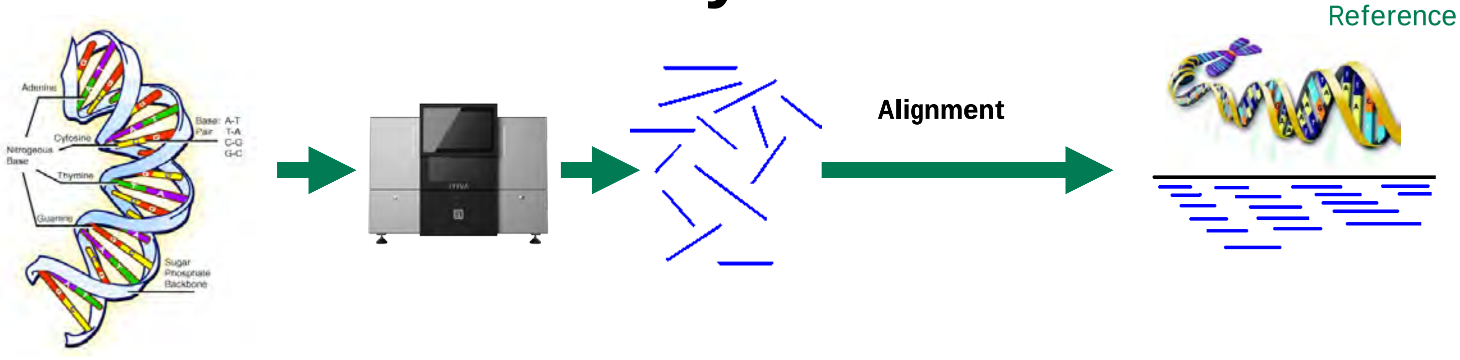
Somatic Variant Calling

Tobias Rausch

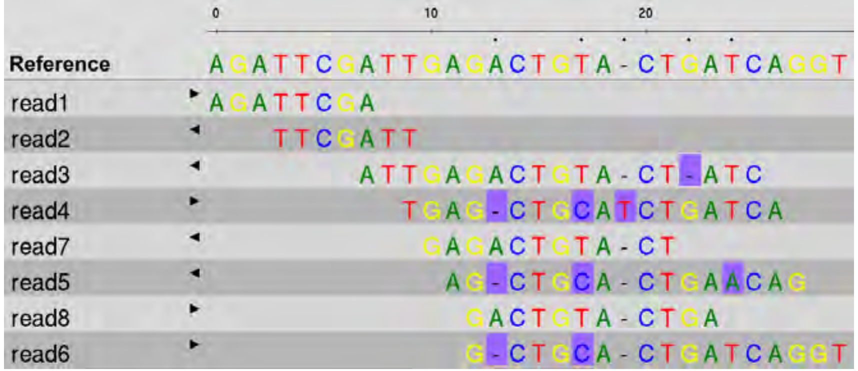
European Molecular Biology Laboratory (EMBL)

5 February 2026

Genome Variation Discovery



Alignment



```
            0         10        20
Reference   AGATTCGATTGAGACTGTA-CTGATCAGGT
read1     ▸ AGATTCGA
read2     ◂    TTCGATT
read3     ◂        ATTGAGACTGTA-CT-ATC
read4     ▸          TGAG-CTGCATCTGATCA
read7     ◂           GAGACTGTA-CT
read5     ◂            AG-CTGCA-CTGAACAG
read8     ▸             GACTGTA-CTGA
read6     ▸             G-CTGCA-CTGATCAGGT
```

Variants

CHR	POS	ID	REF	ALT	GT
chr1	12	.	GA	G	0/1
chr1	17	rs123	T	C	0/1

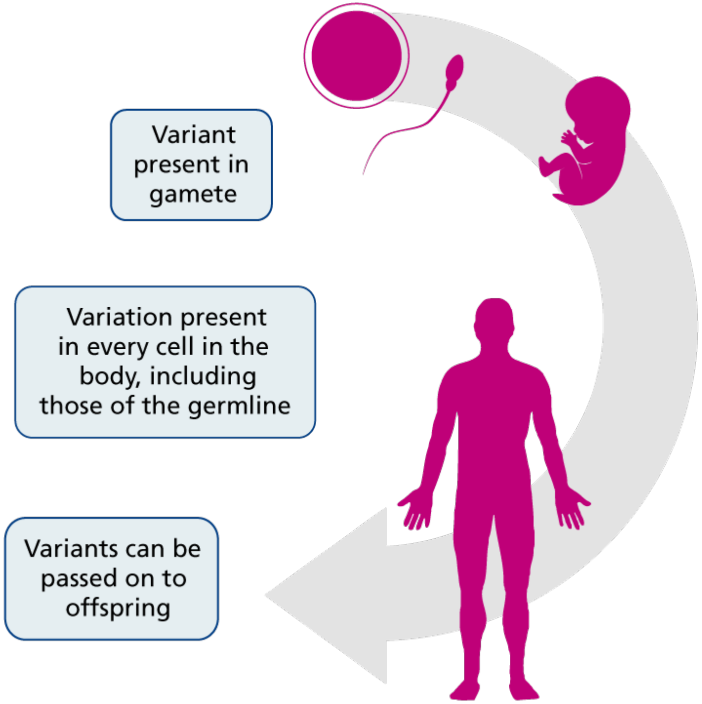
Genotype (GT):

0/0: Homozygous reference
0/1: Heterozygous
1/1: Homozygous alternative

EMBL

Germline and somatic variants

Germline Variants

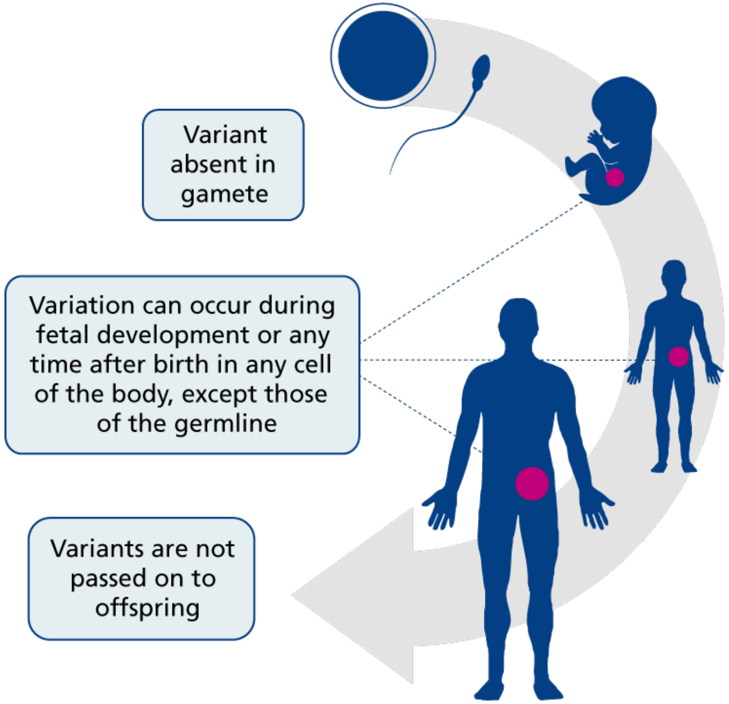


Variant present in gamete

Variation present in every cell in the body, including those of the germline

Variants can be passed on to offspring

Somatic Variants



Variant absent in gamete

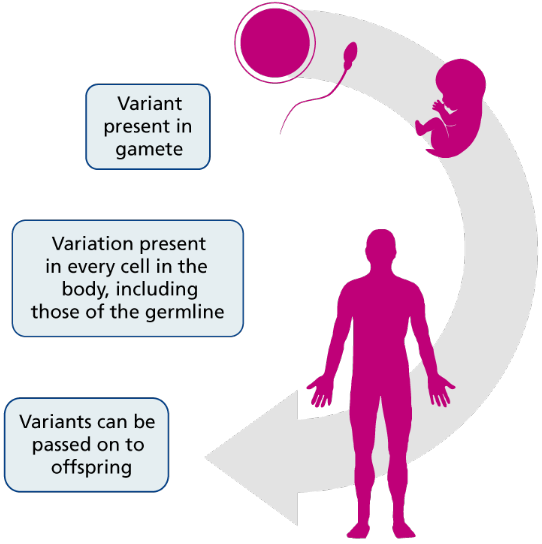
Variation can occur during fetal development or any time after birth in any cell of the body, except those of the germline

Variants are not passed on to offspring

Source: https://www.genomicseducation.hee.nhs.uk/

Germline and somatic variants

Germline Variants

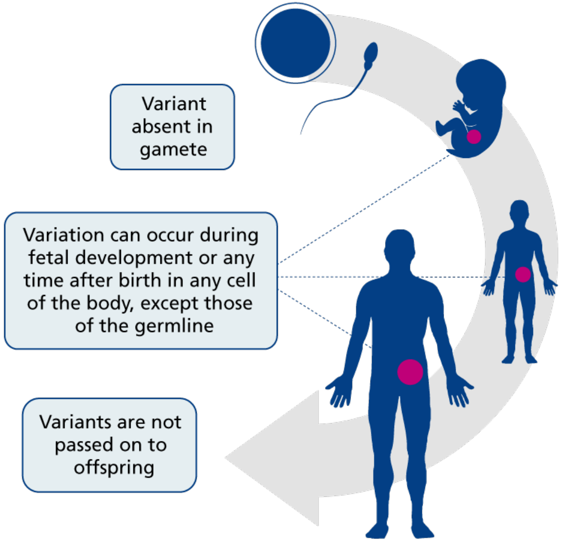


Variant present in gamete

Variation present in every cell in the body, including those of the germline

Variants can be passed on to offspring

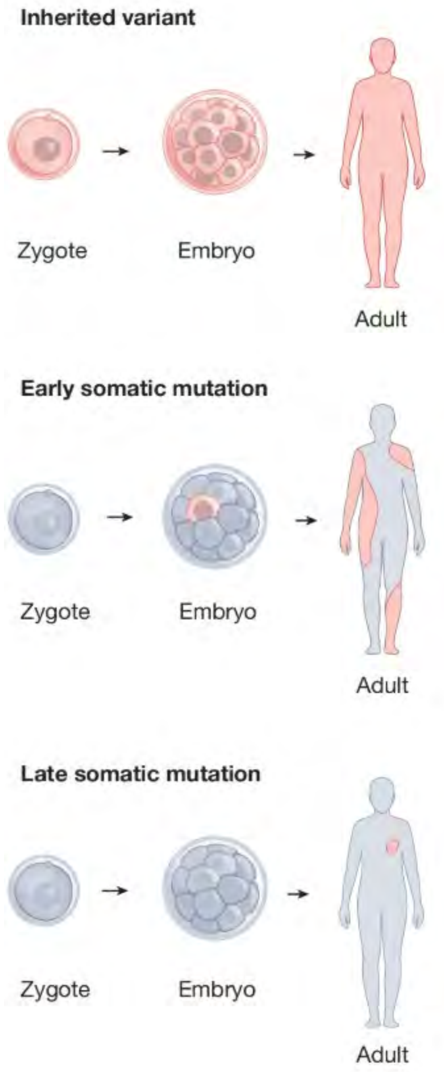
Somatic Variants



Variant absent in gamete

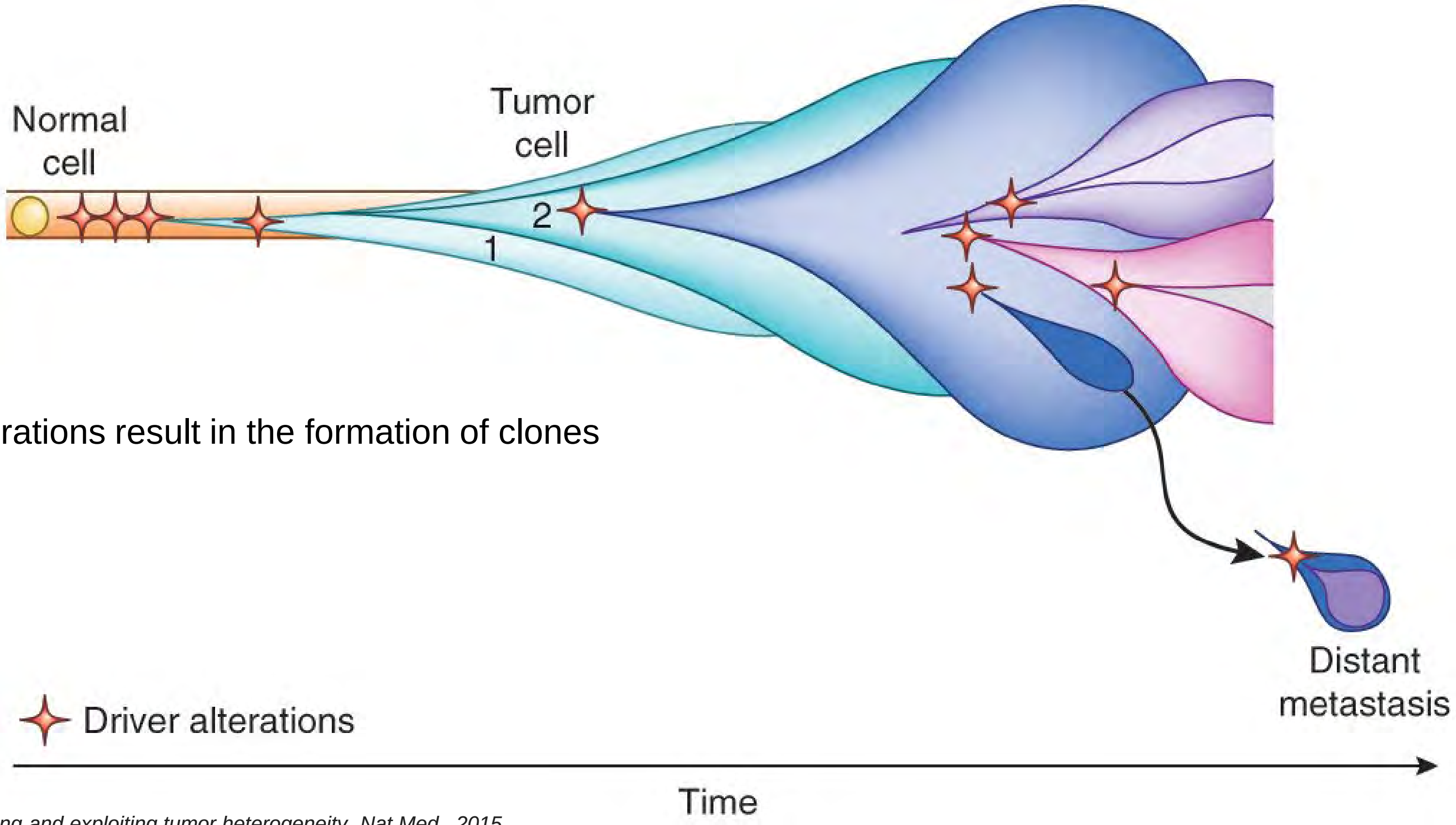
Variation can occur during fetal development or any time after birth in any cell of the body, except those of the germline

Variants are not passed on to offspring



Source: Coorens et al., The Somatic Mosaicism across Human Tissues Network. Nature. 2025

Cancer as a disease of the genome

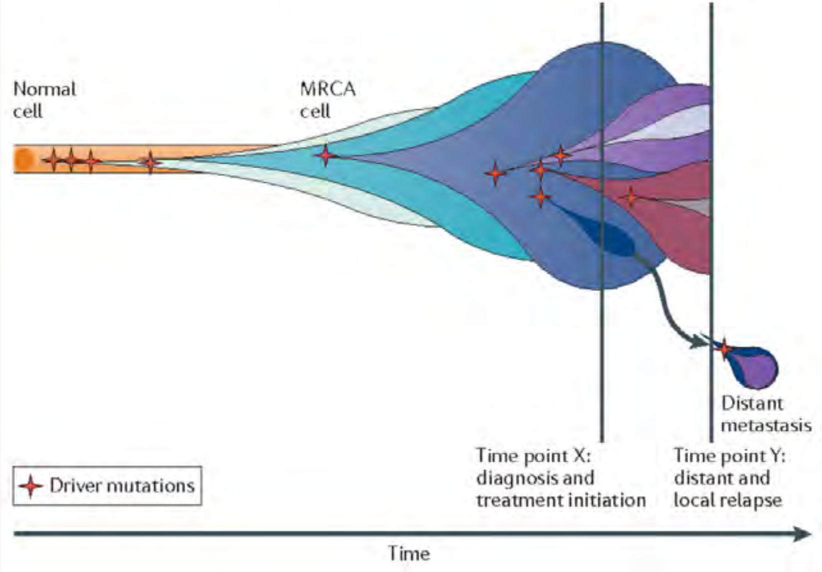


Tumor evolution
Somatic driver alterations result in the formation of clones

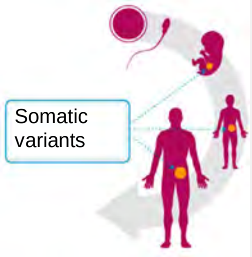
Alizadeh et al. Toward understanding and exploiting tumor heterogeneity. Nat Med., 2015.

Agenda: Somatic variants as the driver of cancer

Tumor Evolution

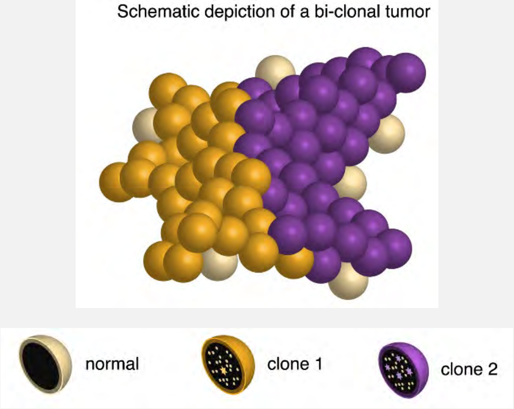

Normal cell
MRCA cell
Distant metastasis
Driver mutations
Time point X: diagnosis and treatment initiation
Time point Y: distant and local relapse
Time

Germline and somatic variants

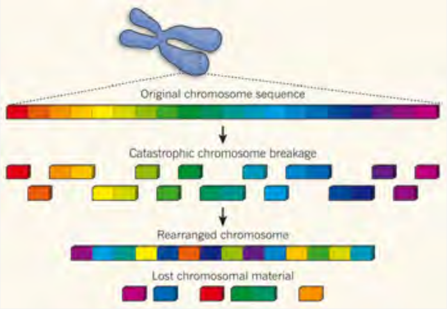


- Per genome
  - ~3-5 million germline variants
  - 100-100,000 somatic variants

Tumor heterogeneity



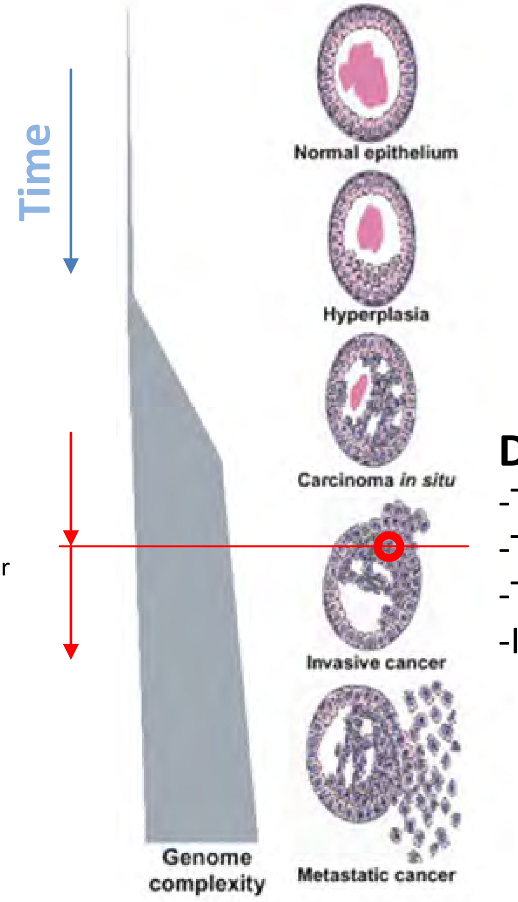
Complex Rearrangements


Original chromosome sequence
Catastrophic chromosome breakage
Rearranged chromosome
Lost chromosomal material

Sources:
Genomic landscape of non-small cell lung cancer in smokers and never-smokers. Govindan et al., Cell. 2012 Sep 14;150(6):1121-34.
Evolution of the cancer genome. Yates and Campbell, Nat Rev Genet. 2012 Nov;13(11):795-806.
Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. Rausch et al., Cell. 2012 Jan 20;148(1-2):59-71.

Cancer Genomics

Sequencing provides a snapshot in time and space



Tumor Phylogeny & Tumor Evolution

DNA-Sequencing

-Tumor cell content (tumor purity)

-Tumor heterogeneity (subclonality)

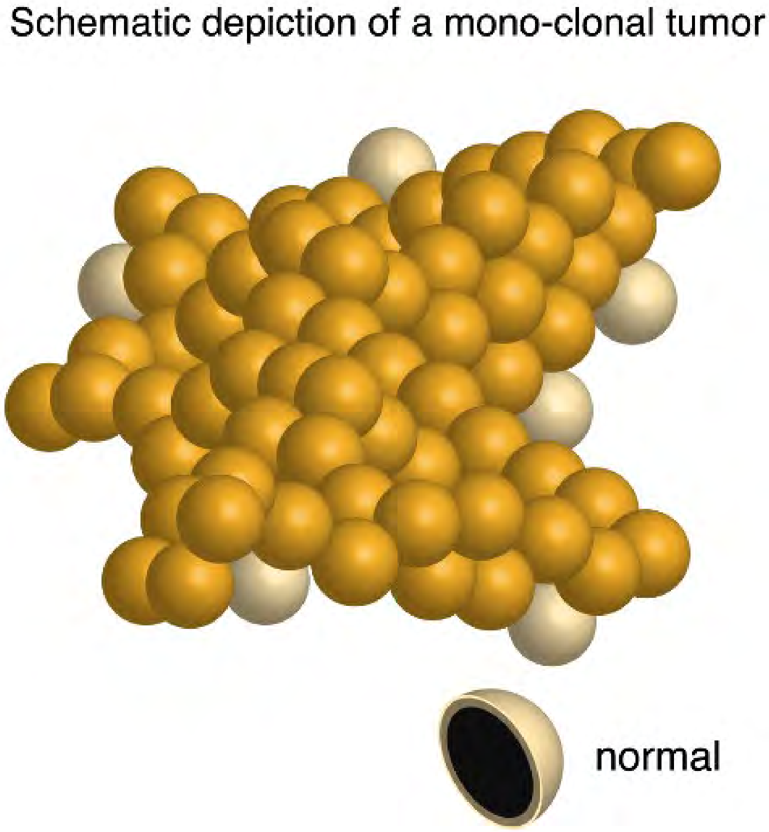
-Tumor ploidy

-Interplay of somatic & germline mutations

Source: Chromosome aberrations in solid tumors. Albertson et al., Nat Genet. 2003 Aug;34(4):369-76.

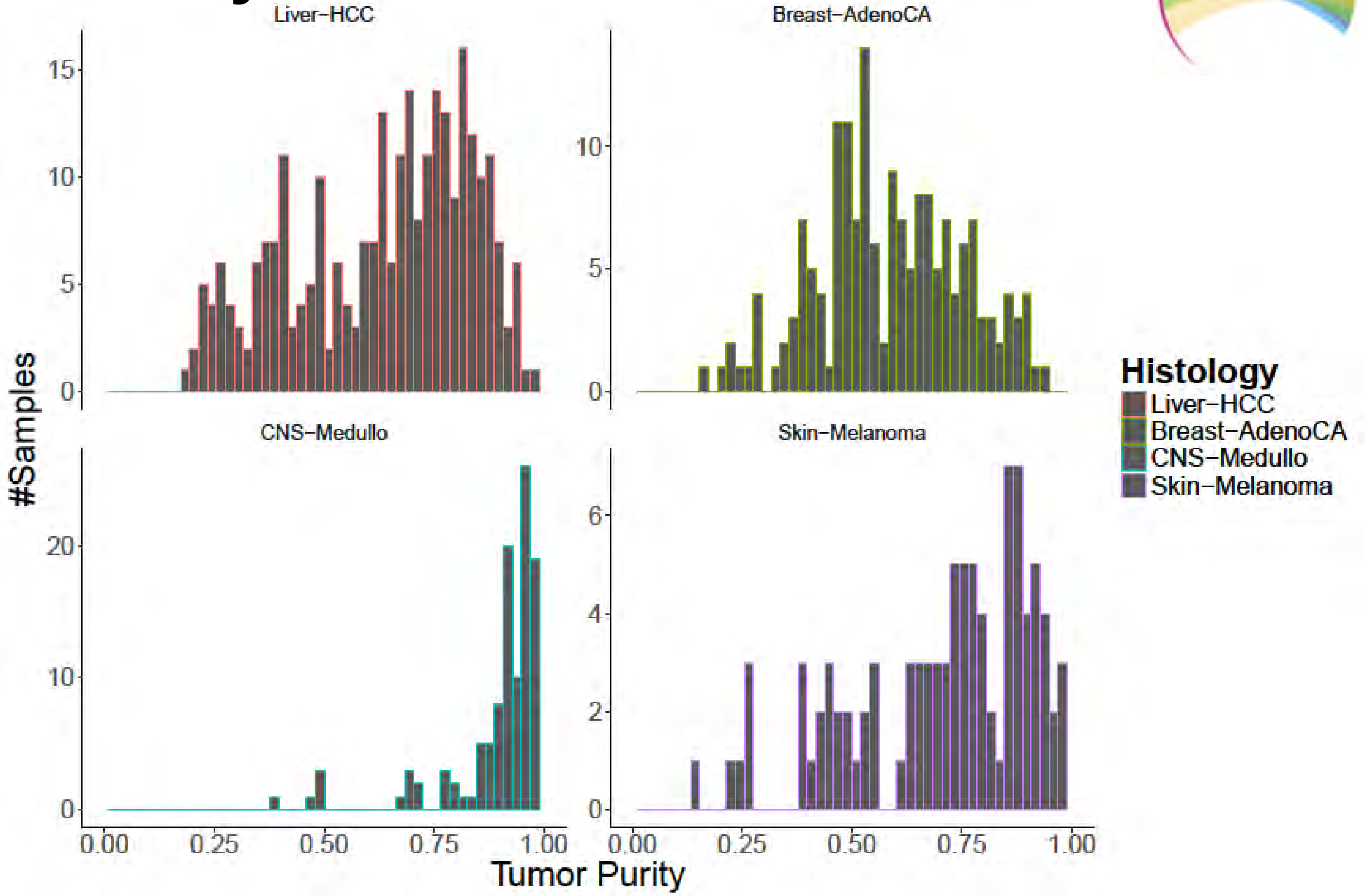
EMBL

Tumor Purity



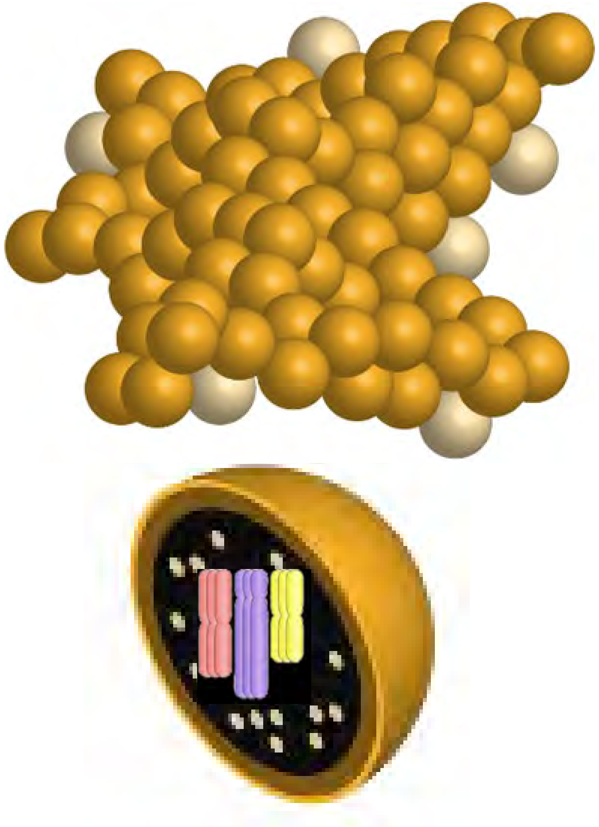
Source: Genomic landscape of non-small cell lung cancer in smokers and never-smokers. Govindan et al., Cell. 2012 Sep 14;150(6):1121-34.

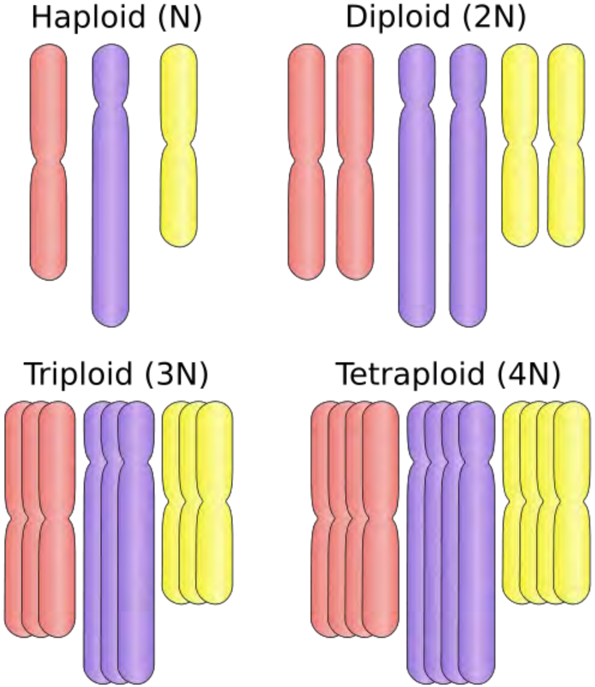
Tumor Purity



Source: Pan-cancer analysis of whole genomes. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Nature volume 578, pages82–93 (2020)

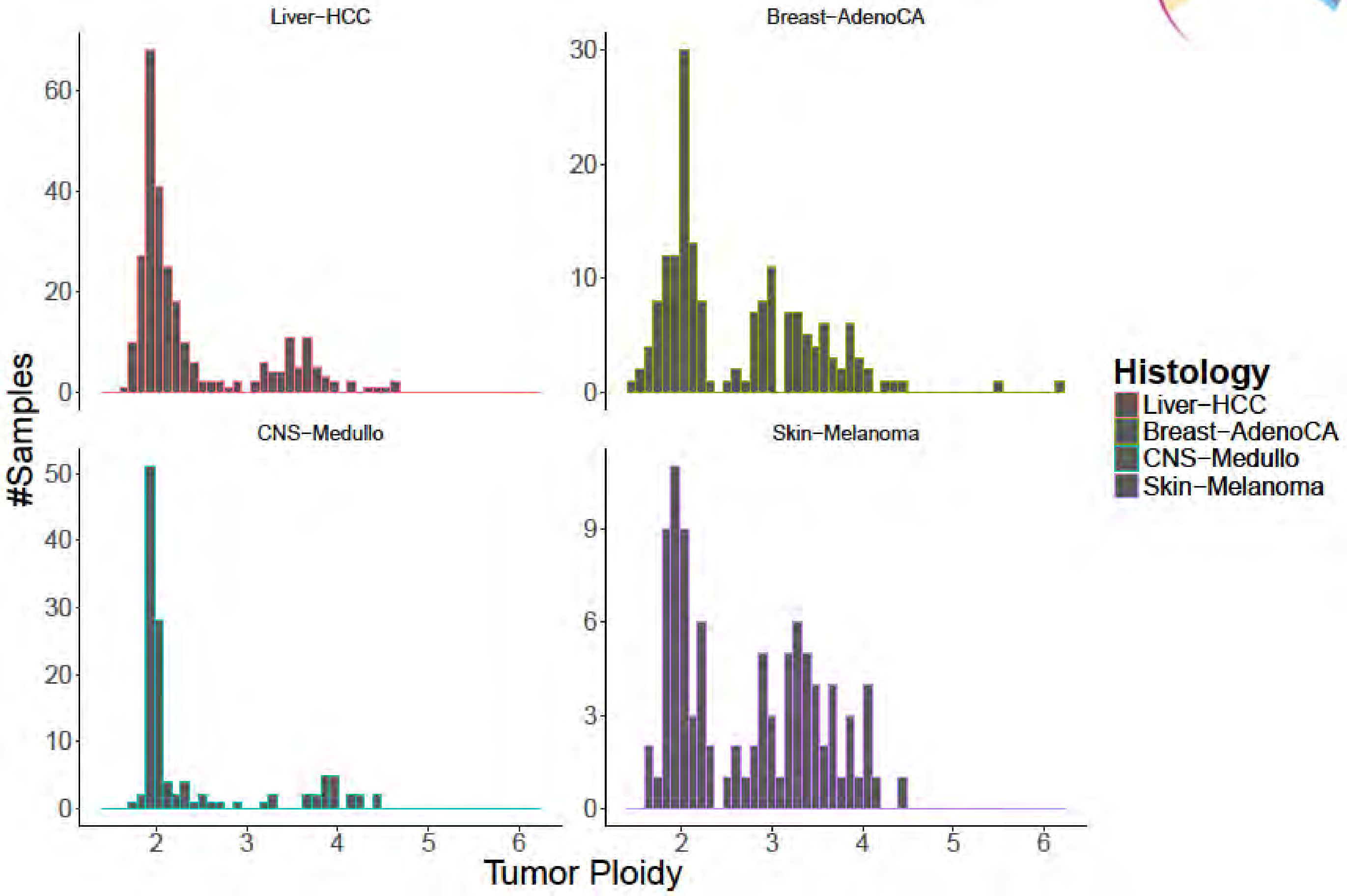
Tumor Ploidy





EMBL

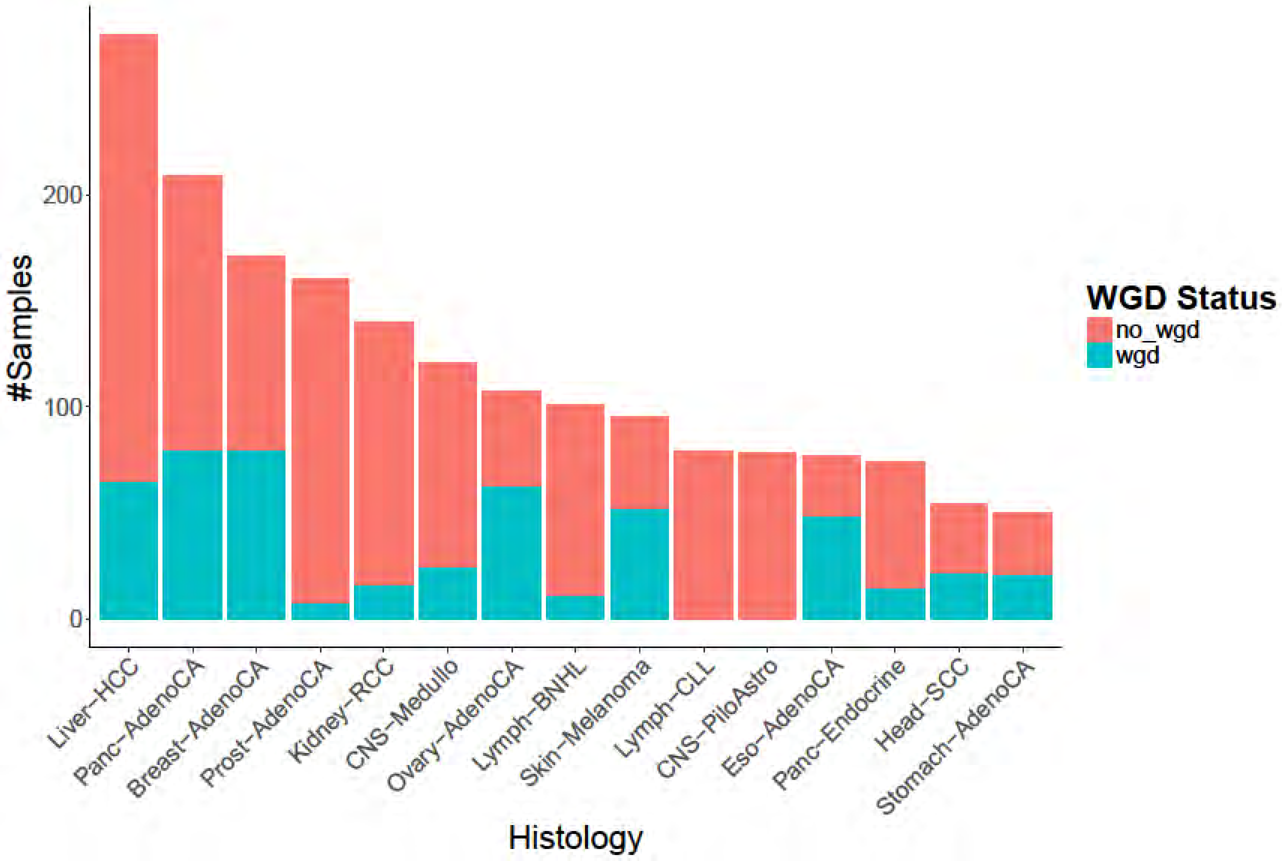
Tumor Ploidy



PCAWG
PanCancer Analysis of WHOLE GENOMES

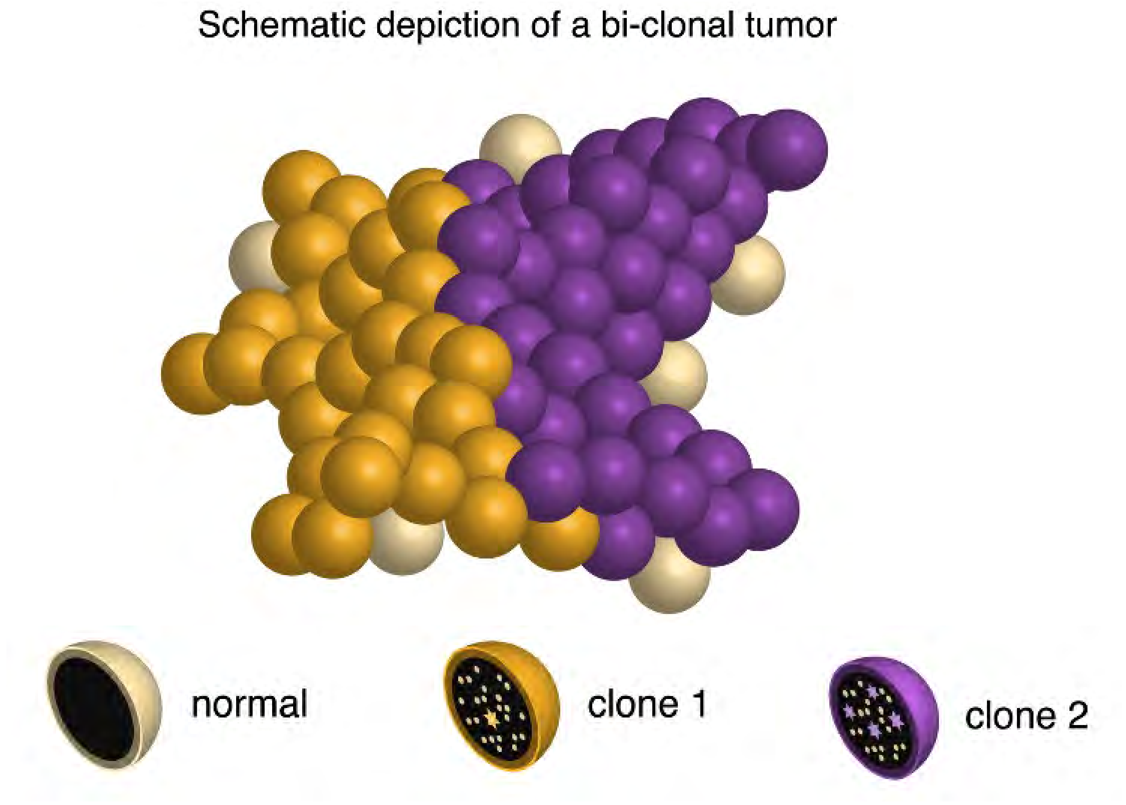
EMBL

Inferred Whole-Genome Duplication (WGD) Status



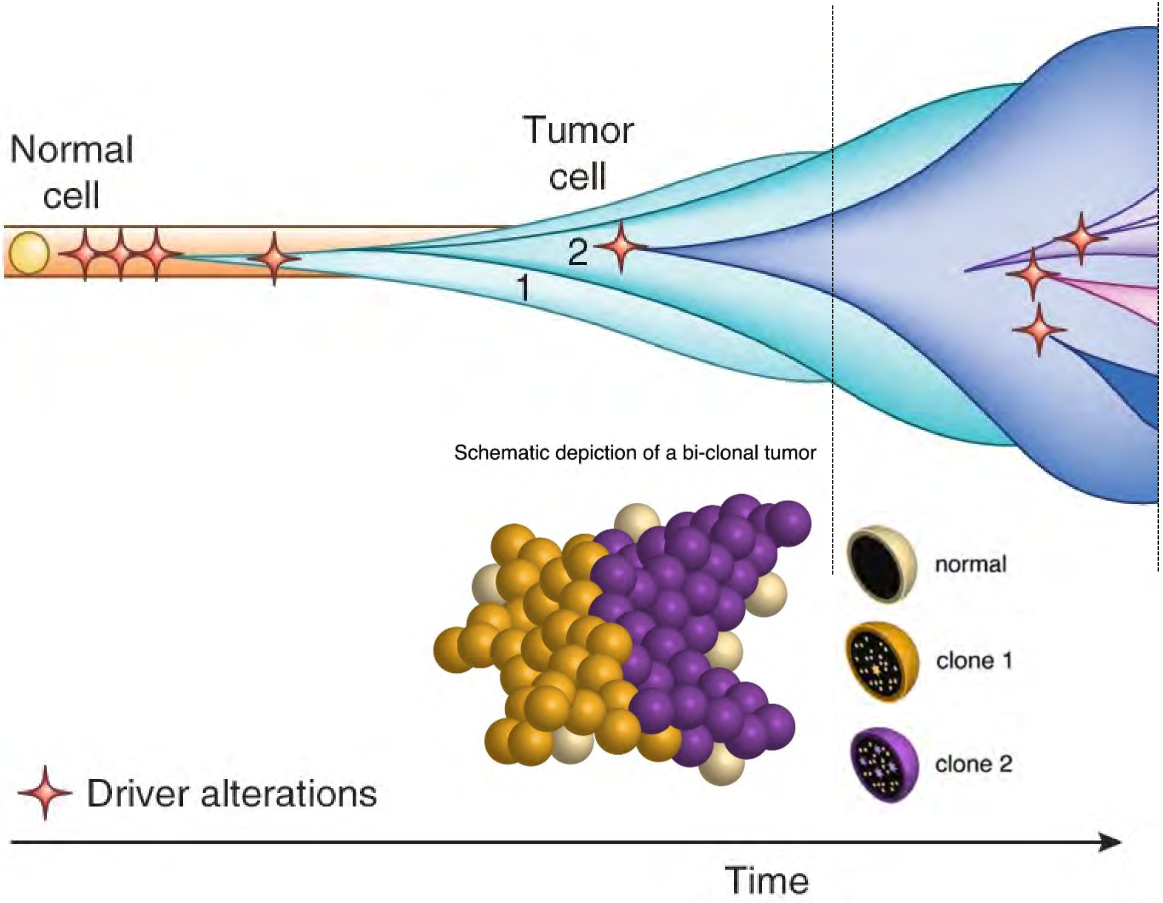
EMBL

Tumor Clonality/Heterogeneity



Source: Genomic landscape of non-small cell lung cancer in smokers and never-smokers. Govindan et al., Cell. 2012 Sep 14;150(6):1121-34.

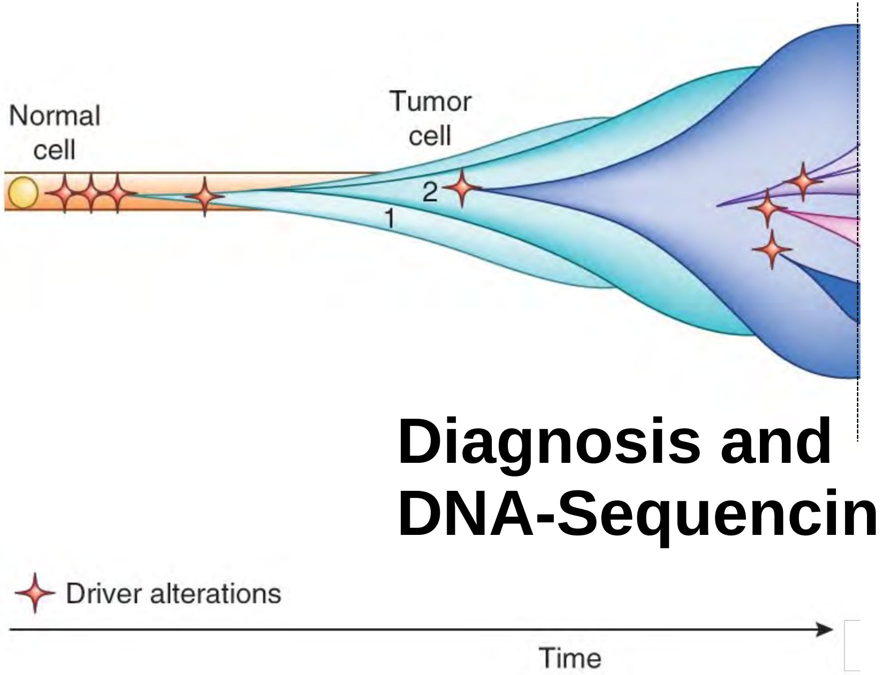
Tumor multi-clonality



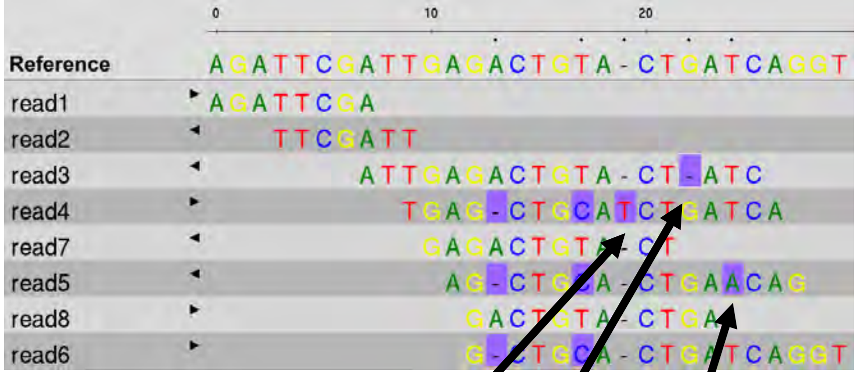
Diagnosis and DNA-Sequencing

Source: *Alizadeh, A., Aranda, V., Bardelli, A. et al. Toward understanding and exploiting tumor heterogeneity. Nat Med 21, 846–853 (2015).*

Tumor multi-clonality



Diagnosis and DNA-Sequencing

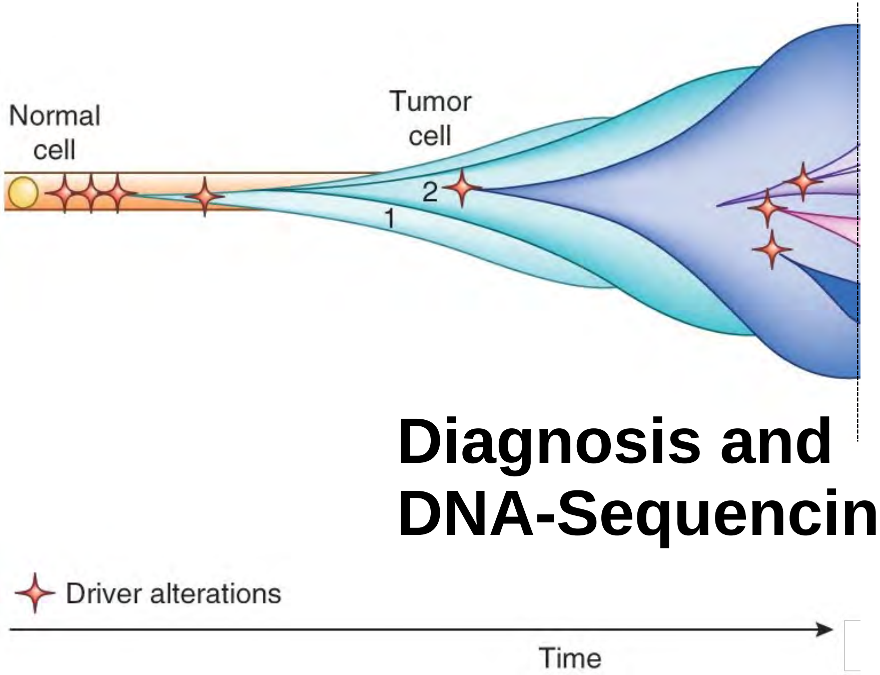


```
                0         10        20
Reference    AGATTCGATTGAGACTGTA-CTGATCAGGT
read1     ▶  AGATTCGA
read2     ◀     TTCGATT
read3     ◀         ATTGAGACTGTA-CT-ATC
read4     ▶           TGAG-CTGCATCTGATCA
read7     ◀            GAGACTGTA-CT
read5     ◀             AG-CTGCA-CTGAACAG
read8     ▶              GACTGTA-CTGA
read6     ▶              G-CTGCA-CTGATCAGGT
```

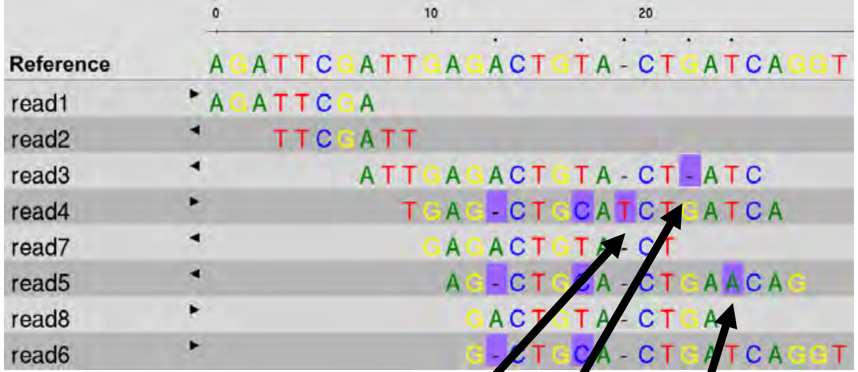
Sequencing errors?

Source: *Alizadeh, A., Aranda, V., Bardelli, A. et al. Toward understanding and exploiting tumor heterogeneity. Nat Med 21, 846–853 (2015).*

Tumor multi-clonality



Diagnosis and DNA-Sequencing



```
               0         10        20
Reference      AGATTCGATTGAGACTGTA-CTGATCAGGT
read1       ►  AGATTCGA
read2       ◄     TTCGATT
read3       ◄         ATTGAGACTGTA-CT-ATC
read4       ►           TGAG-CTGCATCTGATCA
read7       ◄            GAGACTGTA-CT
read5       ◄             AG-CTGCA-CTGAACAG
read8       ►              GACTGTA-CTGA
read6       ►              G-CTGCA-CTGATCAGGT
```

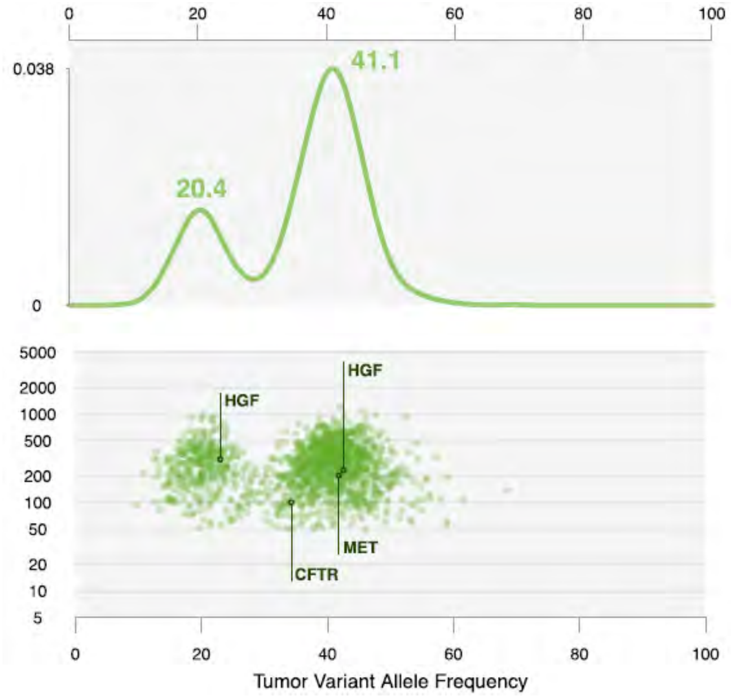
Sequencing errors?

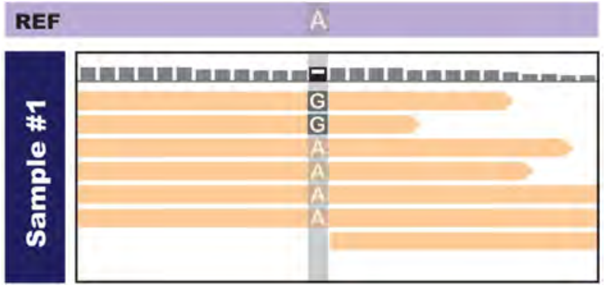
→ Cancer genomes are often sequenced to >60x because of tumor purity, tumor heterogeneity and many chromosomal aberrations

Source: Alizadeh, A., Aranda, V., Bardelli, A. et al. Toward understanding and exploiting tumor heterogeneity. Nat Med 21, 846–853 (2015).

Tumor Clonality/Heterogeneity

- Requires high sequencing depth
  - Typically done with WES data (≥500x coverage)





Variant Allele Frequency = 2/6

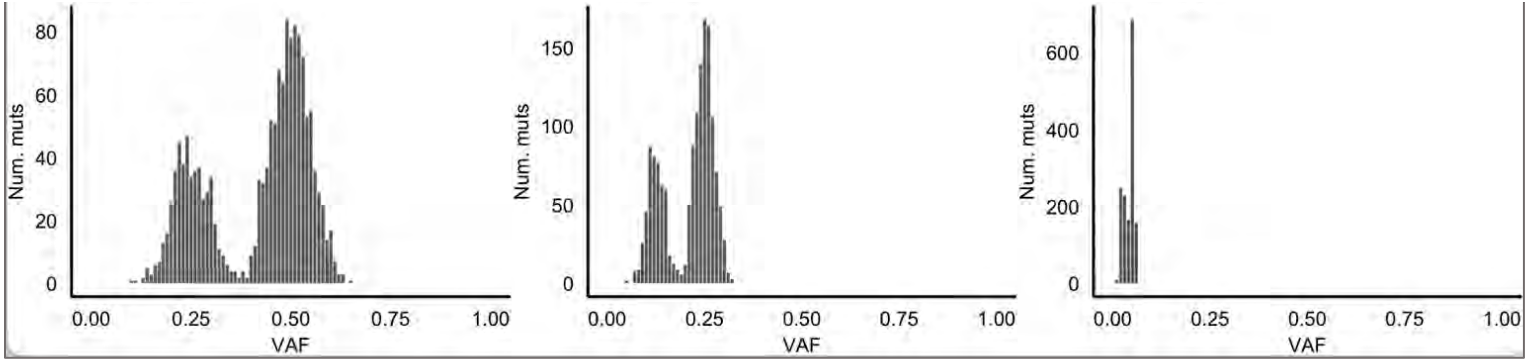
Bi-clonal

82.2% tumor purity

Source: Genomic landscape of non-small cell lung cancer in smokers and never-smokers. Govindan et al., Cell. 2012 Sep 14;150(6):1121-34.

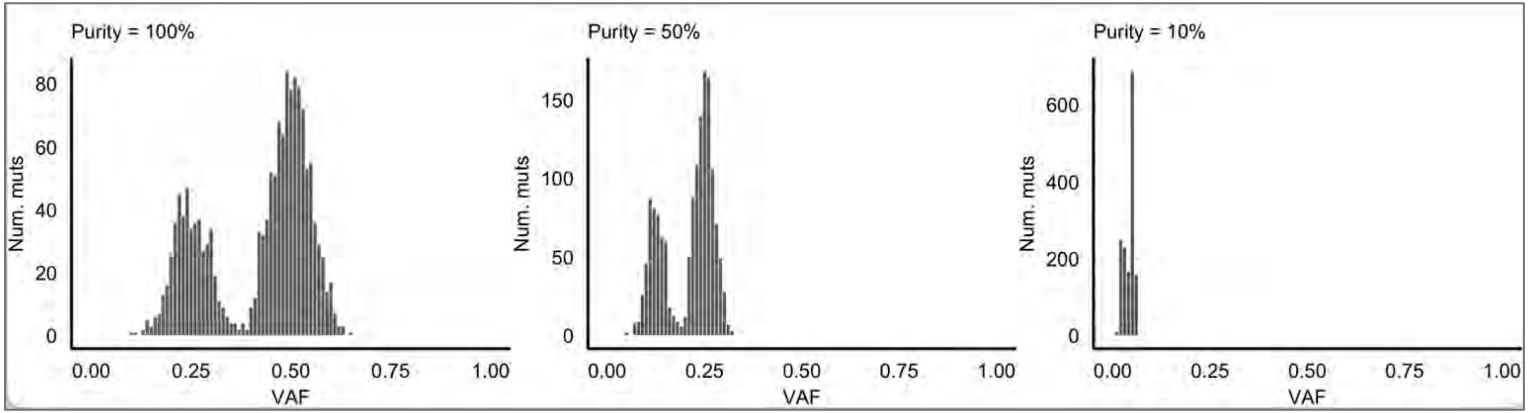
Tumor purity

- Below are 3 tumor samples with varying levels of tumor cell content
- Can you guess the tumor purity based on the somatic VAF?

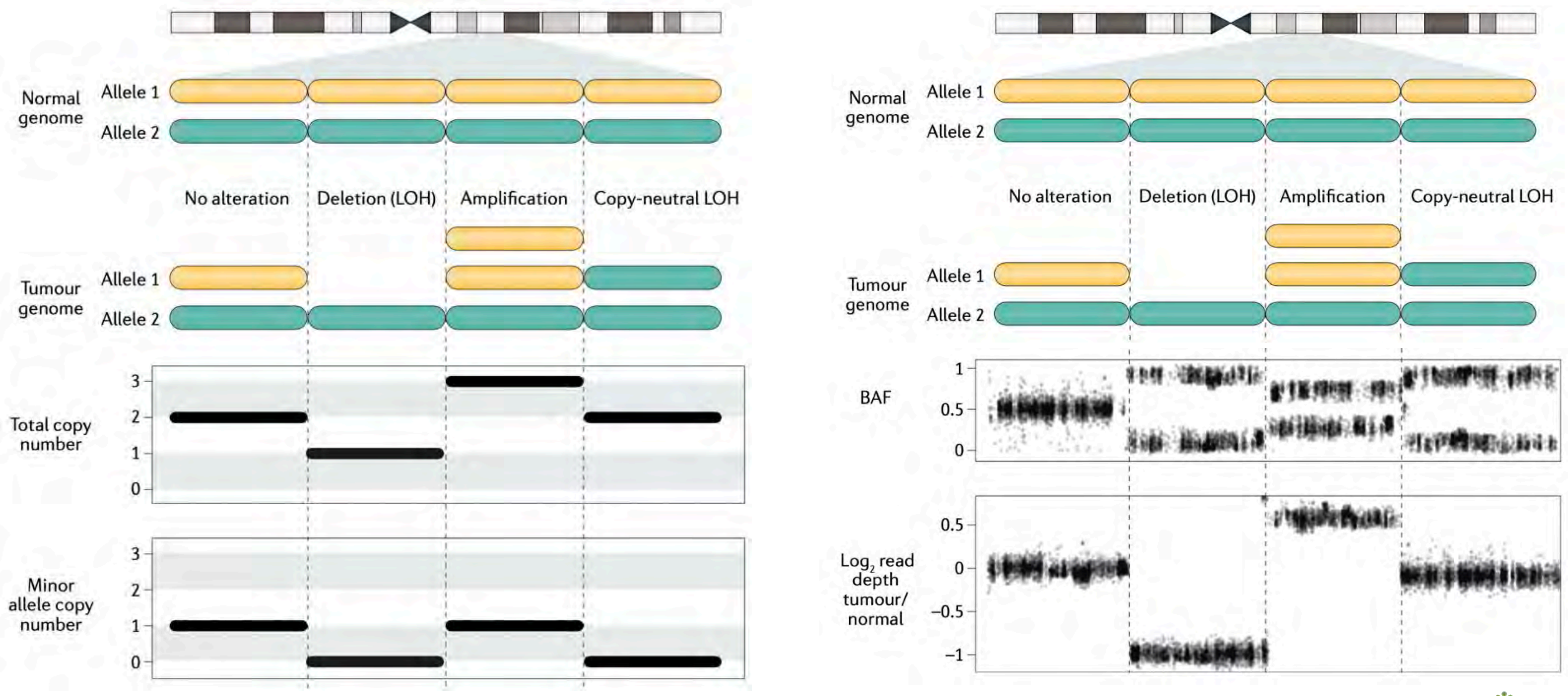


Tumor purity

- Below are 3 tumor samples with varying levels of tumor cell content
- Can you guess the tumor purity based on the somatic VAF?



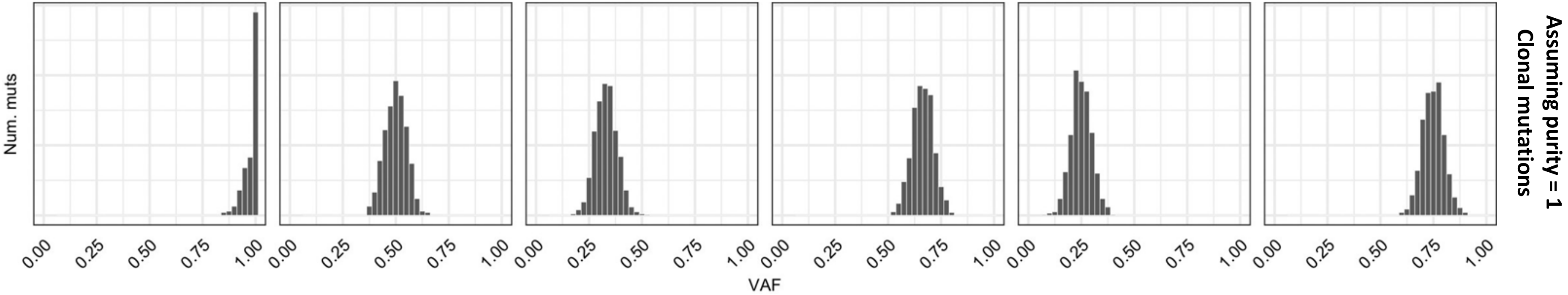
Somatic copy-number variation (SCNA)



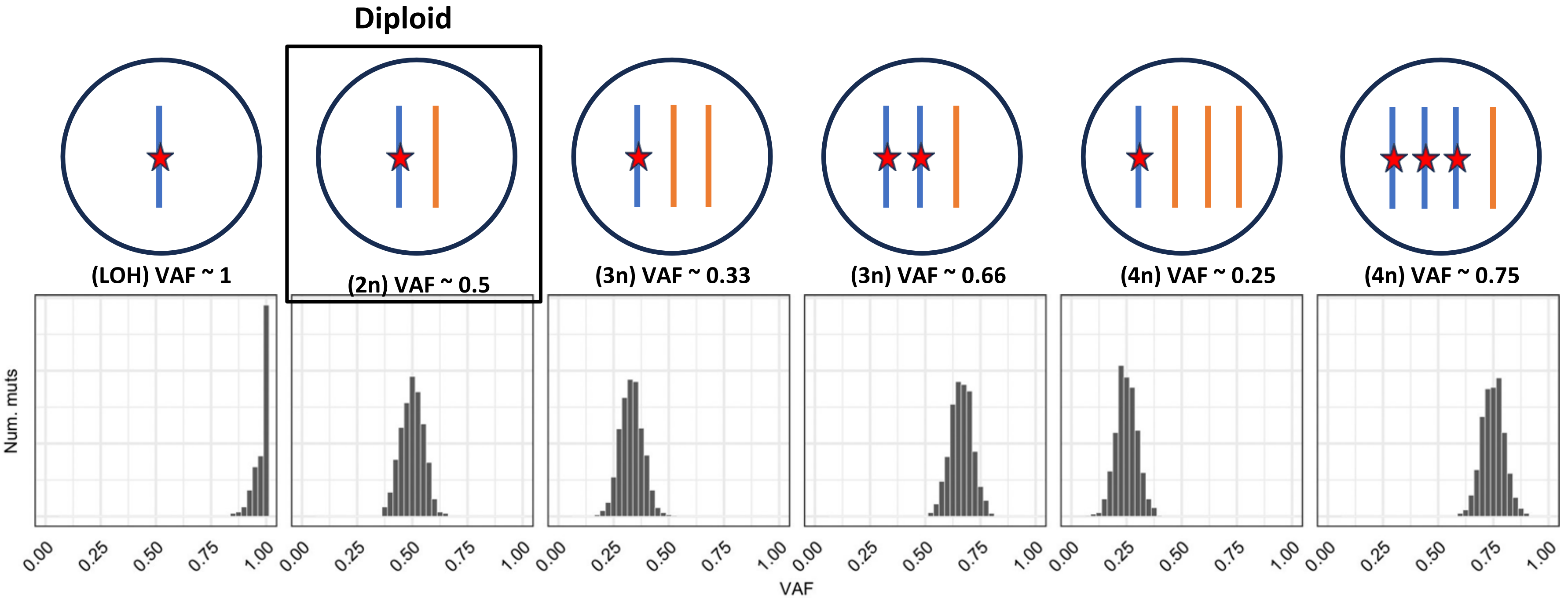
EMBL

Copy number changes affect the VAF distribution

- Assuming a 100% tumor purity and only clonal mutations
- Can you guess the total copy-number?



Copy number changes affect the VAF distribution



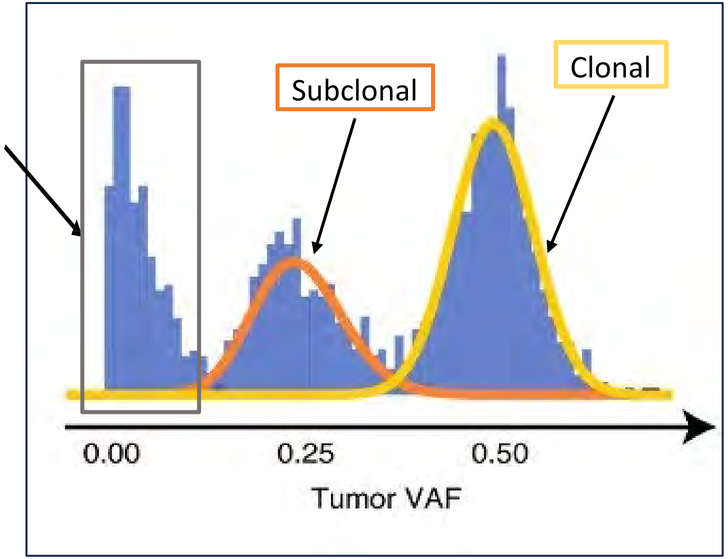
Assuming purity = 1
Clonal mutations

Subclonal heterogeneity

- Bulk sequencing data provides an aggregated view of the genetic information
- Subclonal mutations that are present at low frequencies are difficult to detect
  - → Underestimation of clonal diversity is common

Tail (not all from the same clone)

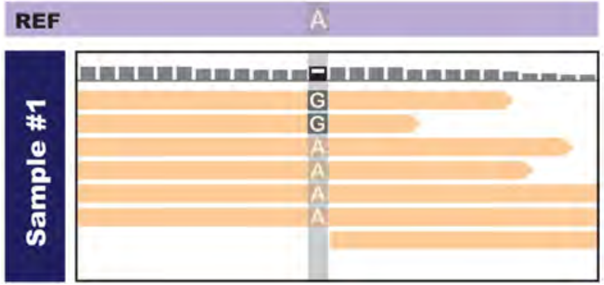
- Very hard to distinguish clones
- New mutations (majority passenger) occurring in every cell division
- The accuracy of the reconstruction **massively depends on the depth of coverage**
- Subclonal mutations at low VAF **heavily affected by sequencing errors**



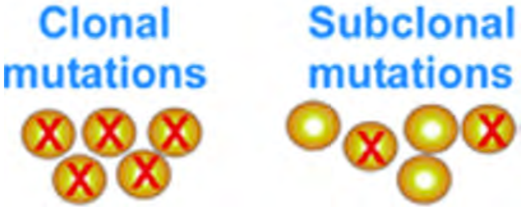
Caravagna et al, Nature Genetics, 2020

Variant vs. Population Allele Frequency

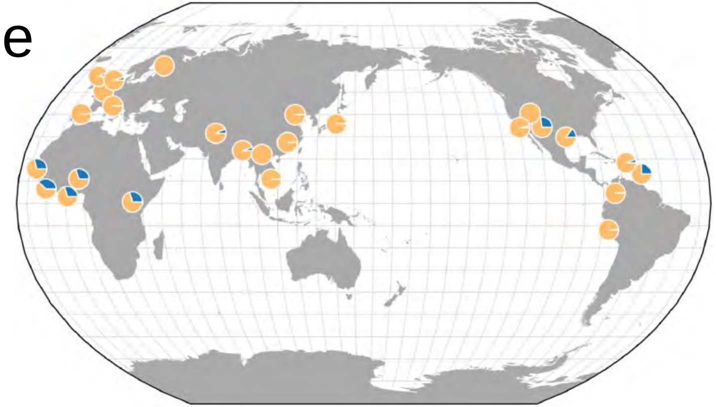
- Somatic variant allele frequency



VAF = 2/6

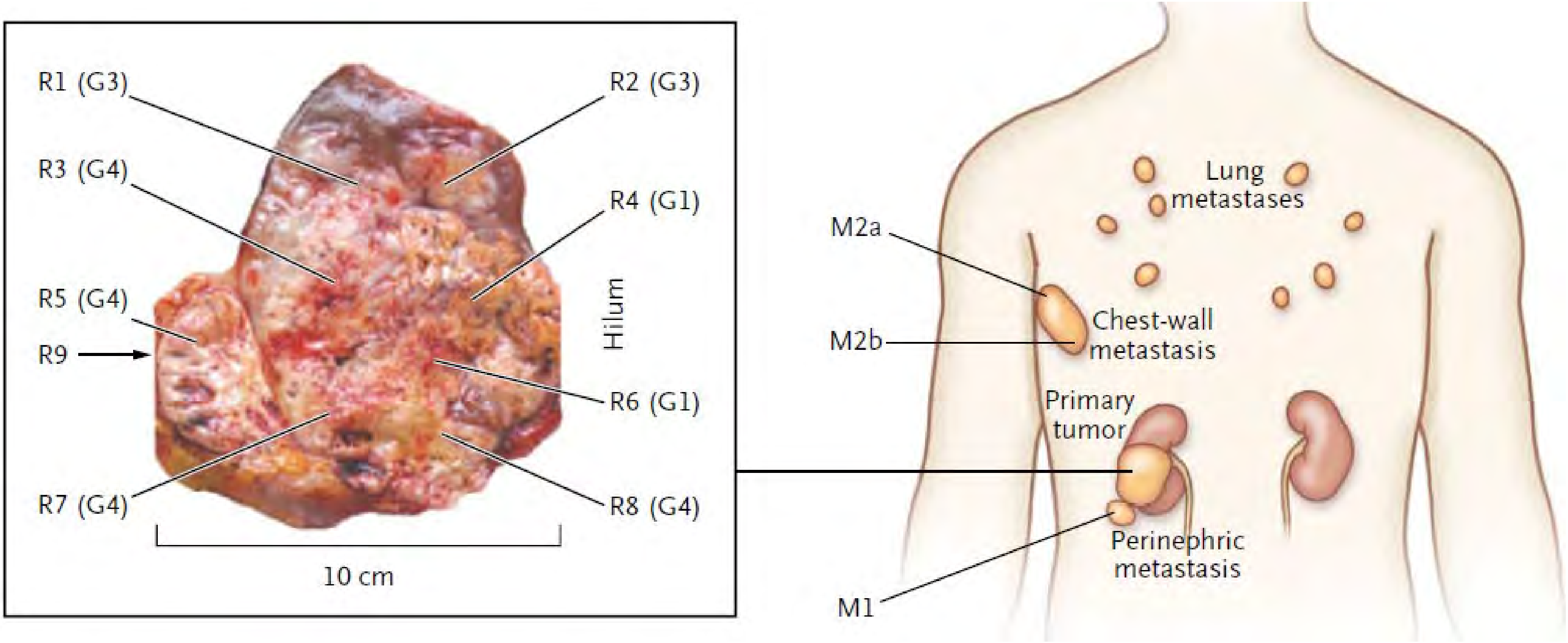


- Population allele frequency of (potentially predisposing) germline variants
  - rs6602666, A>G
    - African: ~26%
    - Finnish: 0%



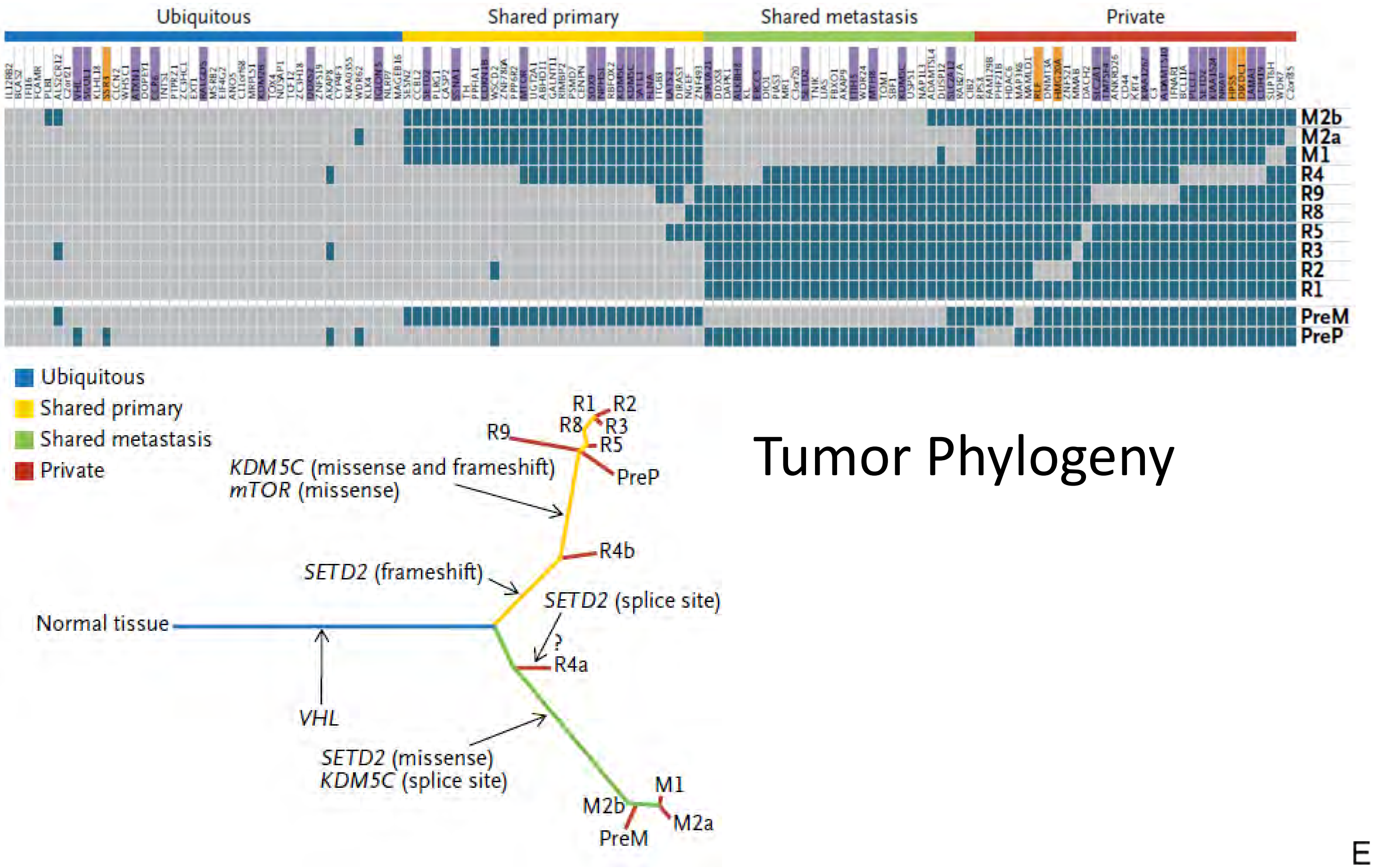
Source: https://doi.org/10.1038/srep44548, https://doi.org/10.1016/j.cell.2013.01.019

Tumor Heterogeneity & Tumor Evolution



Source: Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. Gerlinger et al., N Engl J Med. 2012 Mar 8;366(10):883-92.

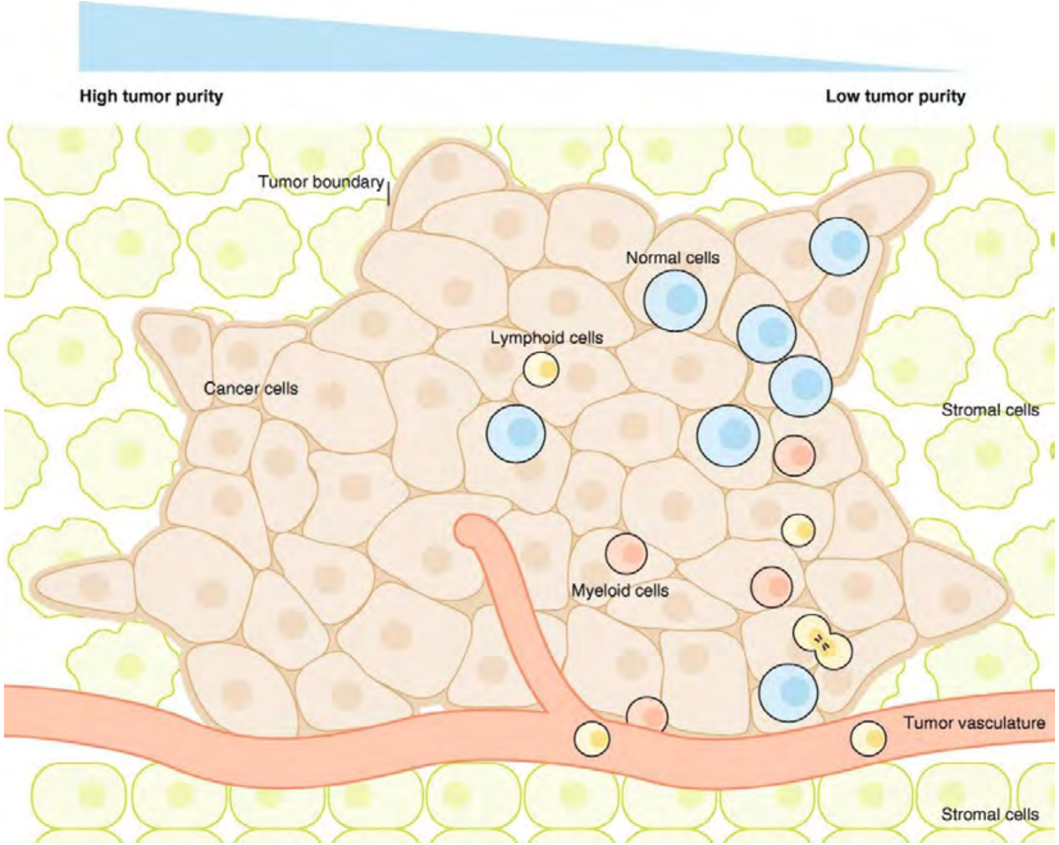
Tumor Heterogeneity & Tumor Evolution



Tumor Phylogeny

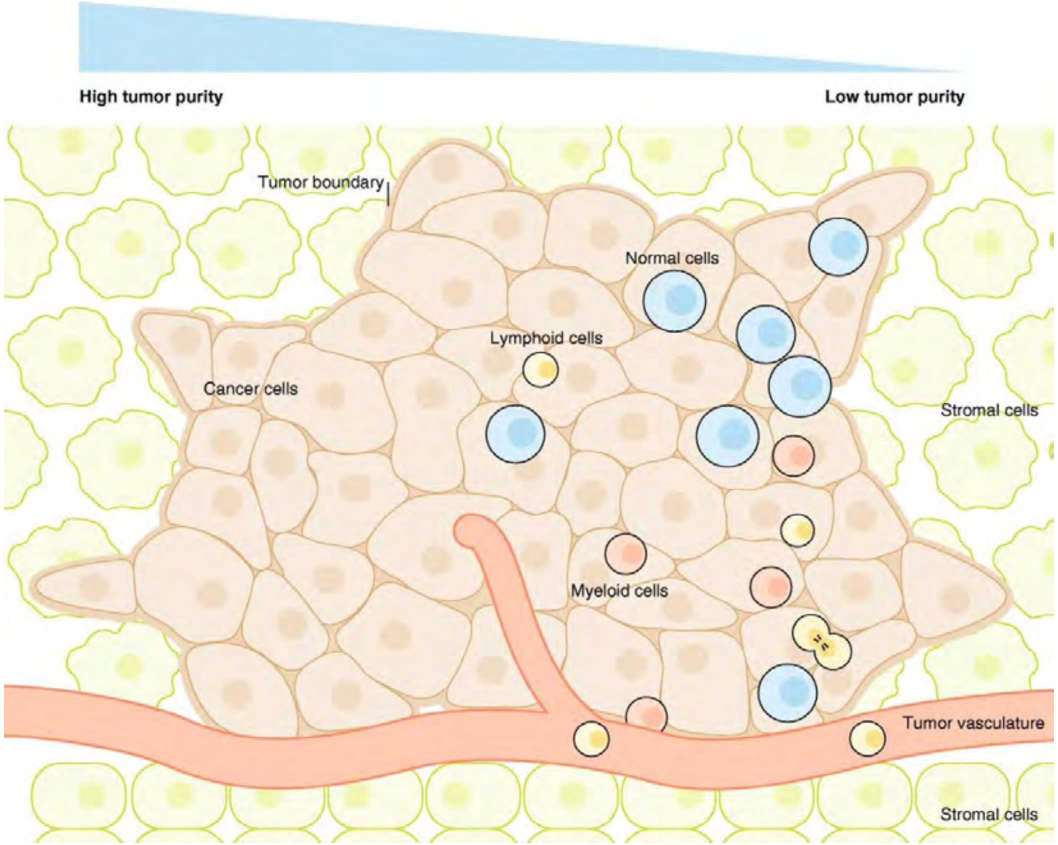
Source: Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. Gerlinger et al., N Engl J Med. 2012 Mar 8;366(10):883-92.

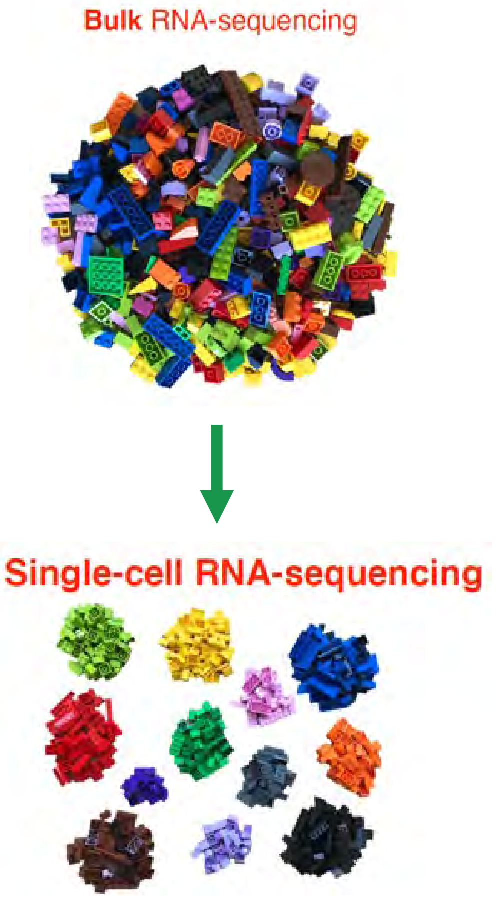
The hidden complexity in bulk sequencing



Source: Understanding tumor ecosystems by single-cell sequencing: promises and limitations. Ren et al., Genome Biol. 2018 Dec 3;19(1):211.

Single-cell sequencing

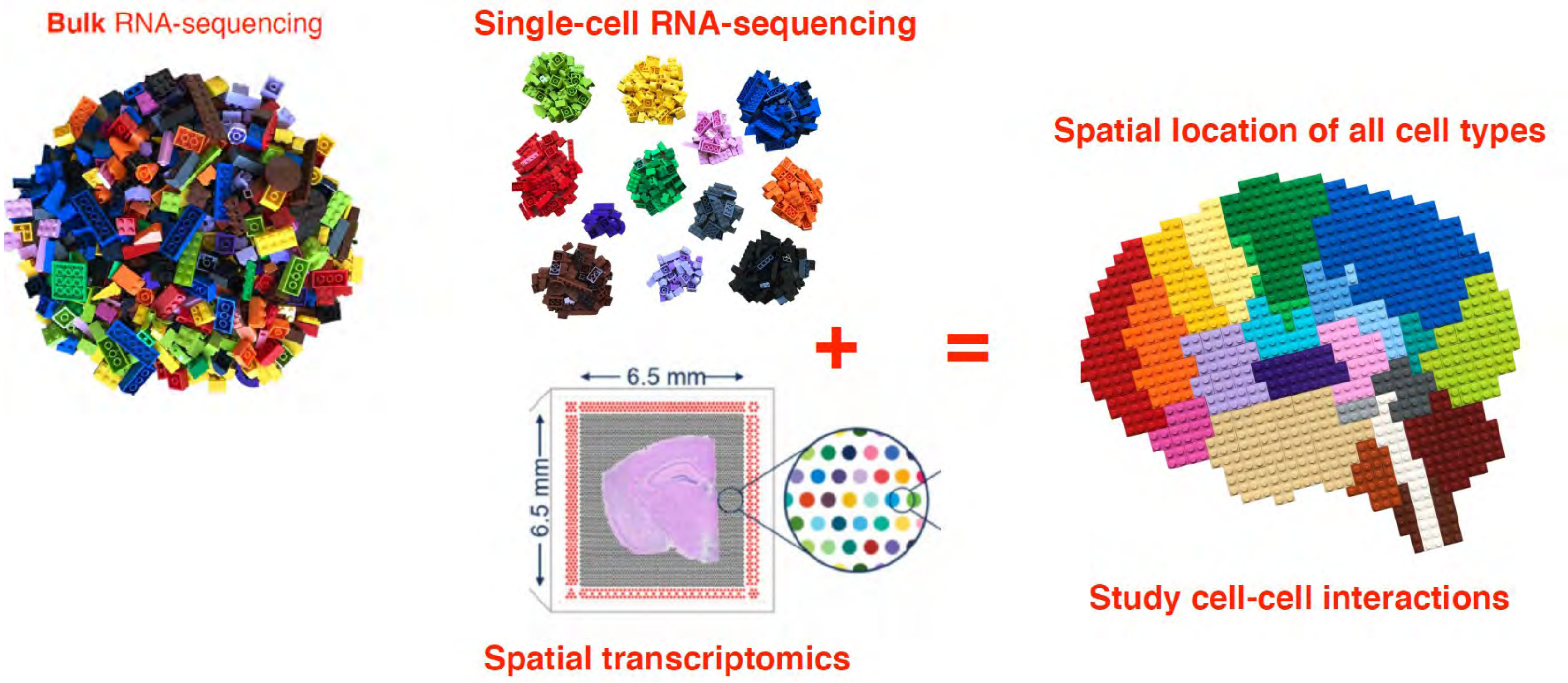




EMBL

Source: Understanding tumor ecosystems by single-cell sequencing: promises and limitations. Ren et al., Genome Biol. 2018 Dec 3;19(1):211.

Single-cell (spatial) transcriptomics



Bulk RNA-sequencing

Single-cell RNA-sequencing

+

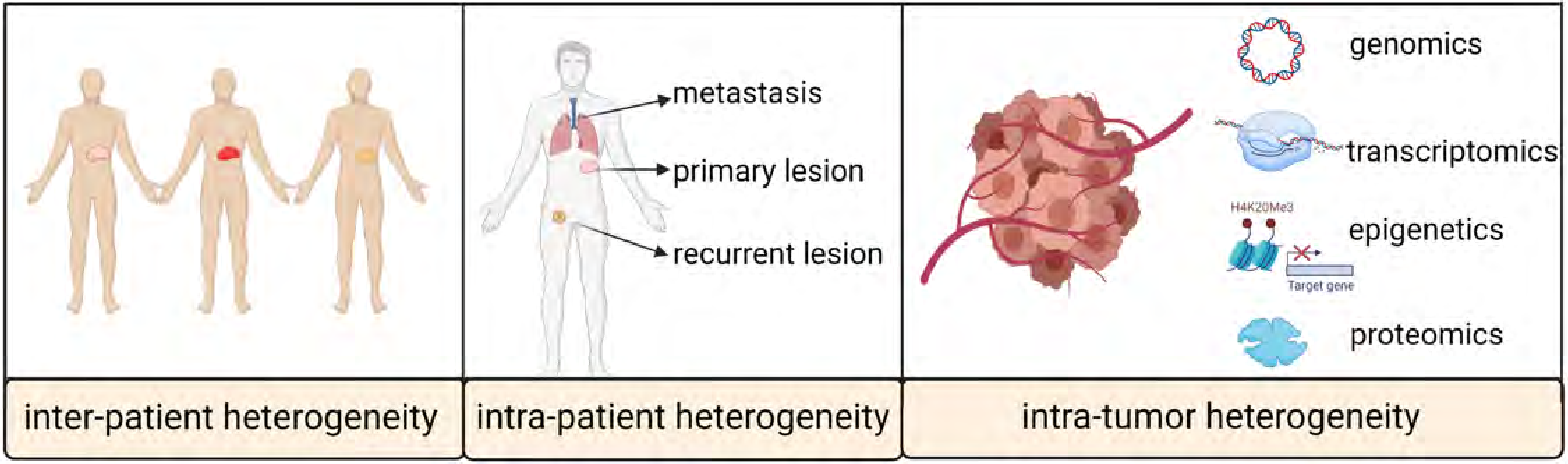
Spatial transcriptomics

=

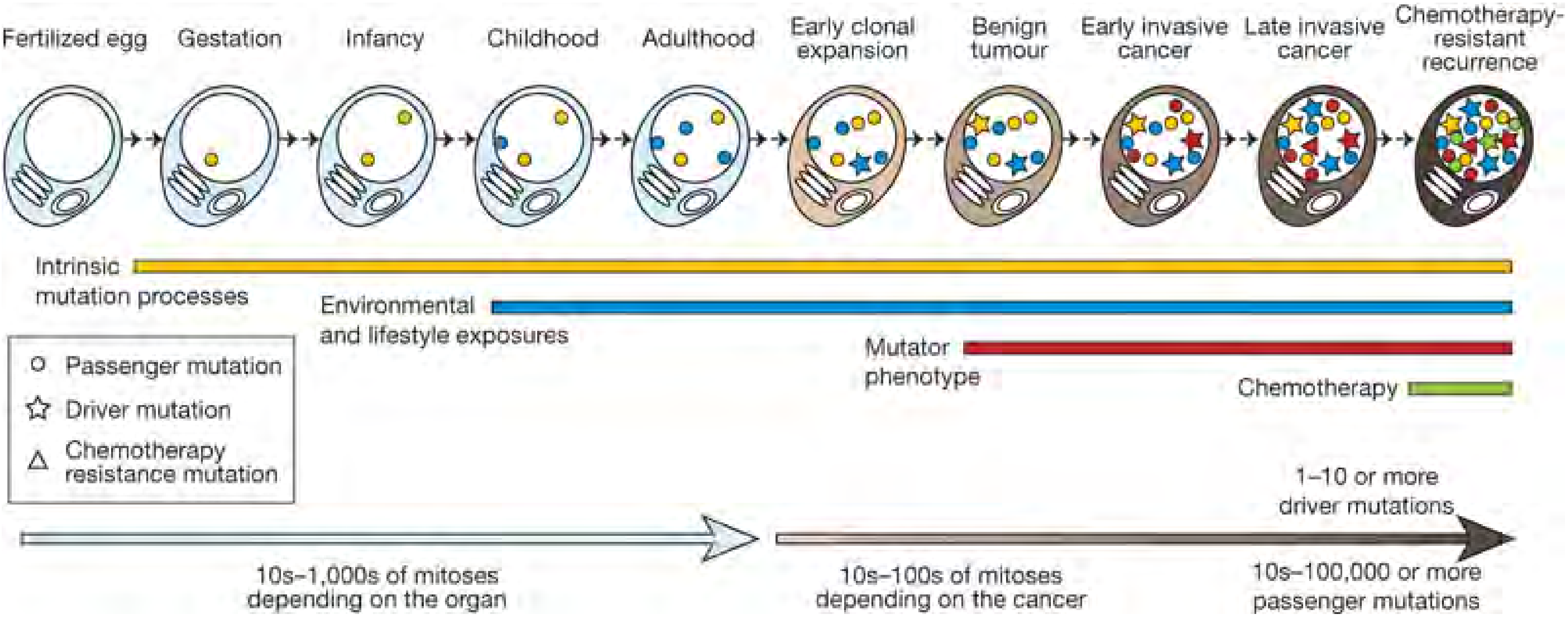
Spatial location of all cell types

Study cell-cell interactions

Tumor heterogeneity at the patient and tumor level



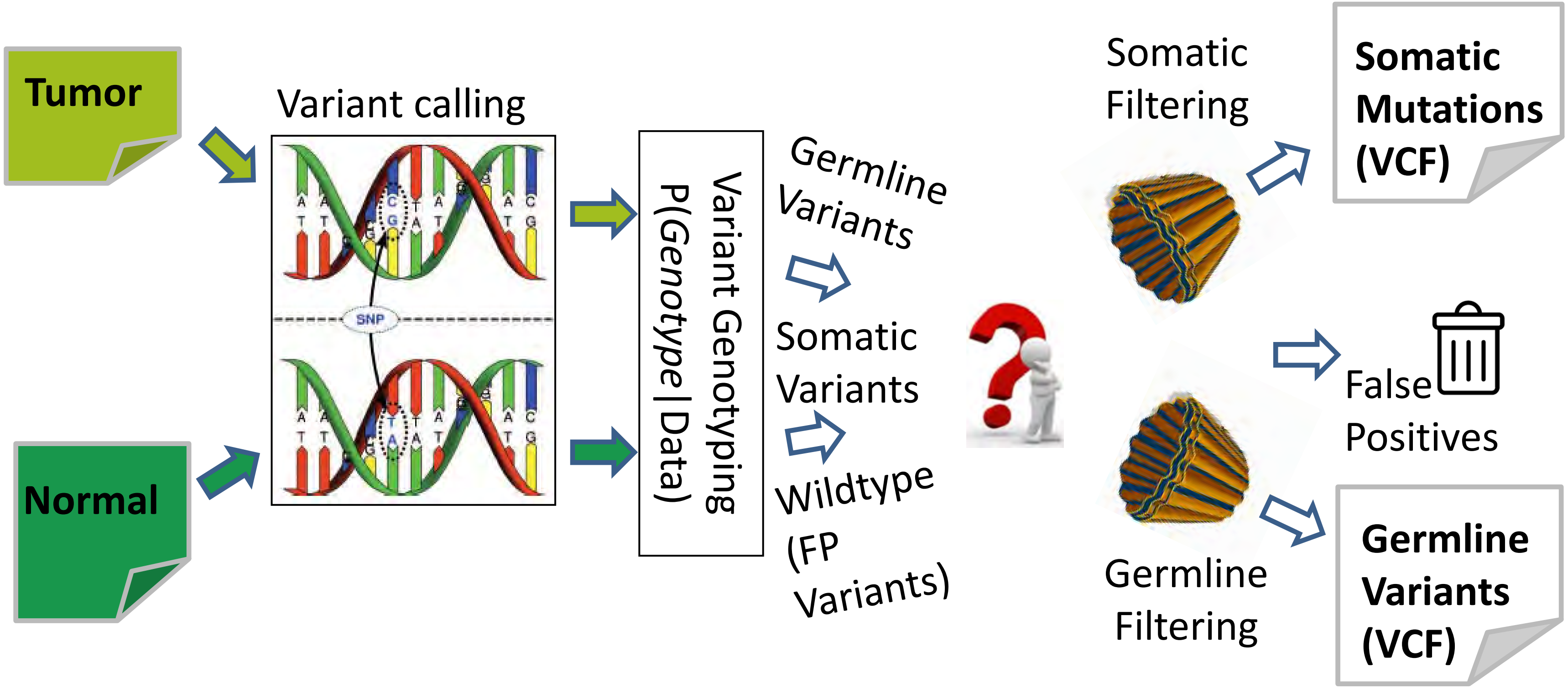
Summary: Cancer as a genetic disease and evolutionary process


Fertilized egg
Gestation
Infancy
Childhood
Adulthood
Early clonal expansion
Benign tumour
Early invasive cancer
Late invasive cancer
Chemotherapy-resistant recurrence
Intrinsic mutation processes
Environmental and lifestyle exposures
Mutator phenotype
Chemotherapy
○ Passenger mutation
☆ Driver mutation
△ Chemotherapy resistance mutation
1–10 or more driver mutations
10s–1,000s of mitoses depending on the organ
10s–100s of mitoses depending on the cancer
10s–100,000 or more passenger mutations

Source: The Cancer Genome, Michael R. Stratton, Peter J. Campbell & P. Andrew Futreal, Nature 458, p 719-724 (9 April 2009)

Cancer Genome Data Analysis

Somatic variant calling



Tumor

Normal

Variant calling

Variant Genotyping
P(Genotype|Data)

Germline Variants

Somatic Variants

Wildtype (FP Variants)

Somatic Filtering

Germline Filtering

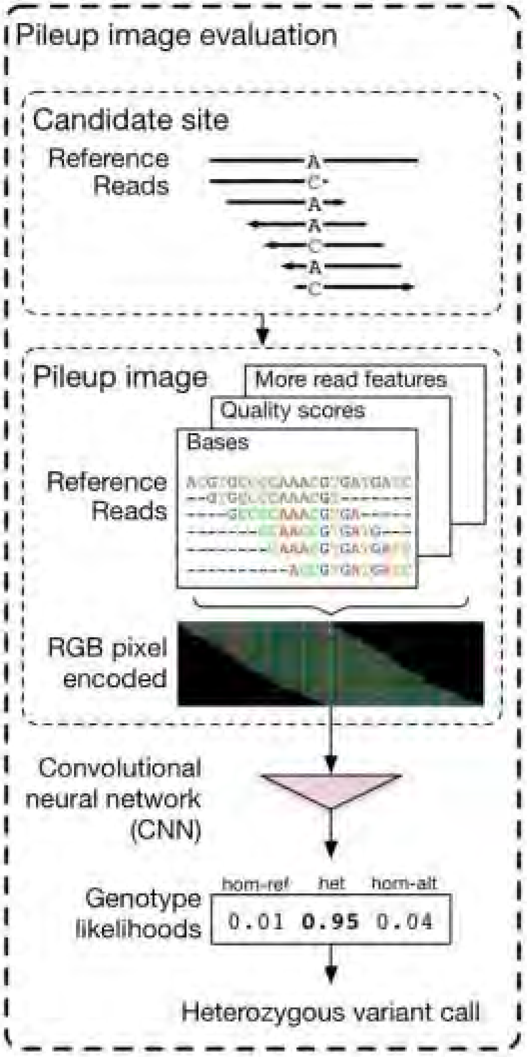
Somatic Mutations (VCF)

False Positives

Germline Variants (VCF)

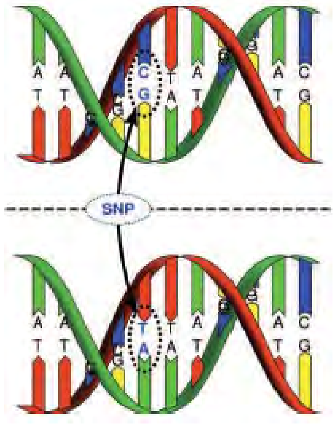
Variant Calling Methods

- Four broad categories
  - Heuristic Methods
    - Hardly used anymore
  - Probabilistic methods
    - Bayesian methods
  - Machine Learning methods
    - Deep Learning methods, e.g., DeepVariant
  - Graph-based methods
    - Pangenome methods



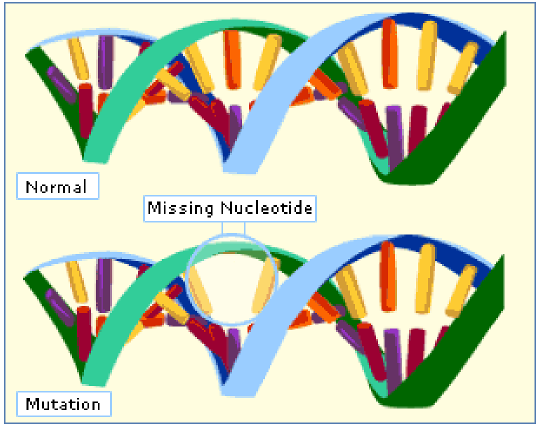
Types of Variants

Single-nucleotide variants (SNVs)



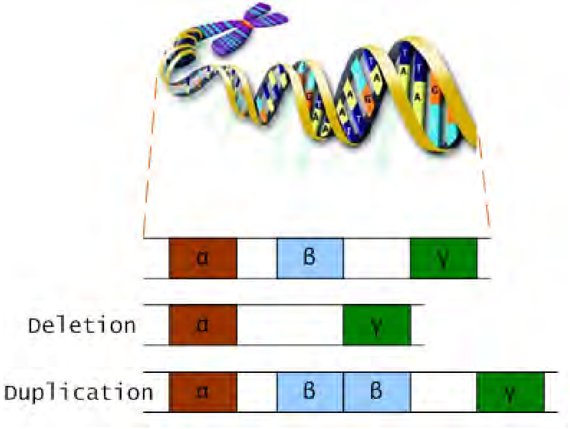
Size: 1bp

Short insertions & deletions (InDels)



1-50bp

Copy-Number Variants (CNVs) & Structural Variants (SVs)

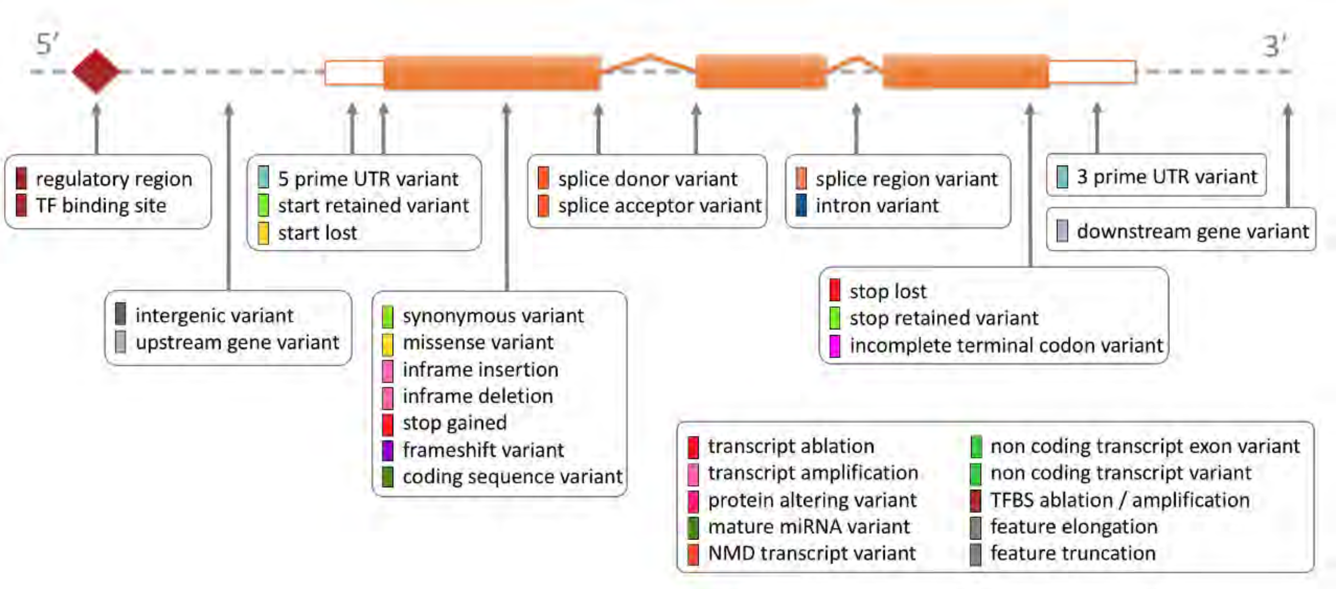


>50bp

Different methods are used to discover and genotype SNVs, InDels, SVs and CNVs.

Further information:
Methods in Genomic Variant Calling: https://www.youtube.com/watch?v=zO9WCOaq3aQ

Interpreting Genomic Variants



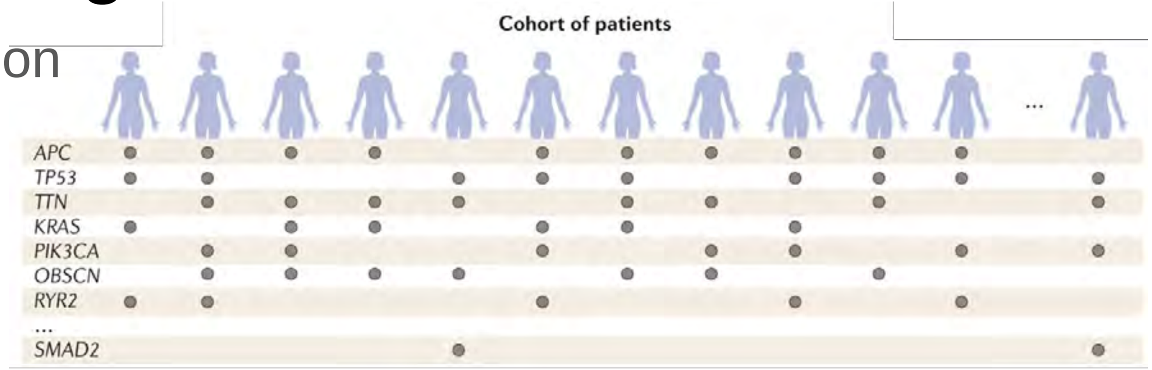
Popular Tools: VEP, Annovar, snpEff

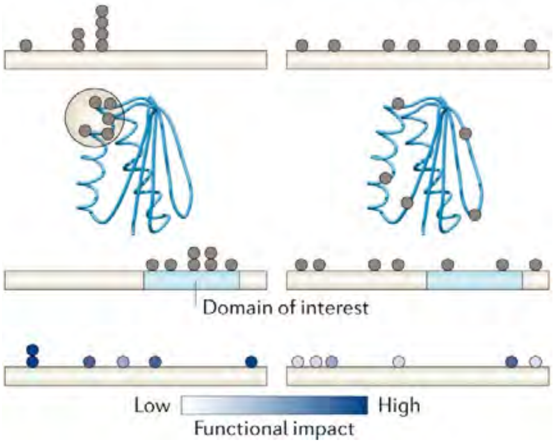
Mediocre support for annotation of copy-number and structural variants

Source: https://m.ensembl.org/info/genome/variation/prediction/predicted_data.html

Driver versus Passenger

- Signals of positive selection
  - Mutation Recurrence
  - Hotspot mutations
  - Protein domain clustering
  - Functional impact bias

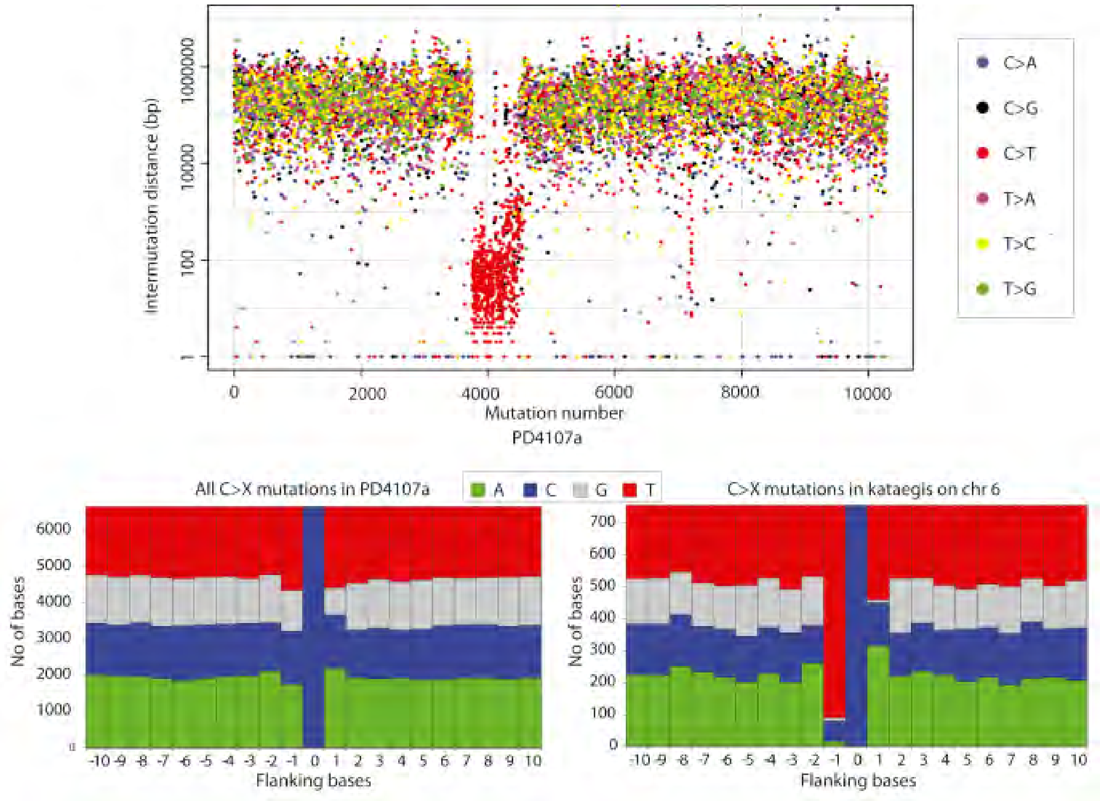




Source: Martínez-Jiménez, F., Muiños, F., Sentís, I. et al. A compendium of mutational cancer driver genes. Nat Rev Cancer 20, 555–572 (2020)

Mutational Signatures

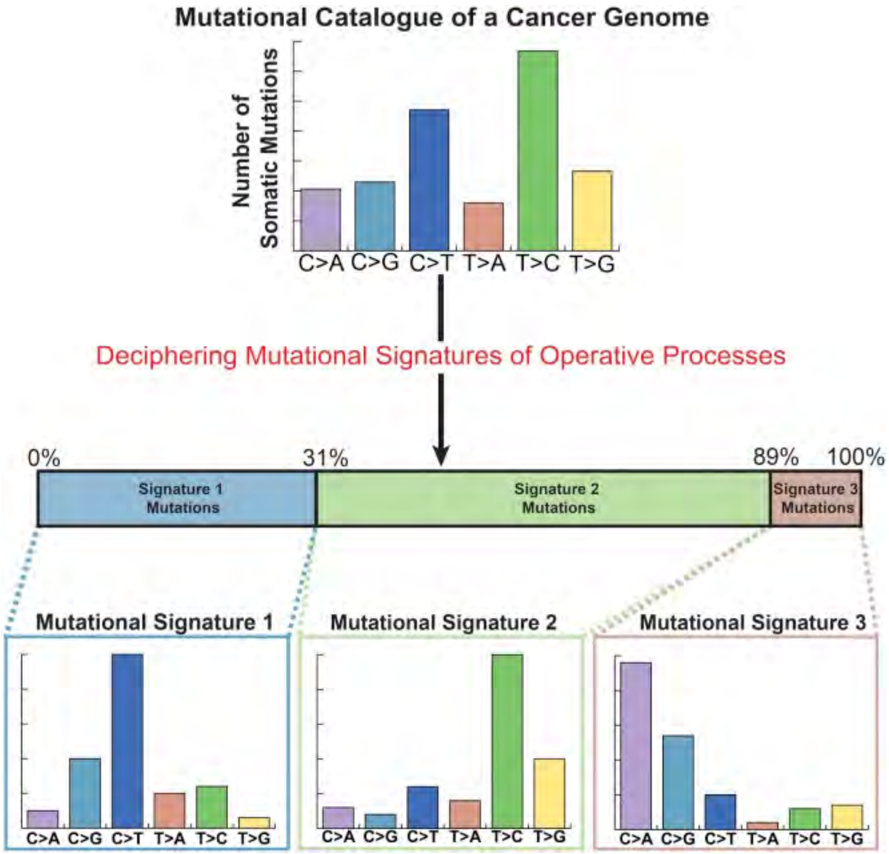
Mutational Signatures



Tri-Nucleotide Context TpCpX

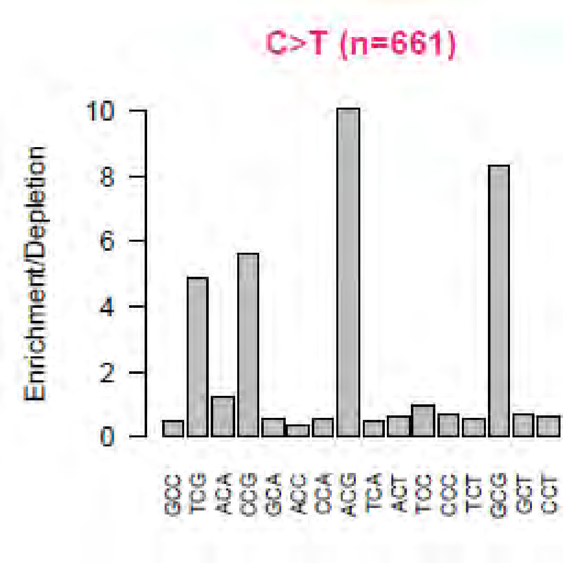
Source: Mutational processes molding the genomes of 21 breast cancers. Nik-Zainal et al., Cell. 2012 May 25;149(5):979-93.

Mutational Signatures

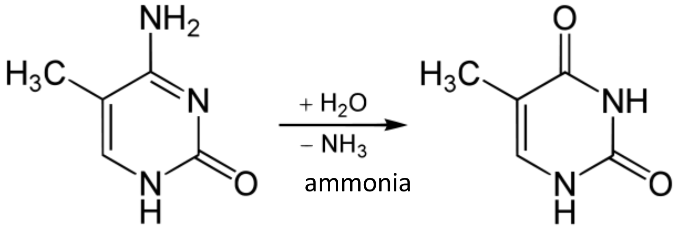


Source: Deciphering signatures of mutational processes operative in human cancer. Alexandrov et al., Cell Rep. 2013 Jan 31;3(1):246-59.

Tri-Nucleotide Sequence Context



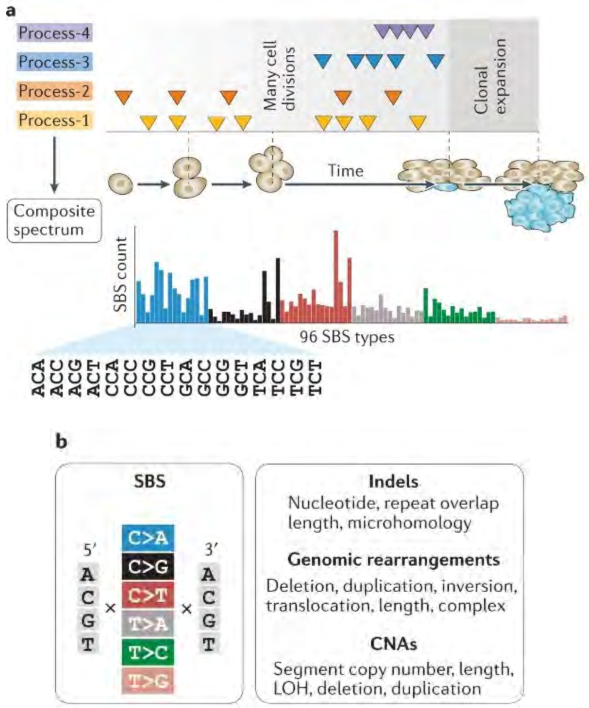
- Elevated **C>T** mutation rate at **XpCpG** trinucleotides
- Deamination of methylated cytosines to thymine (usually at XpCpGs)



EMBL

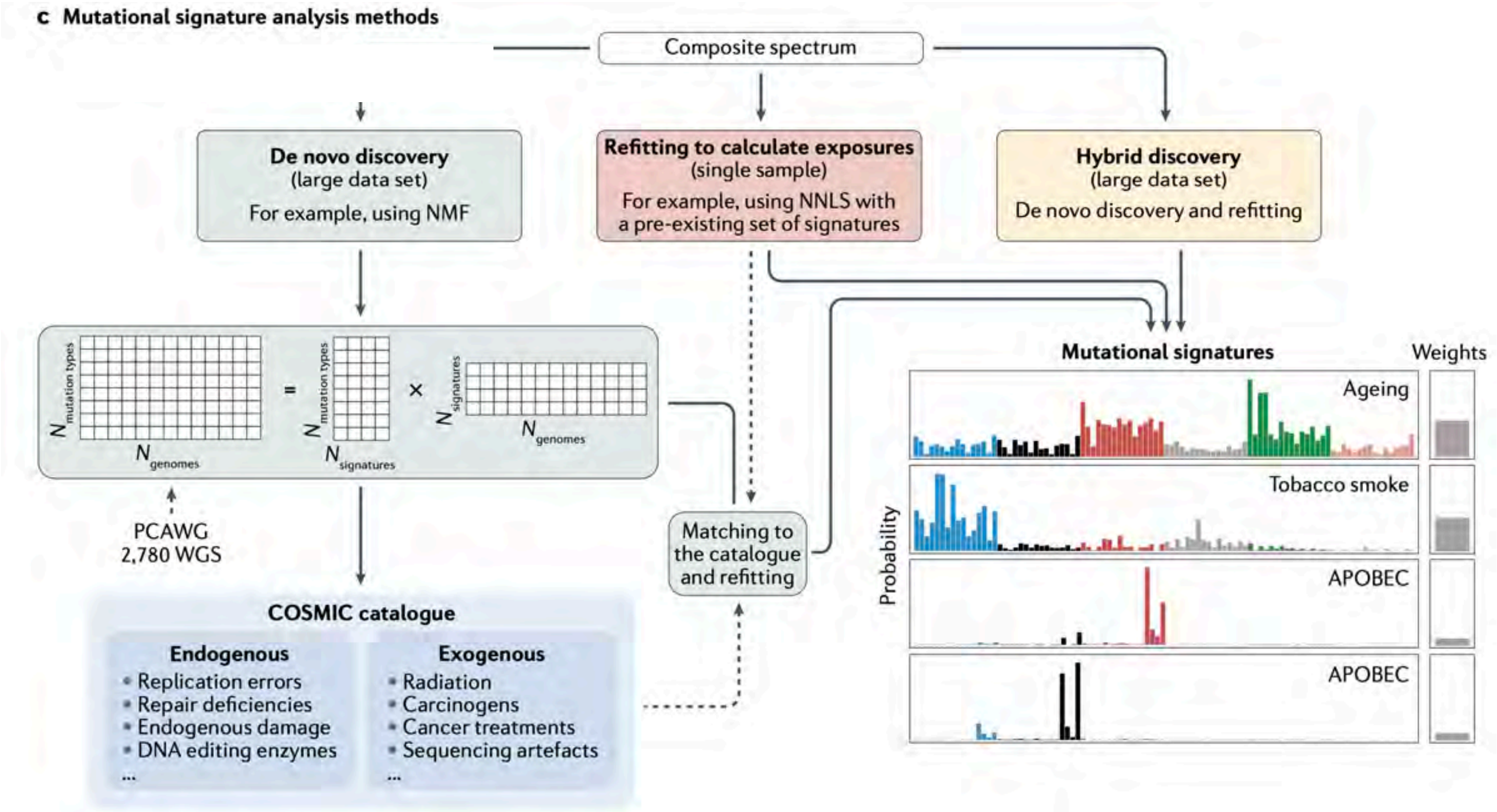
Mutational Signature

- Characterize patterns of DNA mutations
- These patterns can arise from various mutational processes that lead to changes in DNA sequences
- Mutations can occur due to a variety of factors
  - DNA replication errors
  - Exposure to mutagenic agents (e.g., ultraviolet radiation, chemical carcinogens)
  - Defects in DNA repair mechanisms.



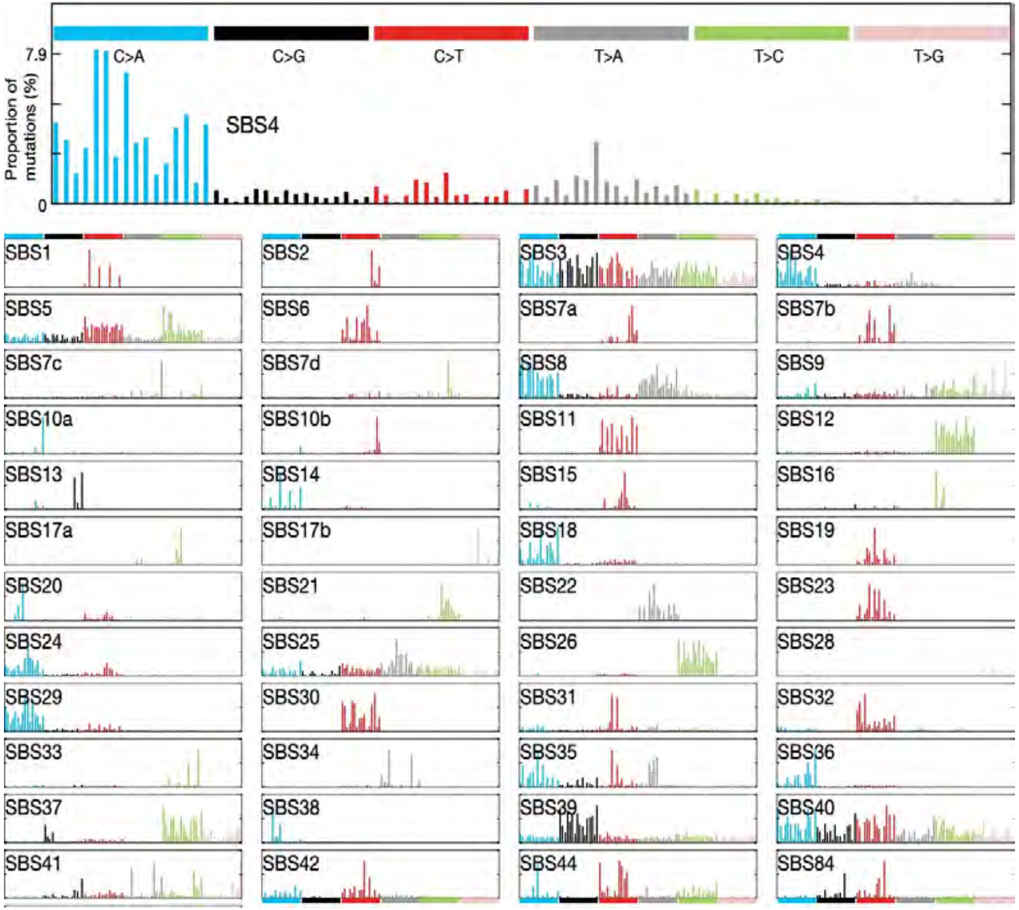
Cortes-Ciriano et al 2022

De novo discovery or re-fitting

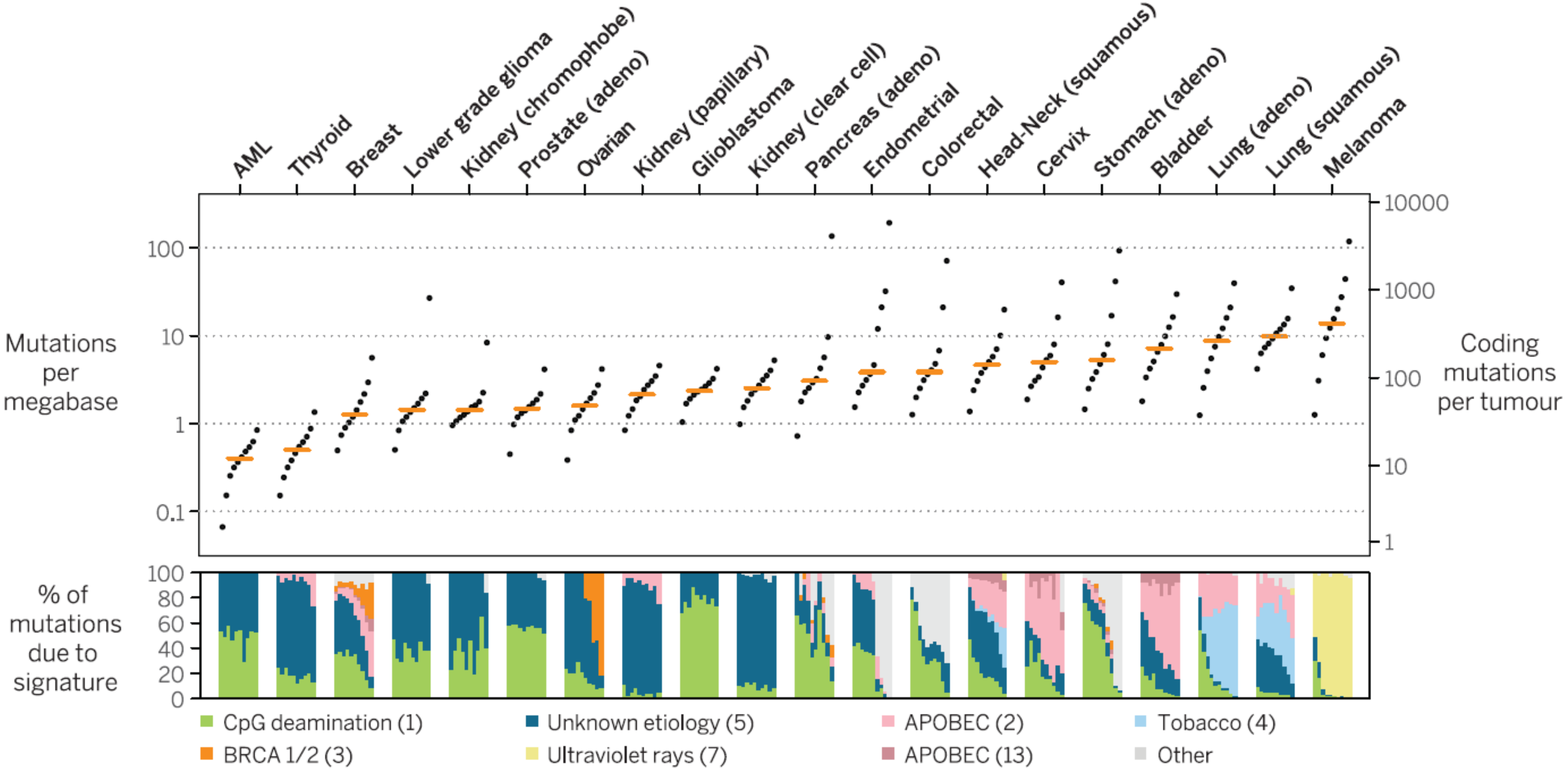


Source: Cortés-Ciriano, I., Gulhan, D.C., Lee, J.JK. et al. Computational analysis of cancer genome sequencing data. Nat Rev Genet **23**, 298–314 (2022)

Mutational signature catalog



Mutational Signatures across Human Cancer Types

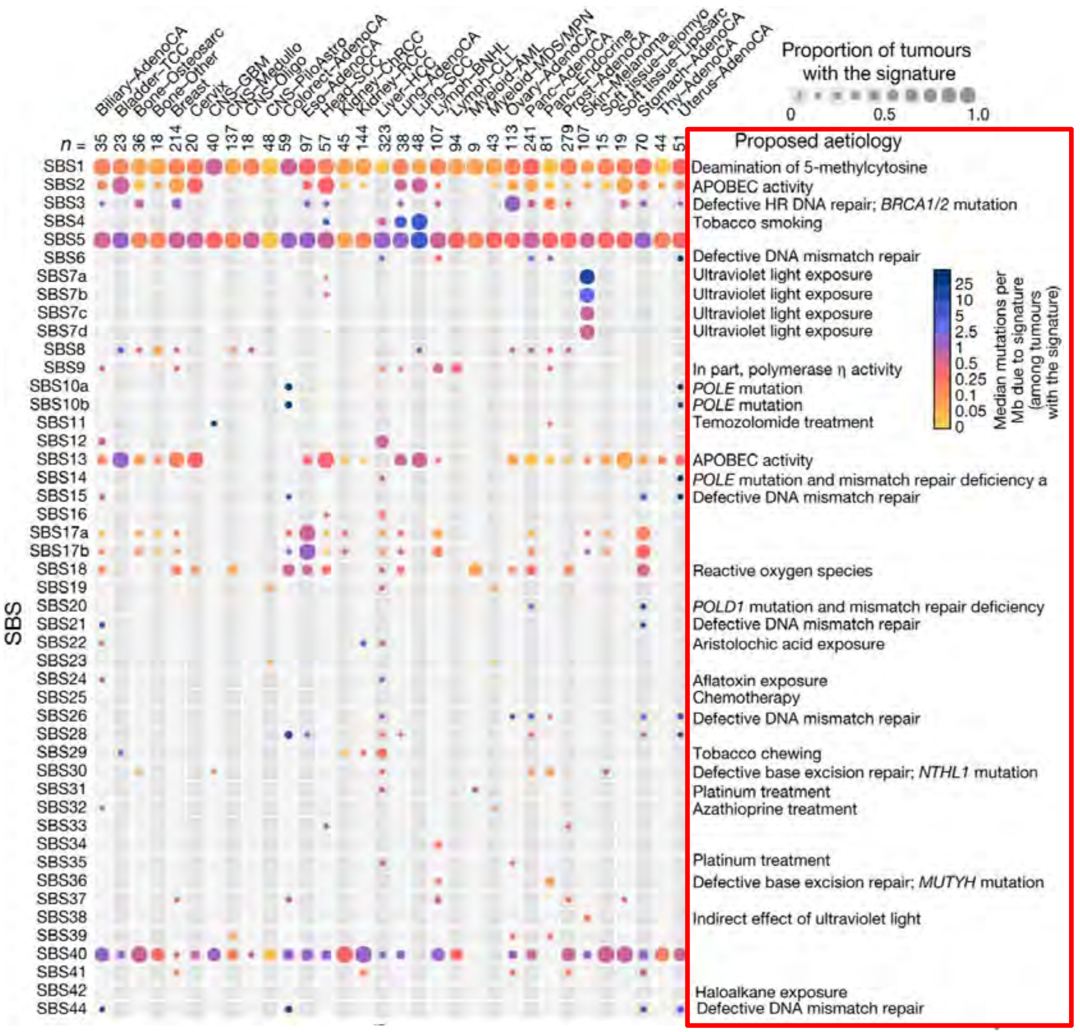


Source: Somatic mutation in cancer and normal cells. Martincorena I and Campbell PJ, Science. 2015 Sep 25;349(6255):1483-9.

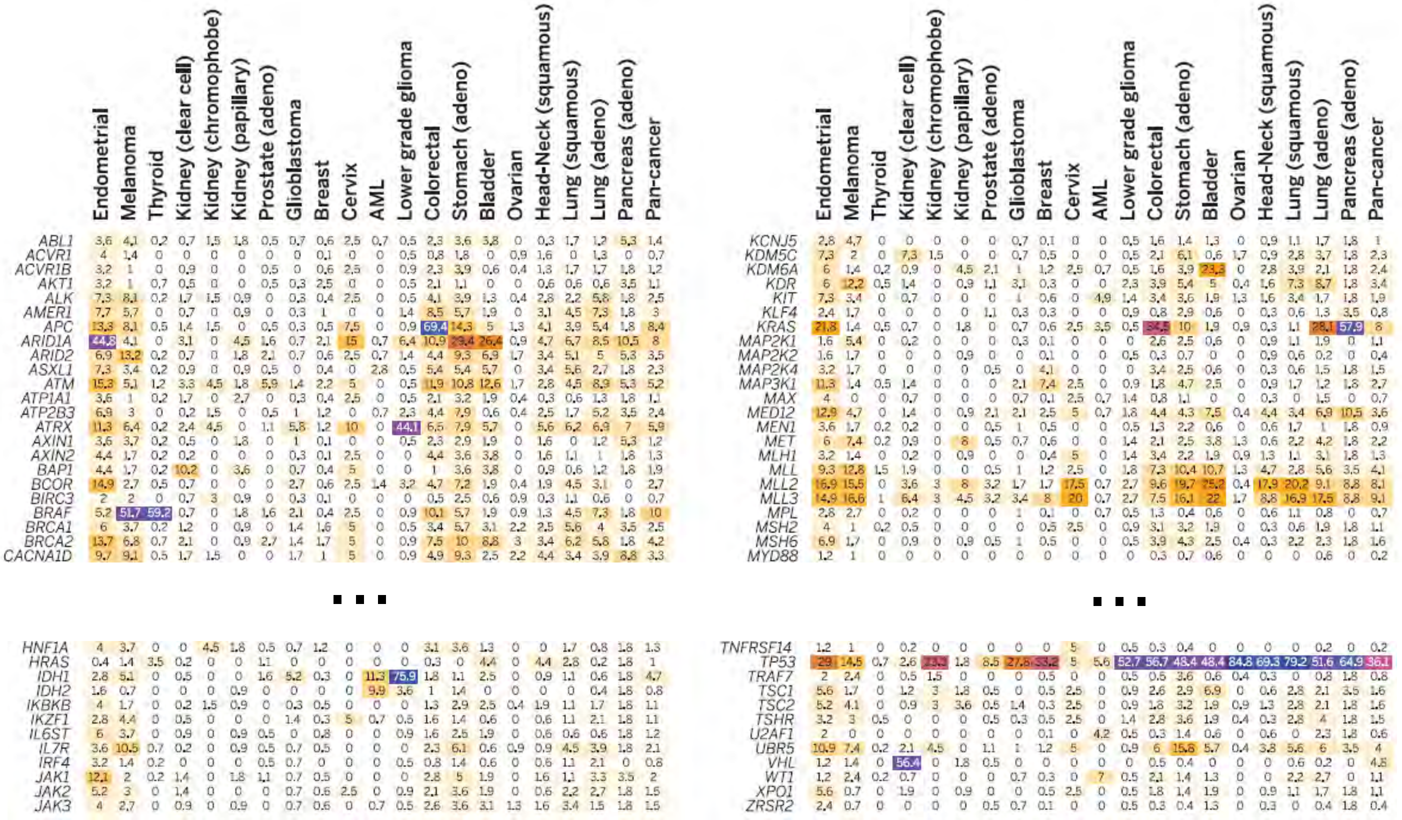
Classification of mutational signatures based on biological processes

- Cancer Aetiology Investigation
  - Mutational signatures provide insights into the underlying causes
  - https://cancer.sanger.ac.uk/signatures/

COSMIC
Catalogue Of Somatic Mutations In Cancer



Recurrently Mutated Cancer Genes



	Endometrial	Melanoma	Thyroid	Kidney (clear cell)	Kidney (chromophobe)	Kidney (papillary)	Prostate (adeno)	Glioblastoma	Breast	Cervix	AML	Lower grade glioma	Colorectal	Stomach (adeno)	Bladder	Ovarian	Head-Neck (squamous)	Lung (squamous)	Lung (adeno)	Pancreas (adeno)	Pan-cancer
ABL1	3.6	4.1	0.2	0.7	1.5	1.8	0.5	0.7	0.6	2.5	0.7	0.5	2.3	3.6	3.8	0	0.3	1.7	1.2	5.3	1.4
ACVR1	4	1.4	0	0	0	0	0	0	0.1	0	0	0.5	0.8	1.8	0	0.9	1.6	0	1.3	0	0.7
ACVR1B	3.2	1	0	0.9	0	0	0.5	0	0.6	2.5	0	0.9	2.3	3.9	0.6	0.4	1.3	1.7	1.7	1.8	1.2
AKT1	3.2	1	0.7	0.5	0	0	0.5	0.3	2.5	0	0	0.5	2.1	1.1	0	0	0.6	0.6	0.6	3.5	1.1
ALK	7.3	8.1	0.2	1.7	1.5	0.9	0	0.3	0.4	2.5	0	0.5	4.1	3.9	1.3	0.4	2.8	2.2	5.8	1.8	2.5
AMER1	7.7	5.7	0	0.7	0	0.9	0	0.3	1	0	0	1.4	8.5	5.7	1.9	0	3.1	4.5	7.3	1.8	3
APC	13.3	8.1	0.5	1.4	1.5	0	0.5	0.3	0.5	7.5	0	0.9	69.4	14.3	5	1.3	4.1	3.9	5.4	1.8	8.4
ARID1A	44.8	4.1	0	3.1	0	4.5	1.6	0.7	2.1	15	0.7	6.4	10.9	29.4	26.4	0.9	4.7	6.7	8.5	10.5	8
ARID2	6.9	13.2	0.2	0.7	0	1.8	2.1	0.7	0.6	2.5	0.7	1.4	4.4	9.3	6.9	1.7	3.4	5.1	5	5.3	3.5
ASXL1	7.3	3.4	0.2	0.9	0	0.9	0.5	0	0.4	0	2.8	0.5	5.4	5.4	5.7	0	3.4	5.6	2.7	1.8	2.3
ATM	15.3	5.1	1.2	3.3	4.5	1.8	5.9	1.4	2.2	5	0	0.5	11.9	10.8	12.6	1.7	2.8	4.5	8.9	5.3	5.2
ATP1A1	3.6	1	0.2	1.7	0	2.7	0	0.3	0.4	2.5	0	0.5	2.1	3.2	1.9	0.4	0.3	0.6	1.3	1.8	1.1
ATP2B3	6.9	3	0	0.2	1.5	0	0.5	1	1.2	0	0.7	2.3	4.4	7.9	0.6	0.4	2.5	1.7	5.2	3.5	2.4
ATRX	11.3	6.4	0.2	2.4	4.5	0	1.1	5.8	1.2	10	0	44.1	6.5	7.9	5.7	0	5.6	6.2	6.9	7	5.9
AXIN1	3.6	3.7	0.2	0.5	0	1.8	0	1	0.1	0	0	0.5	2.3	2.9	1.9	0	1.6	0	1.2	5.3	1.2
AXIN2	4.4	1.7	0.2	0.2	0	0	0	0.3	0.1	2.5	0	0	4.4	3.6	3.8	0	1.6	1.1	1	1.8	1.3
BAP1	4.4	1.7	0.2	10.2	0	3.6	0	0.7	0.4	5	0	0	1	3.6	3.8	0	0.9	0.6	1.2	1.8	1.9
BCOR	14.9	2.7	0.5	0.7	0	0	0	2.7	0.6	2.5	1.4	3.2	4.7	7.2	1.9	0.4	1.9	4.5	3.1	0	2.7
BIRC3	2	2	0	0.7	3	0.9	0	0.3	0.1	0	0	0	0.5	2.5	0.6	0.9	0.3	1.1	0.6	0	0.7
BRAF	5.2	51.7	59.2	0.7	0	1.8	1.6	2.1	0.4	2.5	0	0.9	10.1	5.7	1.9	0.9	1.3	4.5	7.3	1.8	10
BRCA1	6	3.7	0.2	1.2	0	0.9	0	1.4	1.6	5	0	0.5	3.4	5.7	3.1	2.2	2.5	5.6	4	3.5	2.5
BRCA2	13.7	6.8	0.7	2.1	0	0.9	2.7	1.4	1.7	5	0	0.9	7.5	10	8.8	3	3.4	6.2	5.8	1.8	4.2
CACNA1D	9.7	9.1	0.5	1.7	1.5	0	0	1.7	1	5	0	0.9	4.9	9.3	2.5	2.2	4.4	3.4	3.9	8.8	3.3
...																					
HNF1A	4	3.7	0	0	4.5	1.8	0.5	0.7	1.2	0	0	0	3.1	3.6	1.3	0	0	1.7	0.8	1.8	1.3
HRAS	0.4	1.4	3.5	0.2	0	0	1.1	0	0	0	0	0	0.3	0	4.4	0	4.4	2.8	0.2	1.8	1
IDH1	2.8	5.1	0	0.5	0	0	1.6	5.2	0.3	0	11.3	75.9	1.8	1.1	2.5	0	0.9	1.1	0.6	1.8	4.7
IDH2	1.6	0.7	0	0	0	0.9	0	0	0	0	9.9	3.6	1	1.4	0	0	0	0	0.4	1.8	0.8
IKBKB	4	1.7	0	0.2	1.5	0.9	0	0.3	0.5	0	0	0	1.3	2.9	2.5	0.4	1.9	1.1	1.7	1.8	1.1
IKZF1	2.8	4.4	0	0.5	0	0	0	1.4	0.3	5	0.7	0.5	1.6	1.4	0.6	0	0.6	1.1	2.1	1.8	1.1
IL6ST	6	3.7	0	0.9	0	0.9	0.5	0	0.8	0	0	0.9	1.6	2.5	1.9	0	0.6	0.6	0.6	1.8	1.2
IL7R	3.6	10.5	0.7	0.2	0	0.9	0.5	0.7	0.5	0	0	0	2.3	6.1	0.6	0.9	0.9	4.5	3.9	1.8	2.1
IRF4	3.2	1.4	0.2	0	0	0	0.5	0.7	0	0	0	0.5	0.8	1.4	0.6	0	0.6	1.1	2.1	0	0.8
JAK1	12.1	2	0.2	1.4	0	1.8	1.1	0.7	0.5	0	0	0	2.8	5	1.9	0	1.6	1.1	3.3	3.5	2
JAK2	5.2	3	0	1.4	0	0	0	0.7	0.6	2.5	0	0.9	2.1	3.6	1.9	0	0.6	2.2	2.7	1.8	1.5
JAK3	4	2.7	0	0.9	0	0.9	0	0.7	0.6	0	0.7	0.5	2.6	3.6	3.1	1.3	1.6	3.4	1.5	1.8	1.5
KCNJ5	2.8	4.7	0	0	0	0	0	0.7	0.1	0	0	0.5	1.6	1.4	1.3	0	0.9	1.1	1.7	1.8	1
KDM5C	7.3	2	0	7.3	1.5	0	0	0.7	0.5	0	0	0.5	2.1	6.1	0.6	1.7	0.9	2.8	3.7	1.8	2.3
KDM6A	6	1.4	0.2	0.9	0	4.5	2.1	1	1.2	2.5	0.7	0.5	1.6	3.9	23.3	0	2.8	3.9	2.1	1.8	2.4
KDR	6	12.2	0.5	1.4	0	0.9	1.1	3.1	0.3	0	0	2.3	3.9	5.4	5	0.4	1.6	7.3	8.7	1.8	3.4
KIT	7.3	3.4	0	0.7	0	0	0	1	0.6	0	4.9	1.4	3.4	3.6	1.9	1.3	1.6	3.4	1.7	1.8	1.9
KLF4	2.4	1.7	0	0	0	0	1.1	0.3	0.3	0	0	0.9	0.8	2.9	0.6	0	0.3	0.6	1.3	3.5	0.8
KRAS	21.8	1.4	0.5	0.7	0	1.8	0	0.7	0.6	2.5	3.5	0.5	34.5	10	1.9	0.9	0.3	1.1	28.1	57.9	8
MAP2K1	1.6	5.4	0	0.2	0	0	0	0.3	0.1	0	0	0	2.6	2.5	0.6	0	0.9	1.1	1.9	0	1.1
MAP2K2	1.6	1.7	0	0	0	0.9	0	0	0.1	0	0	0.5	0.3	0.7	0	0	0.9	0.6	0.2	0	0.4
MAP2K4	3.2	1.7	0	0	0	0	0.5	0	4.1	0	0	0	3.4	2.5	0.6	0	0.3	0.6	1.5	1.8	1.5
MAP3K1	11.3	1.4	0.5	1.4	0	0	0	2.1	7.4	2.5	0	0.9	1.8	4.7	2.5	0	0.9	1.7	1.2	1.8	2.7
MAX	4	0	0	0.7	0	0	0	0.7	0.1	2.5	0.7	1.4	0.8	1.1	0	0	0.3	0	1.5	0	0.7
MED12	12.9	4.7	0	1.4	0	0.9	2.1	2.1	2.5	5	0.7	1.8	4.4	4.3	7.5	0.4	4.4	3.4	6.9	10.5	3.6
MEN1	3.6	1.7	0.2	0.2	0	0	0.5	1	0.5	0	0	0.5	1.3	2.2	0.6	0	0.6	1.7	1	1.8	0.9
MET	6	7.4	0.2	0.9	0	8	0.5	0.7	0.6	0	0	1.4	2.1	2.5	3.8	1.3	0.6	2.2	4.2	1.8	2.2
MLH1	3.2	1.4	0	0.2	0	0.9	0	0	0.4	5	0	1.4	3.4	2.2	1.9	0.9	1.3	1.1	3.1	1.8	1.3
MLL	9.3	12.8	1.5	1.9	0	0	0.5	1	1.2	2.5	0	1.8	7.3	10.4	10.7	1.3	4.7	2.8	5.6	3.5	4.1
MLL2	16.9	15.5	0	3.6	3	8	3.2	1.7	1.7	17.5	0.7	2.7	9.6	19.7	25.2	0.4	17.9	20.2	9.1	8.8	8.1
MLL3	14.9	16.6	1	6.4	3	4.5	3.2	3.4	8	20	0.7	2.7	7.5	16.1	22	1.7	8.8	16.9	17.5	8.8	9.1
MPL	2.8	2.7	0	0.2	0	0	0	1	0.1	0	0.7	0.5	1.3	0.4	0.6	0	0.6	1.1	0.8	0	0.7
MSH2	4	1	0.2	0.5	0	0	0	0	0.5	2.5	0	0.9	3.1	3.2	1.9	0	0.3	0.6	1.9	1.8	1.1
MSH6	6.9	1.7	0	0.9	0	0.9	0.5	1	0.5	0	0	0.5	3.9	4.3	2.5	0.4	0.3	2.2	2.3	1.8	1.6
MYD88	1.2	1	0	0	0	0	0	0	0	0	0	0	0.3	0.7	0.6	0	0	0	0.6	0	0.2
...																					
TNFRSF14	1.2	1	0	0.2	0	0	0	0	0	5	0	0.5	0.3	0.4	0	0	0	0	0.2	0	0.2
TP53	29	14.5	0.7	2.6	33.3	1.8	8.5	27.8	33.2	5	5.6	52.7	56.7	48.4	48.4	84.8	69.3	79.2	51.6	64.9	36.1
TRAF7	2	2.4	0	0.5	1.5	0	0	0	0.5	0	0	0.5	0.5	3.6	0.6	0.4	0.3	0	0.8	1.8	0.8
TSC1	5.6	1.7	0	1.2	3	1.8	0.5	0	0.5	2.5	0	0.9	2.6	2.9	6.9	0	0.6	2.8	2.1	3.5	1.6
TSC2	5.2	4.1	0	0.9	3	3.6	0.5	1.4	0.3	2.5	0	0.9	1.8	3.2	1.9	0.9	1.3	2.8	2.1	1.8	1.6
TSHR	3.2	3	0.5	0	0	0	0.5	0.3	0.5	2.5	0	1.4	2.8	3.6	1.9	0.4	0.3	2.8	4	1.8	1.5
U2AF1	2	0	0	0	0	0	0	0	0.1	0	4.2	0.5	0.3	1.4	0.6	0	0.6	0	2.3	1.8	0.6
UBR5	10.9	7.4	0.2	2.1	4.5	0	1.1	1	1.2	5	0	0.9	6	15.6	5.7	0.4	3.8	5.6	6	3.5	4
VHL	1.2	1.4	0	56.4	0	1.8	0.5	0	0	0	0	0	0.5	0.4	0	0	0	0.6	0.2	0	4.8
WT1	1.2	2.4	0.2	0.7	0	0	0	0.7	0.3	0	7	0.5	2.1	1.4	1.3	0	0	2.2	2.7	0	1.1
XPO1	5.6	0.7	0	1.9	0	0.9	0	0	0.5	2.5	0	0.5	1.8	1.4	1.9	0	0.9	1.1	1.7	1.8	1.1
ZRSR2	2.4	0.7	0	0	0	0	0.5	0.7	0.1	0	0	0.5	0.3	0.4	1.3	0	0.3	0	0.4	1.8	0.4

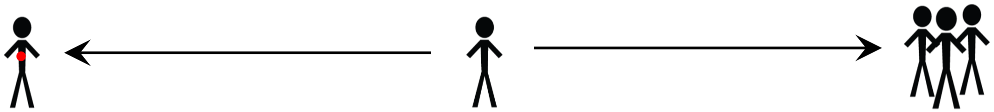
Source: Somatic mutation in cancer and normal cells. Martincorena I and Campbell PJ, Science. 2015 Sep 25;349(6255):1483-9.

Summary: Human Genome Variation

A Cancer Genome
Somatic Variants

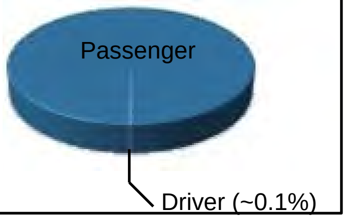
A Typical Genome
Germline Variants

Population of 2,504 peoples
Germline Variants



Origin of Variants

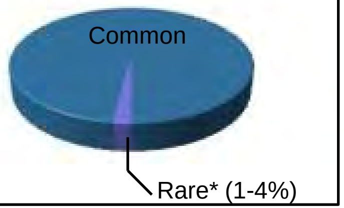
	Coding	Non-coding
Germline	22K	4.1 – 5M
Somatic	~50	5K



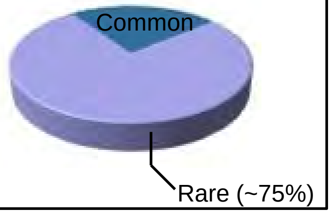
Class of Variants

SNP	3.5 – 4.3M
Indel	550 – 625K
SV	6 – 8K
Total	4.1 – 5M

Prevalence of Variants



SNP	84.7M
Indel	3.6M
SV	60K
Total	88.3M



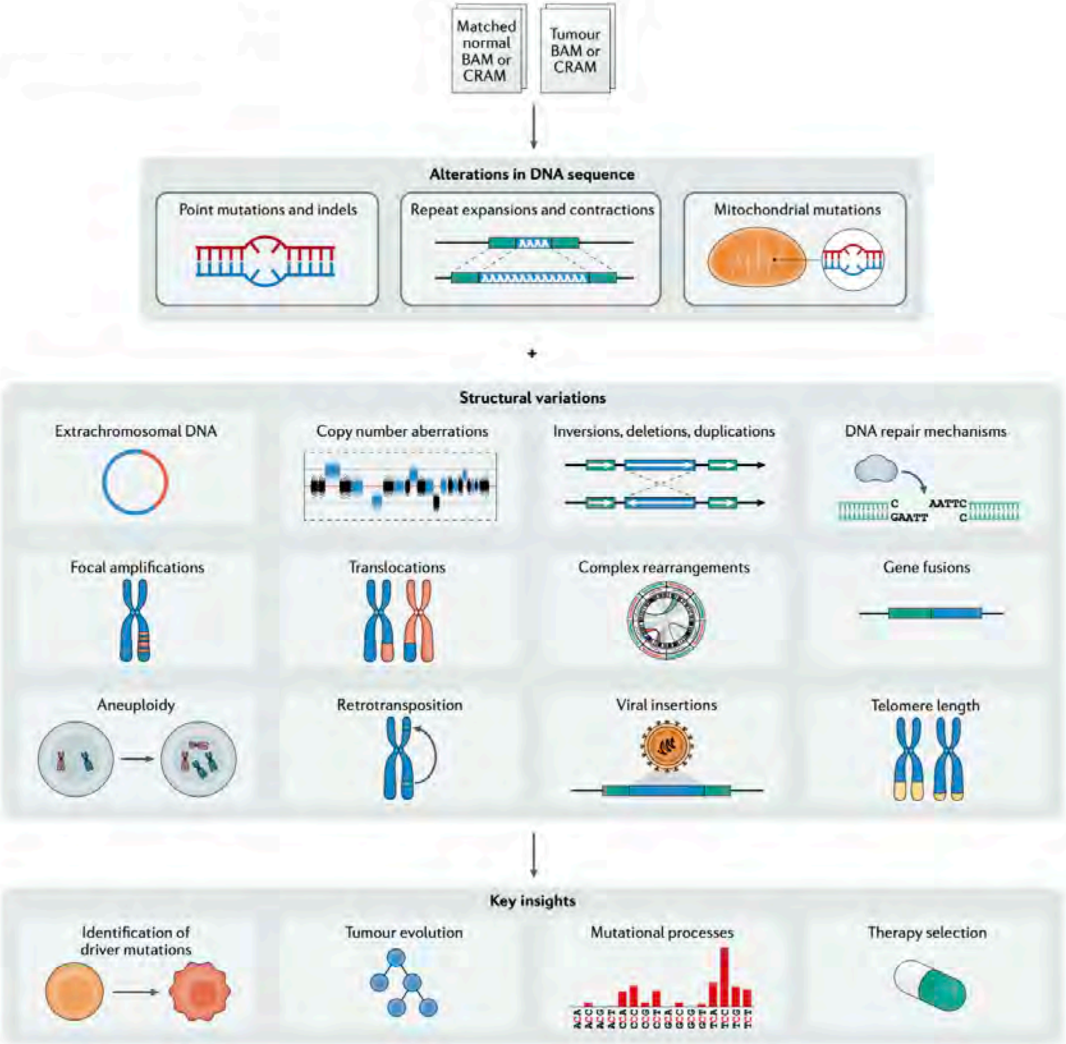
* Variants with allele frequency < 0.5% are considered as rare variants in 1000 genomes project.

Source: Mark Gerstein, http://lectures.gersteinlab.org/summary/, https://doi.org/10.1038/nature15394

Tumor-only sequencing considerations

- Problem: A relatively small number of somatic variants is “hidden” in a large set of germline variants (4-5 Million)
- Study design options
  - Using an unmatched control
    - Rare variants appear as “somatic”
  - Using a panel of normal genomes
    - Fewer rare variants appear as “somatic”
  - Using germline variant catalogs (1000 Genomes, ExAC, gnomAD)
    - Catalogs are often highly curated, i.e., false positive variant calls from your analysis are likely not present and thus, still called as “somatic”
- Tumor-mutational burden (TMB) is often >2-fold over-estimated using tumor-only sequencing

Cancer Genome Data Analysis



Source: Cortés-Ciriano, I., Gulhan, D.C., Lee, J.JK. et al. Computational analysis of cancer genome sequencing data. Nat Rev Genet **23**, 298–314 (2022)

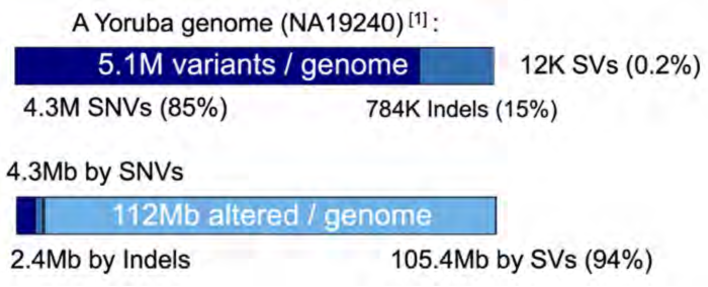
Structural and copy-number variants

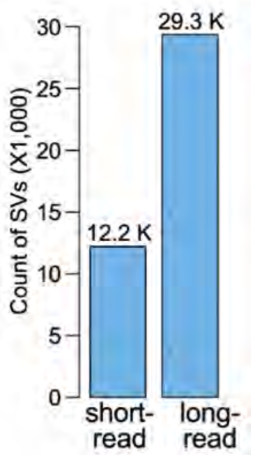
Structural Variants in Numbers

Germline structural variants in Population Genomics

Total numbers:

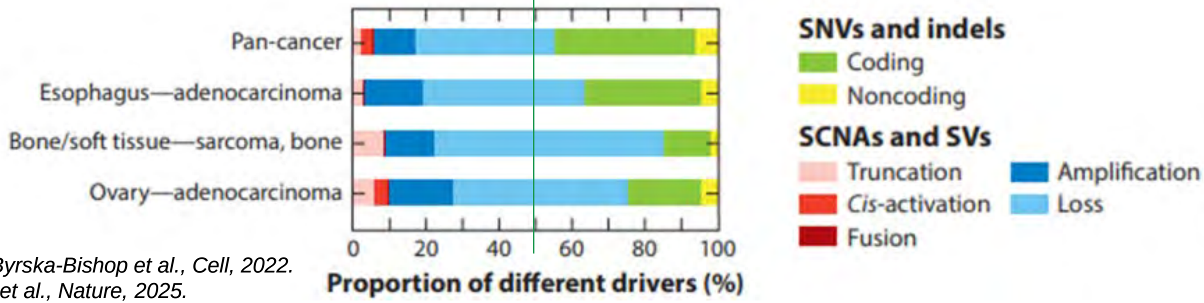
Affected base pairs:





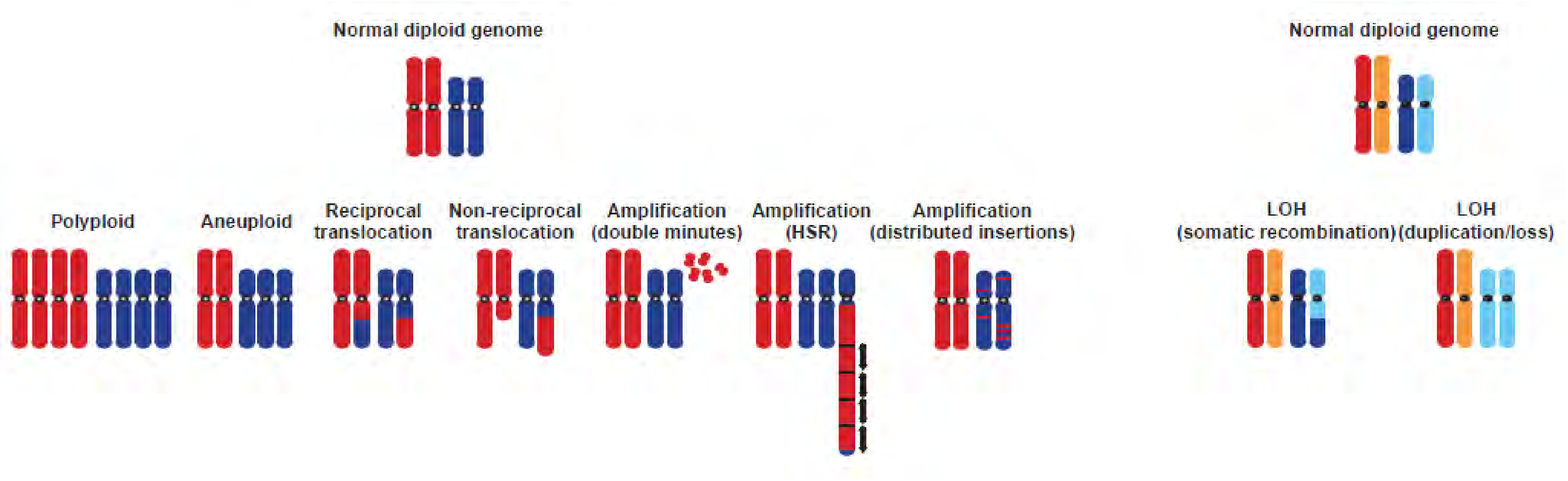
Somatic structural variants in Cancer Genomics

>50% of drivers are SVs or SCNAs



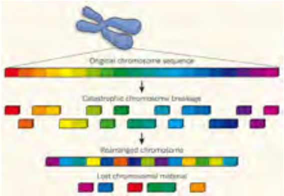
Sources: Pan-cancer analysis of whole genomes, Nature, 2020. Byrska-Bishop et al., Cell, 2022. Cosenza et al., Annu Rev Genomics Hum Genet. 2022. Logsdon et al., Nature, 2025.

Wide Range of Chromosomal Aberrations in Cancer



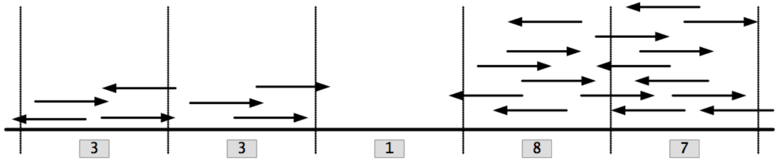
Source: Chromosome aberrations in solid tumors. Albertson et al., Nat Genet. 2003 Aug;34(4):369-76.

Somatic structural variant calling


Original chromosome sequence
Catastrophic chromosome breakage
Rearranged chromosome
Lost chromosomal material

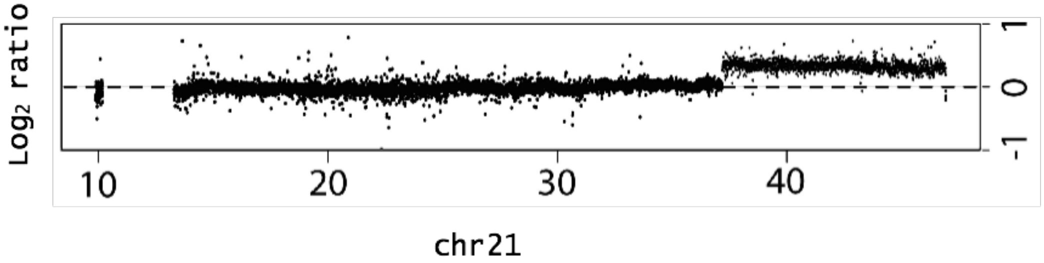
Read-depth

- Read counting in windows for tumor and normal data

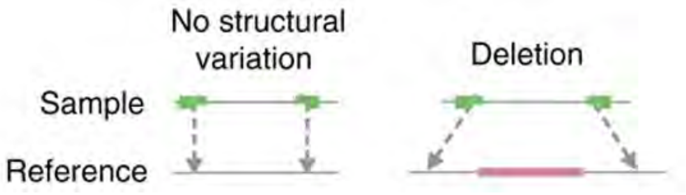


- Log2 ratio for each window
- Chromosome-wide plot

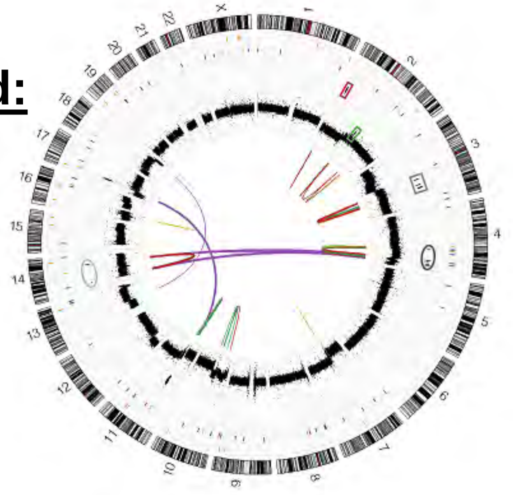
$$\log_2 \frac{\#\text{Reads}_{\textit{Disease}}}{\#\text{Reads}_{\textit{Normal}}}$$



Paired-end / Split-reads

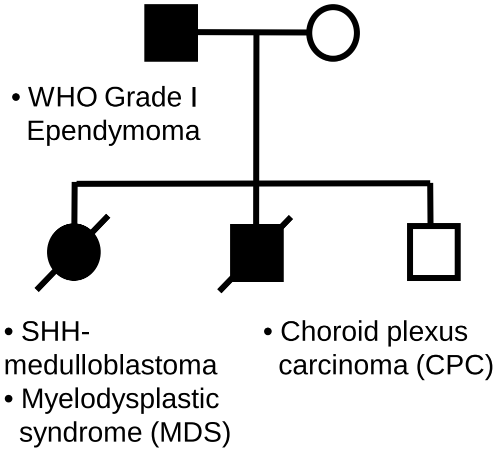


Combined:



Childhood Brain Tumor
Medulloblastoma

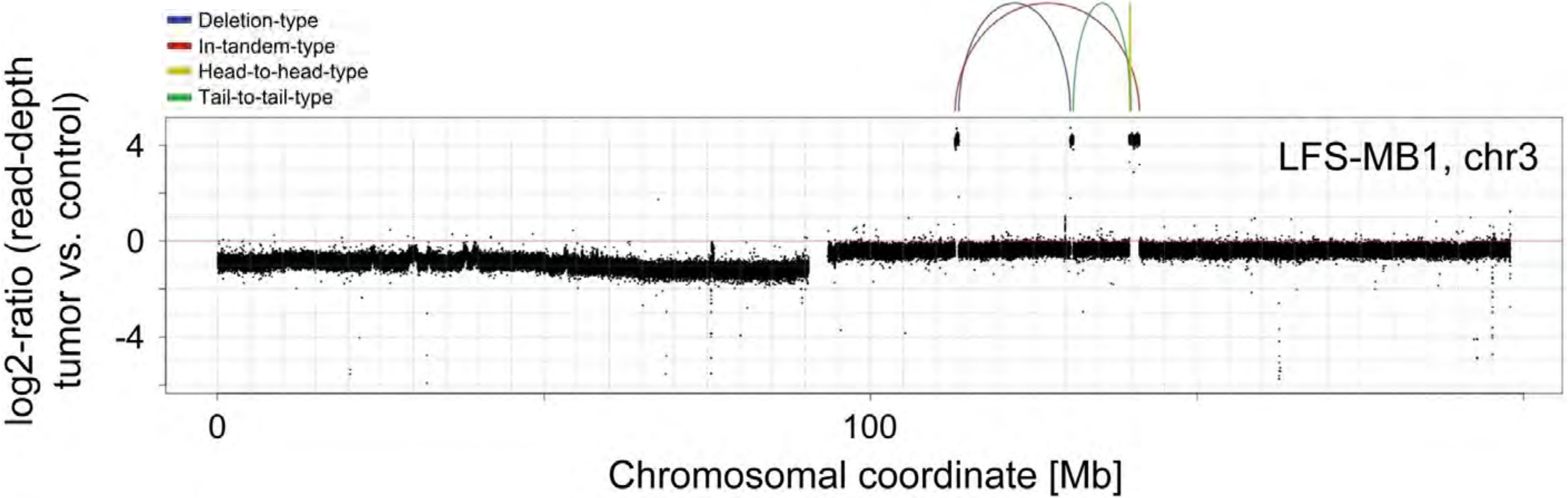
- Li-Fraumeni syndrome
  - Germline TP53 mutation



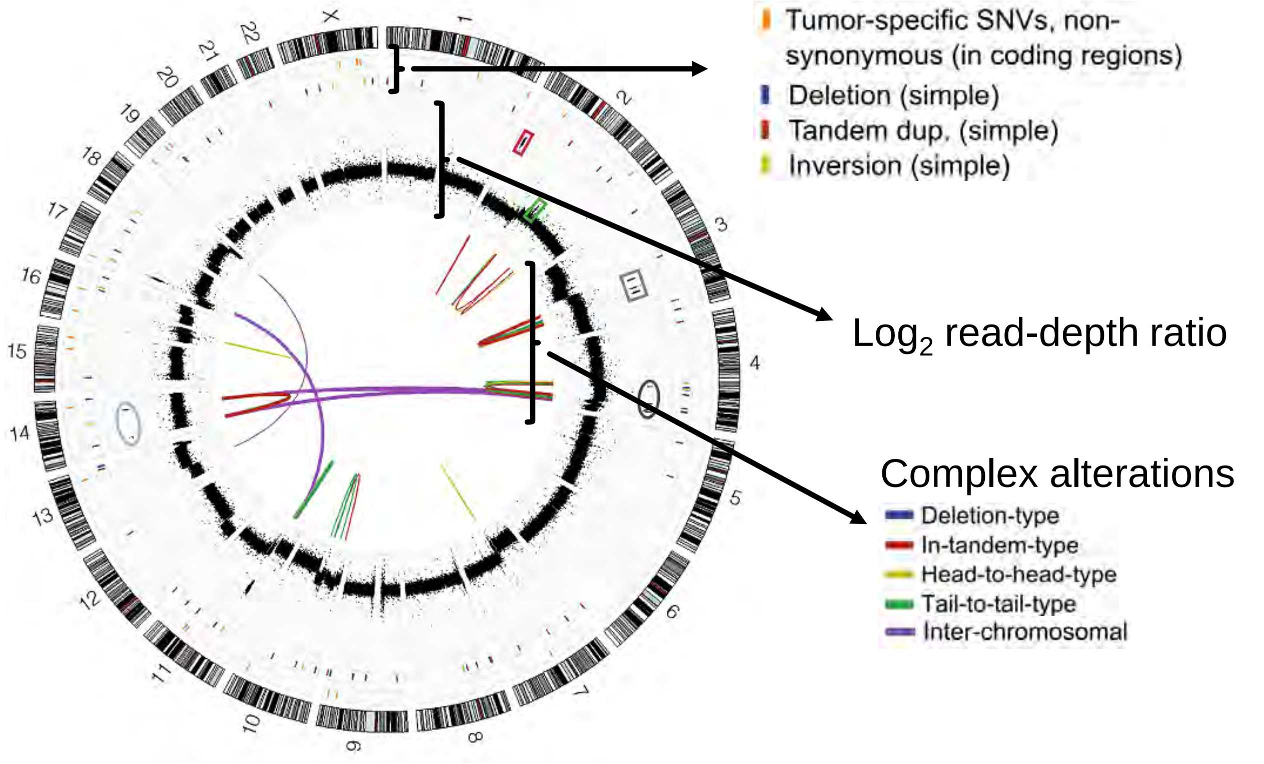
- WHO Grade I Ependymoma

- SHH-medulloblastoma
- Myelodysplastic syndrome (MDS)

- Choroid plexus carcinoma (CPC)



Somatic DNA alterations

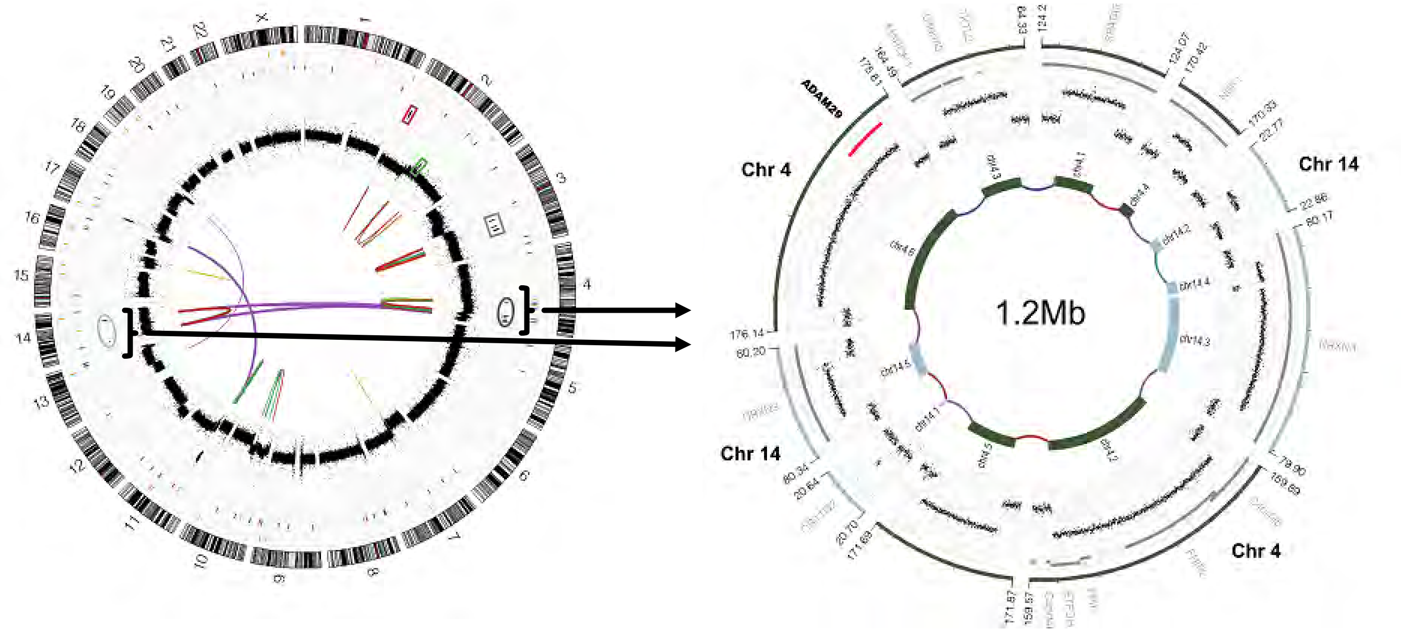


Log_2 read-depth ratio

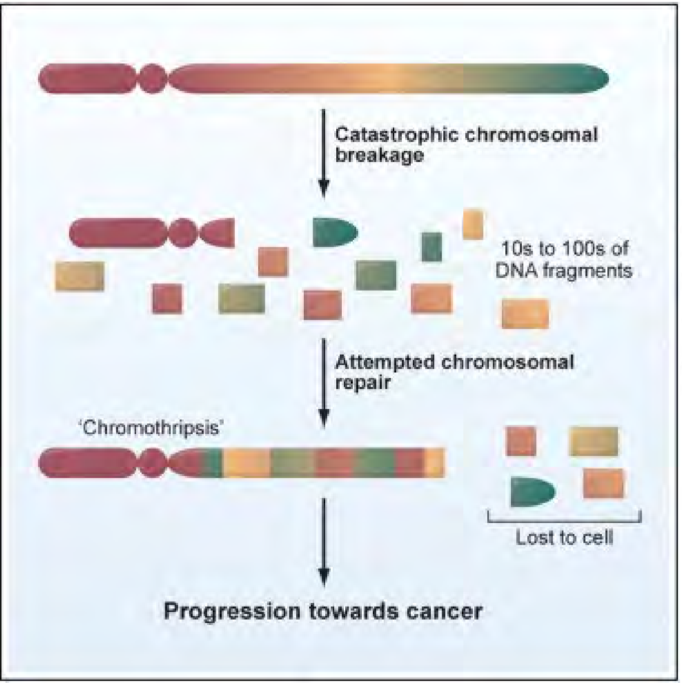
Complex alterations

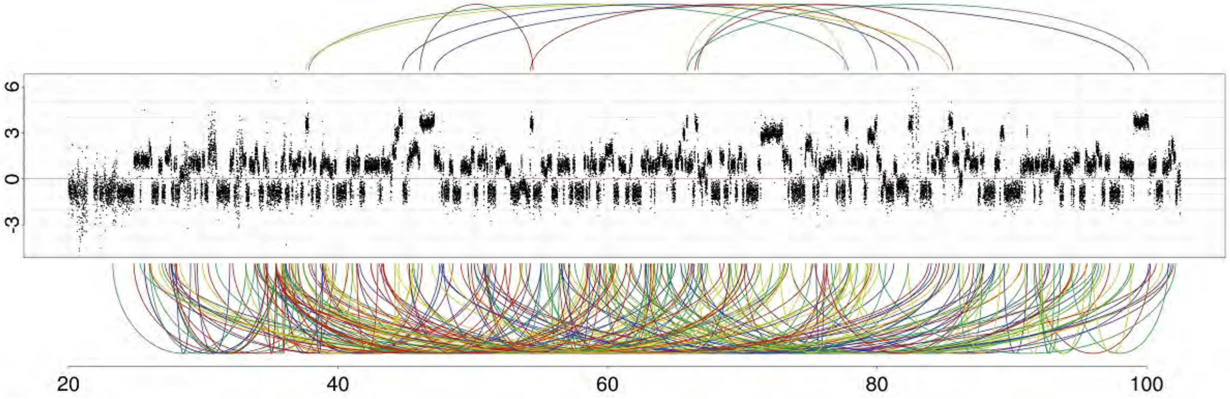
Source: Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. Rausch et al., Cell. 2012 Jan 20;148(1-2):59-71.

Extra-chromosomal DNA (ecDNA)



Chromothripsis

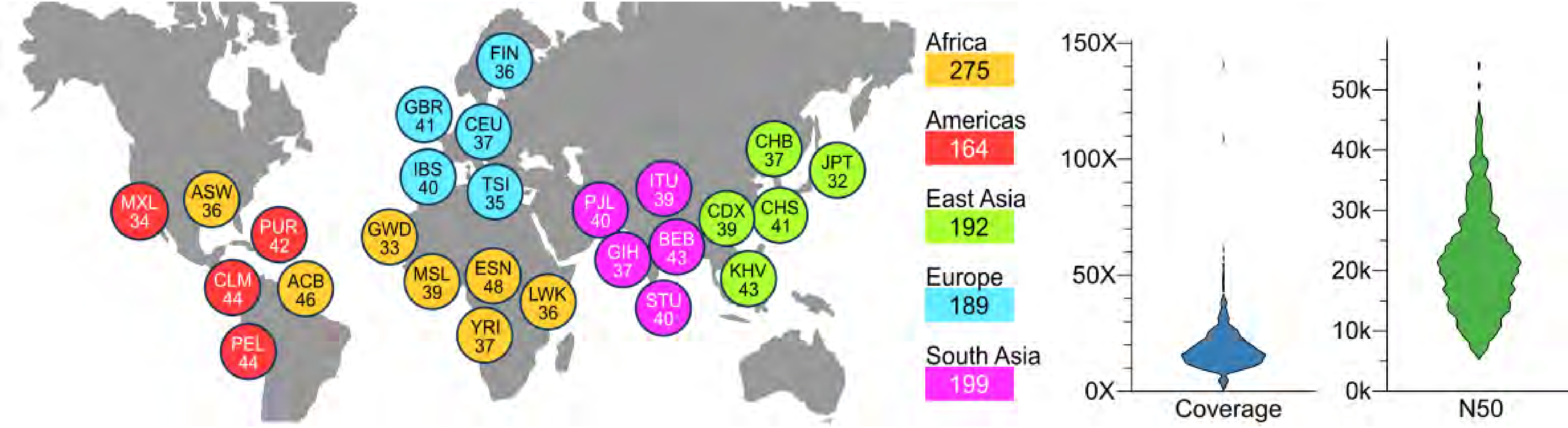

Catastrophic chromosomal breakage
10s to 100s of DNA fragments
Attempted chromosomal repair
'Chromothripsis'
Lost to cell
Progression towards cancer



Source: Stephens et al., Cell, 2011; Rausch et al., Cell, 2012

Structural variant calling using long-reads

ONT Sequencing of 1,019 samples from the 1000 Genomes Project



1,019 samples sequenced with ONT

- ~15x coverage
- **Structural variant calling** using long-reads

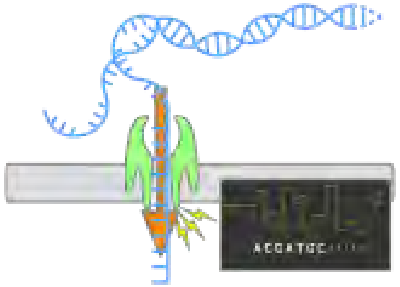
Schloissnig, Pani et al., Nature, 2025.

Pan-genomes and long-reads for SV discovery

Short-reads: 100bp-300bp

2 to 3-fold increase in detected SVs

Long-reads: N50 read length >15,000bp



Nanopore sequencing

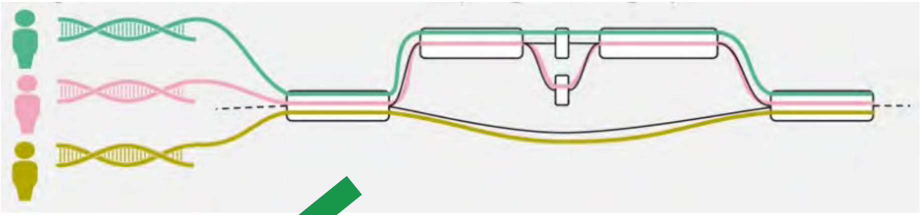
Linear reference genome (GRCh38)

+ ~200Mbp

CHM13 (T2T)

+ ~100Mbp

Graph pan-genome



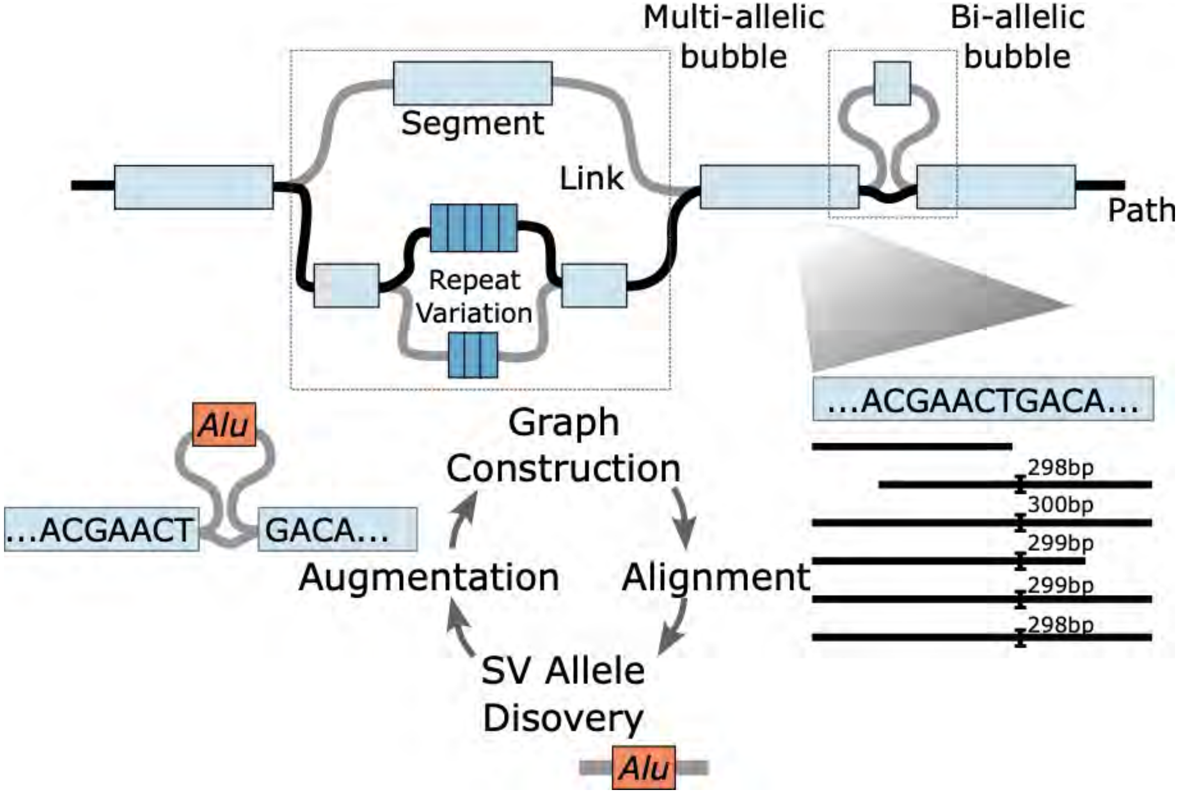
Human Pangenome Reference Consortium 44 samples

?

What have we missed with short-reads and a single reference genome?

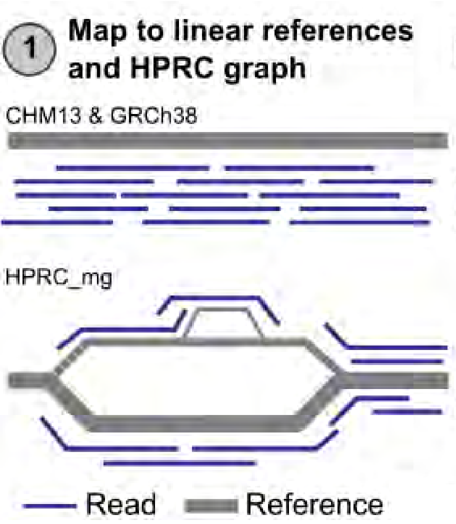
EMBL

SV Analysis by Graph Augmentation (SAGA)

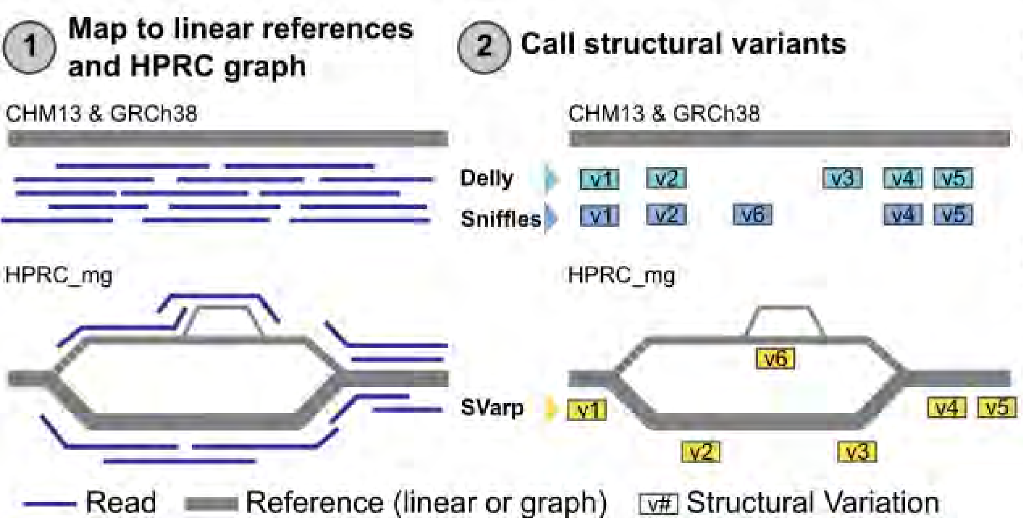


Source: Rausch T, Marschall T, Korbel JO. The impact of long-read sequencing on human population-scale genomics. Genome Res. 2025.

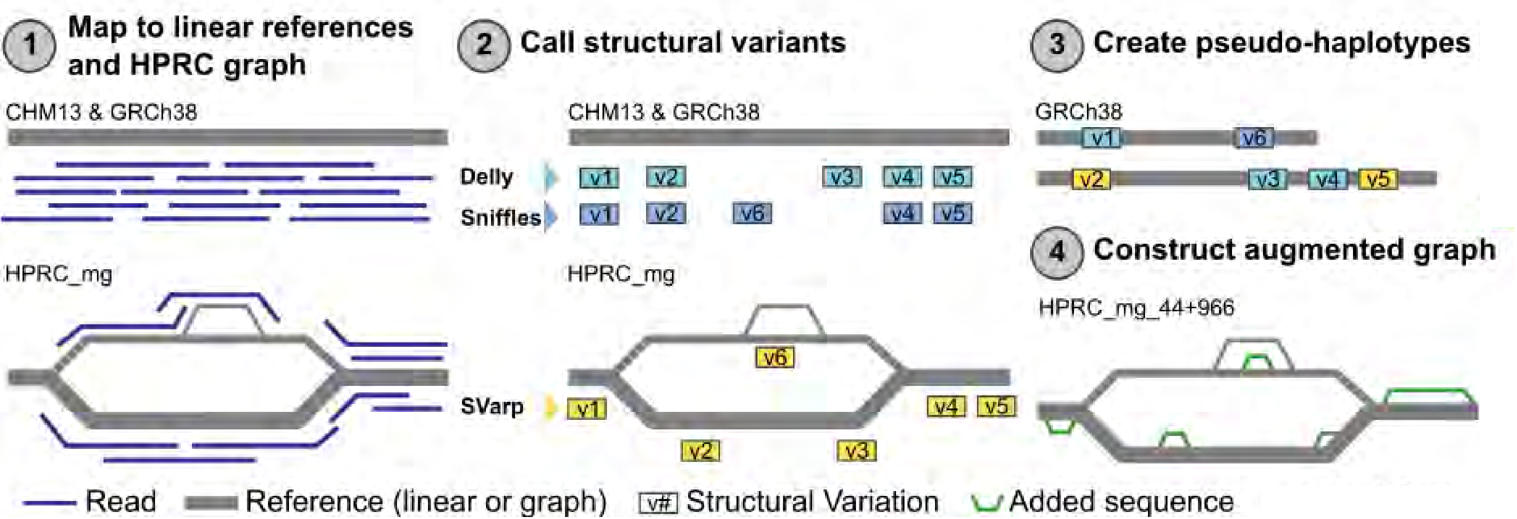
SAGA: SV Analysis by Graph Augmentation



SAGA: SV Analysis by Graph Augmentation

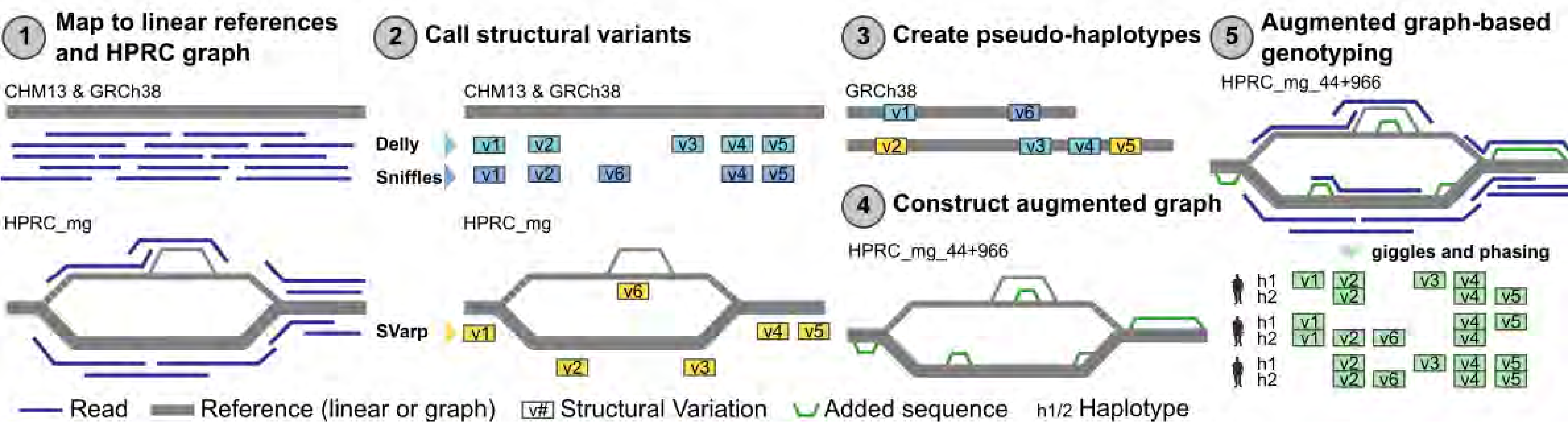


SAGA: SV Analysis by Graph Augmentation



1 Map to linear references and HPRC graph
CHM13 & GRCh38
HPRC_mg
2 Call structural variants
CHM13 & GRCh38
Delly
Sniffles
v1
v2
v3
v4
v5
v6
HPRC_mg
SVarp
3 Create pseudo-haplotypes
GRCh38
4 Construct augmented graph
HPRC_mg_44+966
Read
Reference (linear or graph)
v# Structural Variation
Added sequence

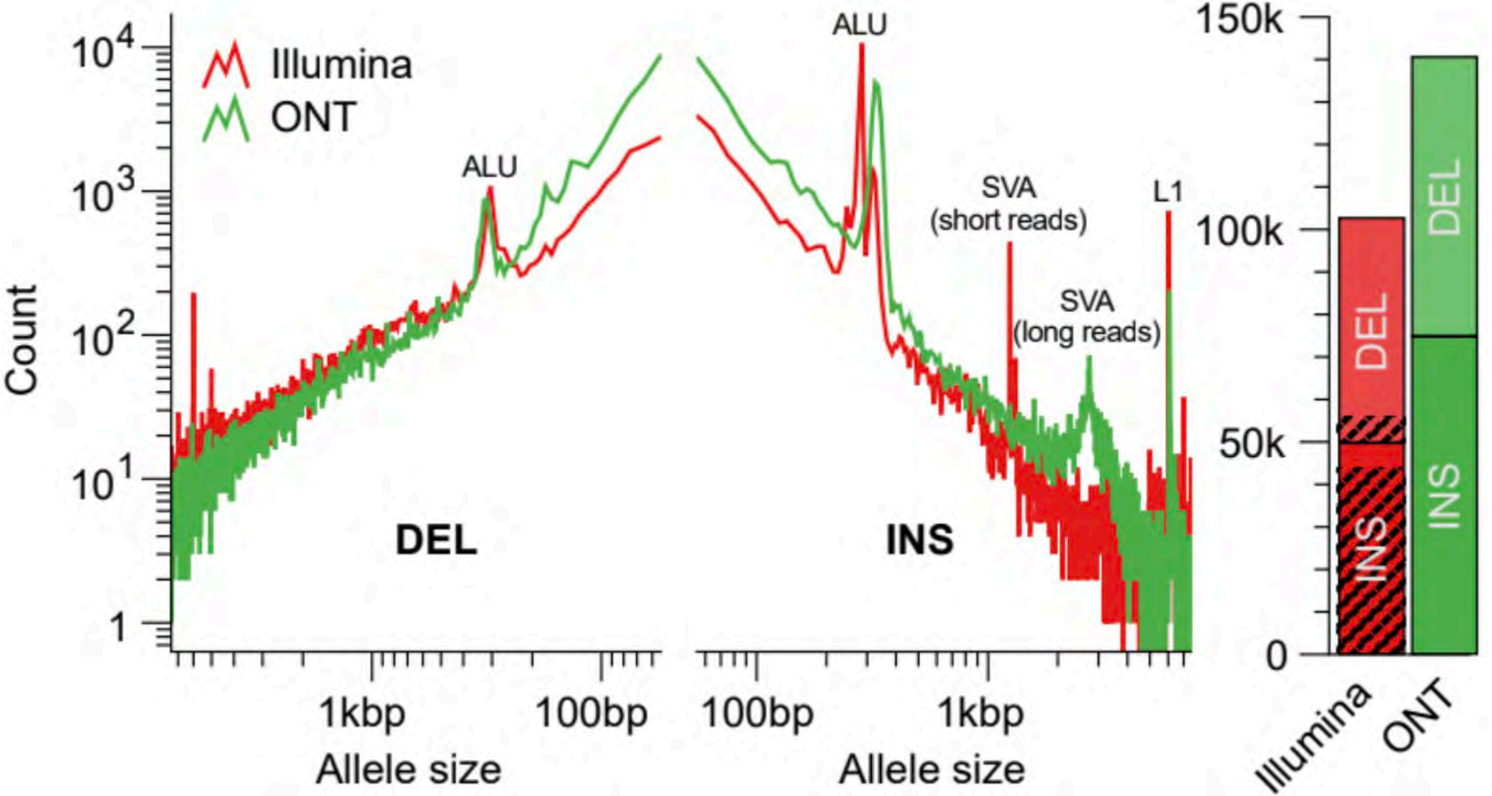
SAGA: SV Analysis by Graph Augmentation



Schloissnig et al., Nature, 2025.

Long-reads facilitate the discovery of sequence-resolved insertions

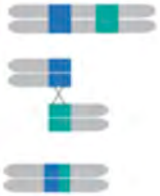
Byrska-Bishop et al. Cell 2022
for comparison

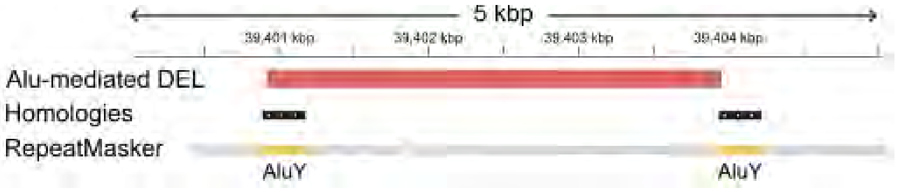


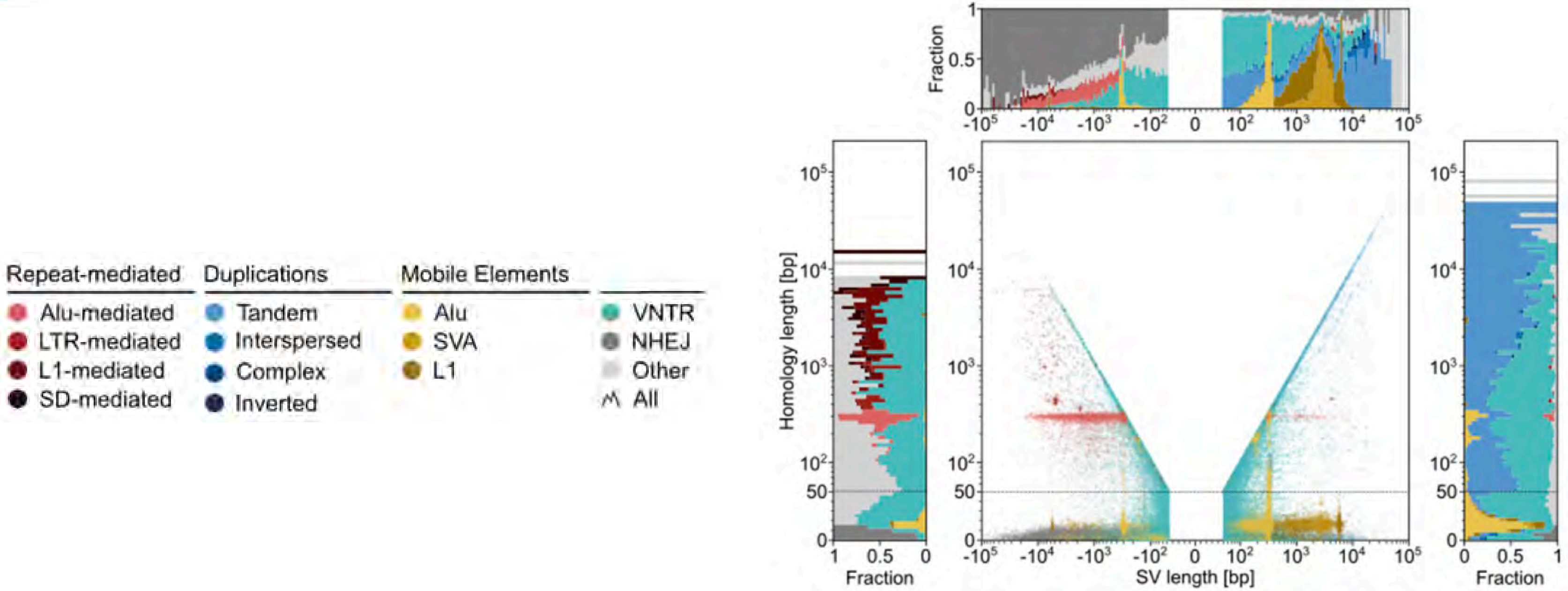
Source: Schloissnig et al., Nature, 2025.

Repeat mediated SVs

- NAHR – Non-allelic homologous recombination







Cancer Genomics using long-reads
Deciphering haplotype-resolved complex rearrangements

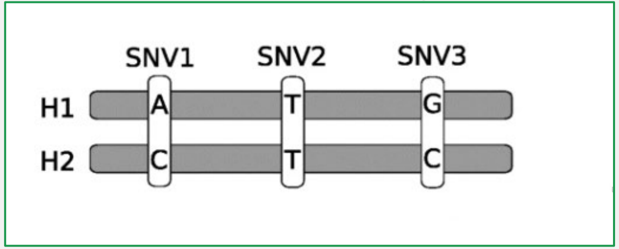
Sample

Li-Fraumeni syndrome
Sonic Hedgehog
Medulloblastoma patient

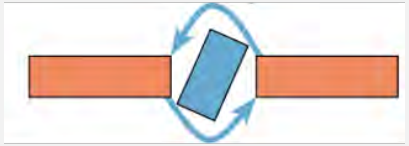
Tissues

Tumor at diagnosis
Tumor post treatment
Blood

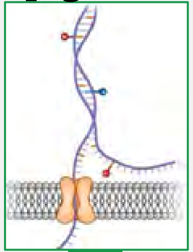
Haplotyping

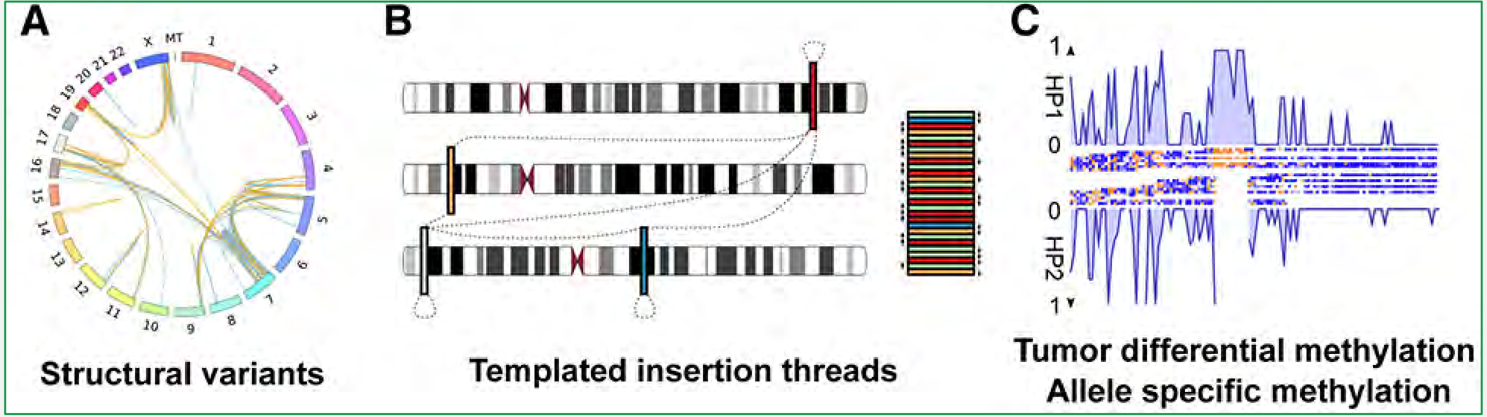


Genetics



Epigenetics





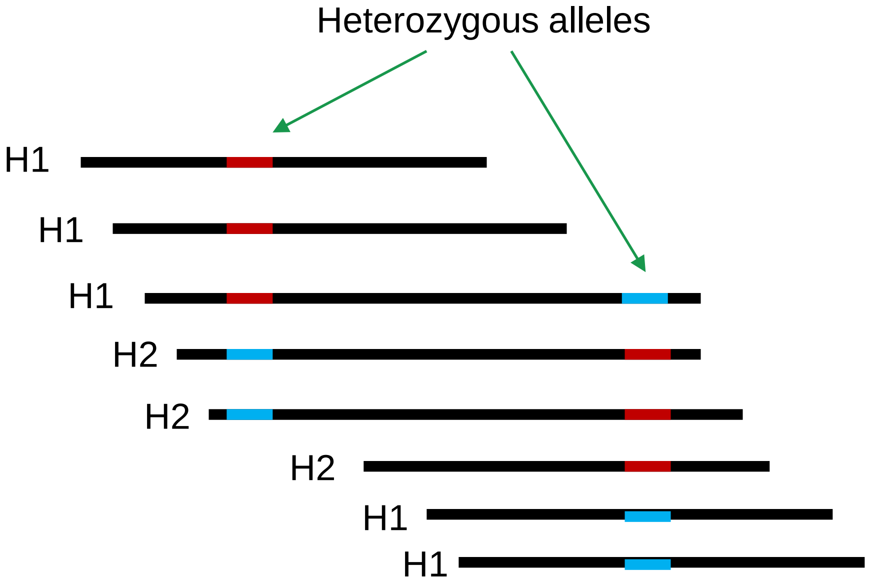
Structural variants

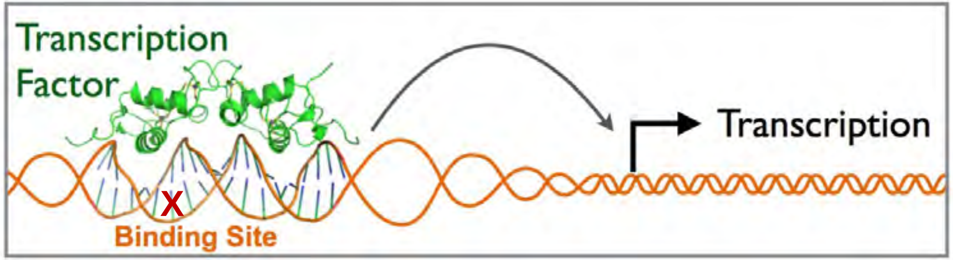
Templated insertion threads

Tumor differential methylation
Allele specific methylation

Source: Rausch et al., Cell Genomics, 2023.

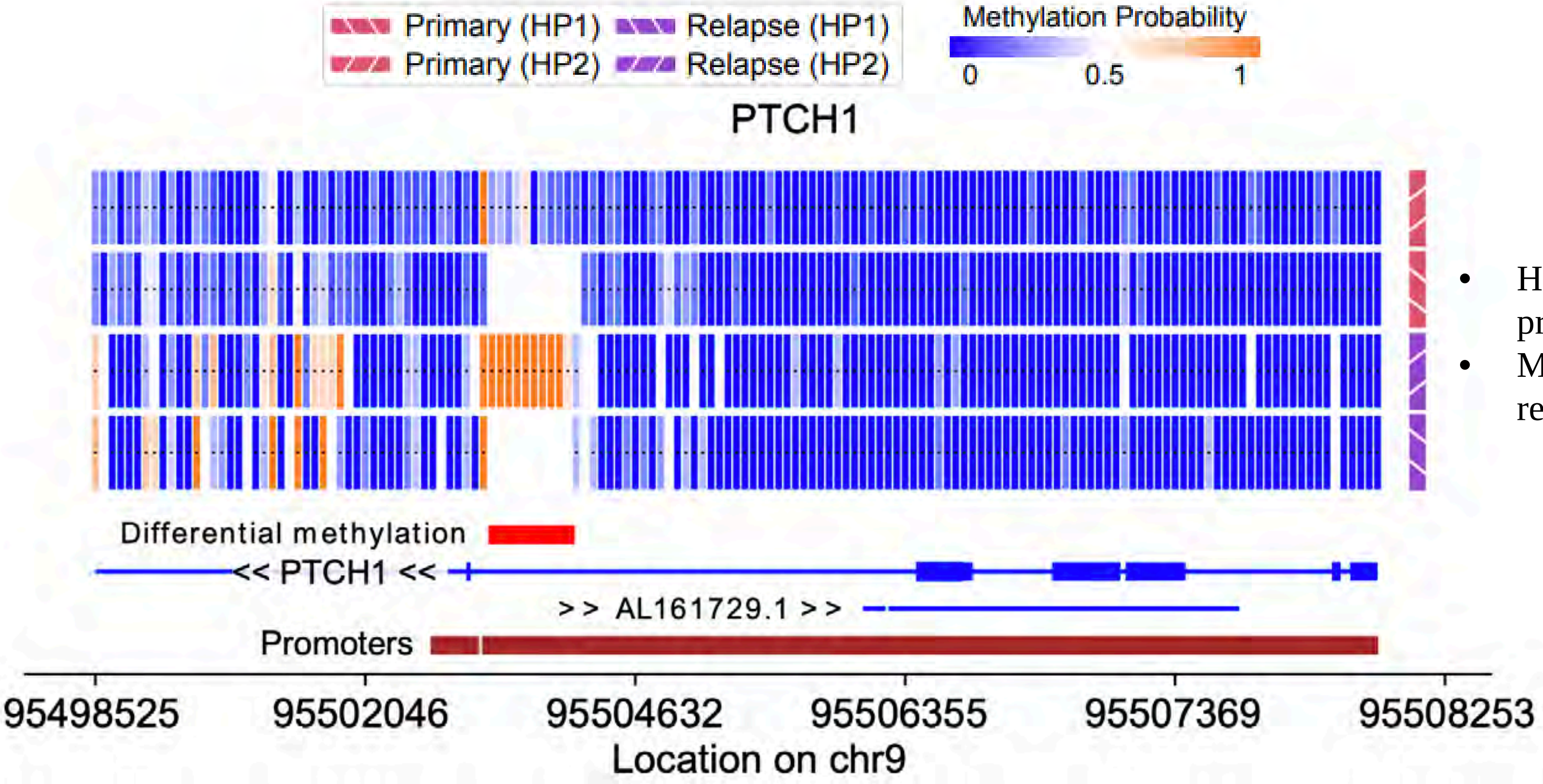
Haplotype-resolved genome analysis





- Applications
  - Analyze compound heterozygotes in rare diseases
  - Measure allele-specific expression, methylation, TF binding, etc.
  - Determine how combinations of variants uniquely situated on each haplotype may affect gene function

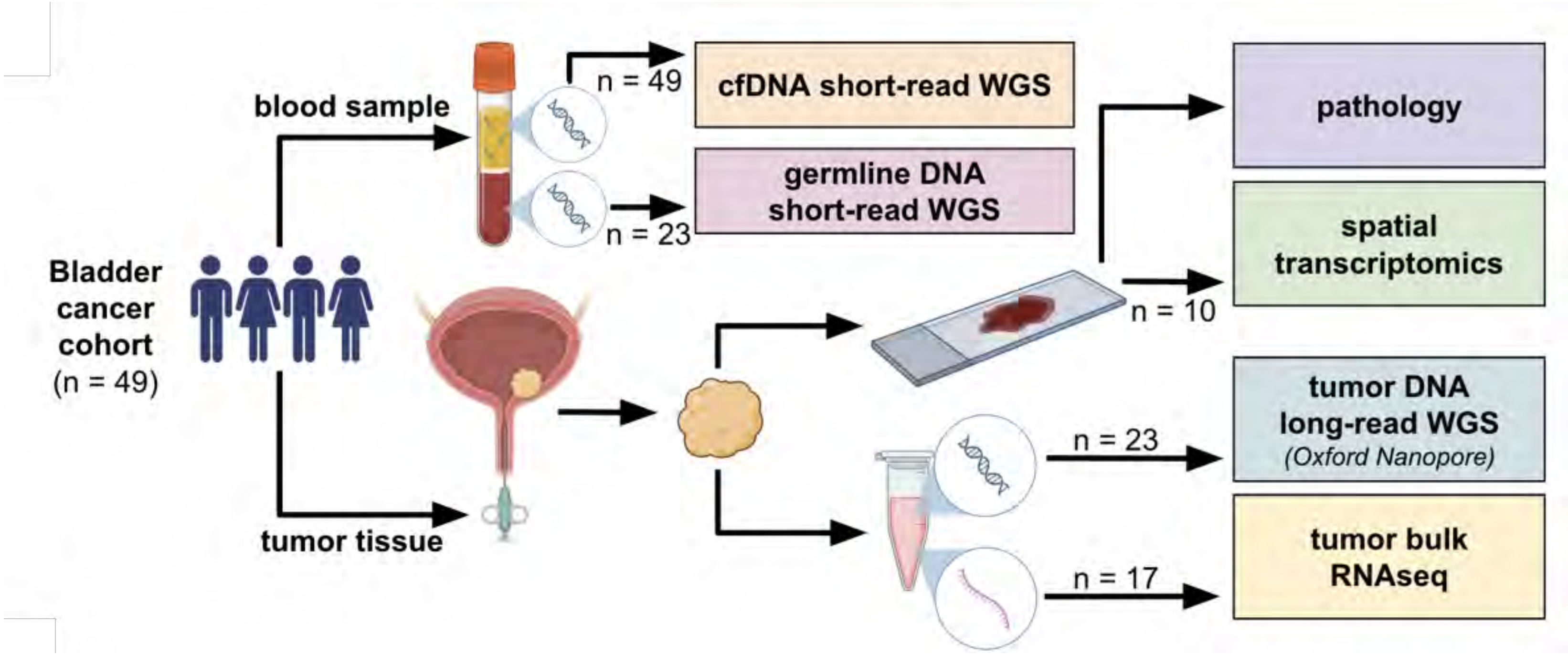
Allele-specific methylation (ASM)



- Het. deletion in the promoter of *PTCH1*
- Methylated in the relapsed tumor

Long-read WGS in bladder cancer patients

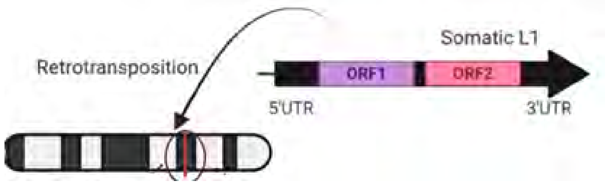
- Multi-omics and spatial analyses

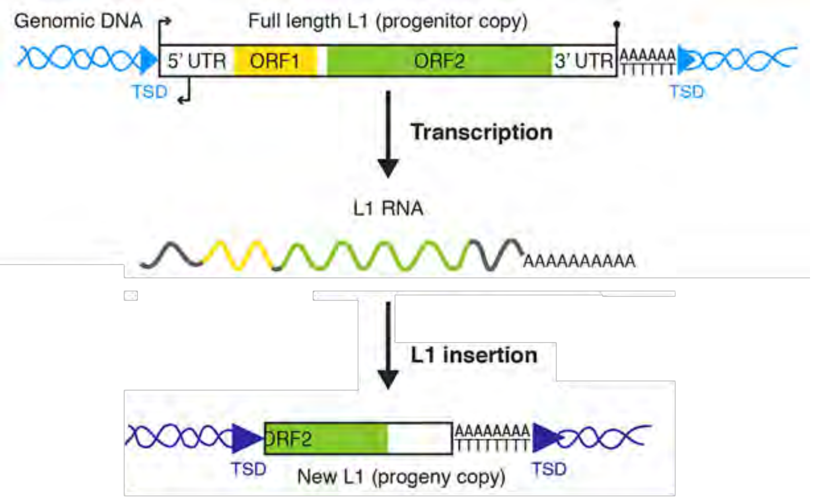


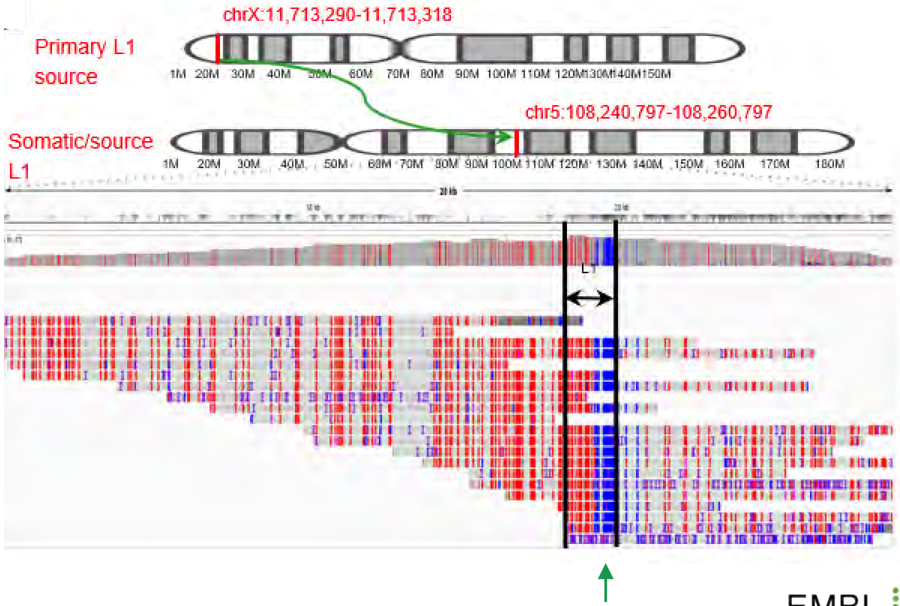
Source: Pribus et al., bioRxiv, 2025.

Frequent somatic LINE-1 (L1) insertions

- Up to >500 somatic L1 insertions per tumor
- Evidence in some samples for L1 multi-jumps

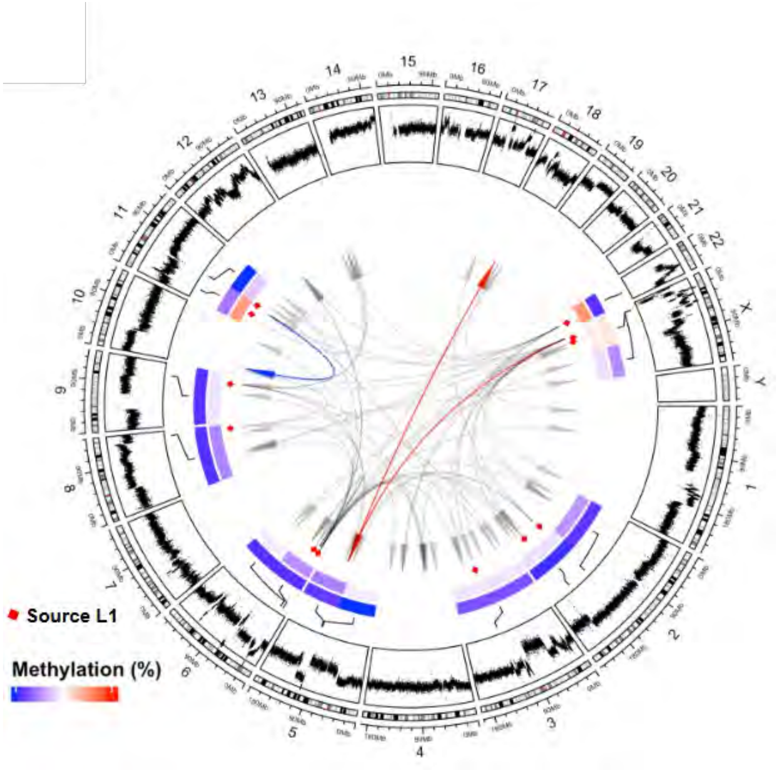


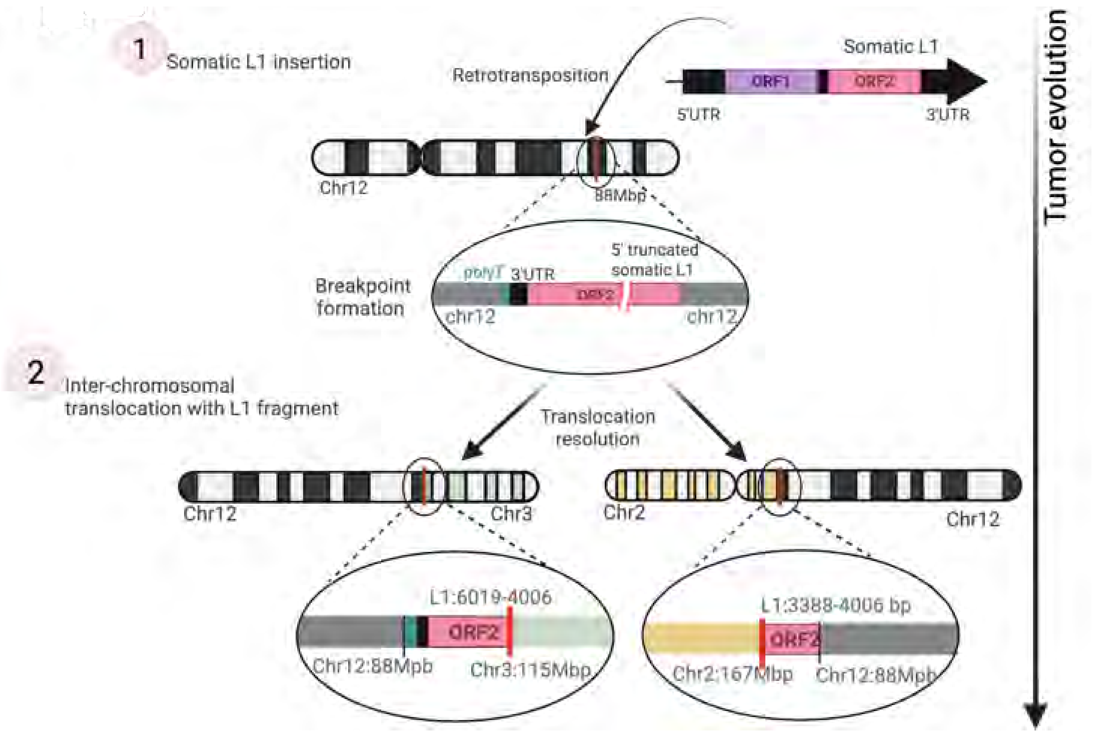



chrX:11,713,290-11,713,318
Primary L1 source
1M 20M 30M 40M 50M 60M 70M 80M 90M 100M 110M 120M130M140M150M
chr5:108,240,797-108,260,797
Somatic/source L1
1M 20M 30M 40M 50M 60M 70M 80M 90M 100M 110M 120M 130M 140M 150M 160M 170M 180M
L1

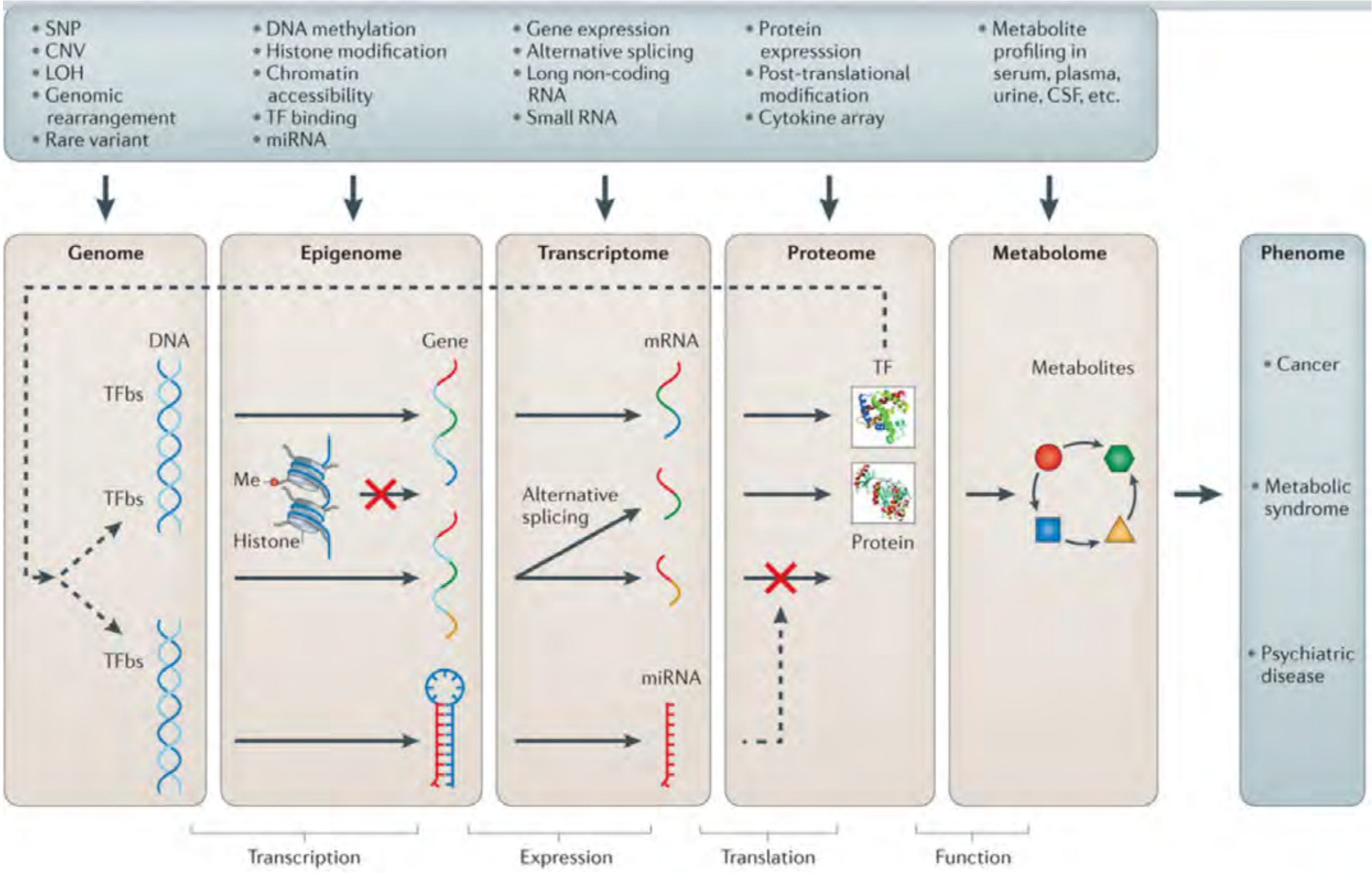
Unmethylated L1 promoter

Somatic L1s are linked with downstream genomic rearrangements and chromosomal instability



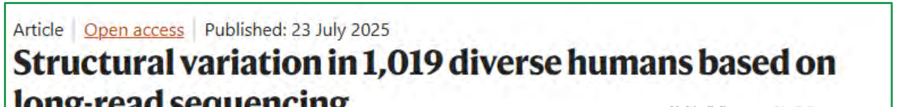


Beyond somatic driver variants



Nature Reviews | Genetics

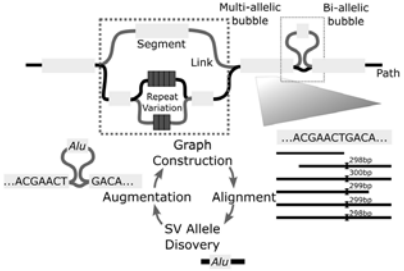
Thank you!

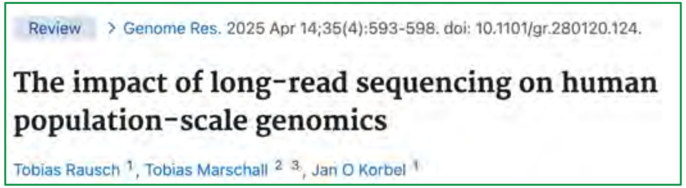


Article | Open access | Published: 23 July 2025

Structural variation in 1,019 diverse humans based on long-read sequencing

Nature **644**, 442–452 (2025) | Cite this article

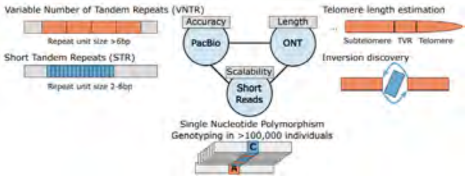




Review > Genome Res. 2025 Apr 14;35(4):593-598. doi: 10.1101/gr.280120.124.

The impact of long-read sequencing on human population-scale genomics

Tobias Rausch [1], Tobias Marschall [2] [3], Jan O Korbel [1]


Variable Number of Tandem Repeats (VNTR)
Repeat unit size >6bp
Short Tandem Repeats (STR)
Repeat unit size 2-6bp
Accuracy
Length
PacBio
ONT
Scalability
Short Reads
Telomere length estimation
Subtelomere
TVR
Telomere
Inversion discovery
Single Nucleotide Polymorphism Genotyping in >100,000 individuals
C
A

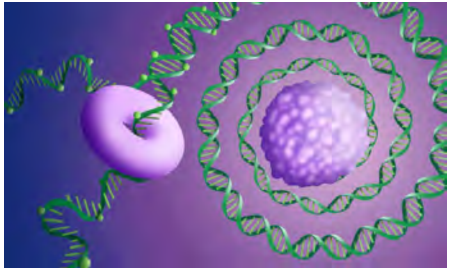
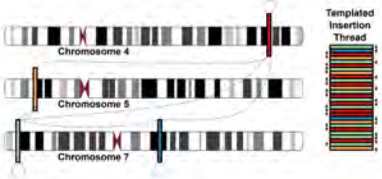
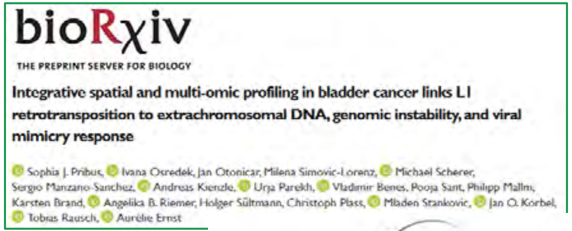


Image: Joana Carvalho/EMBL



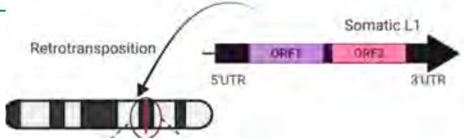


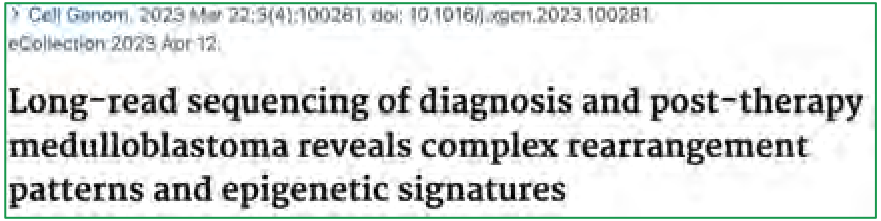
bioRχiv

THE PREPRINT SERVER FOR BIOLOGY

Integrative spatial and multi-omic profiling in bladder cancer links L1 retrotransposition to extrachromosomal DNA, genomic instability, and viral mimicry response

Sophia J. Pribus, Ivana Osredek, Jan Otonicar, Milena Simovic-Lorenz, Michael Scherer, Sergio Manzano-Sanchez, Andreas Kienzle, Urja Parekh, Vladimir Benes, Pooja Sant, Philipp Mallm, Karsten Brand, Angelika B. Riemer, Holger Sültmann, Christoph Plass, Mladen Stankovic, Jan O. Korbel, Tobias Rausch, Aurélie Ernst





> Cell Genom. 2023 Mar 22;3(4):100281. doi: 10.1016/j.xgen.2023.100281. eCollection 2023 Apr 12.

Long-read sequencing of diagnosis and post-therapy medulloblastoma reveals complex rearrangement patterns and epigenetic signatures

Many thanks to all collaborators: Marschall, Ernst, Stegle, Birney, Fröhling, Korbel labs