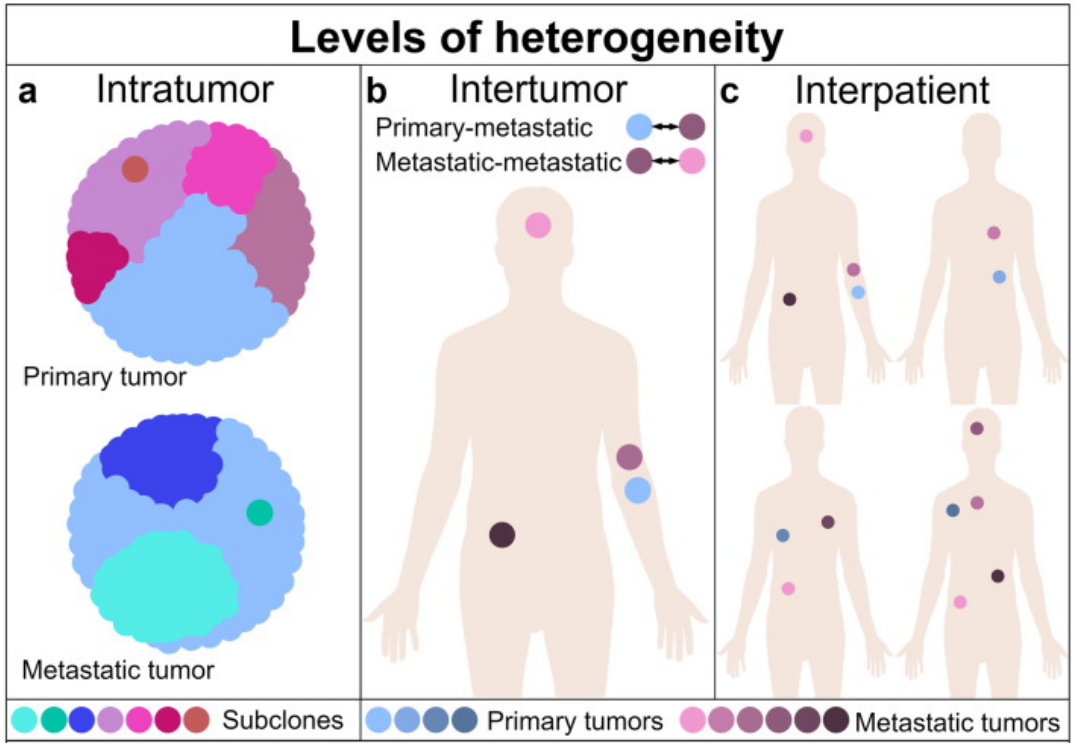
Driver vs passenger



Black & McGranahan *Nat. Reviews Cancer* (2021)

Across patient tumor heterogeneity

- Somatic differences reflect tissue-of-origin
- Somatic alterations define molecular subtypes
- Somatic differences from different stages
- Gender and genetic ancestry

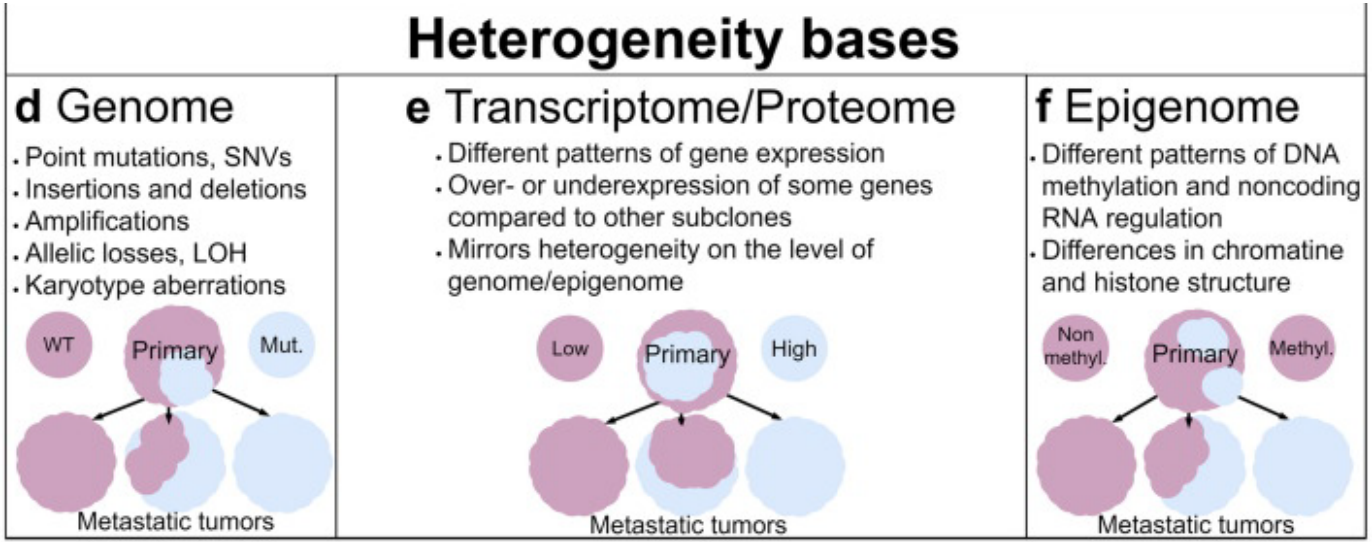


Grzywa et al. Transl Oncol. (2017)

Memorial Sloan Kettering Cancer Center

Within patient tumor heterogeneity

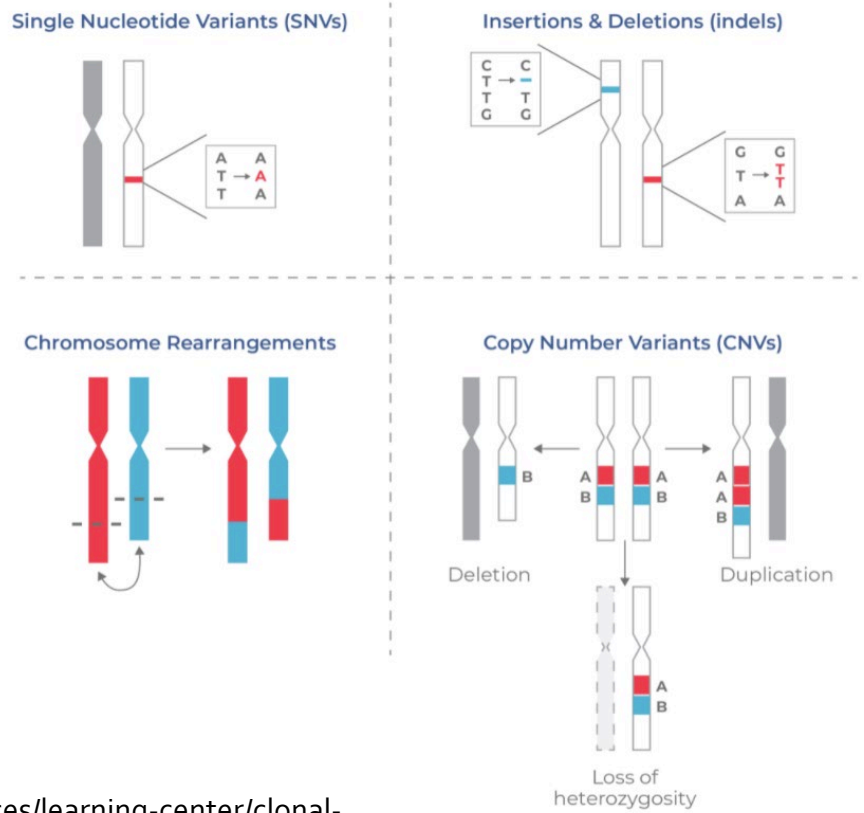
- Inter-tumor: between tumors within a patient
- Intra-tumor heterogeneity: between cells within a tumor lesion (e.g. tumor clones)



Grzywa et al. Transl Oncol. (2017)

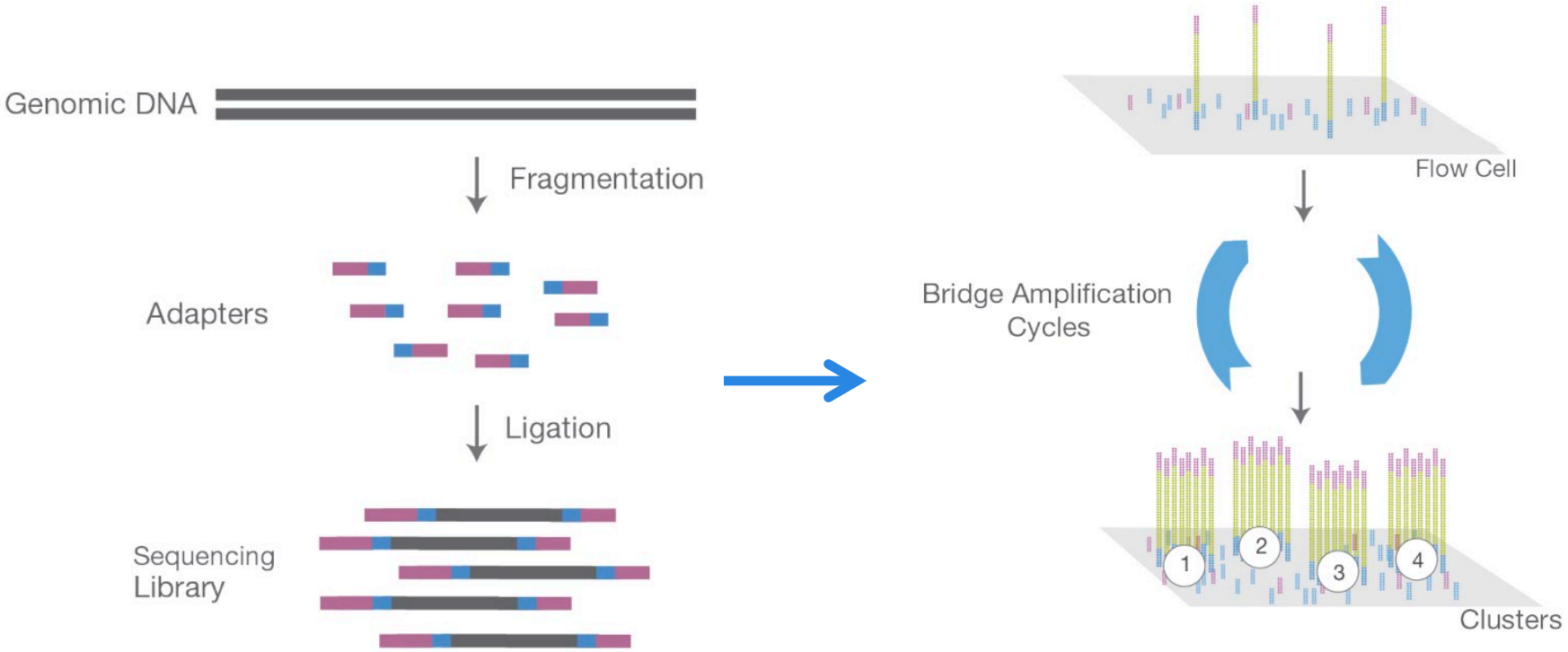
Types of somatic alterations

- Single nucleotide variants
- Insertions and deletions (indel)
- Copy number alterations
- Chromosomal rearrangements



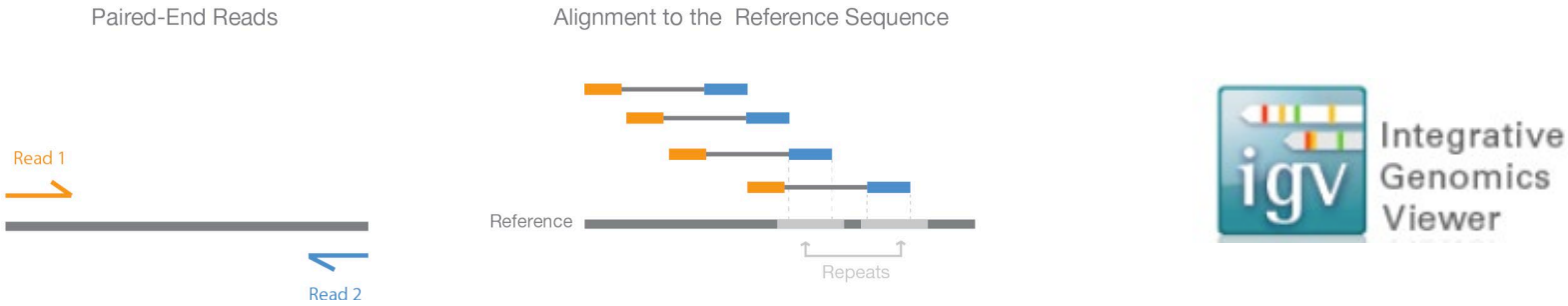
https://missionbio.com/resources/learning-center/clonal-evolution-in-cancer/

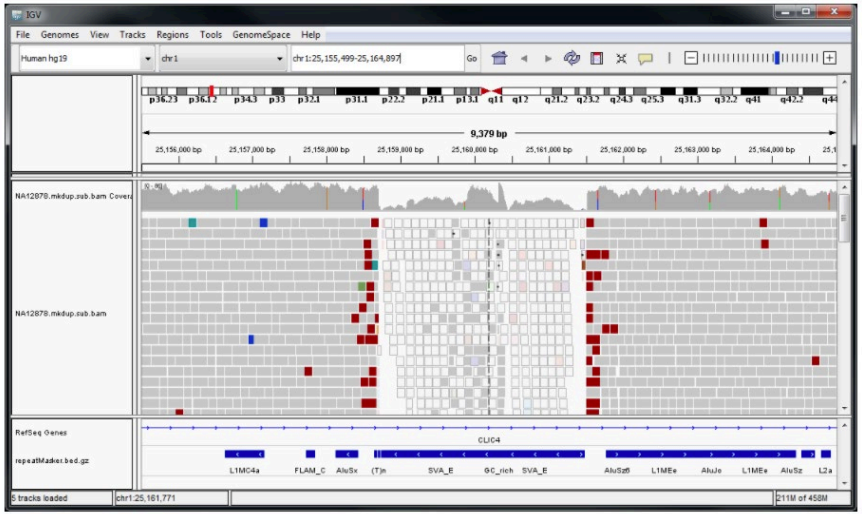
Next-generation sequencing



https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

Next-generation sequencing - pair-end reads alignment





Memorial Sloan Kettering Cancer Center

BWA and Samtools for read alignment and BAM file manipulation

BAM file or CRAM file

```
## @HD  VN:1.0  SO:coordinate
## @SQ  SN:chrX LN:156040895
## @PG  ID:hisat2   PN:hisat2   VN:2.2.0    CL:"/Users/dtang/github/rnaseq/hisat2/../src/hisat2
## @PG  ID:samtools PN:samtools PP:hisat2   VN:1.16 CL:samtools view -b eg/ERR188273_chrX.bam
## @PG  ID:samtools.1   PN:samtools PP:samtools VN:1.16 CL:samtools fillmd -e - genome/chrX.fa
## ERR188273.4711308    73  chrX    21649   0   5S70M   =   21649   0   CGGGT==================
## ERR188273.4711308    133 chrX    21649   0   *   =   21649   0   CTACAGGTGCCCGCCACCATGCCCAG
## ERR188273.4711308    329 chrX    233717  0   5S70M   =   233717  0   CGGGT==================
## ERR188273.14904746   99  chrX    251271  60  75M =   251317  121 ==========================
## ERR188273.14904746   147 chrX    251317  60  75M =   251271  -121    ======C===============
```

Mapping quality

Base quality

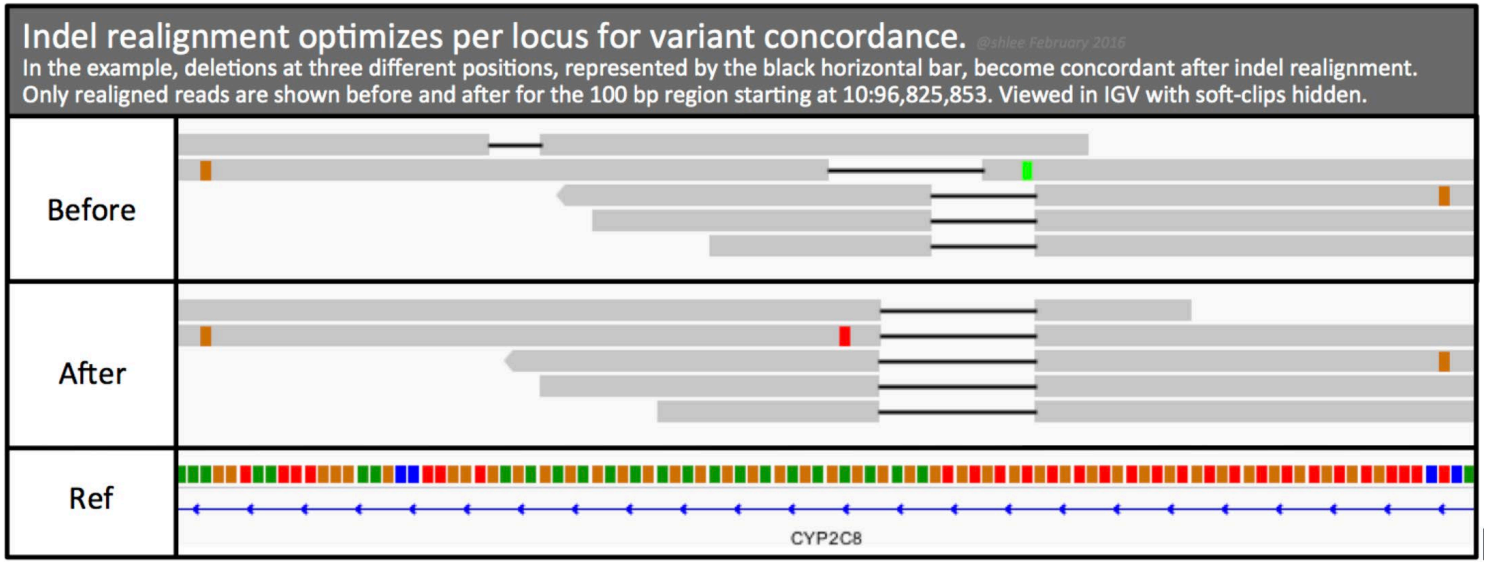
```
============ @@@F=DDFFHGHBHIFFHIGGIFGEGHFHIGIGIFIIIGIGIGGDHIIGIIC@>DGHCHHHGHHFFFFFDEACC@ AS:i:-5
TGTTGGCC CB@FDFFFHHGFHIJJJJIIIIIIIGGGIJGIIJJJJJJFFHIIIIGECHEHHGGHHFF?AACCDDDDDDDDBCD YT:Z:UP
============ @@@F=DDFFHGHBHIFFHIGGIFGEGHFHIGIGIFIIIGIGIGGDHIIGIIC@>DGHCHHHGHHFFFFFDEACC@ AS:i:-5
======== @@<DDDDDFB>HHEGIIGAGIIIBGIIG@FECH<F@GIIFAE=?BCBBCBBB5@<?CBBCCCCAACDCCCCCCCC AS:i:0  XN:
============ #######B?DAHC@EGIIGGEHHGC@GFBFCEGFCIGG@EG@@H<JIEHEF@IGEHGHIIHFGHDDFDDDDD?<B AS:i:-2
```

Re-alignment for improved indel calling

- Genome aligners can only consider each read independently
- Favor alignments with mismatches or soft-clips instead of opening a gap
- May use arbitrary tie-breaking, leading to different, non-parsimonious representations of the event in different reads.

Re-alignment for improved indel calling

- Local realignment considers all reads spanning a given position. This makes it possible to achieve a high-scoring consensus that supports the presence of an indel event.



Overview of somatic alteration detection methods

- SNV and indel detection
- Copy number
- Structural variation

Main problem:

Predict somatic alterations from germline variations accounting for biological and technical uncertainties

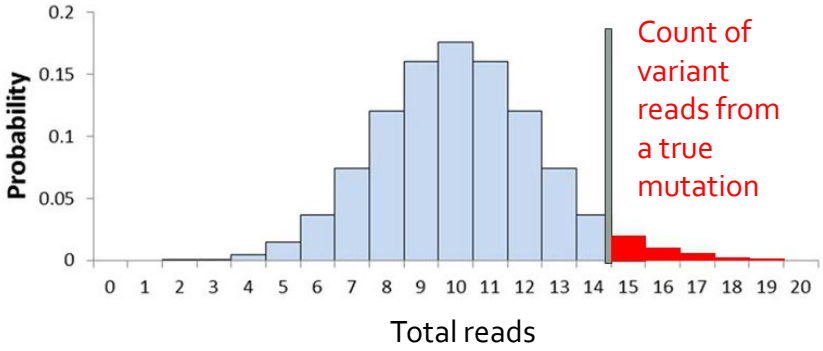
Germline genotyping from NGS

At a given site

Genotypes:	AA,	AB,	BB
	Hom Ref	Het SNP	Hom SNP
Expected allelic fractions:	~0	~0.5	~1

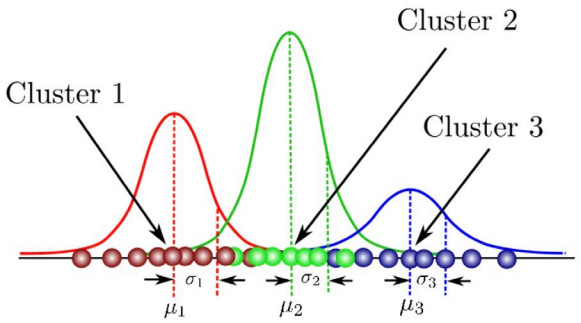
Data from sequencing:

- Depth (total reads)
- Count of reference reads
- Count of variant reads



More advanced algorithms

- Model the prior genotype probability (mutation rate etc)
- Estimate the genotype of a cohort jointly (success probability depends on the genotype and sequencing error rate)



https://towardsdatascience.com/gaussian-mixture-models-explained-6986aaf5a95

Probability of A being true, given B is true

The probability of B being true, given A is true

Likelihood

Prior

The probability A being true

Posterior

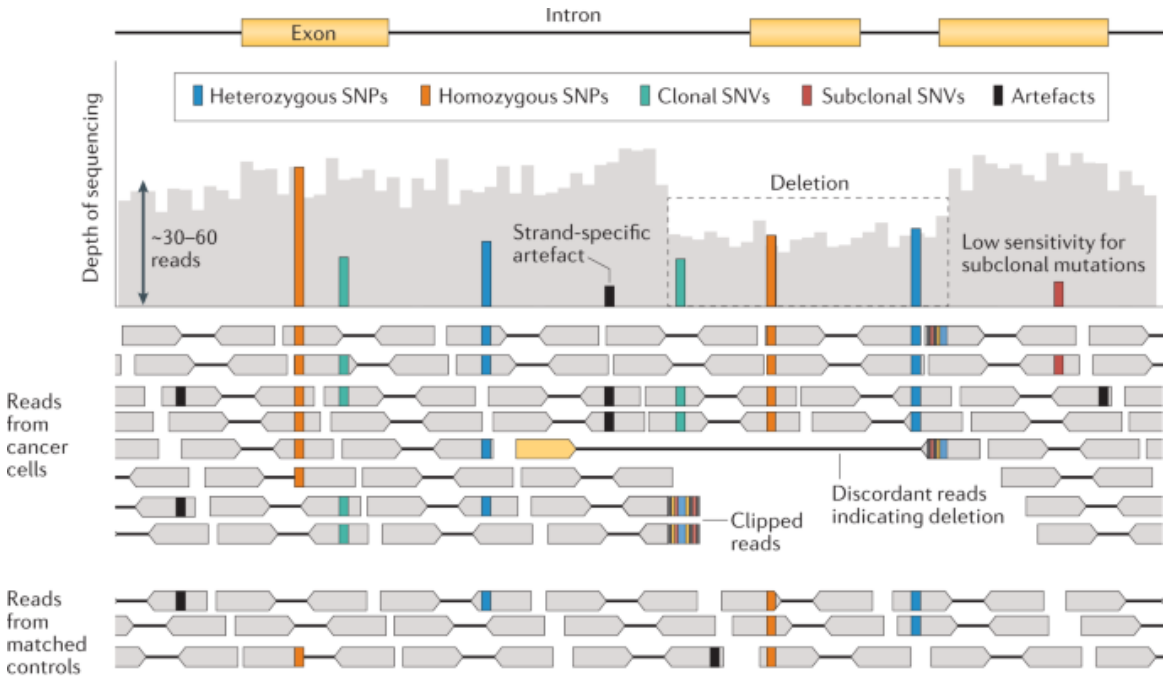
$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Normalization

The probability B being true

Somatic mutation detection is more complicated

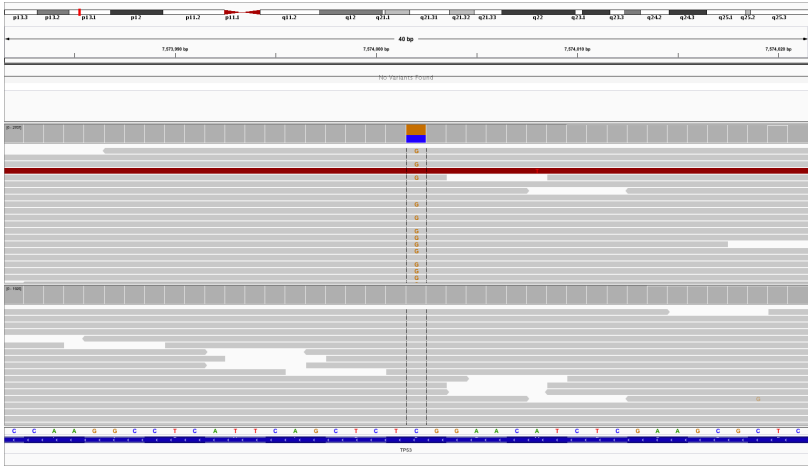
- Tumor purity
- Subclonal mutations
- Local copy numbers



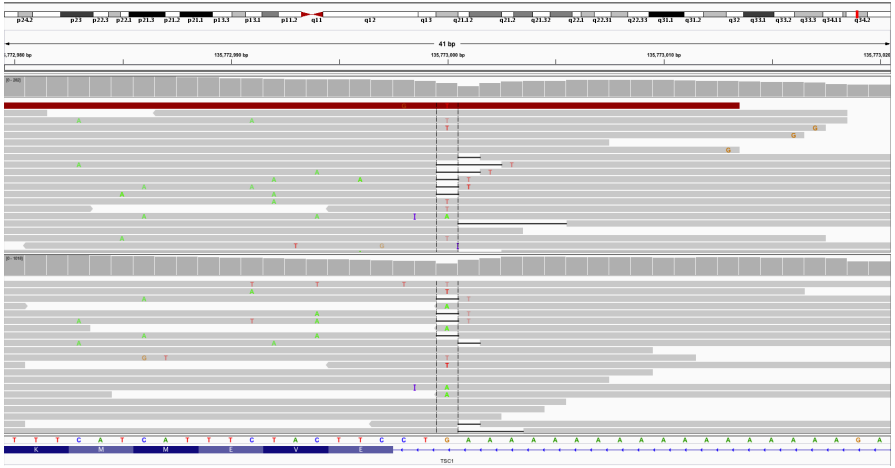
Cortés-Ciriano et al *Nature Reviews Genetics* (2022)

Somatic mutation detection is more complicated

A good call



A bad call



https://best-practices-for-processing-hts-data.readthedocs.io/en/latest/mutect2_pitfalls.html

Memorial Sloan Kettering Cancer Center

Summary of somatic mutation detection

Variant caller	Type of variant	Single-sample mode	Type of core algorithm
BAYSIC [48]	SNV	No	Machine learning (ensemble caller)
CaVEMan [34]	SNV	No	Joint genotype analysis
deepSNV [38]	SNV	No	Allele frequency analysis
EBCall [37]	SNV, indel	No	Allele frequency analysis
FaSD-somatic [31]	SNV	Yes	Joint genotype analysis
FreeBayes [44]	SNV, indel	Yes	Haplotype analysis
HapMuC [42]	SNV, indel	Yes	Haplotype analysis
JointSNVMix2 [30]	SNV	No	Joint genotype analysis
LocHap [43]	SNV, indel	No	Haplotype analysis
LoFreq [36]	SNV, indel	Yes	Allele frequency analysis
LoLoPicker [39]	SNV	No	Allele frequency analysis
MutationSeq [45]	SNV	No	Machine learning
MuSE [40]	SNV	No	Markov chain model
MuTect [35]	SNV	Yes	Allele frequency analysis
SAMtools [8]	SNV, indel	Yes	Joint genotype analysis
Platypus [41]	SNV, indel, SV	Yes	Haplotype analysis

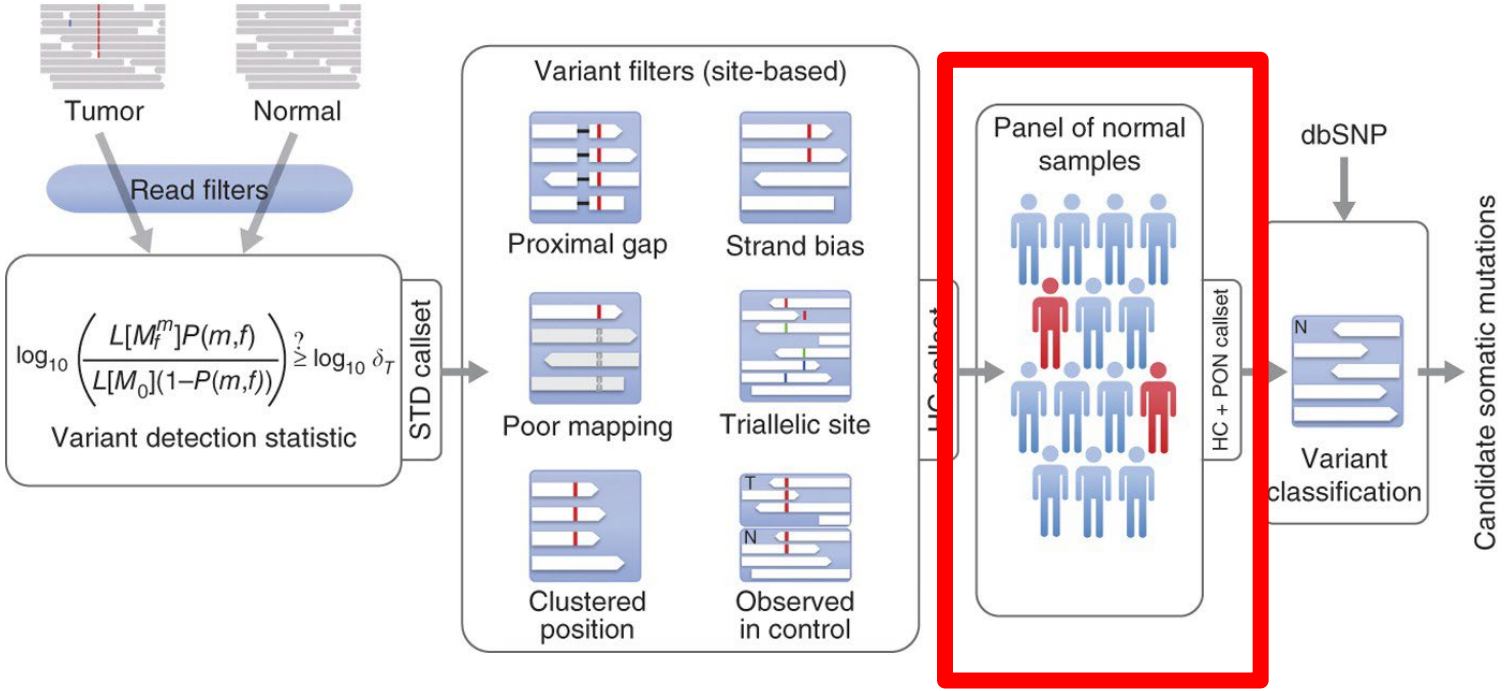
Xu. Comput Struct Biotechnol (2018)

Other things to consider when choosing a tool:

- Type of DNA sample – FFPE
- Type of sequencing data

  WGS
  WES
  Targeted panel

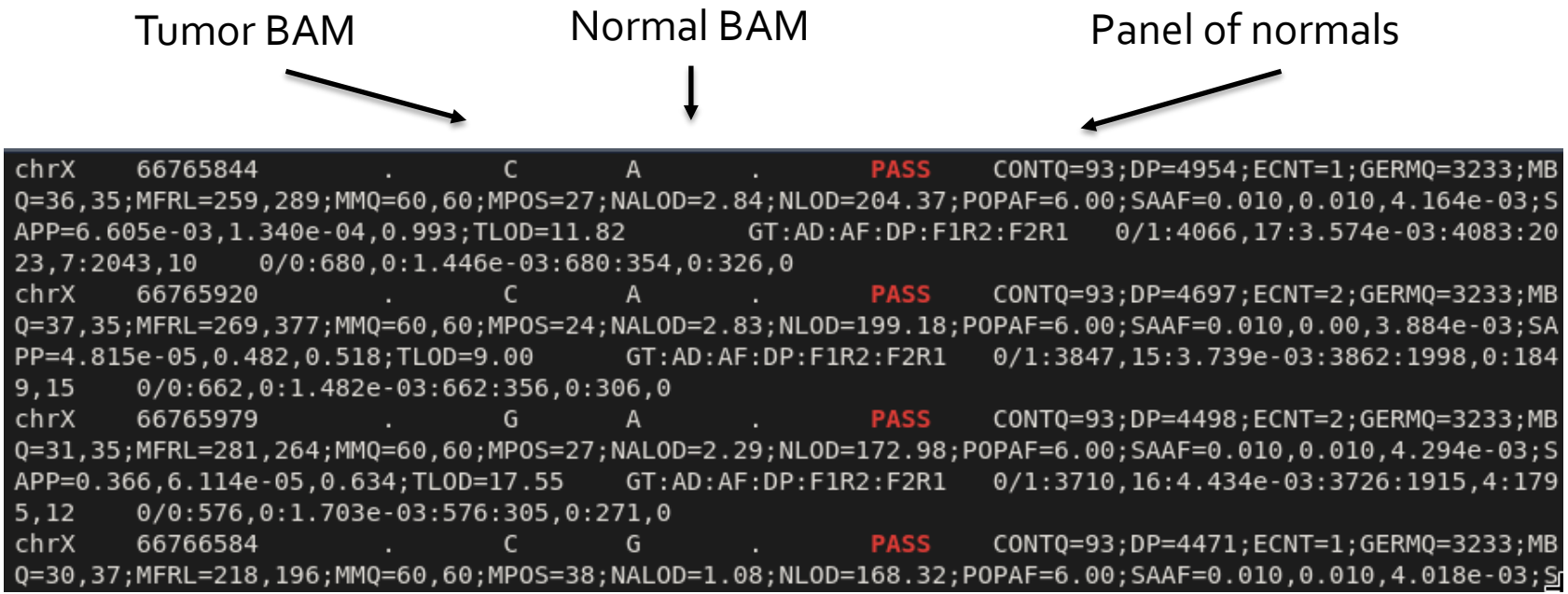
MuTect: a sensitive approach to detect low allelic fraction mutations



M^0 assumes all non-reference reads come from technical artifacts

M^f assumes that variant allele is present at an unknown frequency f.

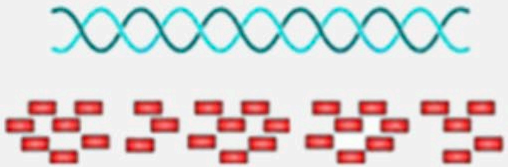
Example output from MuTect2 VCF format




```
chrX    66765844        .       C       A       .       PASS    CONTQ=93;DP=4954;ECNT=1;GERMQ=3233;MB
Q=36,35;MFRL=259,289;MMQ=60,60;MPOS=27;NALOD=2.84;NLOD=204.37;POPAF=6.00;SAAF=0.010,0.010,4.164e-03;S
APP=6.605e-03,1.340e-04,0.993;TLOD=11.82        GT:AD:AF:DP:F1R2:F2R1   0/1:4066,17:3.574e-03:4083:20
23,7:2043,10    0/0:680,0:1.446e-03:680:354,0:326,0
chrX    66765920        .       C       A       .       PASS    CONTQ=93;DP=4697;ECNT=2;GERMQ=3233;MB
Q=37,35;MFRL=269,377;MMQ=60,60;MPOS=24;NALOD=2.83;NLOD=199.18;POPAF=6.00;SAAF=0.010,0.00,3.884e-03;SA
PP=4.815e-05,0.482,0.518;TLOD=9.00      GT:AD:AF:DP:F1R2:F2R1   0/1:3847,15:3.739e-03:3862:1998,0:184
9,15    0/0:662,0:1.482e-03:662:356,0:306,0
chrX    66765979        .       G       A       .       PASS    CONTQ=93;DP=4498;ECNT=2;GERMQ=3233;MB
Q=31,35;MFRL=281,264;MMQ=60,60;MPOS=27;NALOD=2.29;NLOD=172.98;POPAF=6.00;SAAF=0.010,0.010,4.294e-03;S
APP=0.366,6.114e-05,0.634;TLOD=17.55    GT:AD:AF:DP:F1R2:F2R1   0/1:3710,16:4.434e-03:3726:1915,4:179
5,12    0/0:576,0:1.703e-03:576:305,0:271,0
chrX    66766584        .       C       G       .       PASS    CONTQ=93;DP=4471;ECNT=1;GERMQ=3233;MB
Q=30,37;MFRL=218,196;MMQ=60,60;MPOS=38;NALOD=1.08;NLOD=168.32;POPAF=6.00;SAAF=0.010,0.010,4.018e-03;S
```

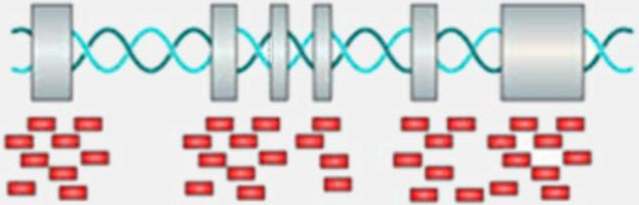
https://yulijia.net/slides/bioinfomatcis_for_medical_students/2019-08-16-A_beginners_guide_to_Call_SNPs_and_indels_Part_III.html#1

Whole genome sequencing



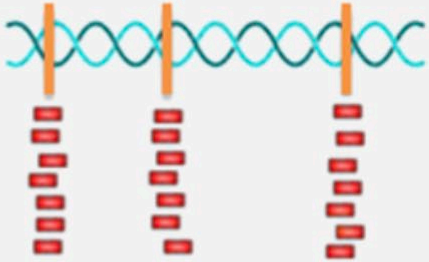
- Sequencing region : whole genome
- Sequencing Depth: >30X
- Covers everything – can identify all kinds of variants including SNPs, INDELs and SV.

Whole exome sequencing



- Sequencing region: whole exome
- Sequencing Depth : >50X ~ 100X
- Identify all kinds of variants including SNPs, INDELs and SV in coding region.
- Cost effective

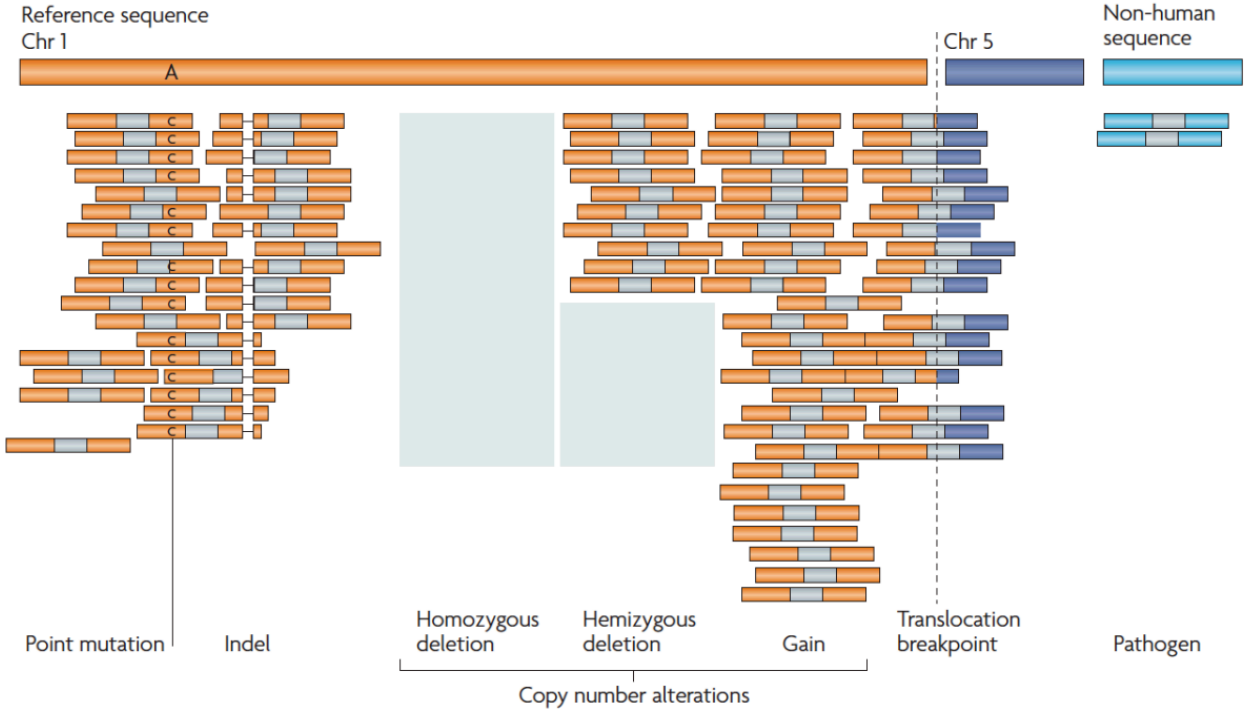
Targeted sequencing



- Sequencing region: specific regions (could be customized)
- Sequencing Depth : >500X
- Identify all kinds of variants including SNPs, INDELs in specific regions
- Most Cost effective

http://www.genomesop.com/somatic-mutations/

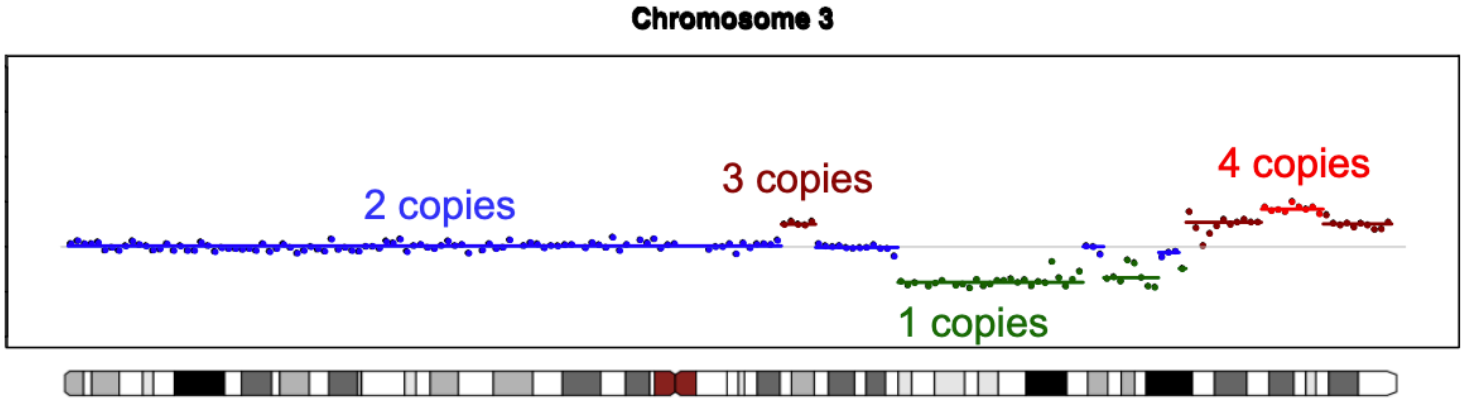
Somatic copy number alterations



Meyerson, Gabriel & Getz. *Nature Review Genetics* (2010)

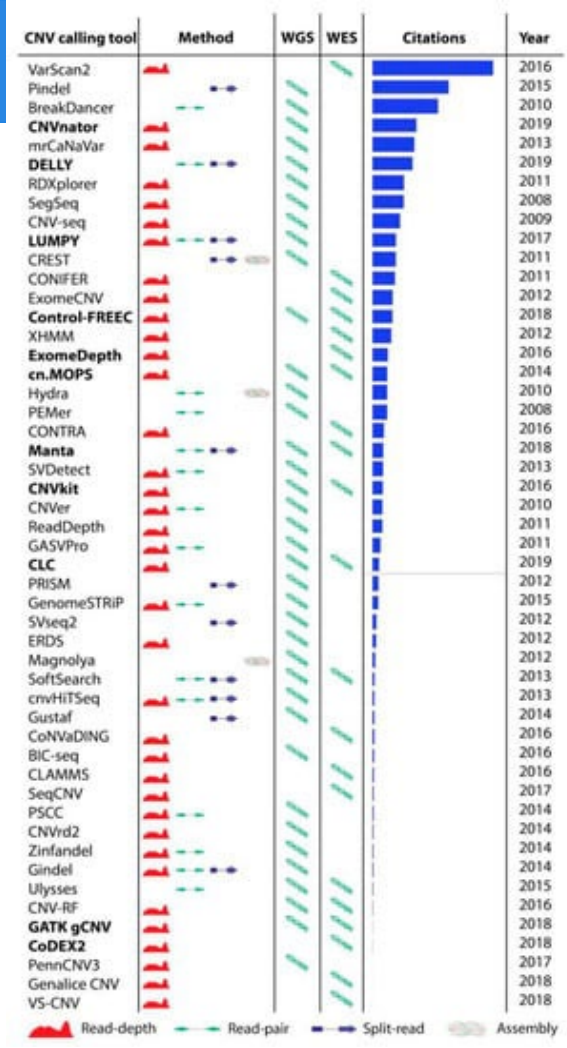
Somatic CNA detection overview

Ratio of tumor/normal coverage → Segmentation of coverage changes → Predicting the number of copies per segment



https://gavinhalab.org/teaching/GS541_sp23/

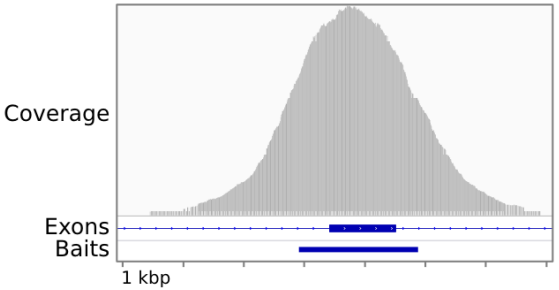
Somatic CNA detection overview



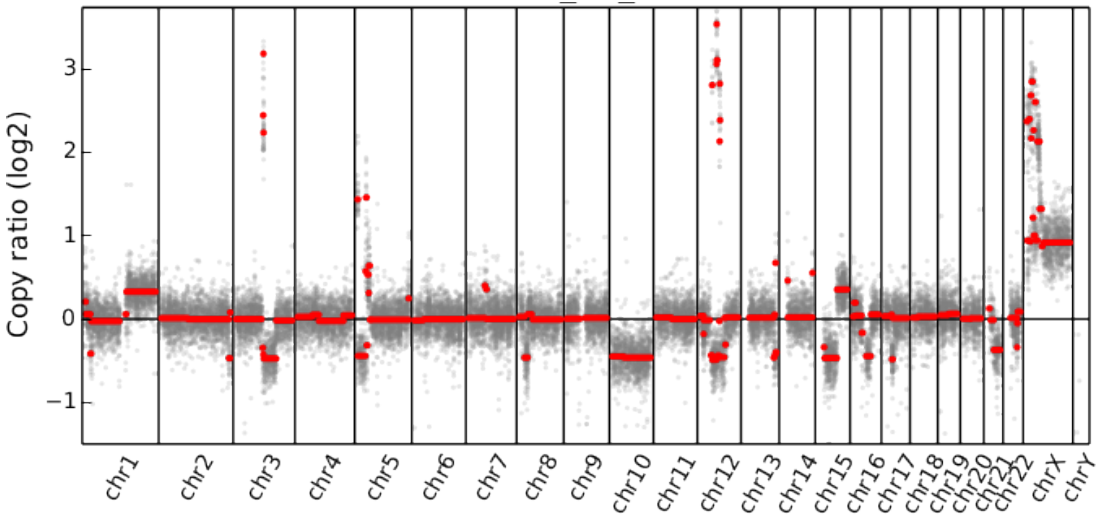
Gabrielaite et al *Cancers* (2021)

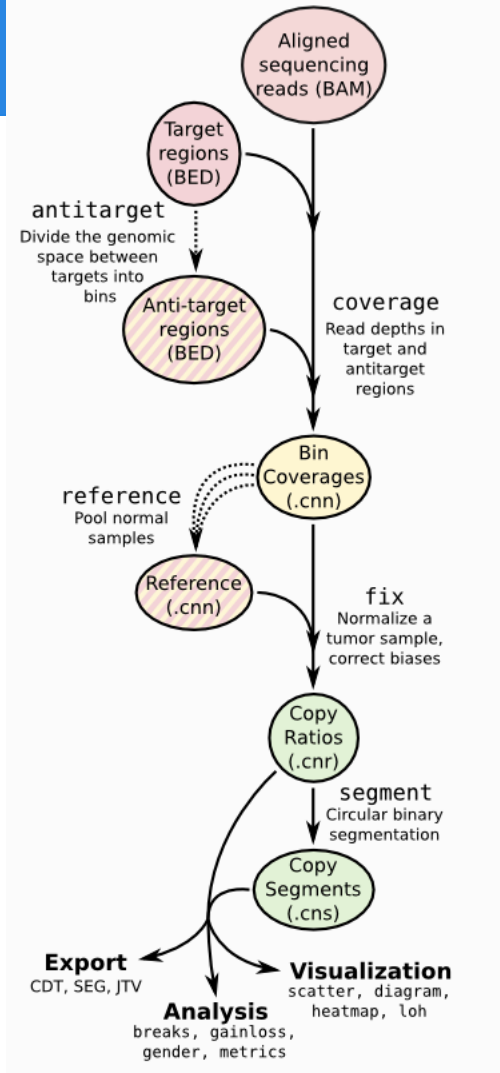
Memorial Sloan Kettering Cancer Center

CNVKit – detection from targeted DNA sequencing

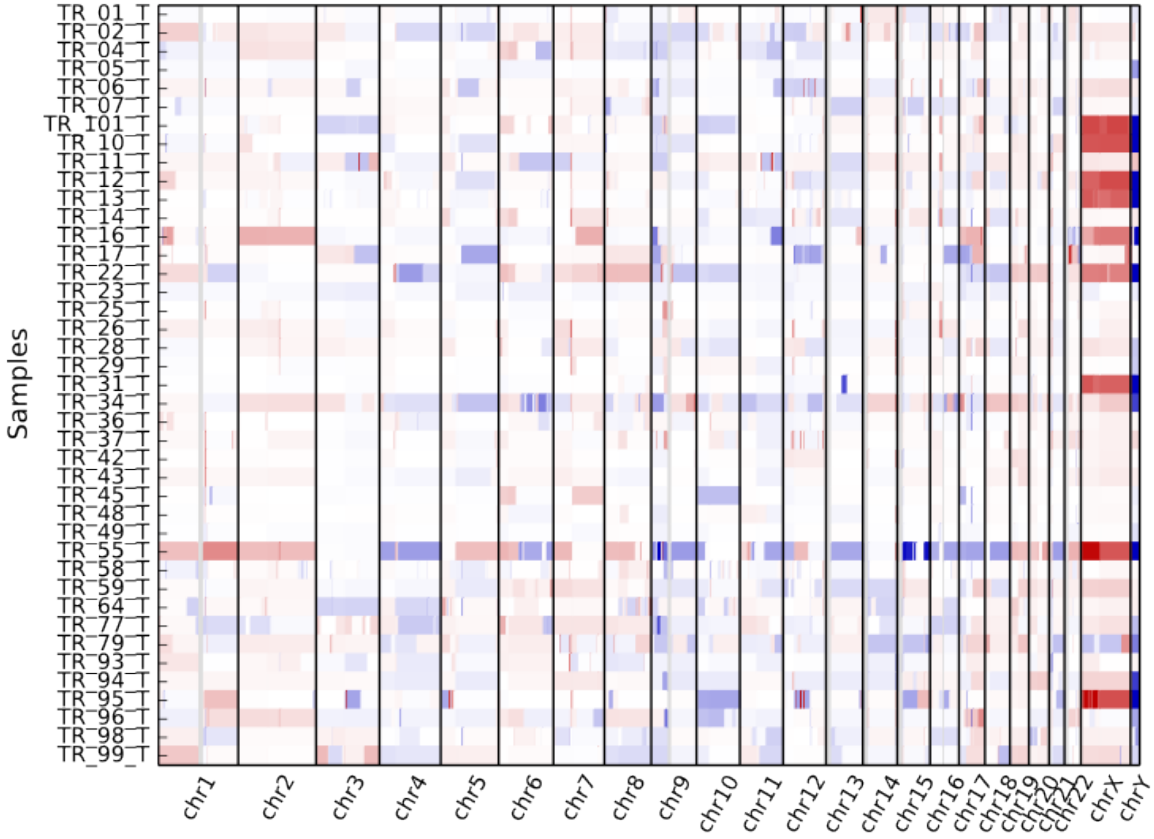


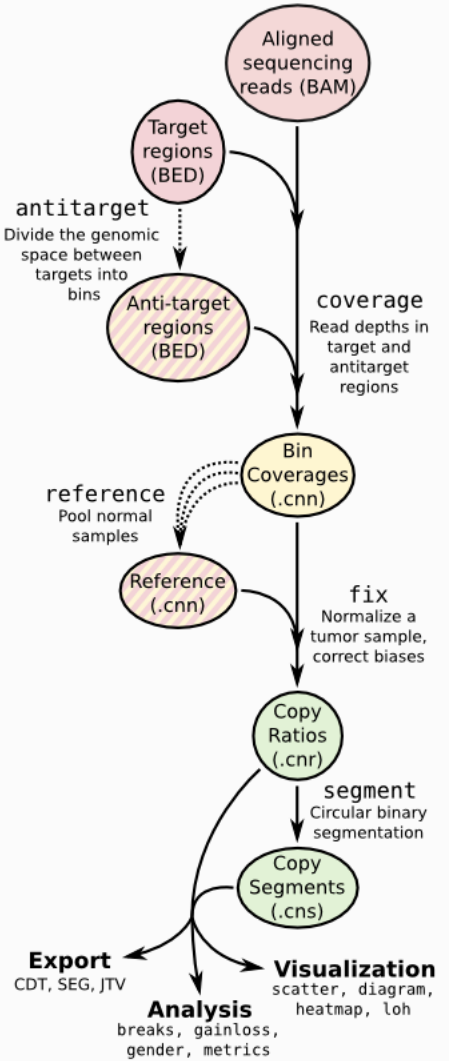
- ✓ Normalization and bias correction can be performed using a panel of normal samples



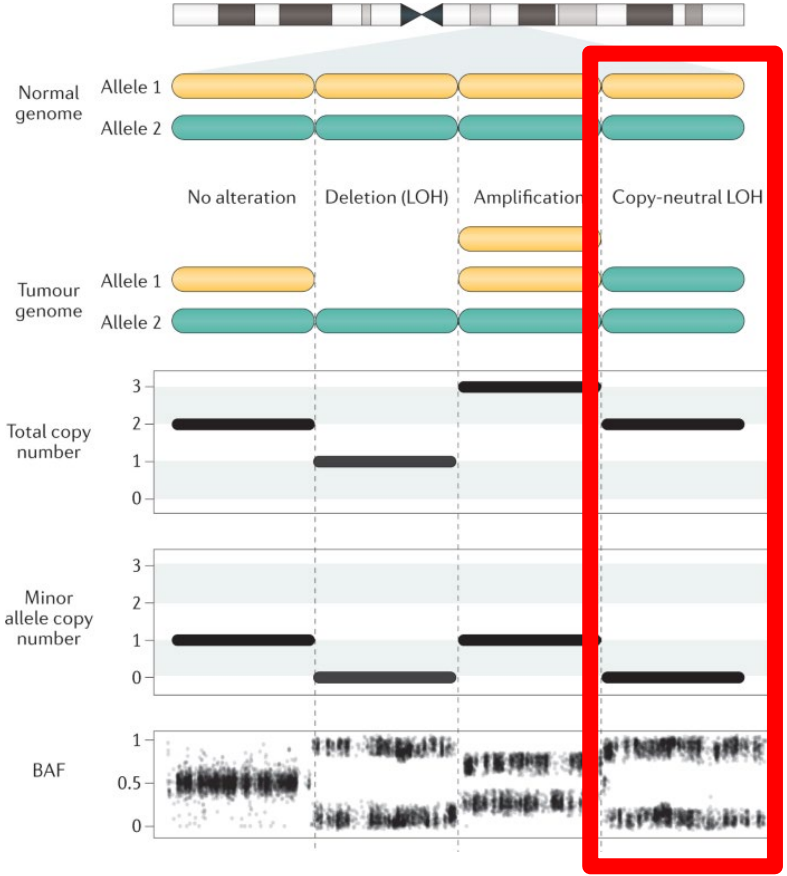


CNVKit – combine on-target and off-target reads





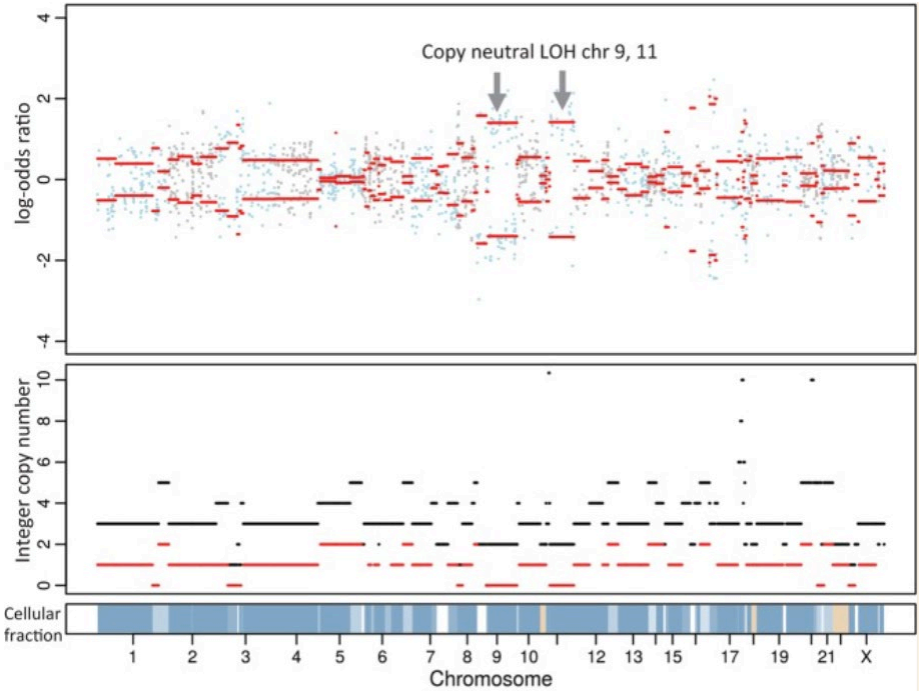
Copy Number Analysis adding the Allelic Features



- Infer fraction of genome altered
- Infer genome-wide LOH

Cortés-Ciriano et al *Nature Reviews Genetics* (2022)

FACETS: allele-specific copy number



- Joint segmentation of read coverage and allelic fraction using HMM

Shen & Seshan *Nucleic Acids Res* (2016)

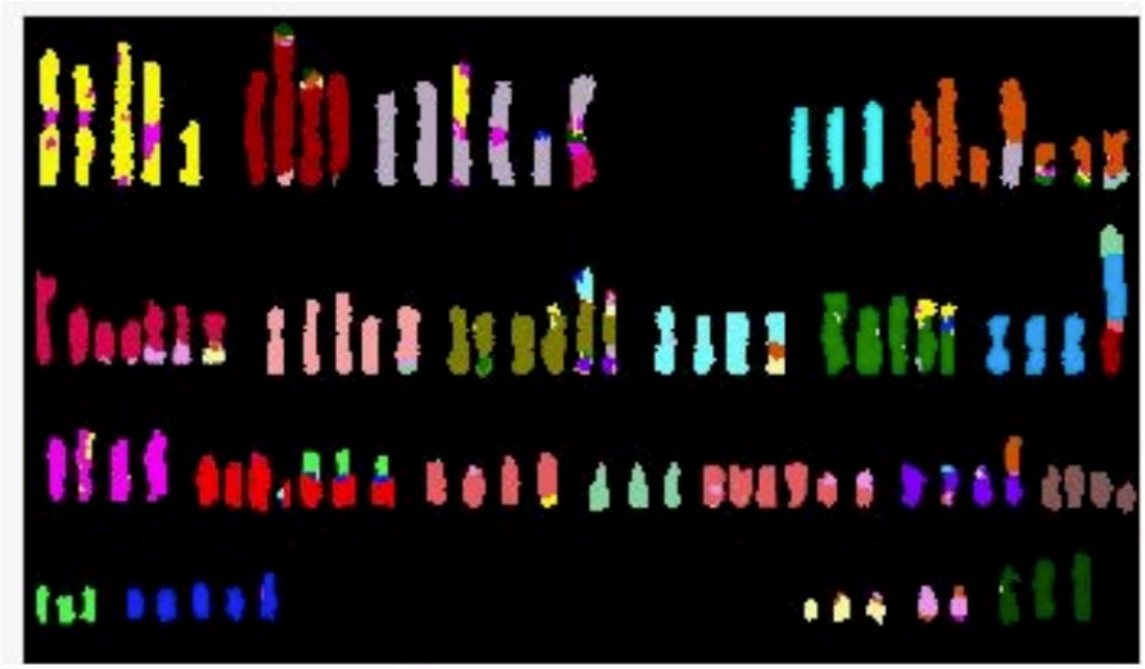
You can also use somatic copy number data to:

- Estimating tumor purity and ploidy (whole-genome doubling) from copy number analysis

- Estimating tumor heterogeneity from copy number analysis

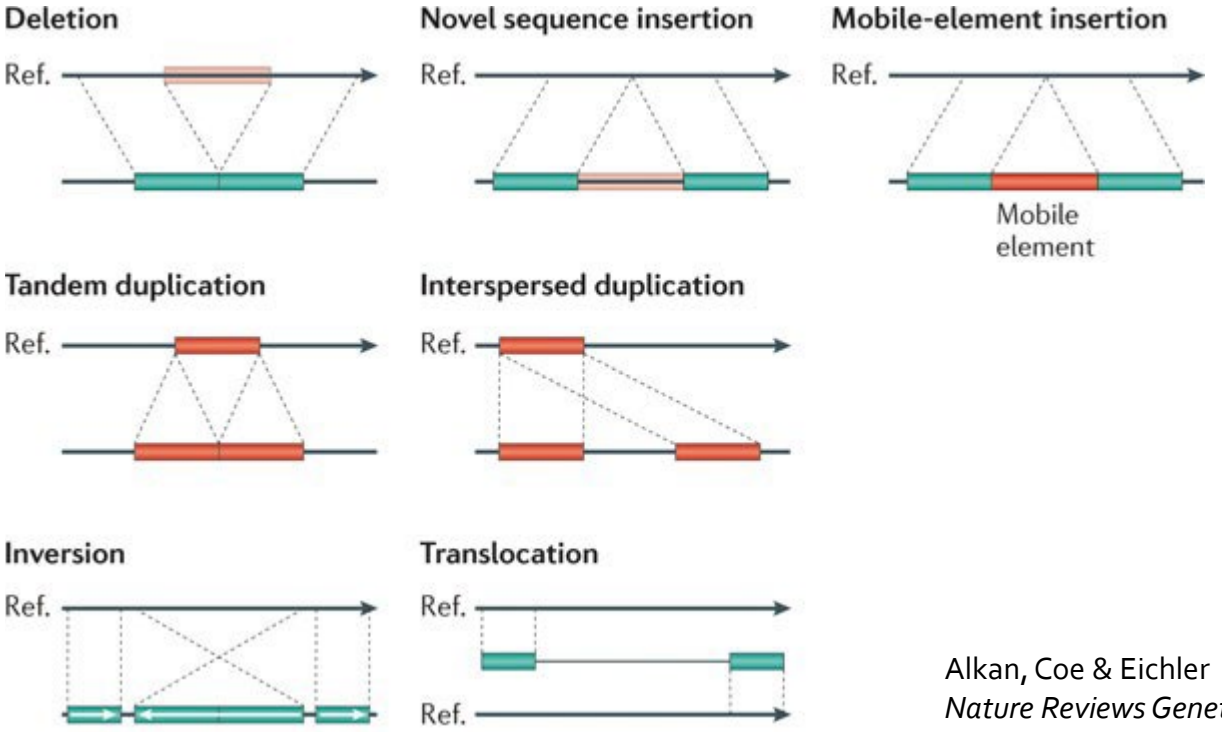
https://gavinhalab.org/teaching/GS541_sp23/

Structural variant detection from short read sequencing



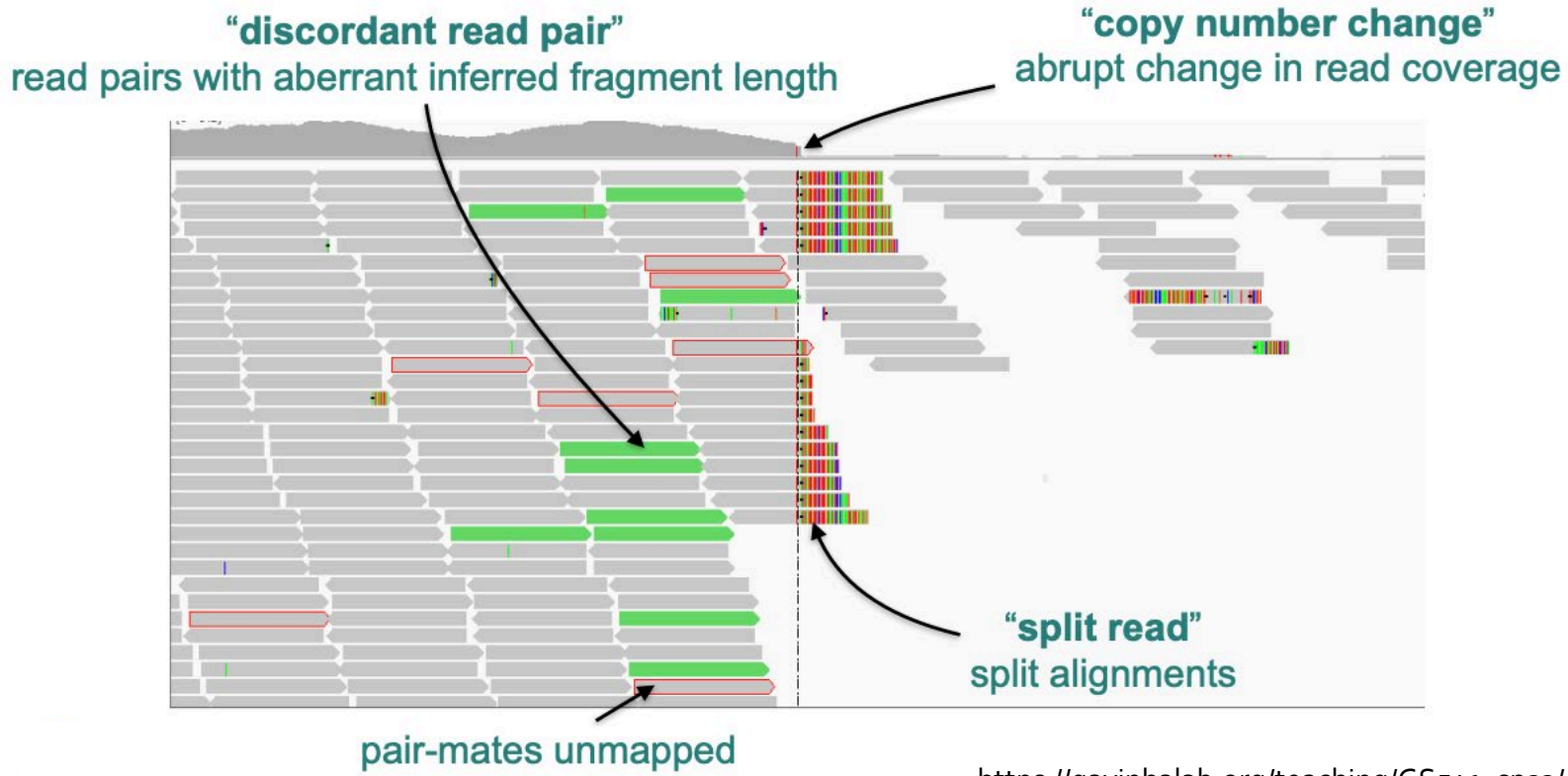
David Huntsman, BC Cancer Agency

Structural variants - genomic alterations >50bp



Alkan, Coe & Eichler
Nature Reviews Genetics (2011)

Sequence features in SVs

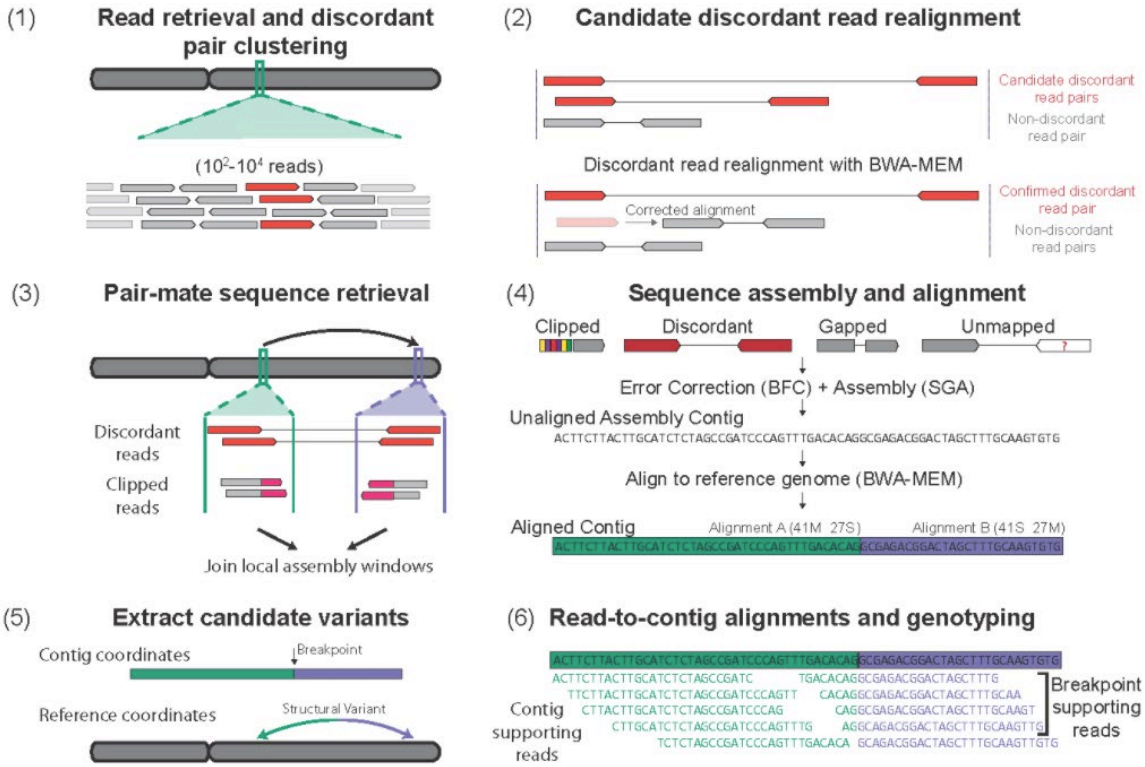


https://gavinhalab.org/teaching/GS541_sp23/

Method for detecting SVs

- Using read-pair information – use the span and orientation of paired reads information
- Using split read information – define breakpoint with high resolution
- Using local sequence assembly – find de novo insertions

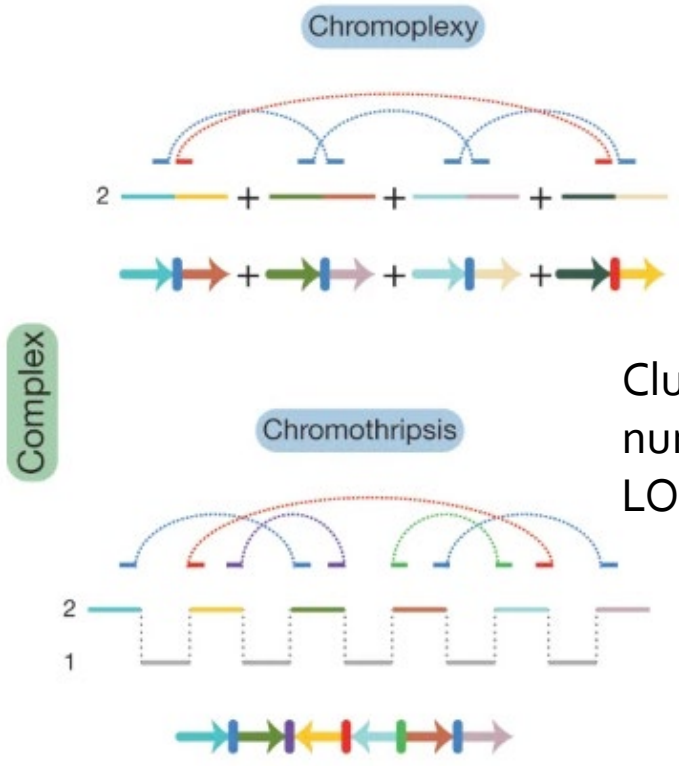
SvABA: using all three methods for SV detection



Wala et al *Genome Research* (2018)

Memorial Sloan Kettering Cancer Center

Complex rearrangements



Chains of balanced translocations

Clustered copy number loss and LOH

Other complex events

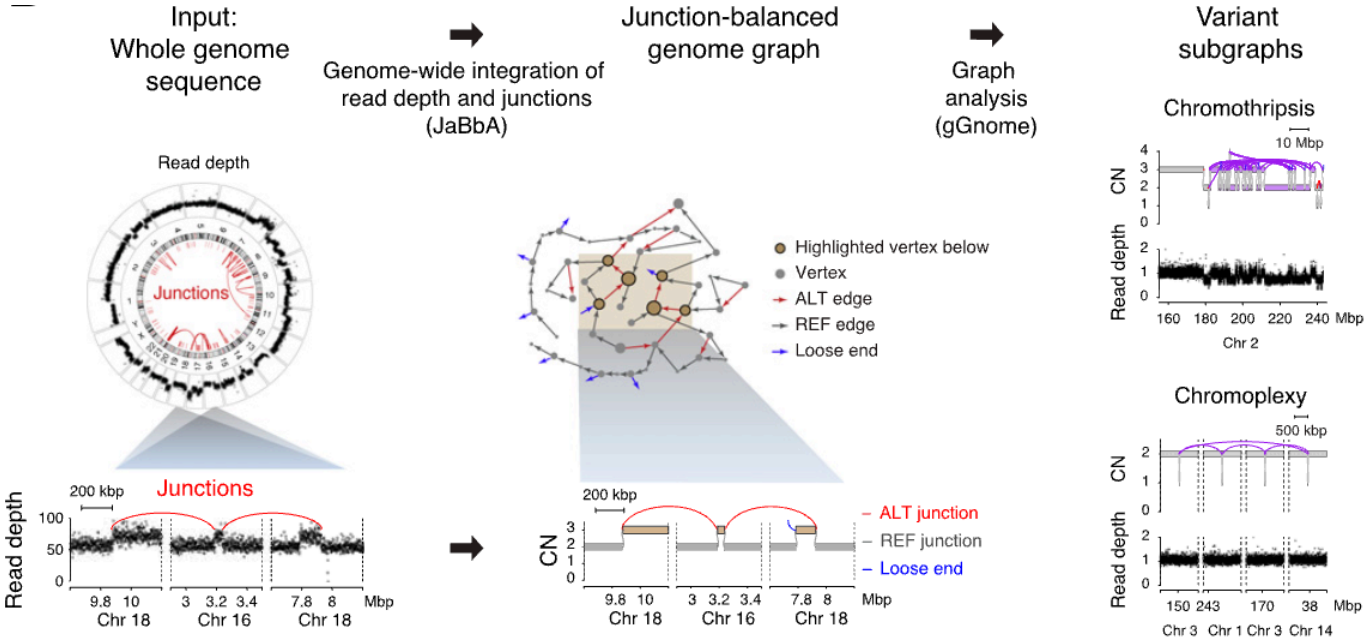
Breakage fusion bridge cycle

Extrachromosomal DNA

Usually involve high-level amplification of oncogenes

Methods to improve SV detection accuracy

- Graph-based classification methods to reconstruct complex SVs



https://github.com/mskilab-org/JaBbA

Memorial Sloan Kettering Cancer Center

Discover cancer genes from tumor sequencing

Passenger mutation

Neutral somatic changes caused by
exposure or genetic instability
that are unrelated to cancer development

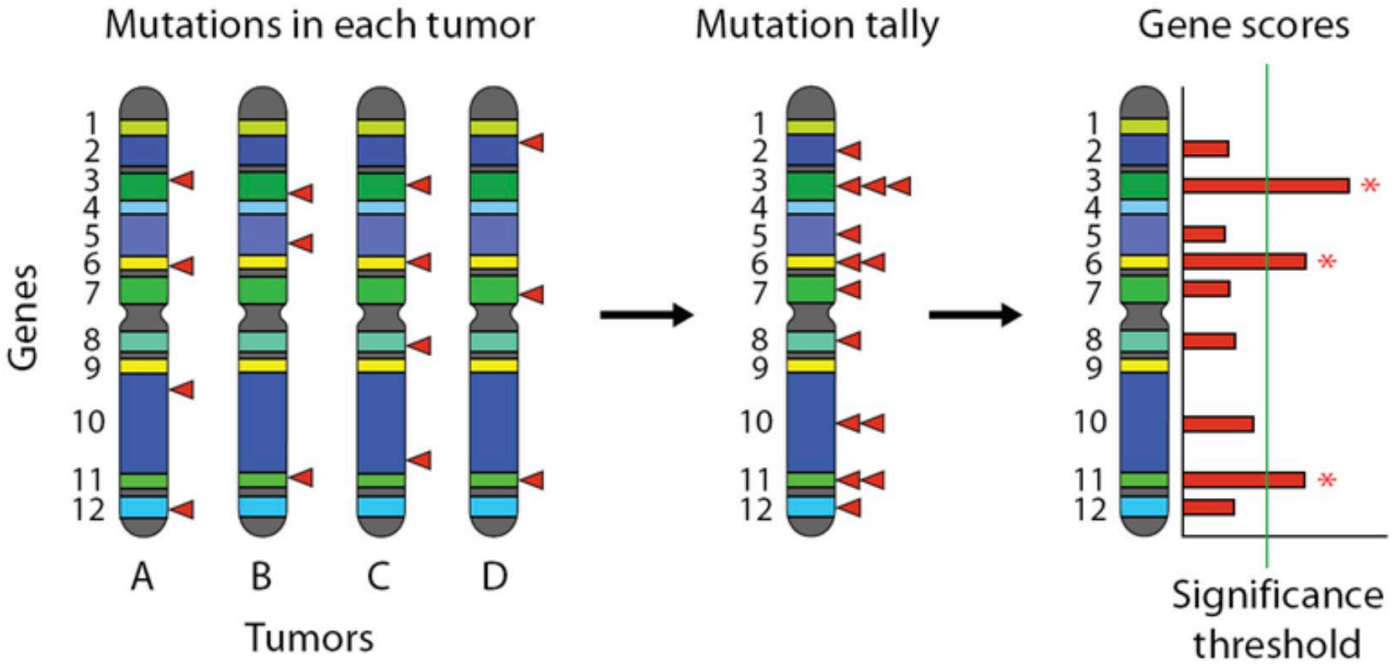
Driver mutation

Somatic changes caused by
exposure or genetic instability
that are essential for cancer development

Identify cancer drivers

- Input data: lists of mutations (and indels) from multiple samples
- Output data: a list of genes significantly mutated in the cohort
- Approach
  - Build a model of the background mutation processes
  - Using statistical methods to find genes mutated more often than expected by chance given background mutation processes

MutSig: identify significantly mutated genes



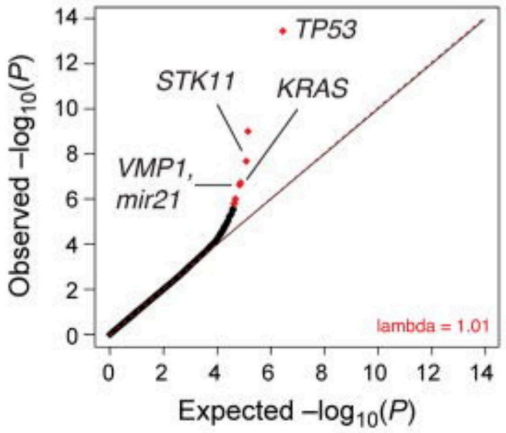
Identify cancer drivers

Additional information to reduce false positives:

- Incorporating variables into the background model to reduce false-positive findings
  - DNA replication time
  - Chromatin state
  - Expression level

fishHook:

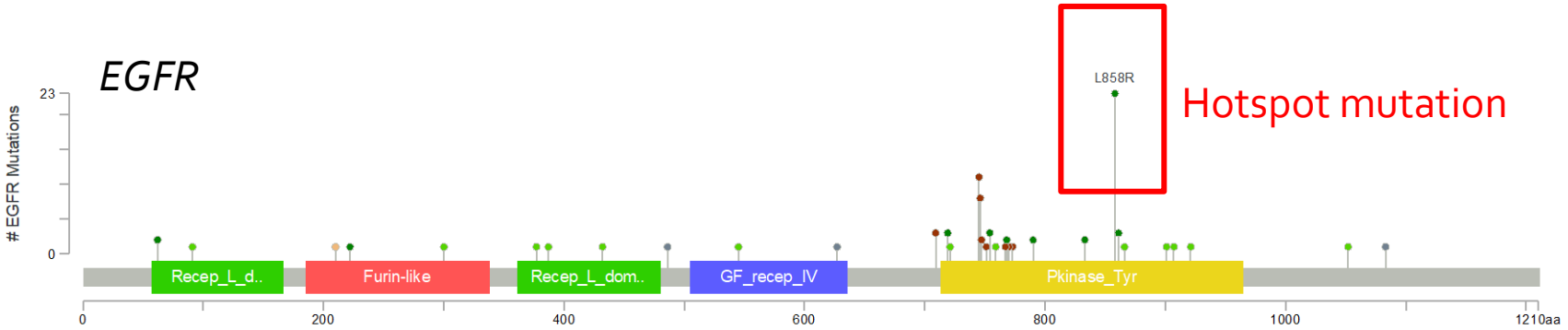
Generalized linear modeling (GLM) of somatic mutation densities and their heterogeneity along the genome



Identify cancer drivers

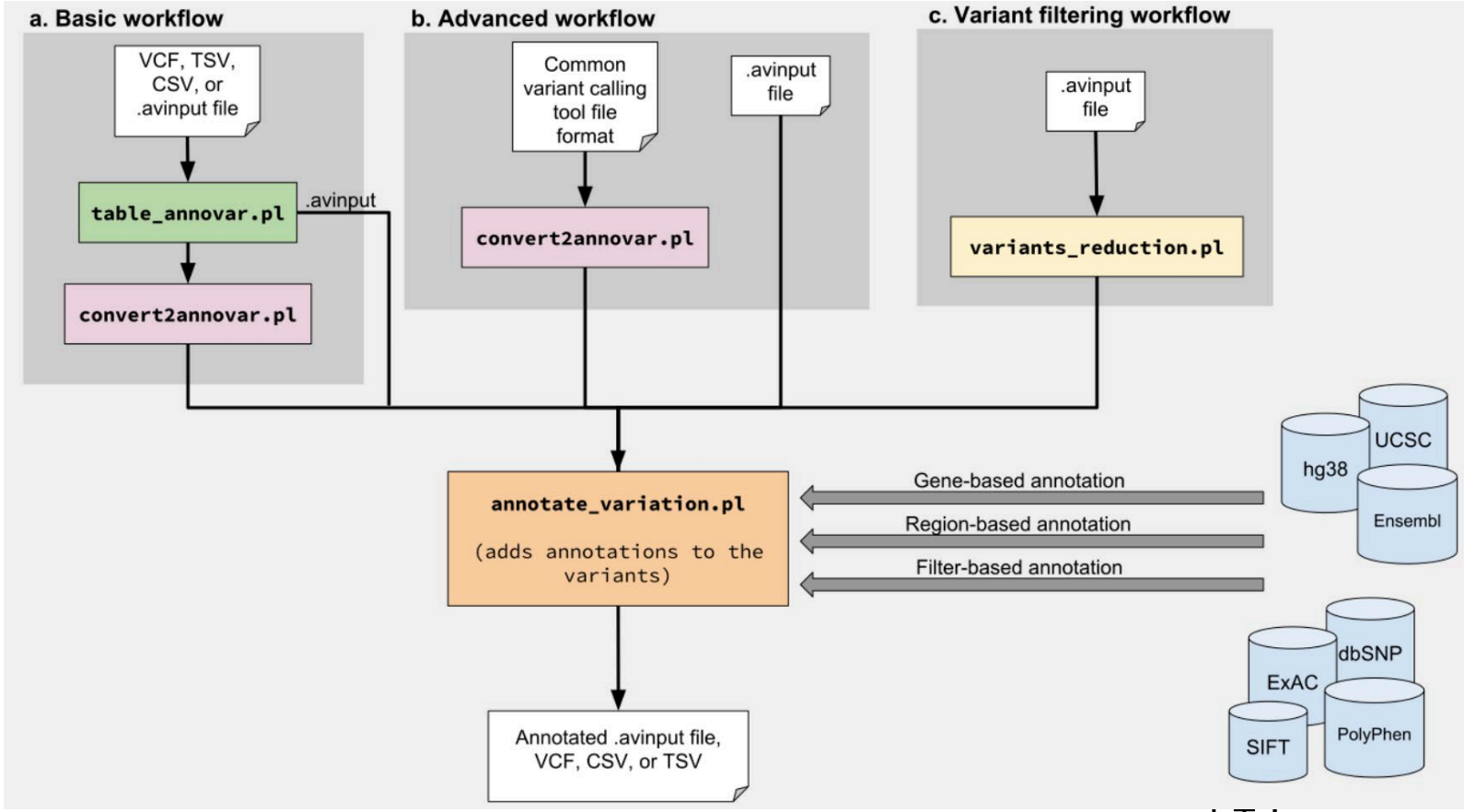
Additional information to reduce false positives:

- Clustered vs sparse



Functional characterization (experimental or computational) is important

ANNOVAR for the interpretation and prioritization of single nucleotide variants



Variant effect prediction (VEP)

For functional prediction of variants in whole-exome data:

- dbnsfp47a: this dataset already includes SIFT, PolyPhen2 HDIV, PolyPhen2 HVAR, LRT, MutationTaster, MutationAssessor, FATHMM, MetaSVM, MetaLR, VEST, CADD, GERP++, DANN, AlphaMissense, fitCons, PhyloP and SiPhy scores, but ONLY on coding variants.

For functional prediction of splice variants:

- dbscsnv11: dbscSNV version 1.1 for splice site prediction by AdaBoost and Random Forest, which score how likely that the variant may affect splicing
- spidex: deep learning based prediction of splice variants. Unlike dbscsnv11, these variants could be far away from canonical splice sites

For disease-specific variants:

- clinvar_20160302: ClinVar database with separate columns (CLINSIG CLNDBN CLNACC CLNDSDB CLNDSDBID) for each variant (Please check the download page for the latest version, or read below for creating your own most updated version)
- cosmic70: the latest COSMIC database with somatic mutations from cancer and the frequency of occurence in each subtype of cancer. For more updated cosmic, see instructions below on how to make them.
- icgc21: International Cancer Genome Consortium version 21 mutations.
- nci60: NCI-60 human tumor cell line panel exome sequencing allele frequency data

RESEARCH ARTICLE | MACHINE LEARNING

Accurate proteome-wide missense variant effect prediction with AlphaMissense

JUN CHENG, GUIDO NOVATI, JOSHUA PAN, CLARE BYCROFT, AKVILĖ ŽEMGULYTĖ, TAYLOR APPLEBAUM, ALEXANDER PRITZEL, LAI HONG WONG, MICHAL ZIELINSKI, [...], AND ŽIGA AVSEC +6 authors Authors Info & Affiliations

SCIENCE • 19 Sep 2023 • Vol 381, Issue 6664 • DOI: 10.1126/science.adg7492

221,665 4

Article | Open access | Published: 10 August 2023

Genome-wide prediction of disease variant effects with a deep protein language model

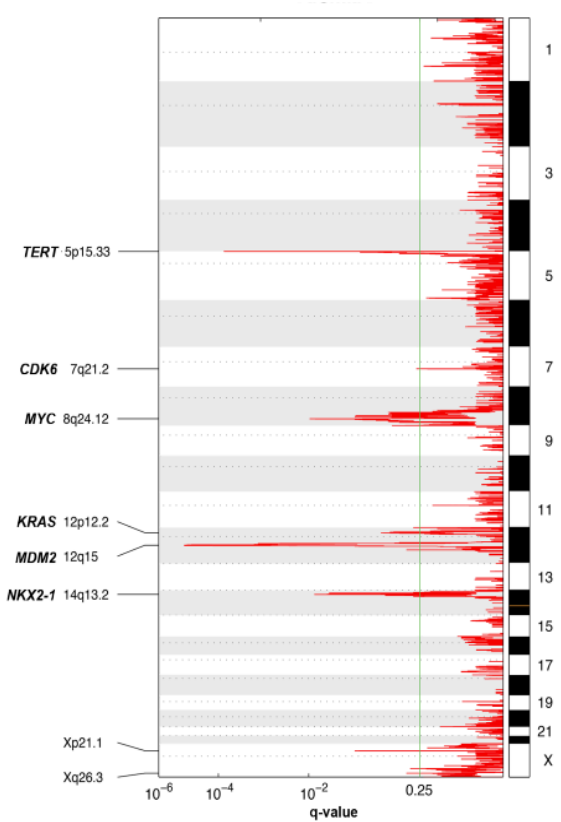
Nadav Brandes, Grant Goldman, Charlotte H. Wang, Chun Jimmie Ye & Vasilis Ntranos

Nature Genetics **55**, 1512–1522 (2023) | Cite this article

70k Accesses | **62** Citations | **186** Altmetric | Metrics

Memorial Sloan Kettering Cancer Center

GISTIC: Identify significantly recurrent SCNAs



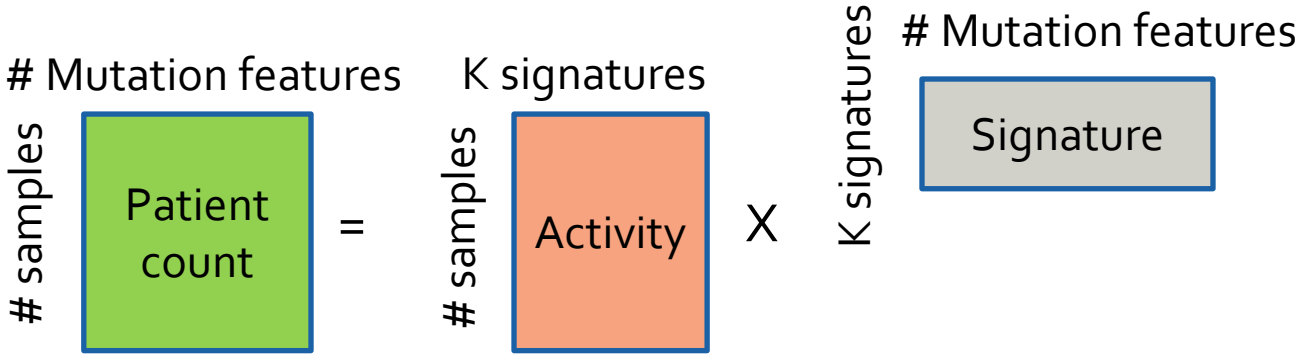
- Identify driver SCNAs by evaluating the frequency and amplitude of observed events

Memorial Sloan Kettering Cancer Center

Identify patterns of somatic alterations

To infer the source of mutational processes

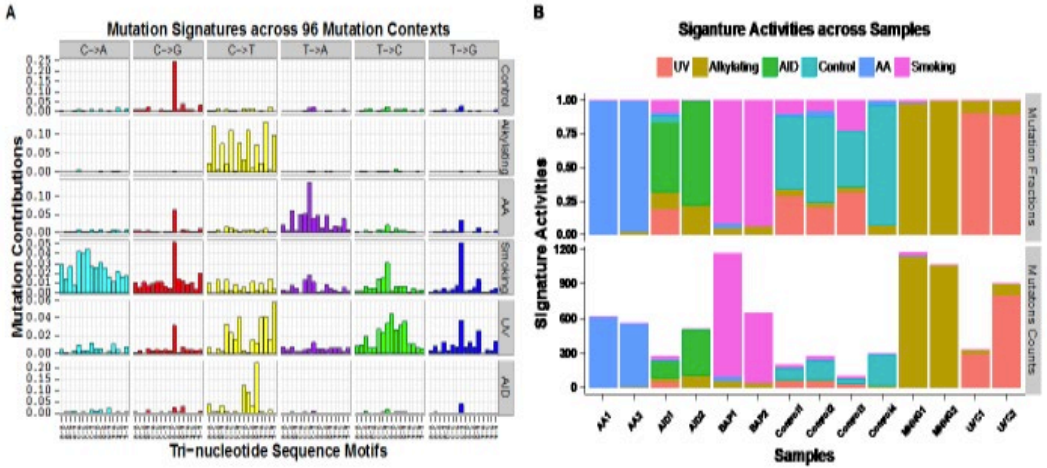
- Non-negative matrix factorization (NMF) is commonly used to detect signatures



Mutational signature

- Input: Counts of 96 base substitutions in tri-nucleotide sequence contexts per patient

- Output:



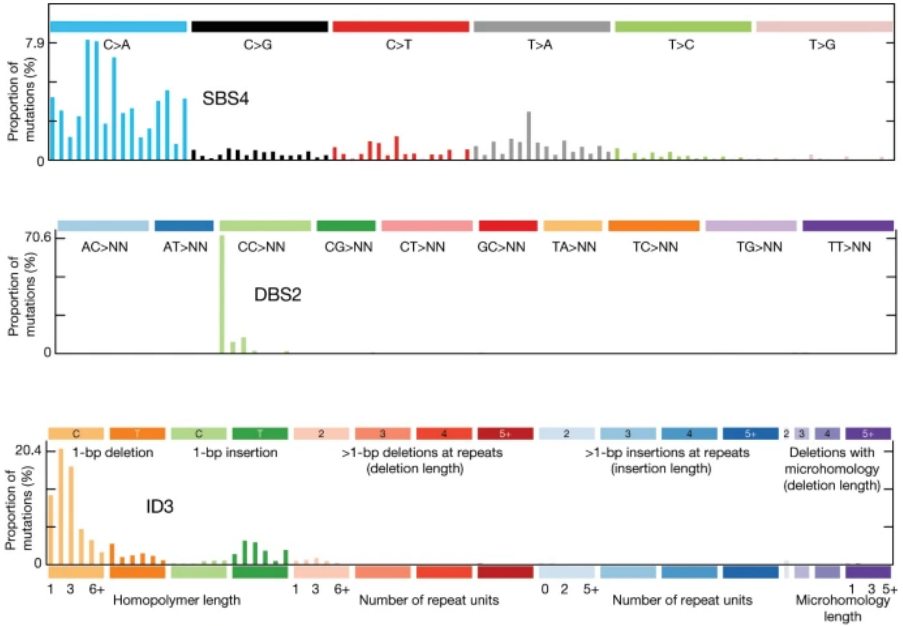
https://software.broadinstitute.org/cancer/cga/msp

Smoking mutational signature

Single-base substitution

Doublet-base substitutions

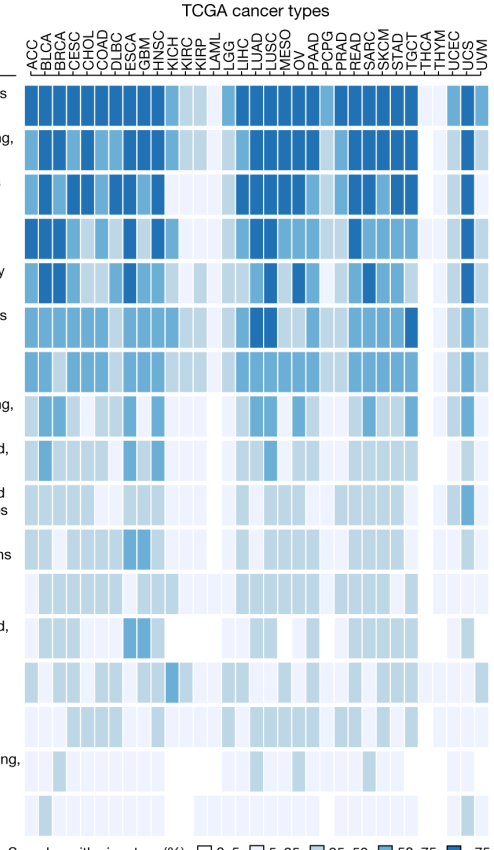
Small indels



Alexandrov et al *Nature* (2020)

Copy number and SV signatures

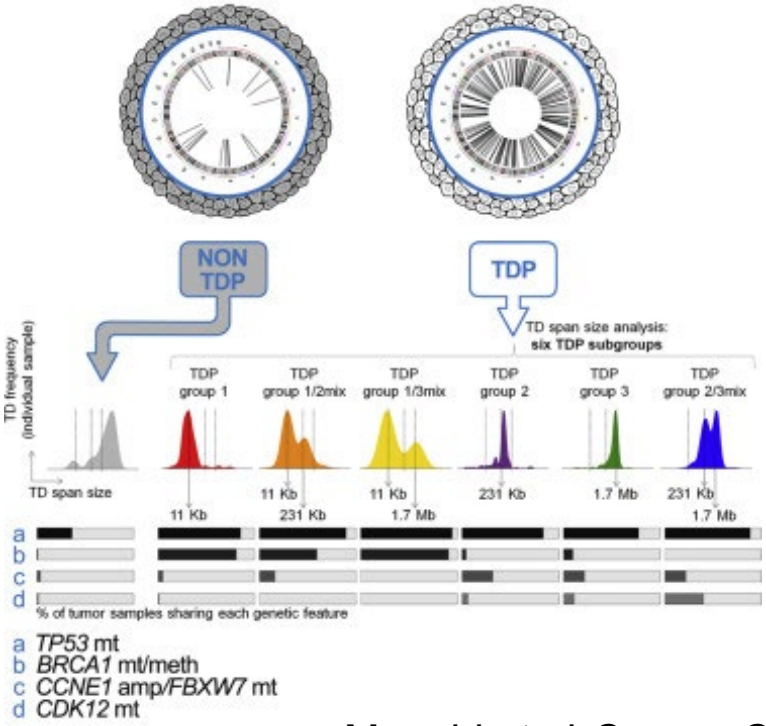
Signature	Pan-cancer prevalence (%)	Putative cause	Confidence	Pattern of change
CX1	80	Chromosome missegregation via defective mitosis and/or telomere dysfunction	★★★	Whole-arm/chromosome changes
CX2	69	IHR	★★★	Short-to-medium-sized, oscillating, single-copy changes
CX3	63	IHR with replication stress and impaired damage sensing	★★★	Long-sized, single-copy changes
CX4	58	PI3K–AKT-mediated toleration of whole-genome duplication	★★★	Medium-sized, clustered and oscillating, two-copy changes
CX5	56	IHR with replication stress	★★★	Medium-sized, two-to-three-copy changes
CX6	51	Chromosome missegregation via defective mitosis	★★★	Whole-arm/chromosome changes
CX7	50	Unknown	☆☆☆	Long-sized, less than one-copy change
CX8	38	Replication stress	★★★	Short-to-medium-sized, oscillating, low-level amplifications
CX9	32	Replication stress	★☆☆	Short-to-medium-sized, clustered, mid-level amplifications
CX10	29	Impaired NHEJ with replication stress	★★☆	Short-to-medium-sized, clustered and oscillating, two-copy changes
CX11	28	Replication stress	★★☆	Short-sized, clustered and oscillating, mid-level amplifications
CX12	27	Unknown	☆☆☆	Very short oscillating changes
CX13	26	Replication stress	★☆☆	Short-to-medium-sized, clustered, high-level amplifications
CX14	26	Chromosome missegregation via defective mitosis	★★☆	Whole-arm/chromosome, less than one-copy changes
CX15	25	Unknown	☆☆☆	Long-sized, less than one-copy changes
CX16	17	Unknown	★☆☆	Very short, clustered and oscillating, two-to-three-copy changes
CX17	16	Unknown	★★☆	Short-to-long-sized, one-to -three-copy changes



Drews et al *Nature* (2022)

Memorial Sloan Kettering Cancer Center

Copy number signatures associated mutations



- *BRCA1*-deficient phenotype — tandem duplicator
- *BRCA2*-deficient — small deletions
- **Genome-wide LOH** — **biomarker for PARP inhibitors in ovarian cancer**

Menghi et al *Cancer Cell* (2018)

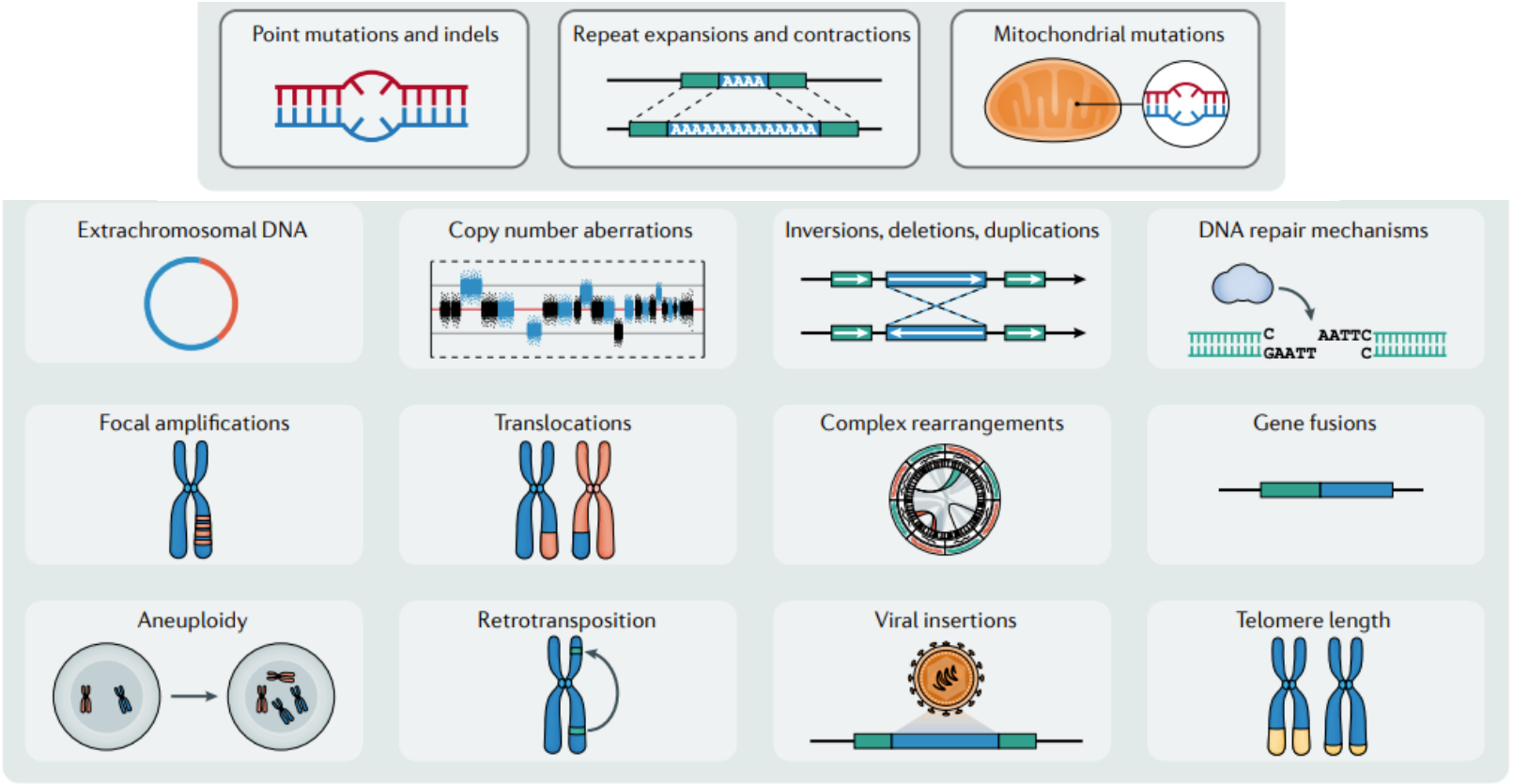
Running significant analysis in practice

- Significant and signature analyses can also be noise detectors
- Filtering false positive mutation or SCNA calls is essential
  - Using consensus calls from **multiple callers**
  - Using a panel of normal
- Multiple algorithms can be used to improve sensitivity/specificity

Create a panel of normal best matching the study cohort

- Examples:
  - Use normal samples from the same sequencing technology (exome, panel, long reads etc.) to model technical artifacts
  - Use normal samples from the same sequencing batch to model batch effects
  - Use normal samples from the same population to filter population-specific germline variations

What else can we detect from tumor sequencing



Cortés-Ciriano et al *Nature Reviews Genetics* (2022)

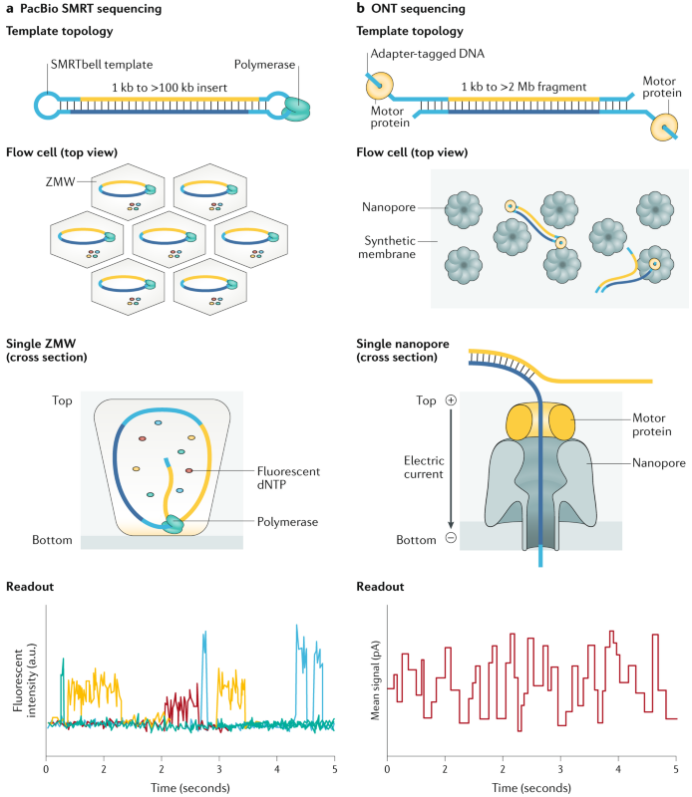
Challenges and future directions:

- In the era of whole-genome analysis, we need better methods to identify significantly mutated non-coding functional elements

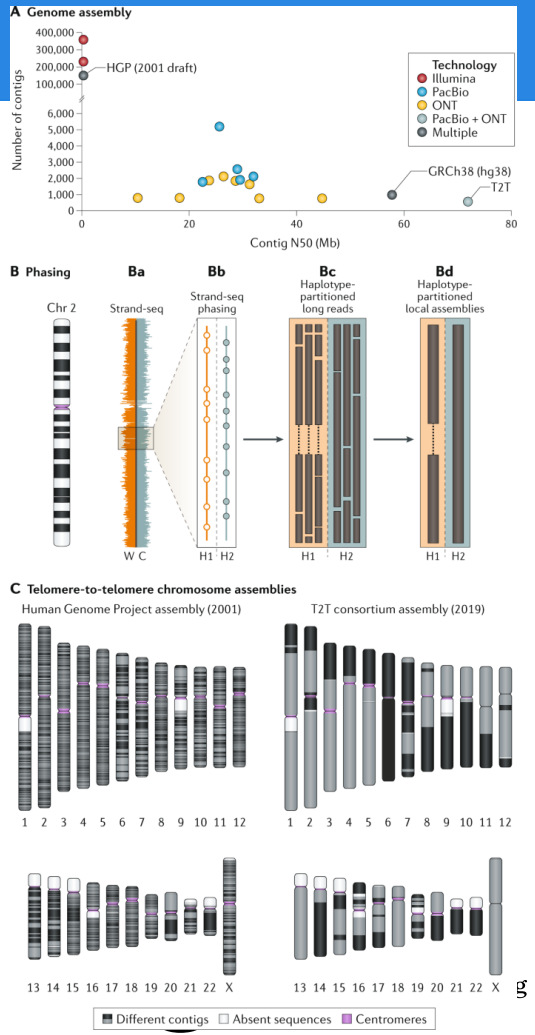
- Methods incorporating genetic and epigenetic data to detect non-coding mutations

- Methods incorporating mutation, copy number and structural variants to detect more cancer genes

Long-read human genome sequencing and application

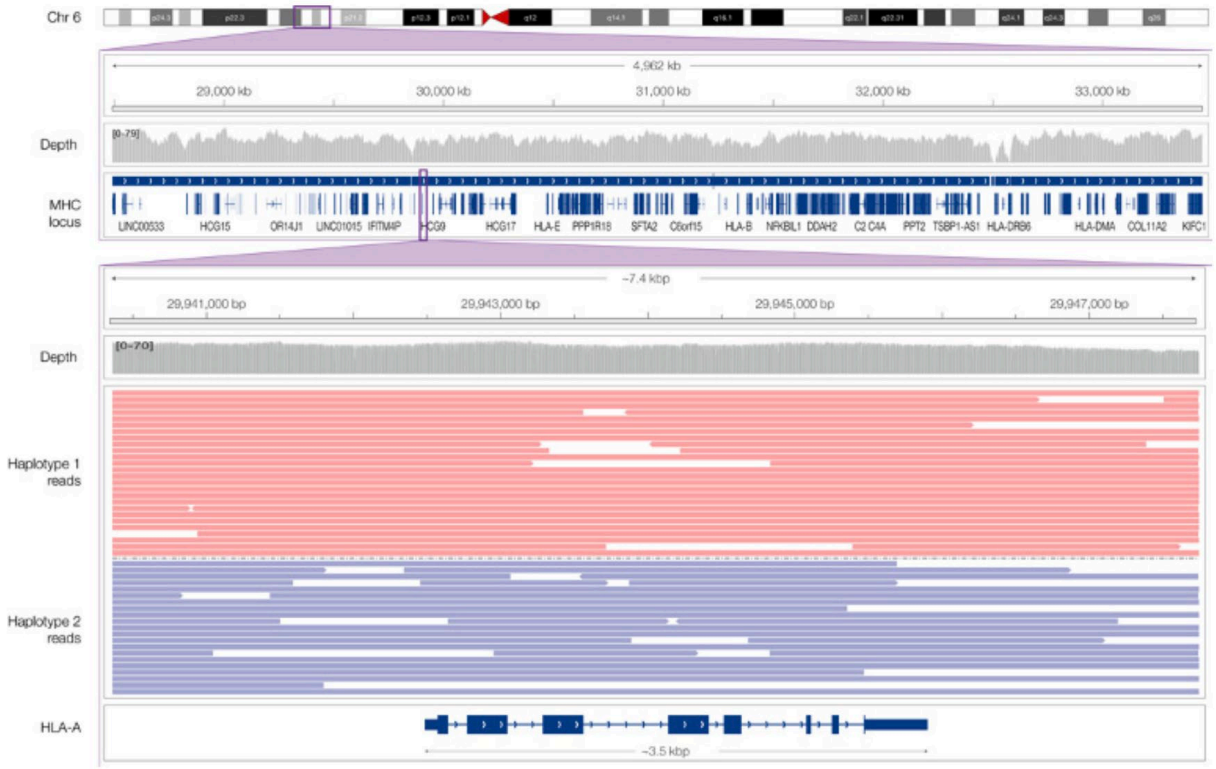


Logsdon, Wollger & Eichler *Nature Reviews Genetics* (2020)

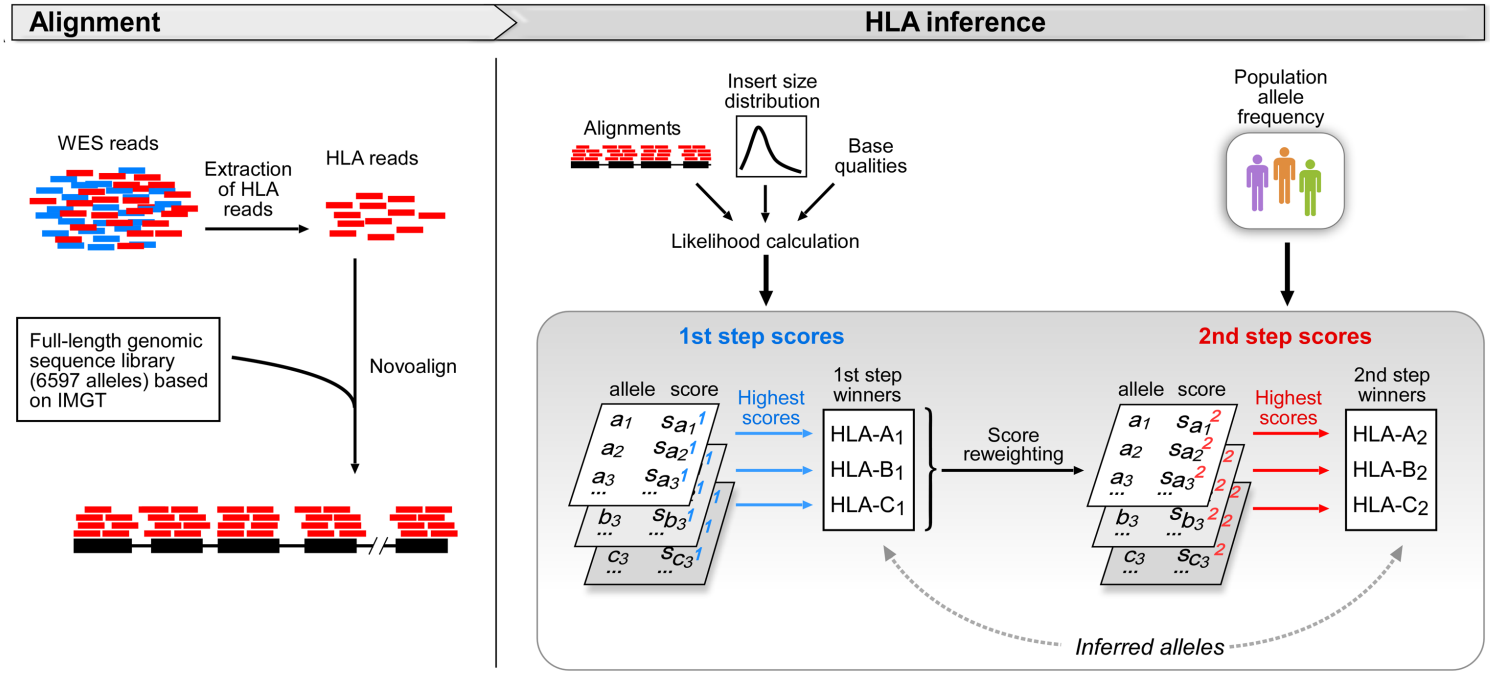


Variant calling from long reads sequencing

- NANOPORE



Polysolver: HLA typing and mutation calling from



Shukla et al *Nat Biotech* (2015)

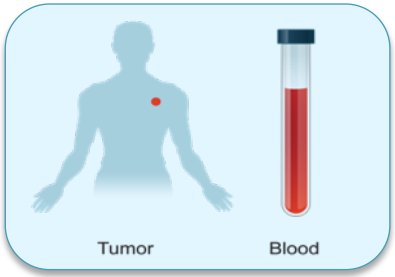
Memorial Sloan Kettering Cancer Center

Clinical cancer gene panel sequencing to identify actionable somatic alterations

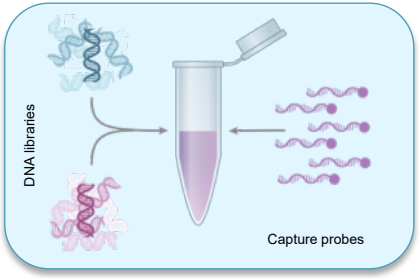
MSK-IMPACT: Comprehensive Cancer Gene Panel

*Deep coverage, (>700x high-sensitivity detection) targeted sequencing of **505 genes** to guide treatment*

1. Patient Consent

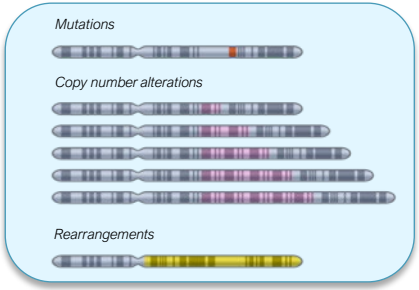


2. Sample Accessioning



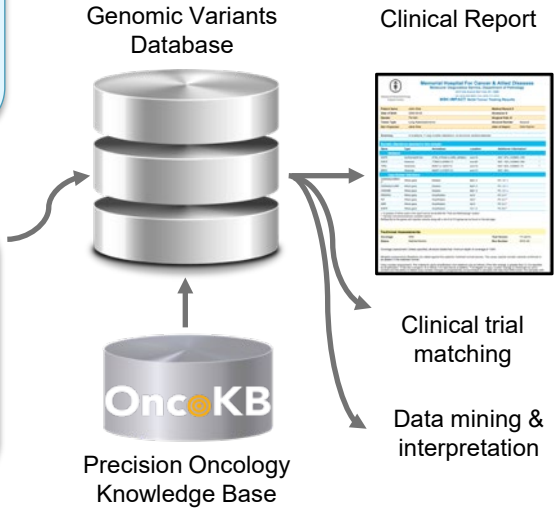
3. Sample Preparation

4. Sequencing



5. Bioinformatics Analysis

6. Case Review and Sign Out

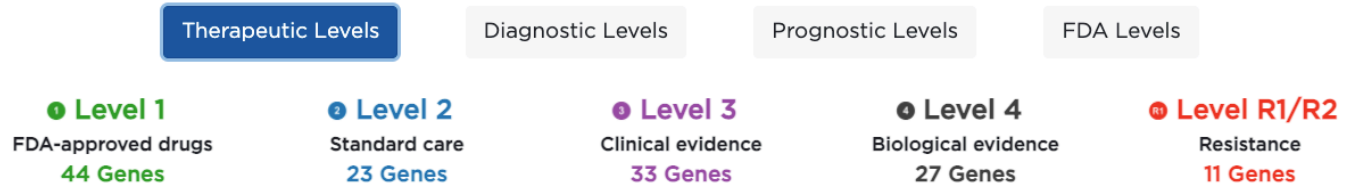


- Matched tumor/normal sequencing
  - Somatic Alterations
- Mutations, copy number alterations, gene fusions, mutational signatures, TMB, germline variants (with additional consent), clonal hematopoiesis

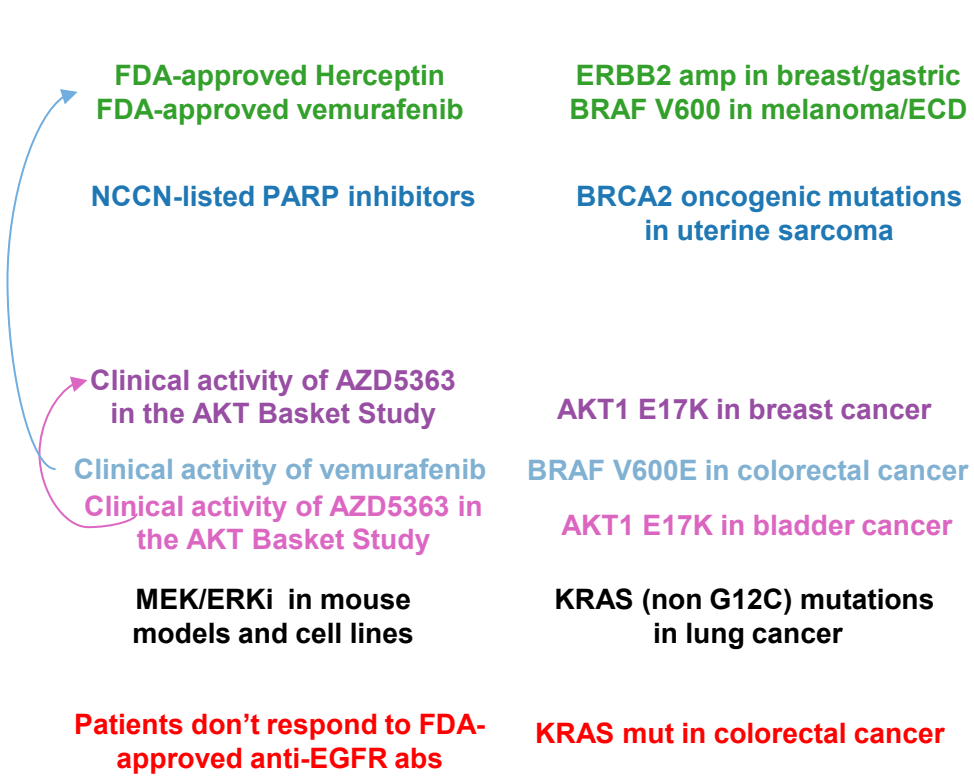
Zehir, Benayed, *et al.*, *Nature Medicine*, 2017

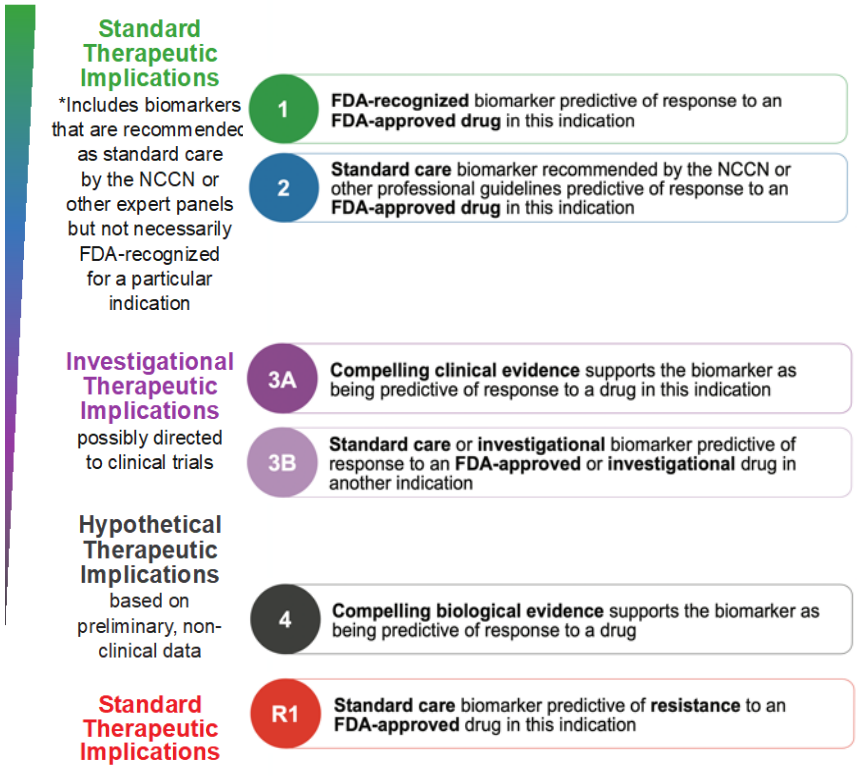
Detect clinically relevant alterations

- Run mutation, indel, SCNA and SV callers
- Calculate TMB, infer fraction of genome altered
- Using OncoKB for functional annotation



OncoKB Ranks Cancer Alterations by their Clinical Actionability





Chakravarty and Solit *Nature Review Genetics* (2021)

Memorial Sloan Kettering Cancer Center

The MSK-IMPACT cohort now includes additional NLP-derived data elements from the Cancer Data Science Initiative (Read more)

MSK Clinical Sequencing Cohort

Targeted sequencing of clinical cases via MSK-IMPACT, MSK-IMPACT-Heme, or MSK-ACCESS with clinical data from various sources, including new NLP-derived data from the Cancer Data Science Initiative (CDSI). Please follow the publication guidelines, including biostatistical review for large analyses, when using these data in abstracts or journal articles. More information and contacts here. These data are available to MSK investigators only and are not to be published or shared with anyone outside of MSK without permission. **Due to a server outage clinical data updates are not currently available. We are working on getting this functionality back.**

Click gene symbols below or enter here | Query

Summary | Clinical Data | CN Segments

Selected: 87,119 patients | 123,441 samples

Custom Selection | Charts | Groups

Cancer Type

	#	Freq
Non-Small Cell Lung Cancer	17,132	13.9%
Breast Cancer	10,155	8.2%
Colorectal Cancer	8,352	6.8%
Mature B-Cell Neoplasms	7,570	6.1%
Leukemia	6,067	4.9%
Pancreatic Cancer	5,922	4.8%
Prostate Cancer	5,497	4.5%
Bladder Cancer	4,382	3.5%
Endometrial Cancer	4,036	3.3%
Soft Tissue Sarcoma	3,896	3.2%
Esophagogastric Cancer	3,515	2.8%

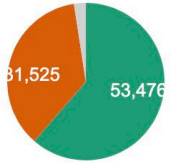
Search...

Genomic Profile Sample Counts

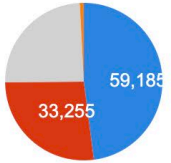
Molecular Profile	#	Freq
Mutations	123,440	99.9%
Structural Variants	122,740	99.4%
Copy Number Alterations (MSK-I…	122,739	99.4%

Search...

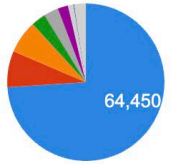
Overall Survival Status



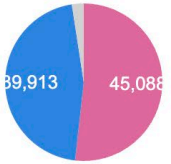
Sample Type



Race

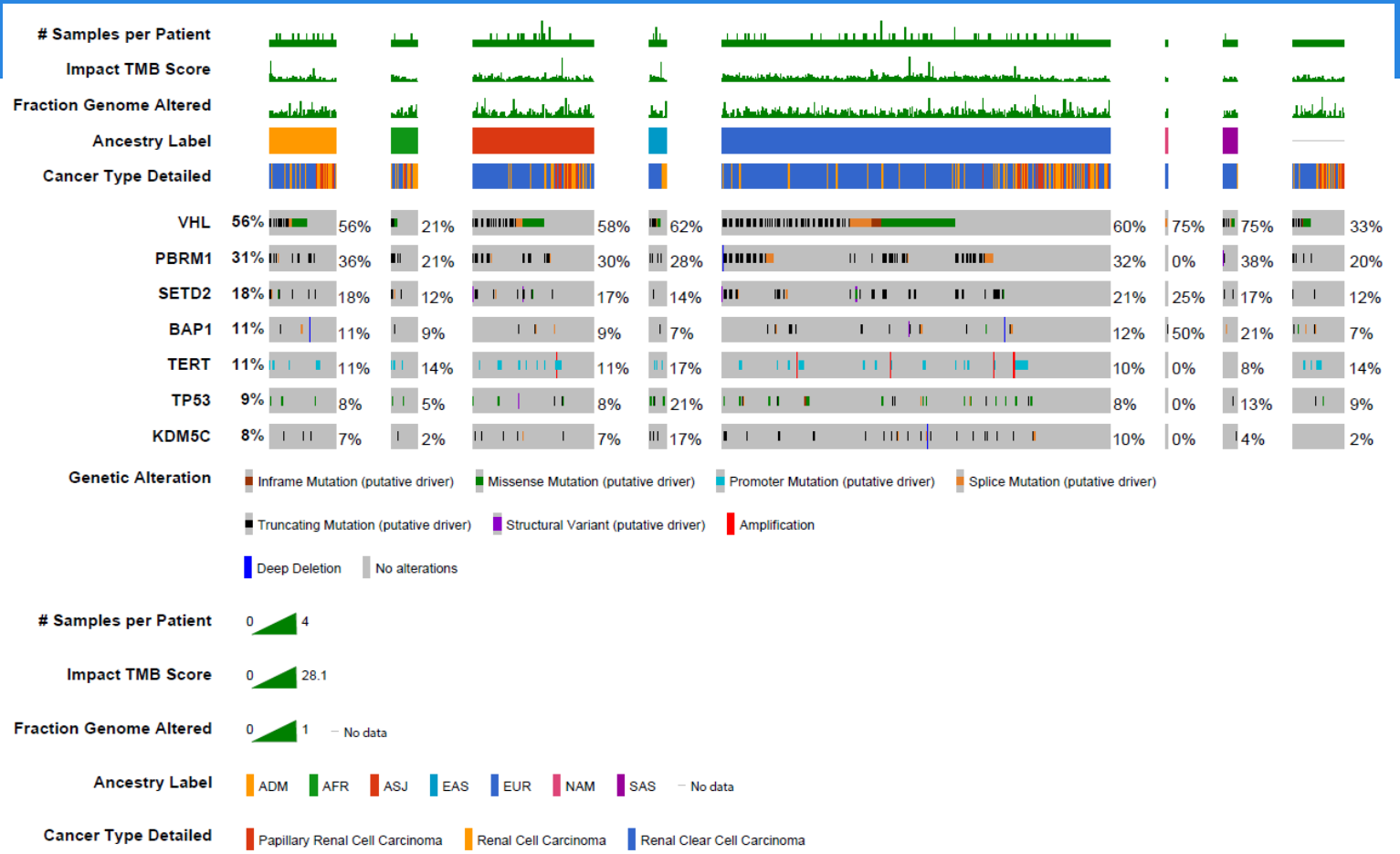


Sex



Stage (Highest Recorded) | Ethnicity | MSI Type | Gene Panel | Somatic Status | Sample Class

Memorial Sloan Kettering Cancer Center



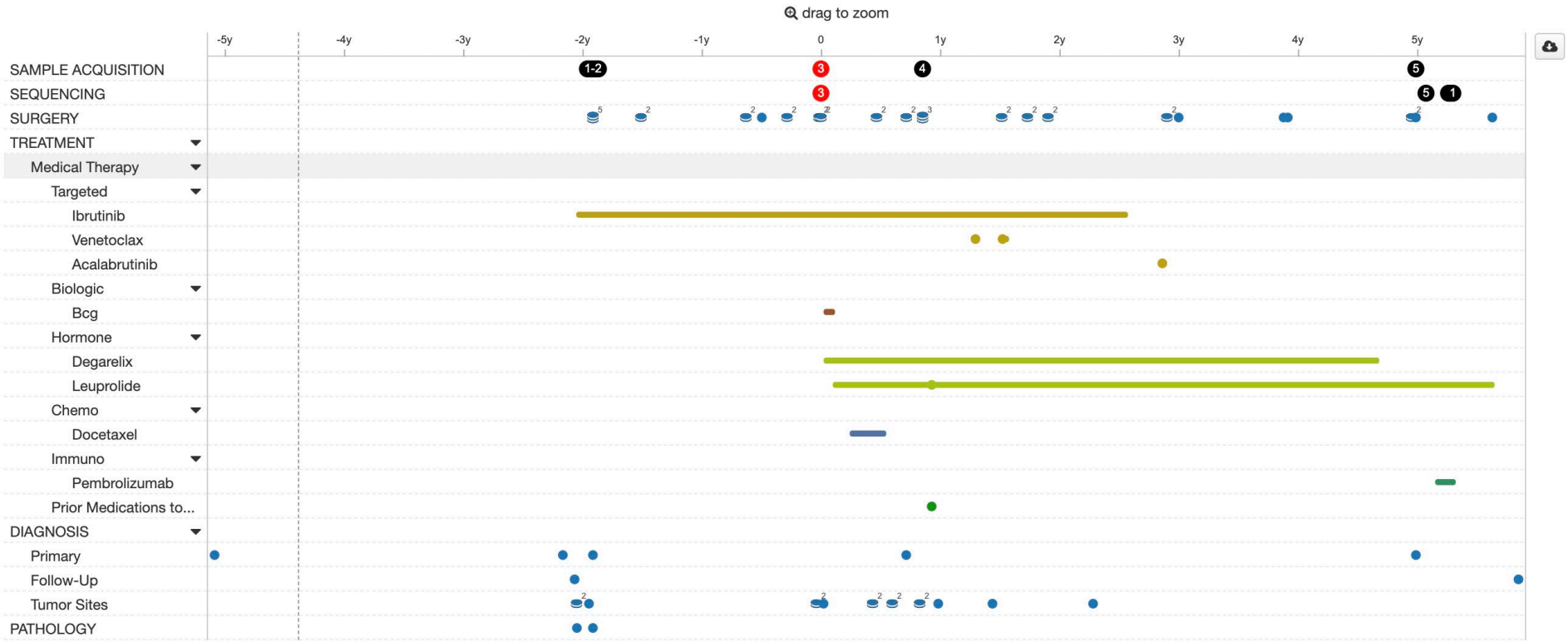
Memorial Sloan Kettering Cancer Center

Patient: P-0029248, MALE, LIVING (70 months)

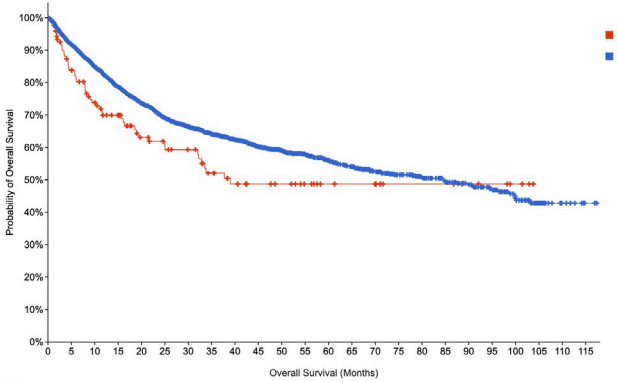
MSK Clinical Sequencing Cohort

Samples: 5 P-0029248-T02-IM7, Primary (Bladder), Bladder Cancer (Bladder Urothelial Carcinoma), MSI-H 1, TMB-H 1 Show all 5 samples

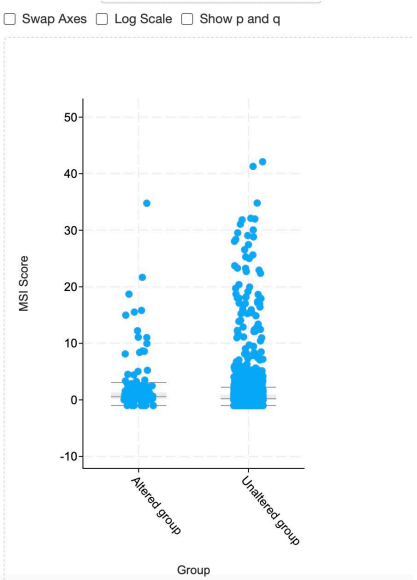
Summary | Pathways | Clinical Data



Memorial Sloan Kettering Cancer Center

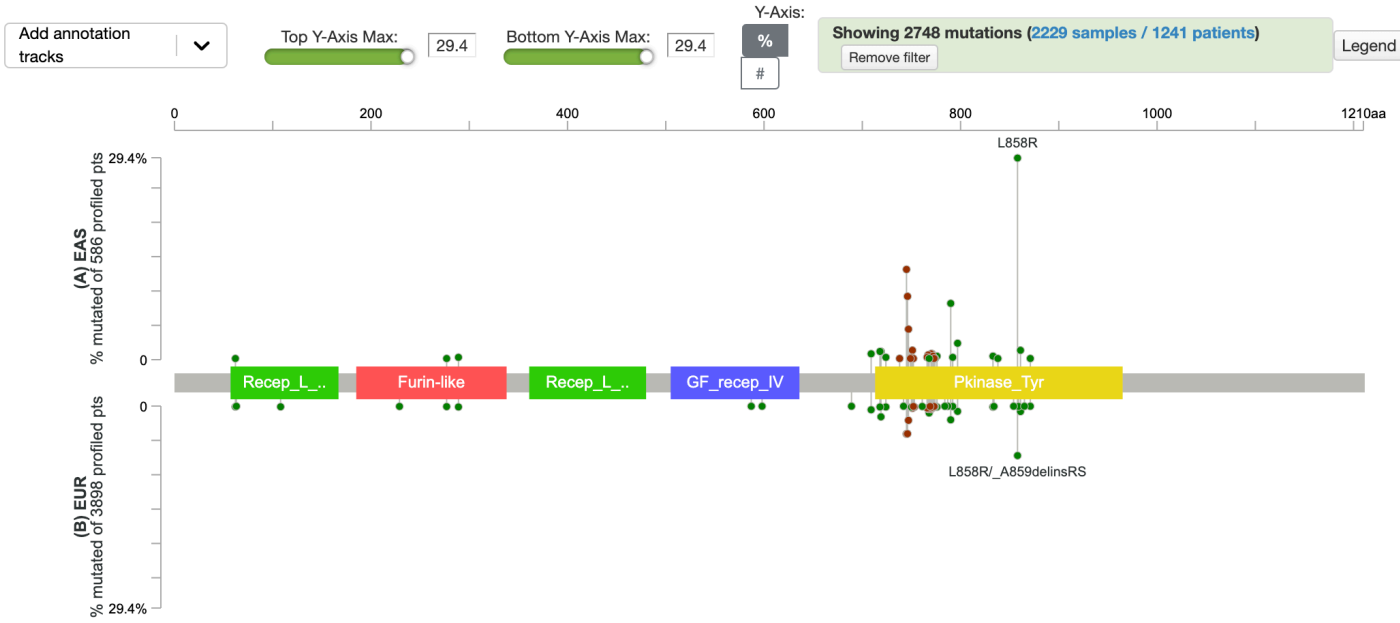


Clinical Attribute	Attribute Type	Statistical Test	p-Value	q-Value ▲
MSI Score	Sample	Wilcoxon Test	**1.72e-10**	**1.08e-8**
Mutation Count	Sample	Wilcoxon Test	**3.64e-10**	**1.15e-8**
Impact TMB Percentile (Within Tumor Type)	Sample	Wilcoxon Test	**5.19e-7**	**8.459e-6**
Impact TMB Percentile (Across All Tumor Types)	Sample	Wilcoxon Test	**5.37e-7**	**8.459e-6**
Impact TMB Score	Sample	Wilcoxon Test	**6.87e-7**	**8.660e-6**
SO comments	Sample	Chi-squared Test	**4.548e-6**	**4.775e-5**
MSI Comment	Sample	Chi-squared Test	**7.083e-6**	**6.374e-5**
Gene Panel	Sample	Chi-squared Test	**1.127e-5**	**8.874e-5**
Sample Class	Sample	Chi-squared Test	**8.396e-5**	**5.877e-4**
Tumor Site: Lymph Node (NLP)	Patient	Chi-squared Test	**1.111e-3**	**6.997e-3**
MSI Type	Sample	Chi-squared Test	**4.256e-3**	**0.0238**
Fraction Genome Altered	Sample	Wilcoxon Test	**4.868e-3**	**0.0238**
Sample Type	Sample	Chi-squared Test	**4.906e-3**	**0.0238**
Tumor Site: CNS/Brain (NLP)	Patient	Chi-squared Test	**0.0106**	**0.0478**
Institute Source	Sample	Chi-squared Test	0.0166	0.0696
Tumor Site: Liver (NLP)	Patient	Chi-squared Test	0.0264	0.0985



Memorial Sloan Kettering Cancer Center

EGFR mutations: (A) EAS vs (B) EUR



NM_005228

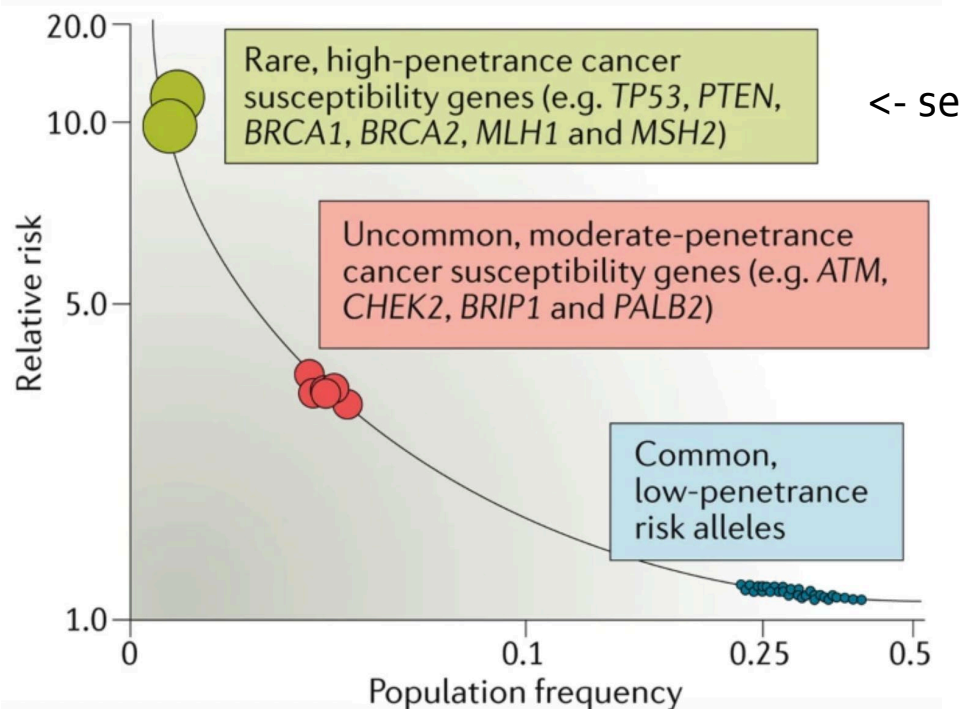
NM_005228 | ENST00000275493
CCDS5514 | EGFR_HUMAN

	Driver	VUS	
	857	40	(A) EAS
Missense	502	33	
Truncating	0	3	
Inframe	355	1	
Splice	0	3	

	Driver	VUS	
	1891	178	(B) EUR
Missense	1048	155	
Truncating	0	8	
Inframe	843	3	
Splice	0	12	

Fisher Exact Two-Sided Test p-value for filtered mutations - (A) EAS vs (B) EUR: $< 10^{-10}$

Memorial Sloan Kettering Cancer Center

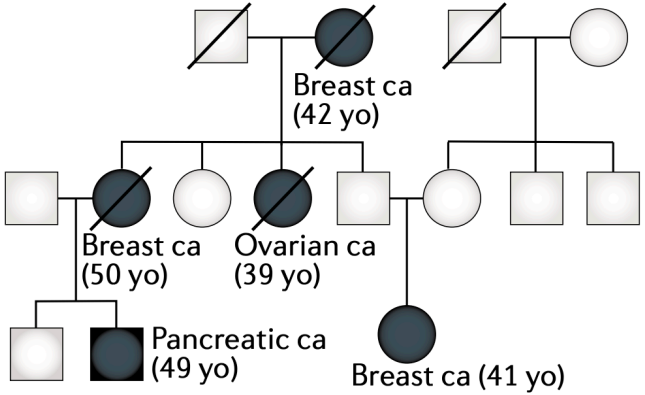


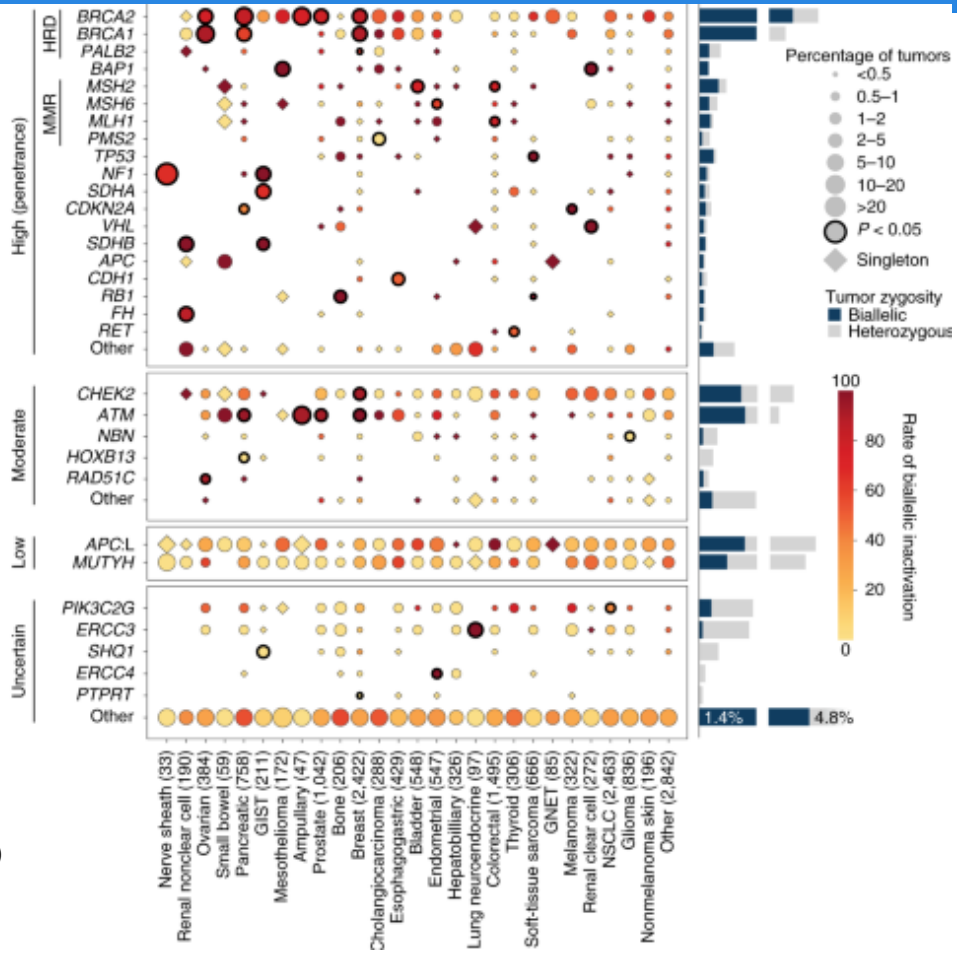
<- sequencing rare families

<- sequencing a lot of high-risk individuals

<- SNP genotyping and imputation

Leveraging matched normal sequencing to study rare germline variants



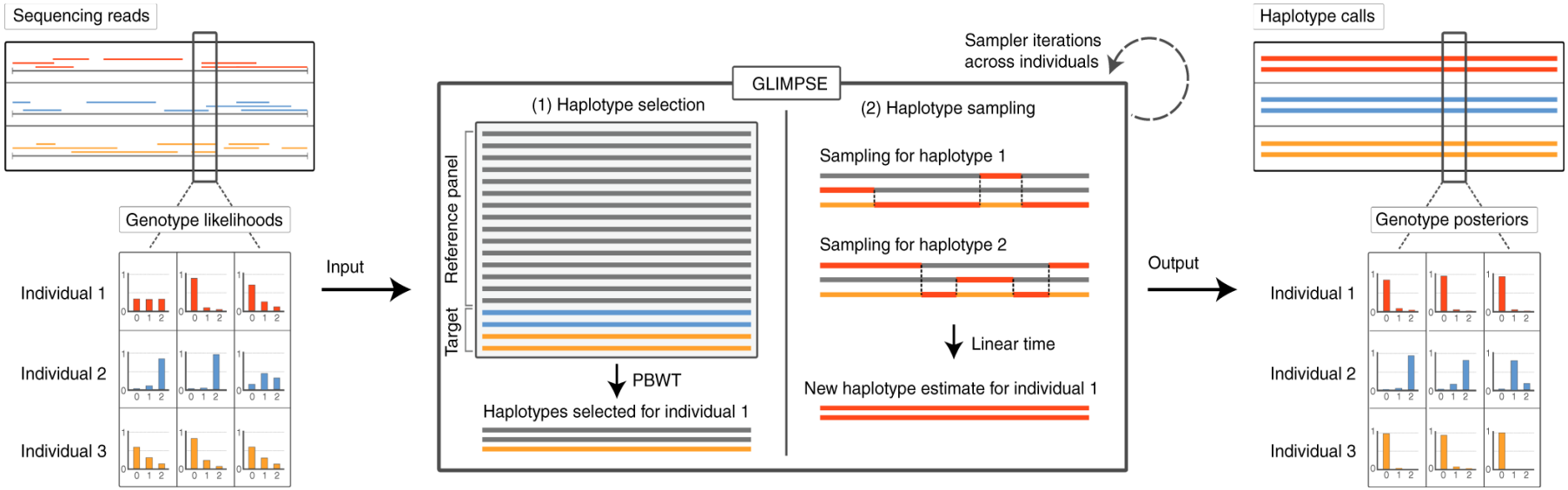


Srinivasan et al *Nature Genetics* (2021)

New methods for analyzing clinical sequencing data are needed

- Lower sensitivity in detecting mutational signatures
- Lower sensitivity in detecting rearrangements
- Limited in detecting genome-wide LOH

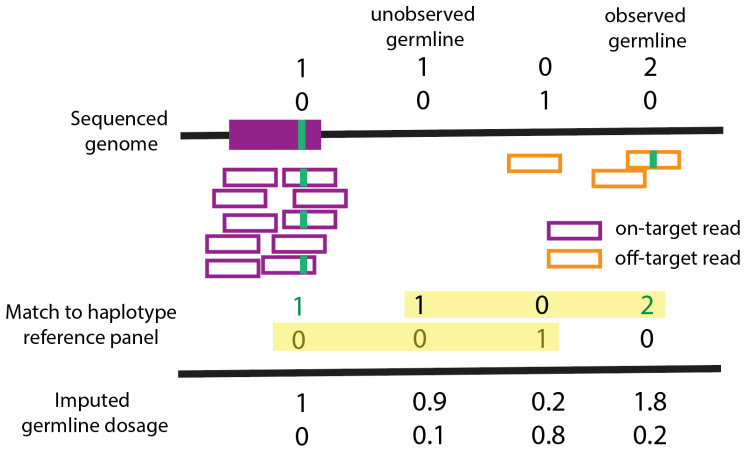
Imputation workflow from GLIMPSE



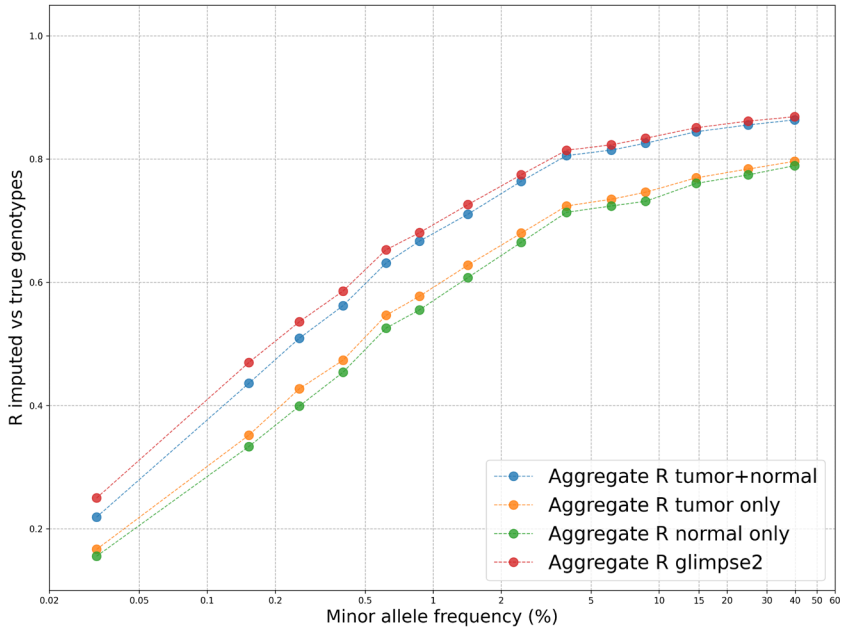
Memorial Sloan Kettering Cancer Center

Leverage off-target reads to infer germline haplotypes

❖ Imputation of common variants using on/off-target reads



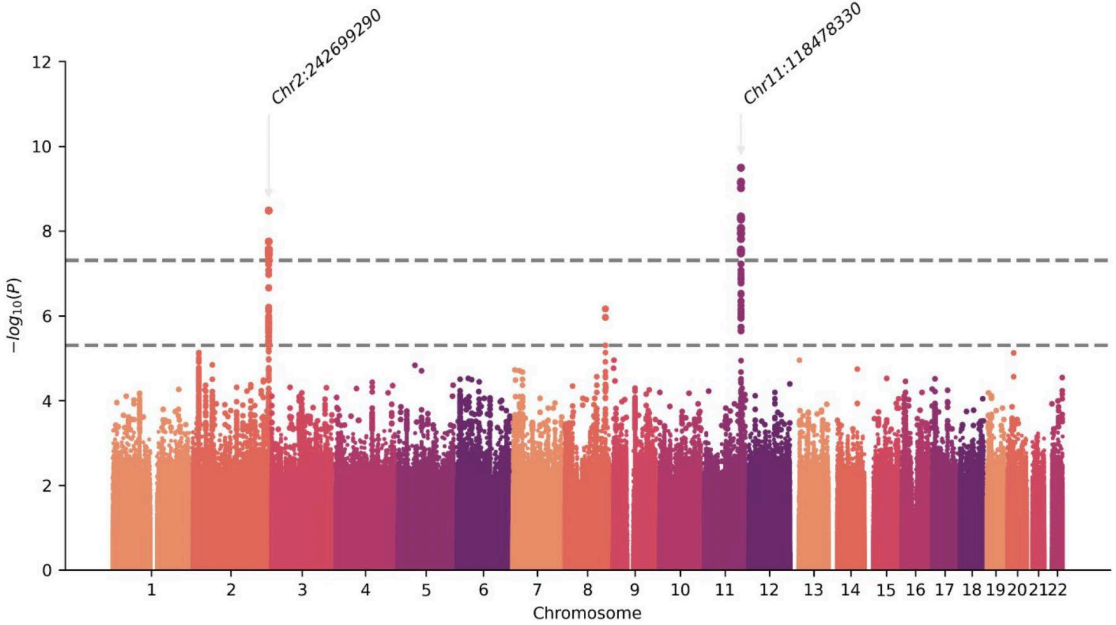
❖ Benchmarking imputation accuracy – comparing genotyping dosage of 50 paired WGS and MSK-IMPACT samples

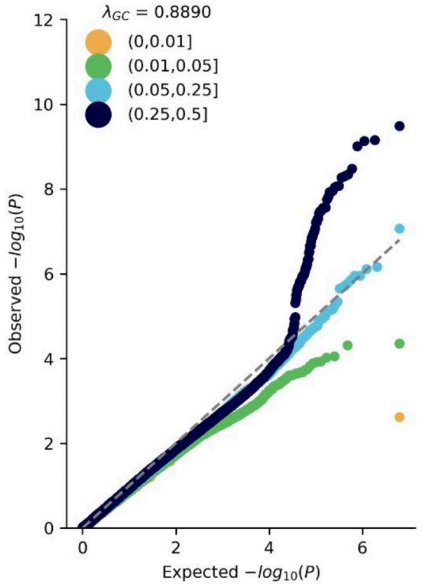


Memorial Sloan Kettering Cancer Center

“driver” GWAS using MSK-IMPACT panel sequencing

- Genome-wide association study (GWAS) of *IDH1*-mutant vs *IDH1*-wildtype GBM tumors replicated known associations

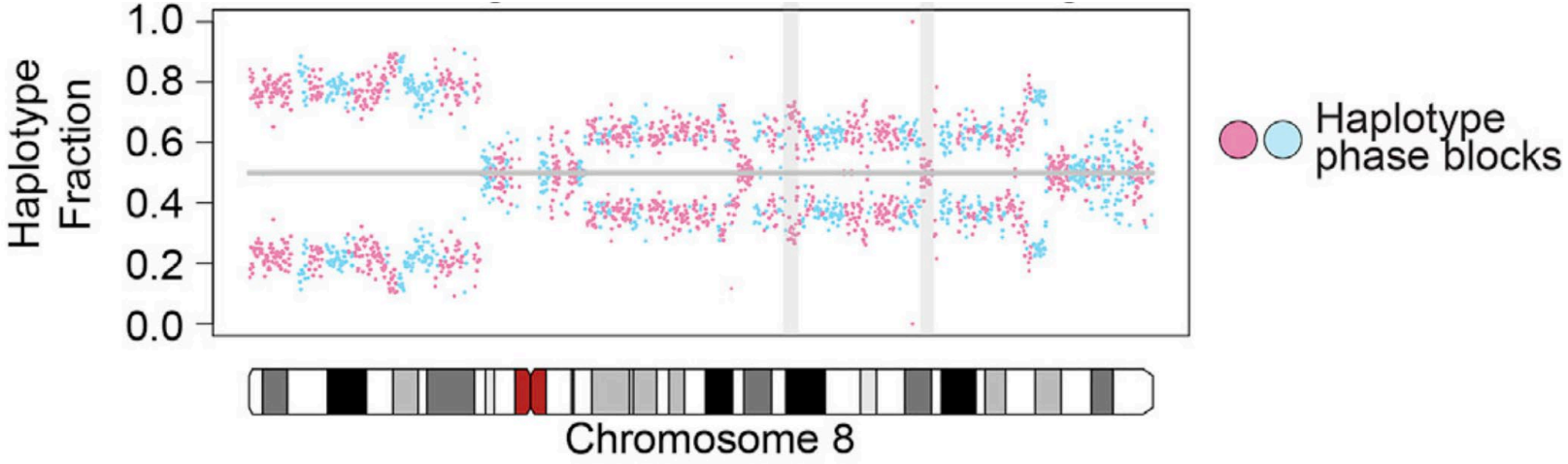




Memorial Sloan Kettering Cancer Center

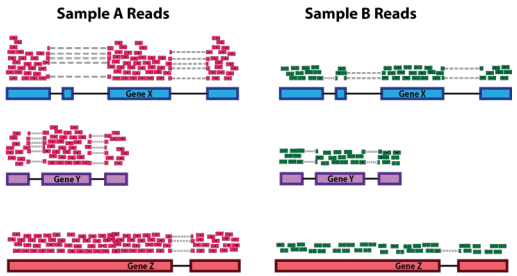
Rotation project

- Use imputed and phased variants for haplotype-based CNA on panel data



Analyzing RNA-seq data

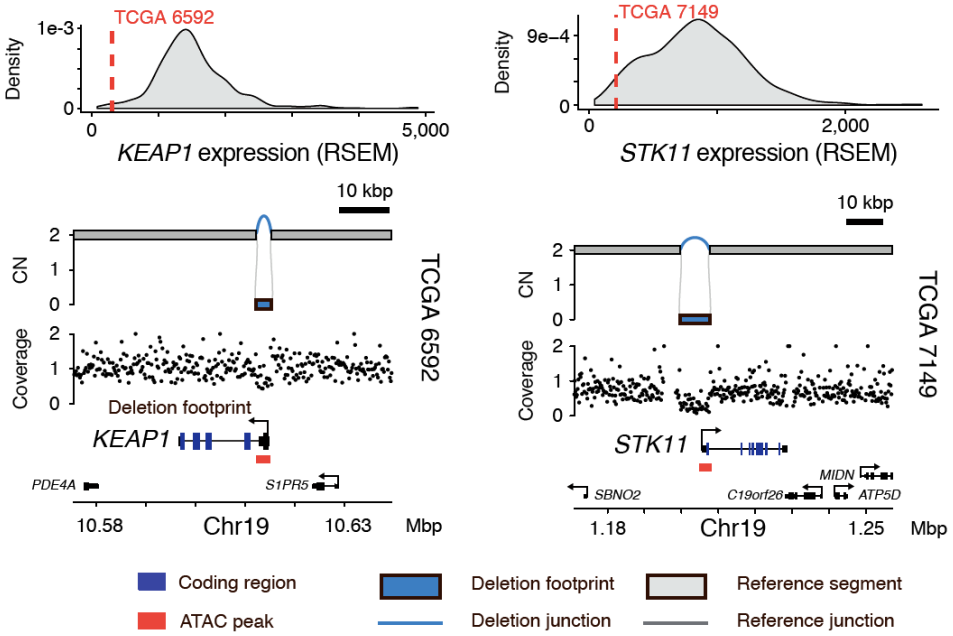
RNA-seq data normalization

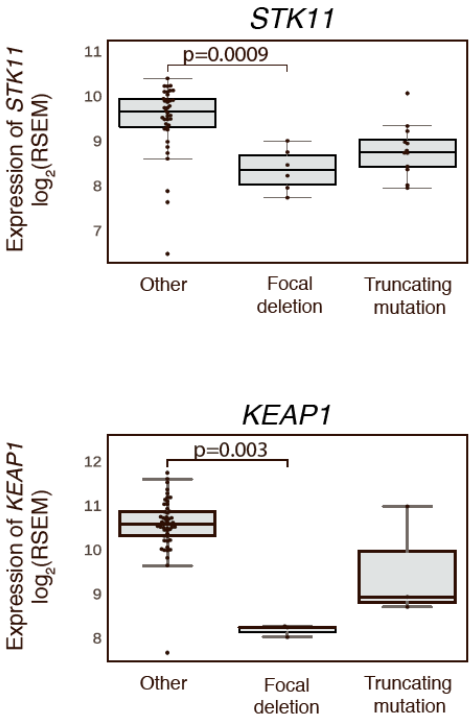


Normalization method	Description	Accounted factors	Recommendations for use
CPM (counts per million)	counts scaled by total number of reads	sequencing depth	gene count comparisons between replicates of the same samplegroup; **NOT for within sample comparisons or DE analysis**
TPM (transcripts per kilobase million)	counts per length of transcript (kb) per million reads mapped	sequencing depth and gene length	gene count comparisons within a sample or between samples of the same sample group; **NOT for DE analysis**
RPKM/FPKM (reads/fragments per kilobase of exon per million reads/fragments mapped)	similar to TPM	sequencing depth and gene length	gene count comparisons between genes within a sample; **NOT for between sample comparisons or DE analysis**

https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html

Examination of gene expression is important to understand SVs



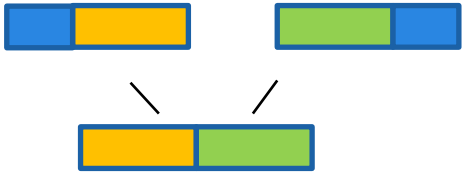


Carrot-Zhang*, Yao*, Devarakonda* et al. *Cell Reports* (2021)

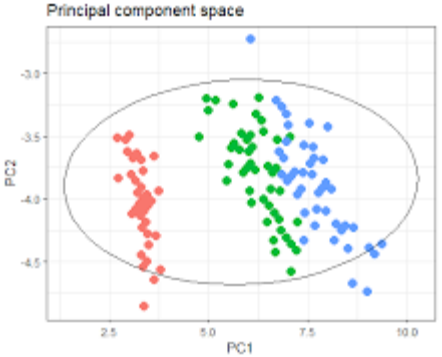
Memorial Sloan Kettering Cancer Center

Why RNA-seq is still needed

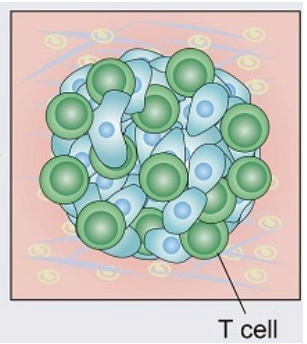
Detect gene fusion



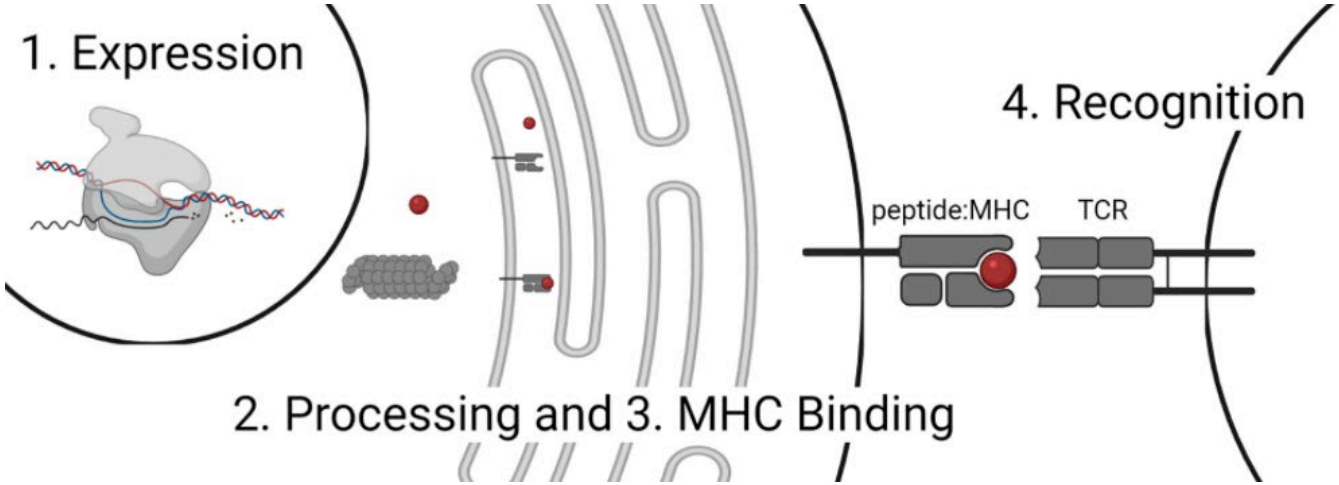
Assist histological classification



Infer tumor microenvironment



Neoantigen prediction



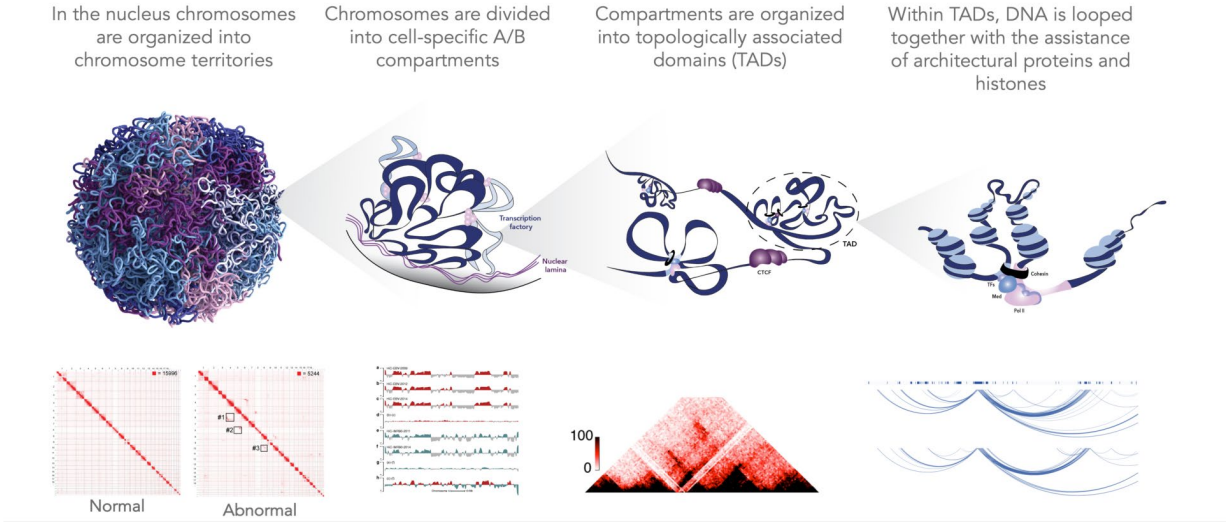
Borden et al Front Oncol (2021)

Why RNA-seq is needed for neoantigen prediction

- Infer T cell/B cell repertoire clonality
- Infer T cell/B cell receptor diversity
- Infer HLA expression

Other genomic and epigenetic profiling tools

- Single cell DNA/RNA-sequencing
- ATAC-seq
- Hi-C and Hi-Chip
  - Enhancer amplification
  - Enhancer hijacking



https://arimagenomics.com/

Resources

- 1000 Genomes Project (https://www.internationalgenome.org/)
- UK10K (https://www.uk10k.org/)
- Exome Aggregation Consortium (ExAC) (http://exac.broadinstitute.org/)
- Genome Aggregation Database (gnomAD) (https://gnomad.broadinstitute.org/)
- ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/)
- The Cancer Genome Atlas (TCGA) (https://portal.gdc.cancer.gov/)
- International Cancer Genome Consortium (ICGC) (https://icgc.org/)
- cBioPortal (cbioportal.mskcc.org)

Popular tools

- BWA - alignment
- Samtools
- MuTect2 – mutation and indel
- Strelka – mutation and indel
- ichorCNA
- Absolute – allelic-specific CNA
- Facets – allelic-specific CNA
- Svaba – SVs
- Jabba – SVs
- ANNOVAR – annotation
- OncoKB – annotation
- MutSig2CV – significantly mutated genes
- GISTIC2 – recurrent CNVs
- Deseq2 – gene expression
- FusionCatcher – gene fusion from RNA-seq
- polySolver – HLA typing
- HLALOH
- MixCR – TCR/BCR
- PyClone – tumor evolution