

wellcome
connecting
science

connecting people with science

Day 3: Driver Gene Identification and Oncoplots

29th October 2025
Khon Kaen, Thailand



Driver Gene Identification and Oncoplots

Nivedita Mukherjee

Research Scholar (PhD)

niveditamukherjee997@gmail.com

Computational Cancer Genomics Lab

Sabarinathan Radhakrishnan, Assistant Professor

National Centre for Biological Sciences (NCBS - TIFR), Bangalore

Materials partially borrowed from:

Patricia Basurto-Lozada, PhD Candidate LIIGH-UNAM

Nyasha Chambwe, Assistant Professor, Institute of Molecular Medicine, Feinstein Institutes for Medical Research

Federico Abascal, PhD Wellcome Sanger Institute



Outline

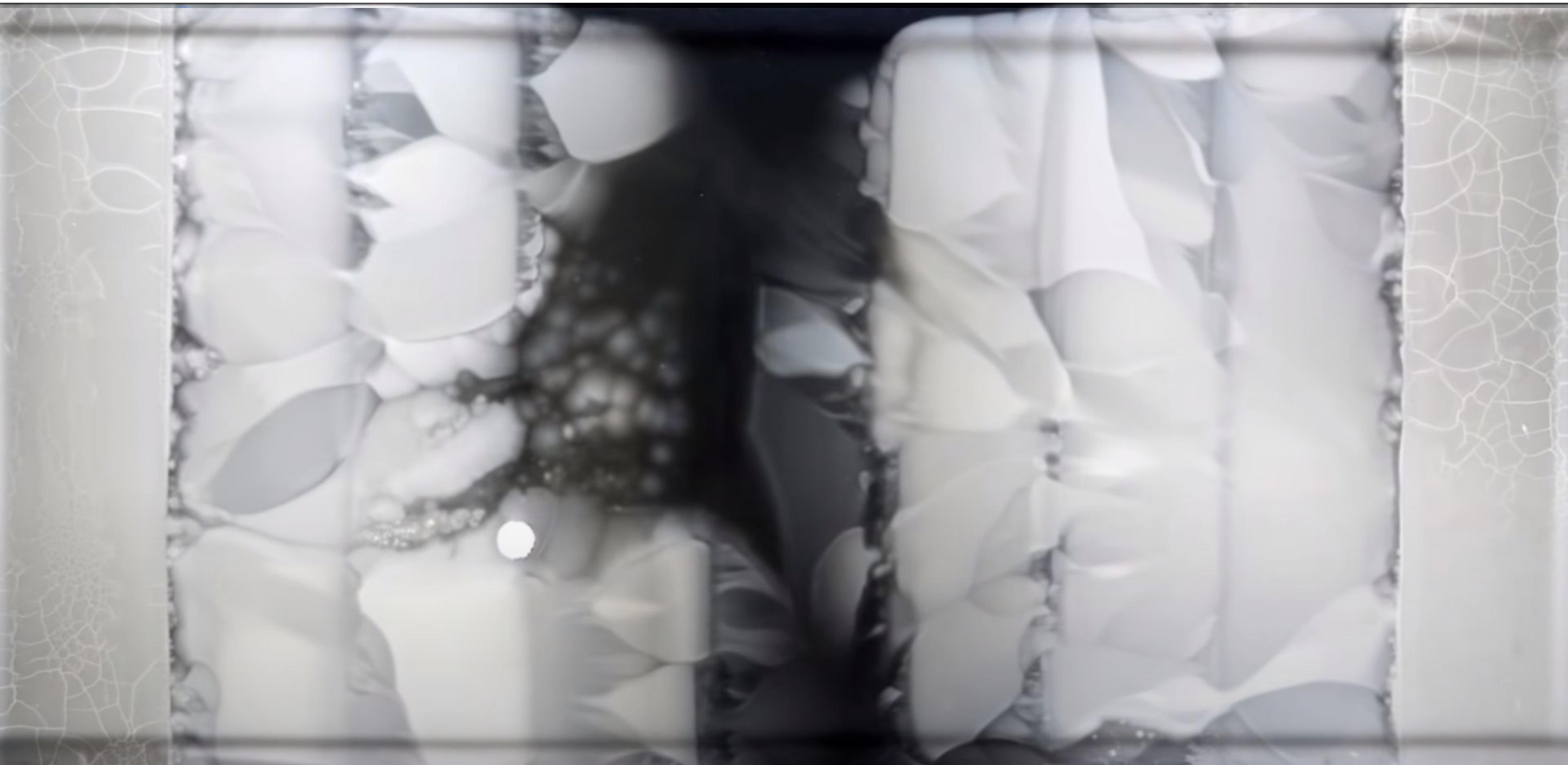
- What is a **cancer driver** gene/mutation?
- **Identification/Discovery** of cancer drivers
 - Cohort-based frequency/recurrence analysis
 - Challenges associated
- Practicals: dNdScv, Oncoplot
- **Clinical relevance** of cancer drivers
- **Timing** driver alterations during clonal evolution

What is a **cancer driver** gene/mutation?

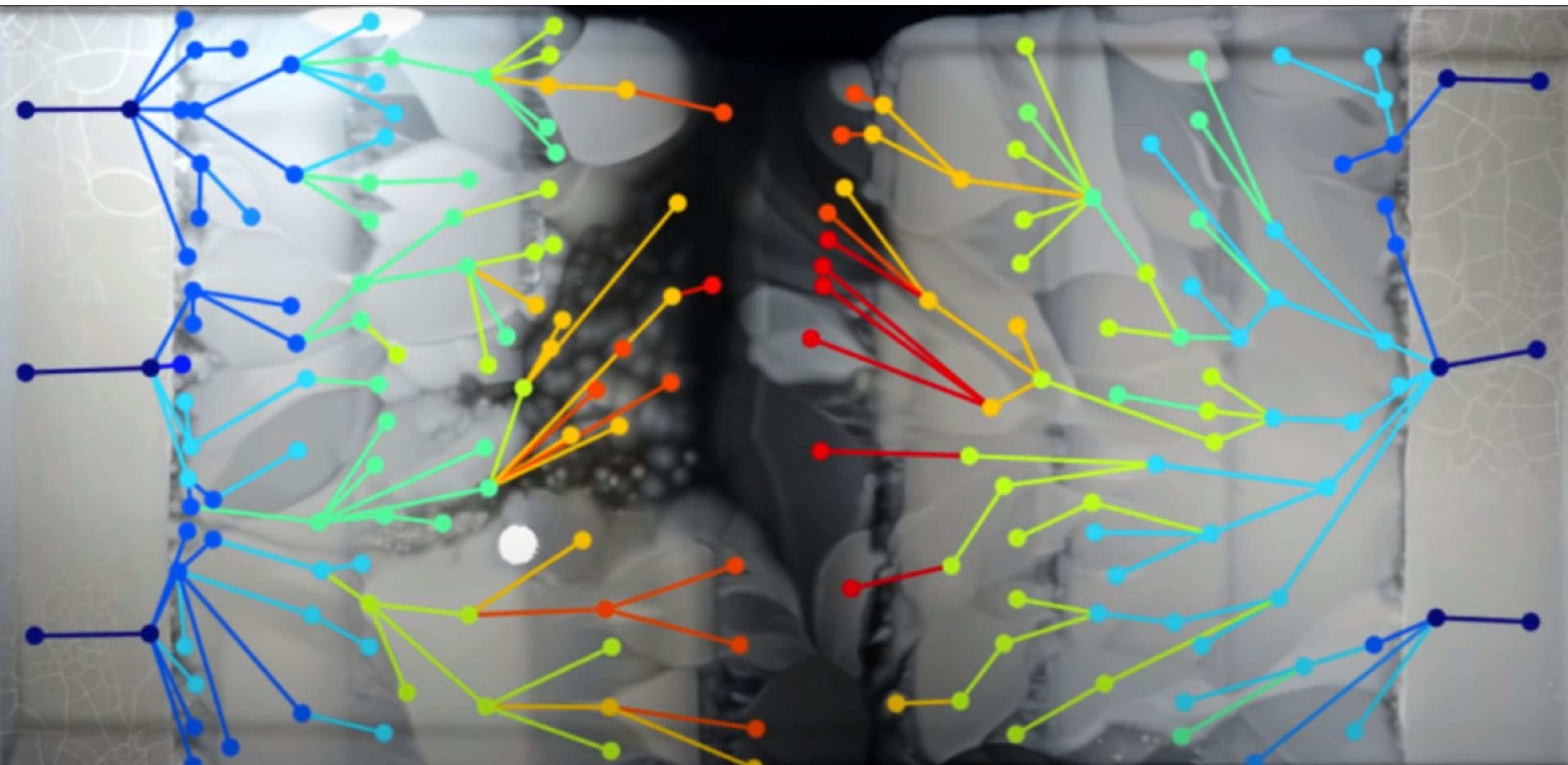
“Nothing in biology makes sense except in the light of evolution.”

– Theodosius G. Dobzhansky

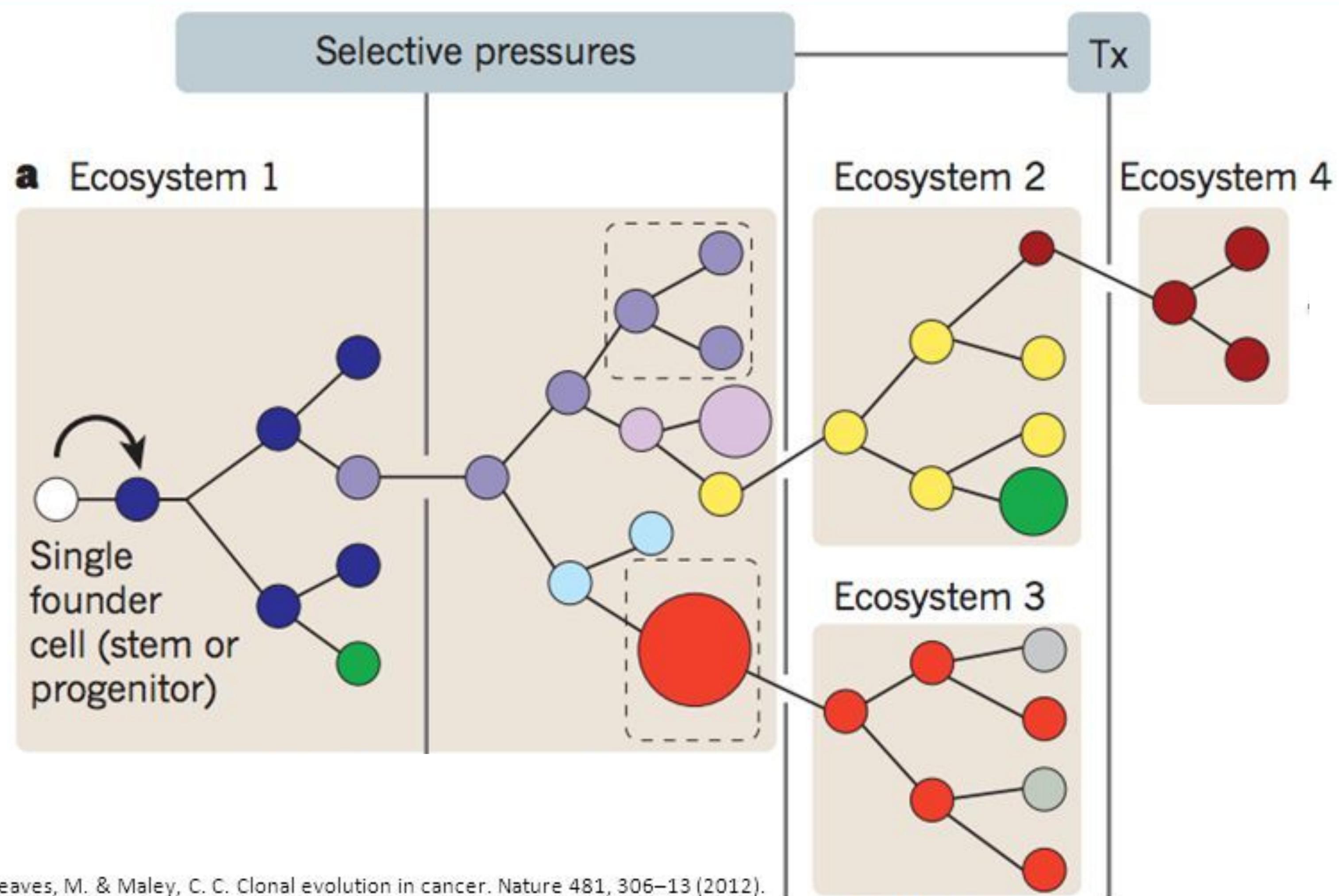
The Evolution of Bacteria on a “Mega-Plate” Petri Dish



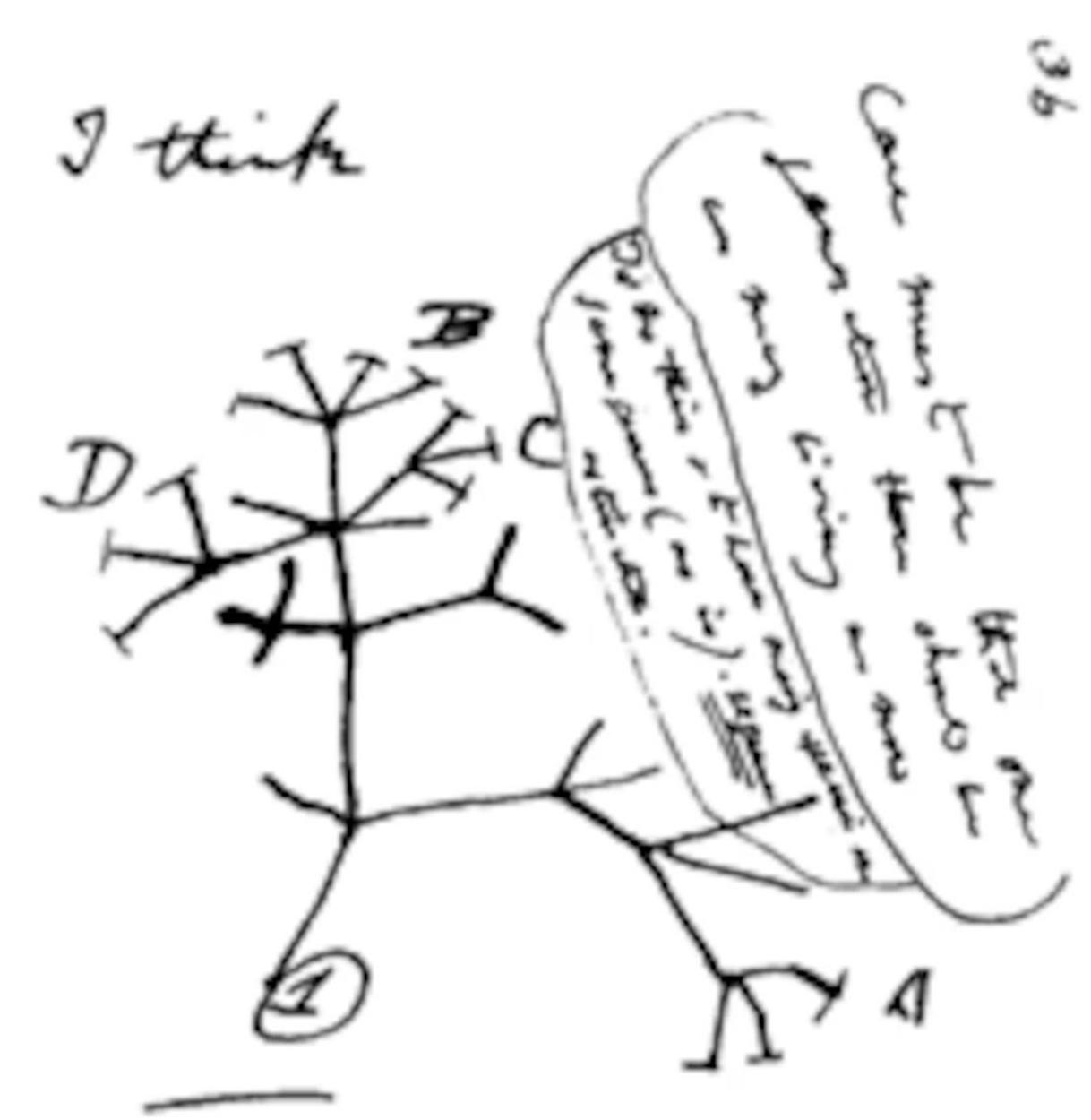
The Evolution of Bacteria on a “Mega-Plate” Petri Dish



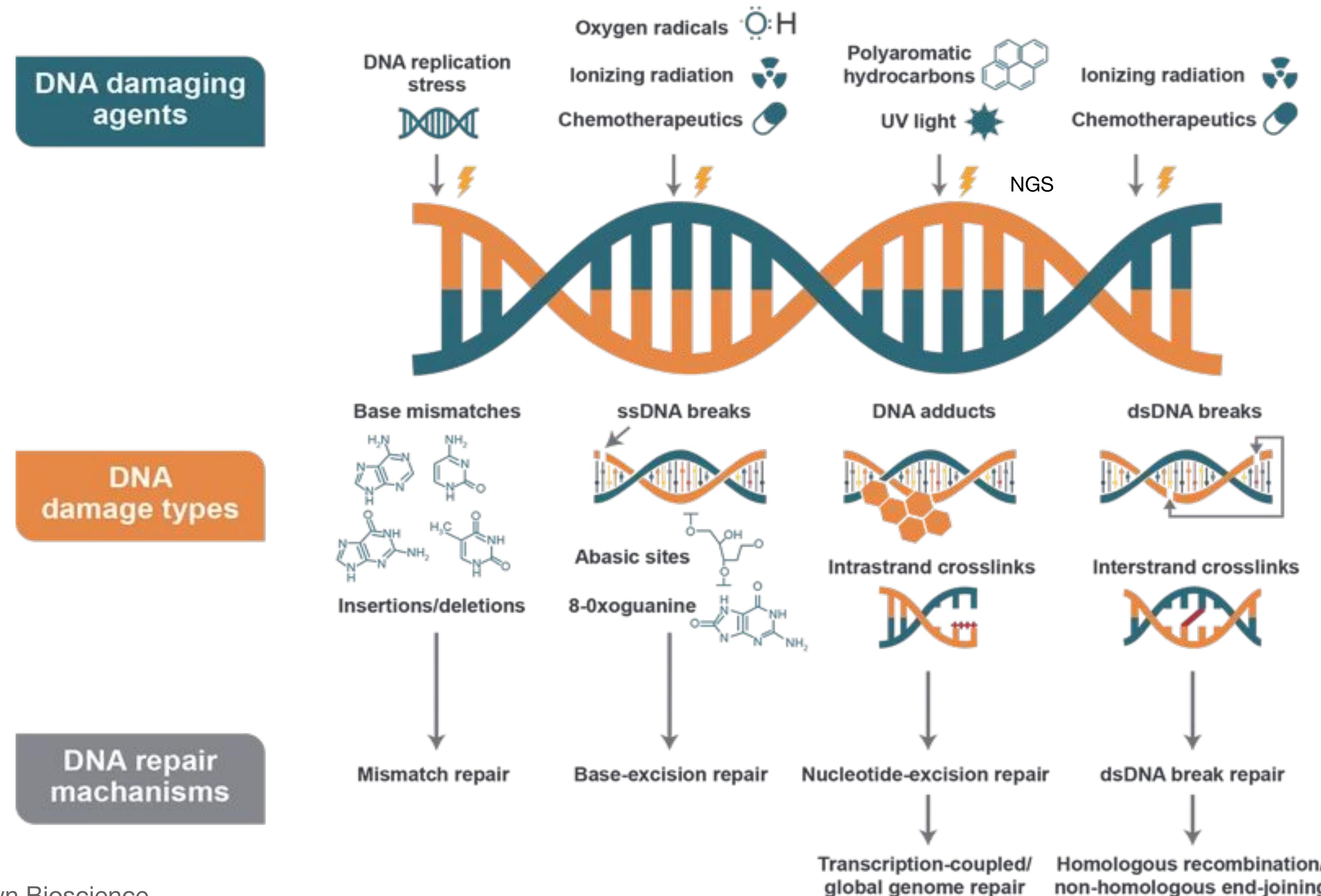
Somatic Evolution



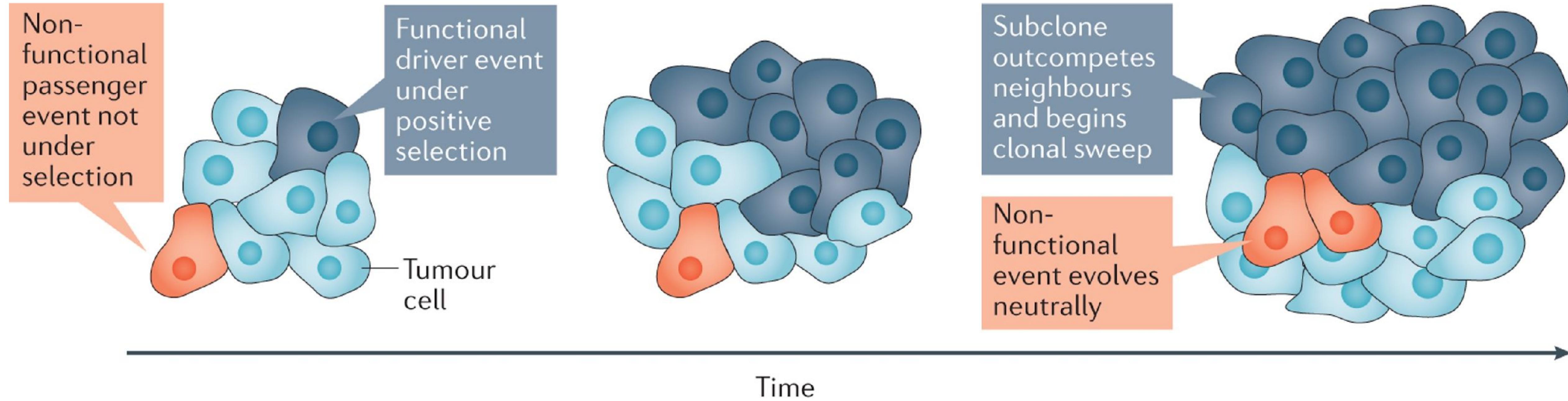
Greaves, M. & Maley, C. C. Clonal evolution in cancer. Nature 481, 306–13 (2012).



Genotypic variation leads to phenotypic variation



Natural selection acts on phenotypic variation

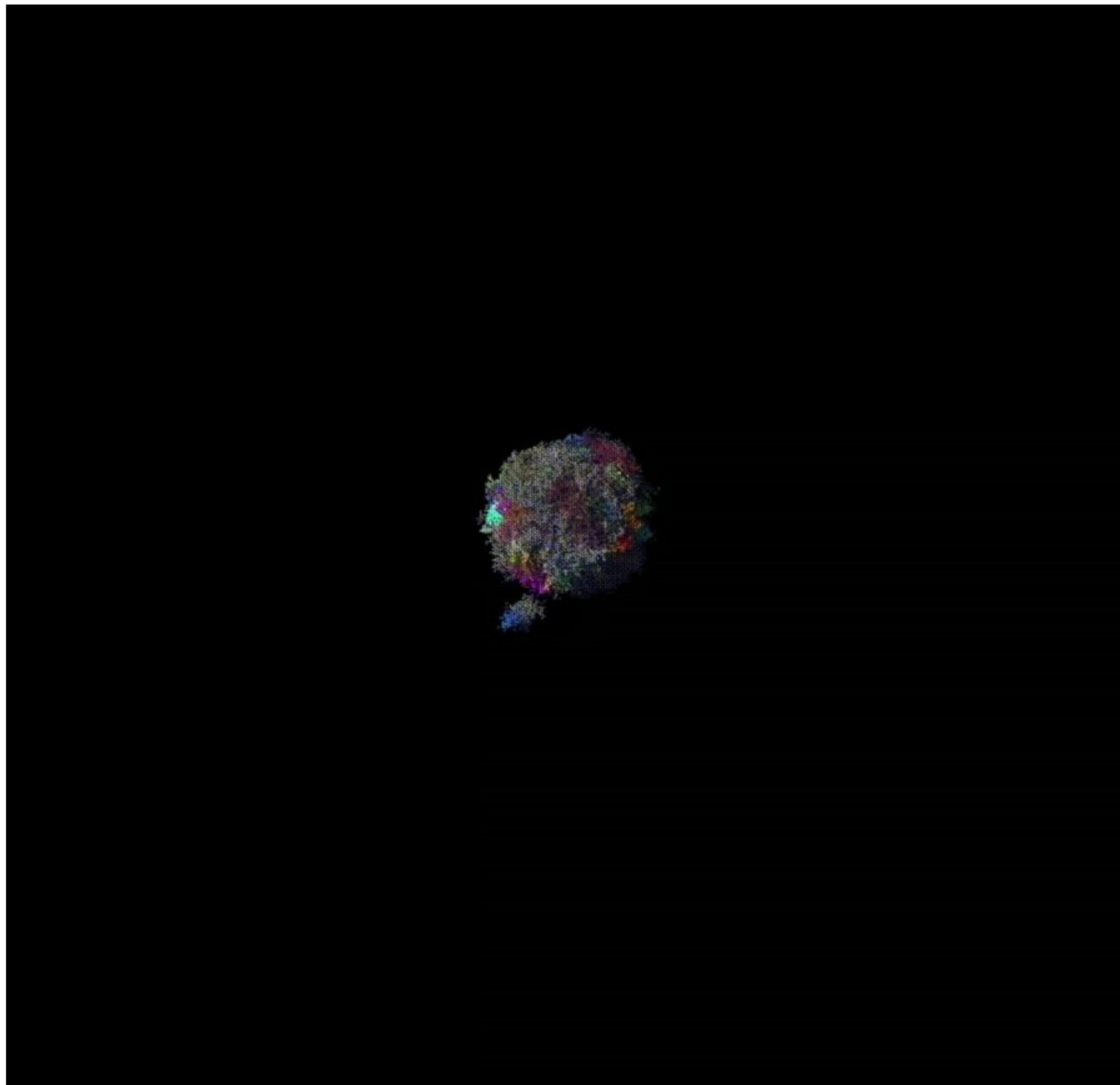


Black JRM, McGranahan N. Nat Rev Cancer. 2021 Jun;21(6):379-392.

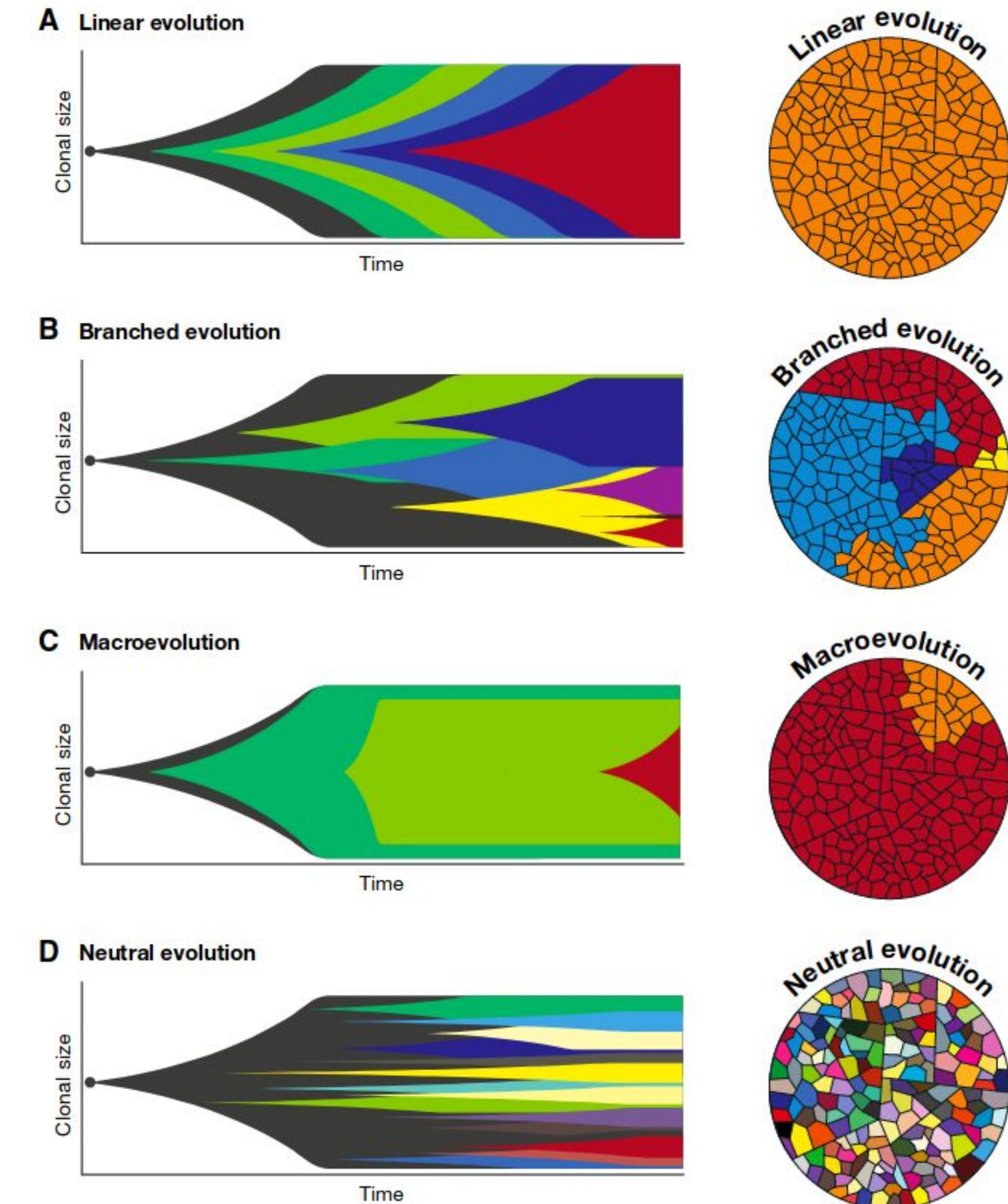
“Cancer is one end-product of somatic evolution, in which a single clonal lineage acquires a complement of *driver* mutations that enables the cells to evade normal constraints on cell proliferation, invade tissues, and spread to other organs.”

Matricrena et al., 2017

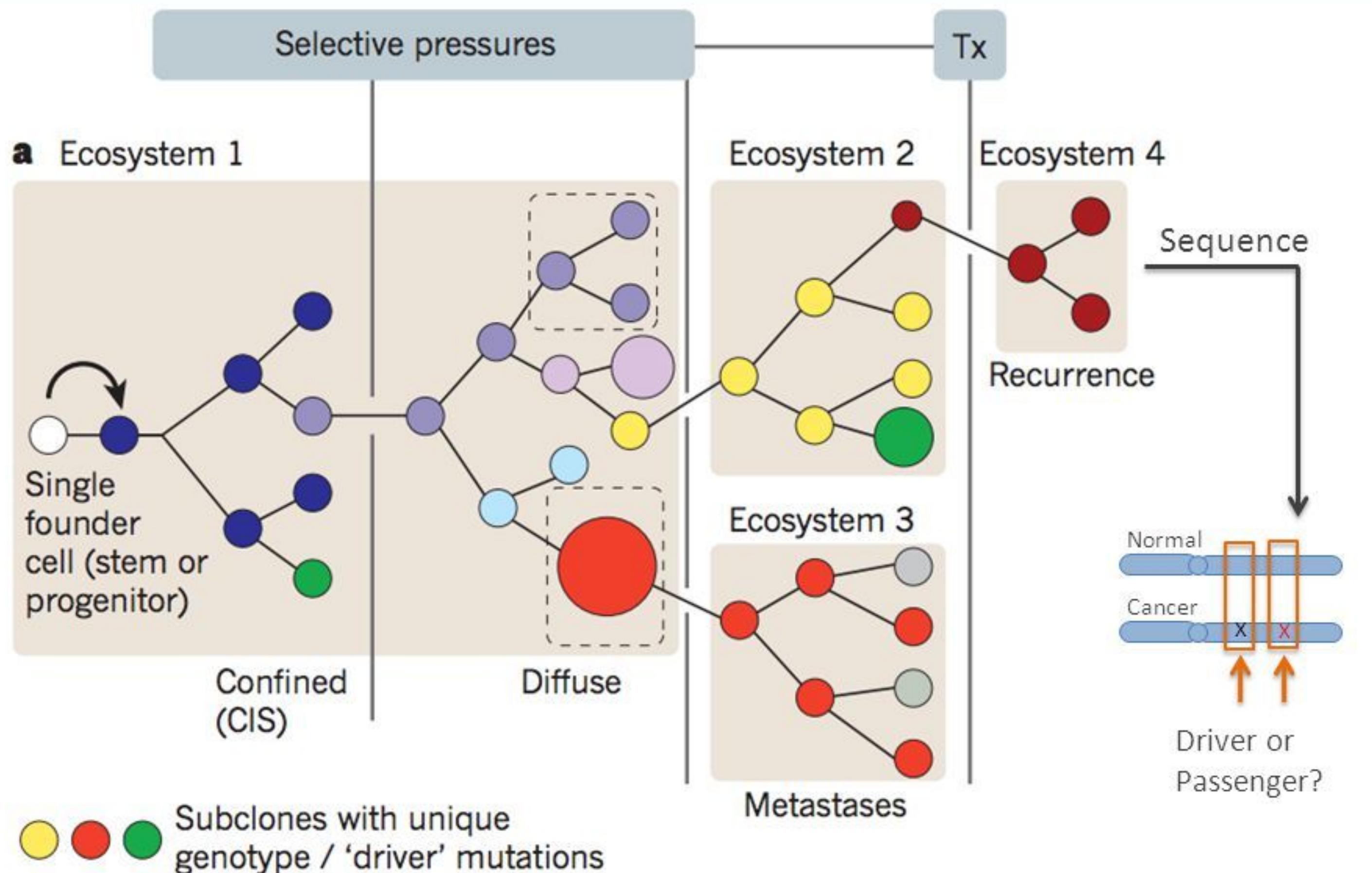
Clonal Evolution in Cancer



DOI:10.1038/nature14971
<https://youtu.be/lpytolxRu0o>



Drivers and Passengers



Drivers evolve under natural selection

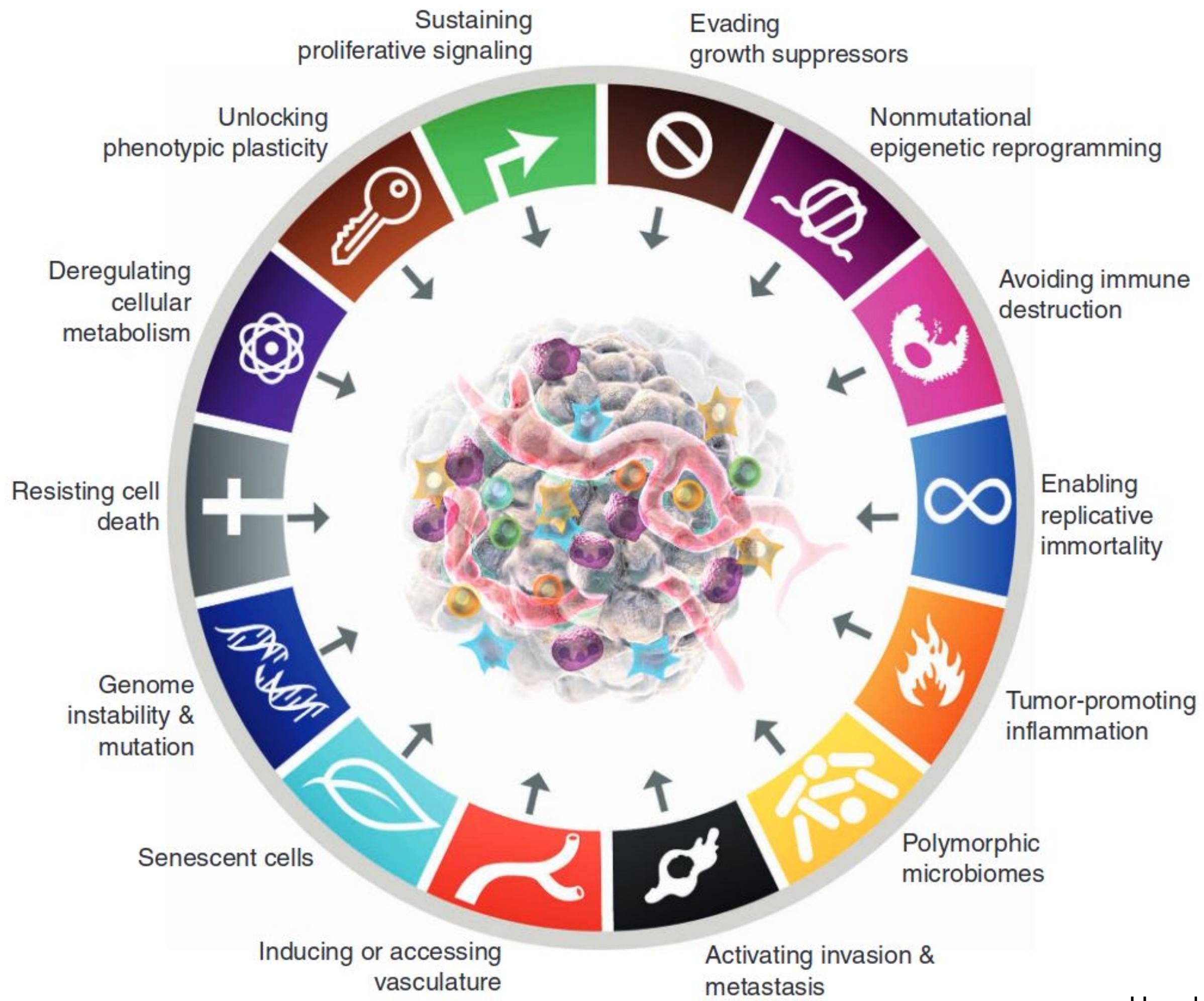
Passengers evolve under genetic drift



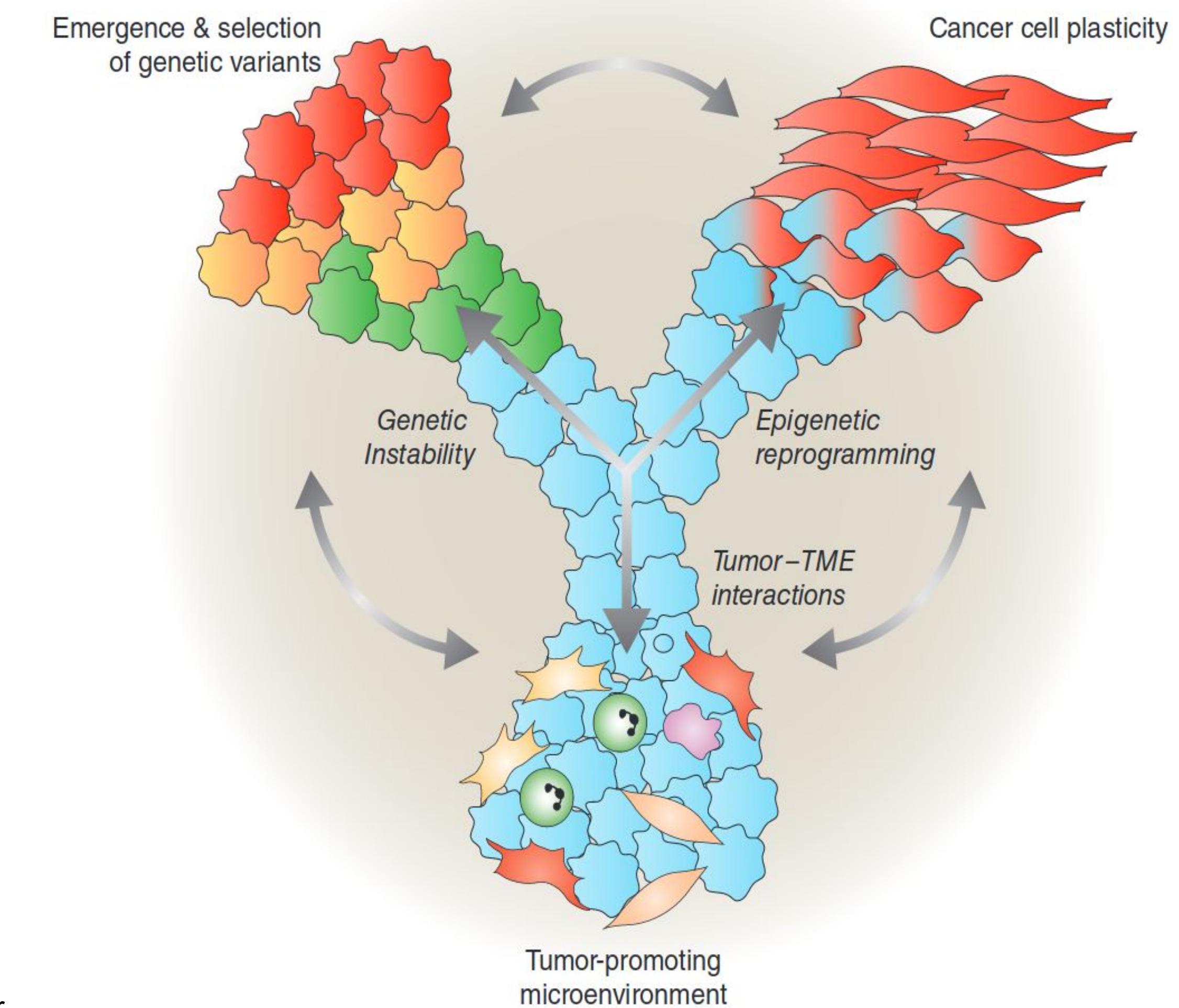
Greaves, M. & Maley, C. C. Clonal evolution in cancer. *Nature* 481, 306–13 (2012).

Drivers and Passengers

Drivers are causal alterations that enable the hallmarks of cancer

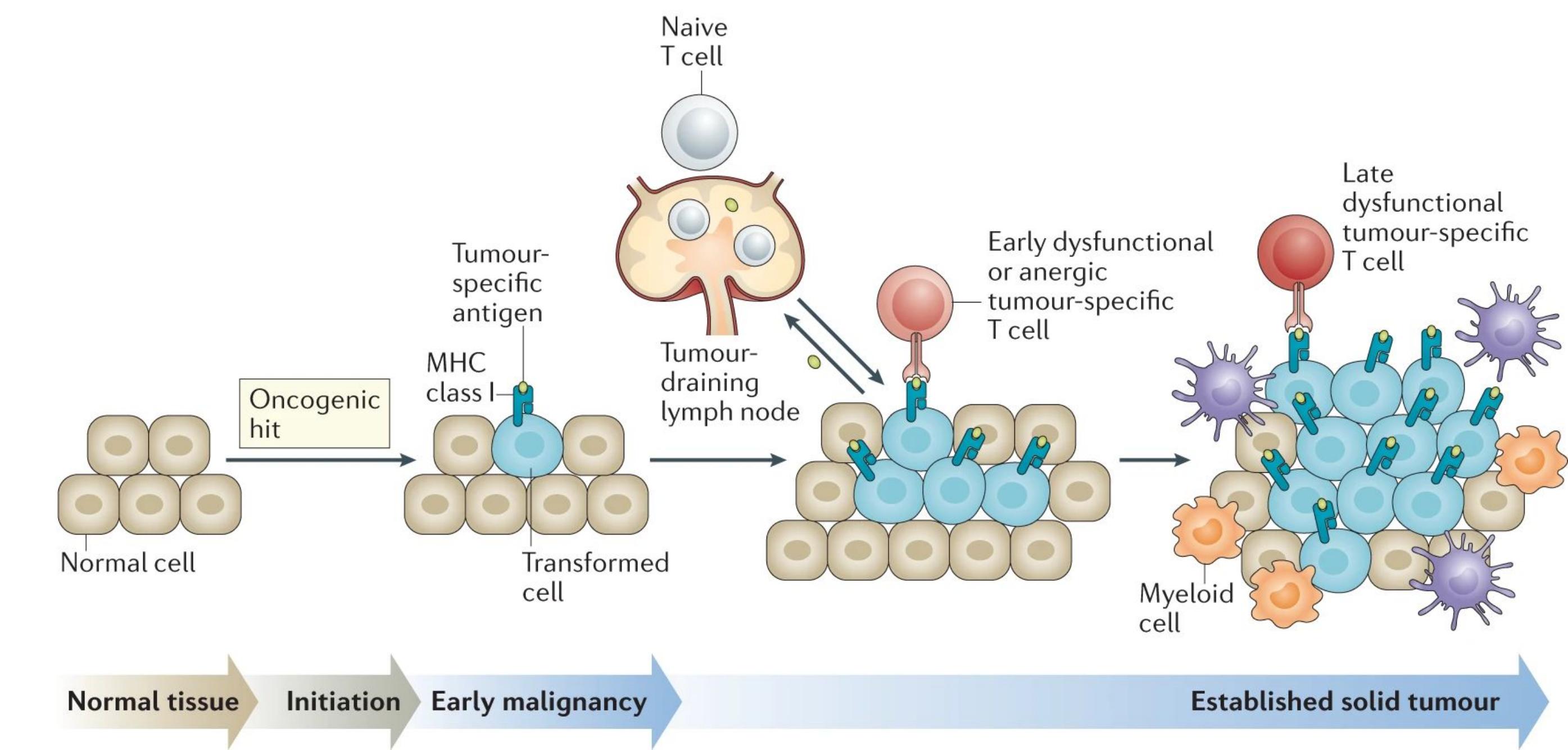
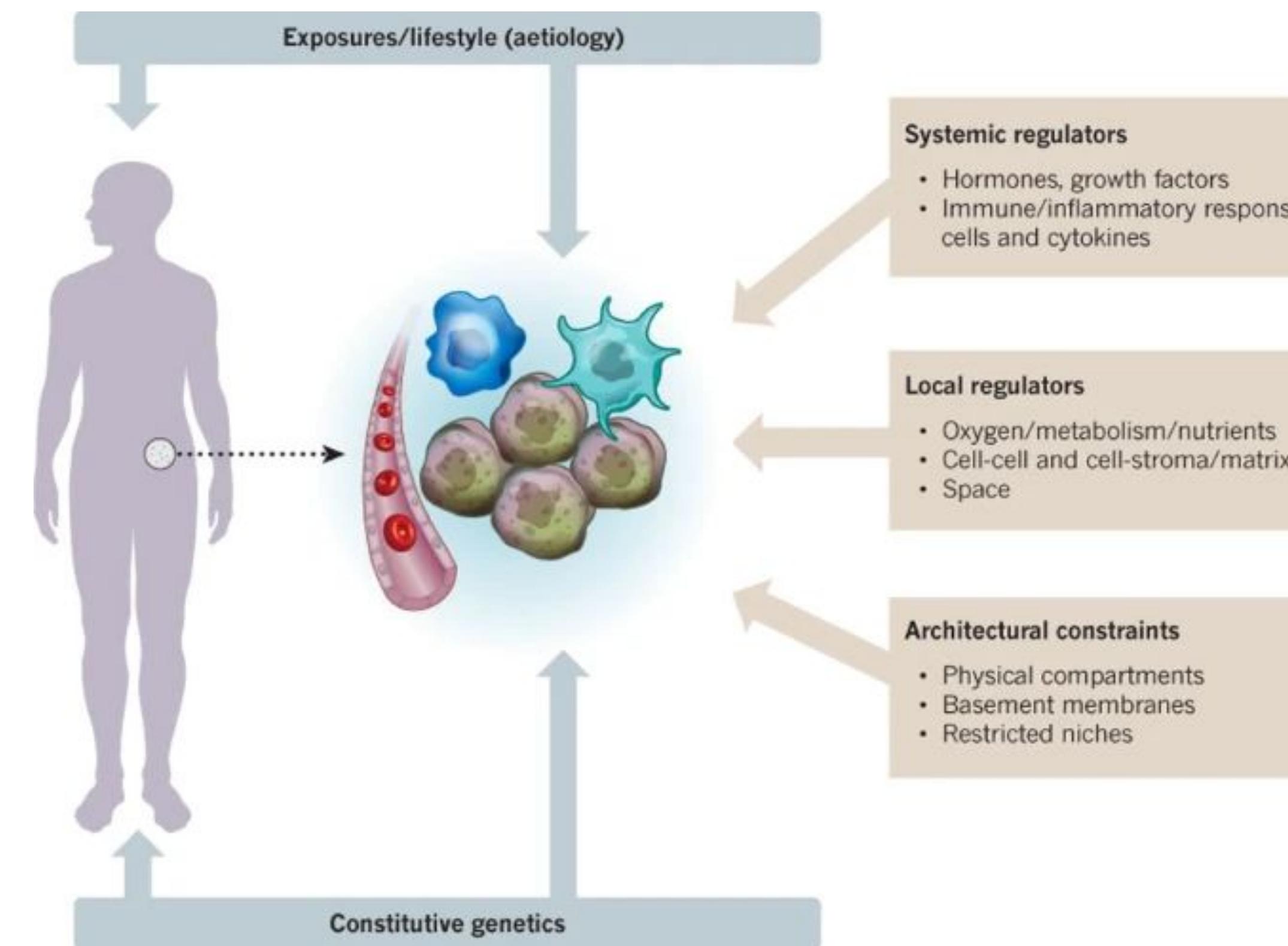


Hanahan D. Cancer Discov. 2022;12(1):31-46

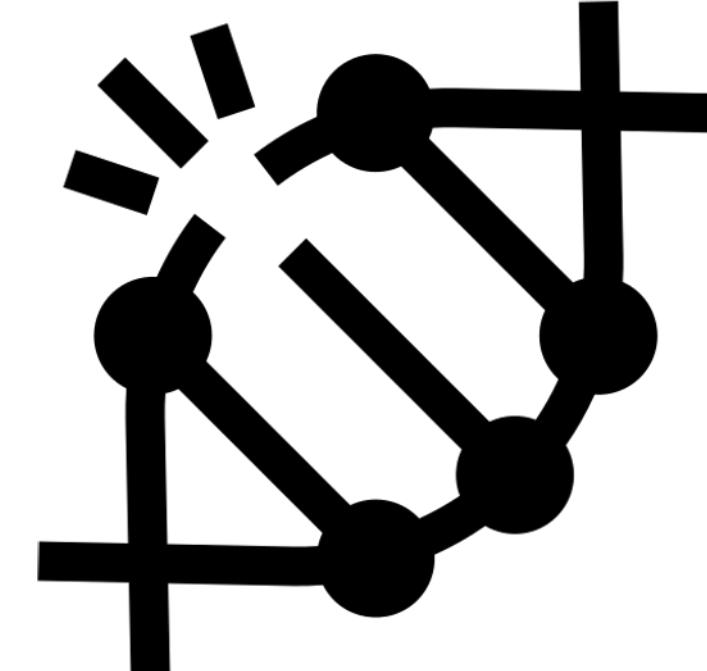


Ciriello et al

The adaptive landscape of tissue ecosystems



Driver mutations and driver genes



- Mutations in cancer can be classified into:
 - Driver mutations
 - Passenger mutations
- **Driver mutations**
 - Provide a selective advantage to the cell.
 - Promote cancer development
- Passenger mutations
 - Neutral mutations.

Genes that harbor driver mutations are called “**cancer driver genes**”.

A genetic basis of cancer: The first identified driver genes

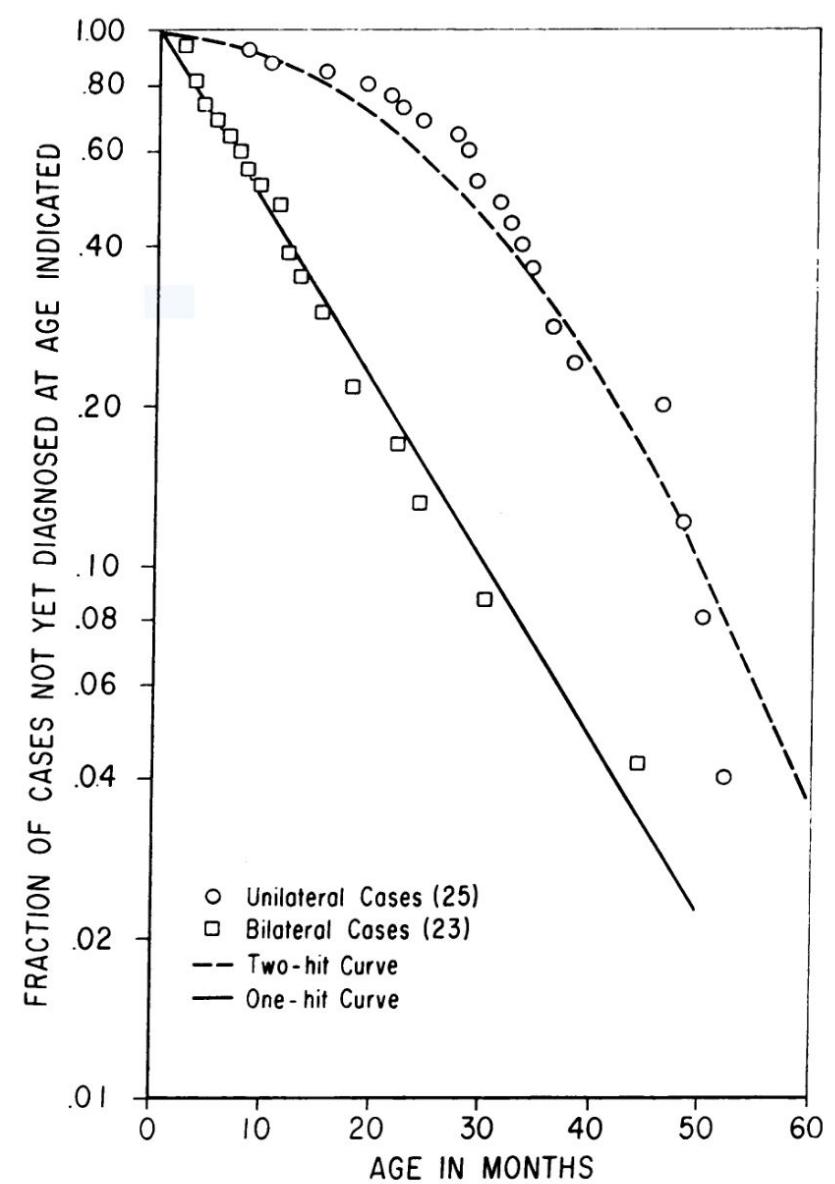
RB1: a tumor suppressor gene

Mutation and Cancer: Statistical Study of Retinoblastoma

ALFRED G. KNUDSON, JR.

Graduate School of Biomedical Sciences and M. D. Anderson Hospital and Tumor Institute,
The University of Texas at Houston, Houston, Texas 77025

Communicated by James V. Neel, February 8, 1971



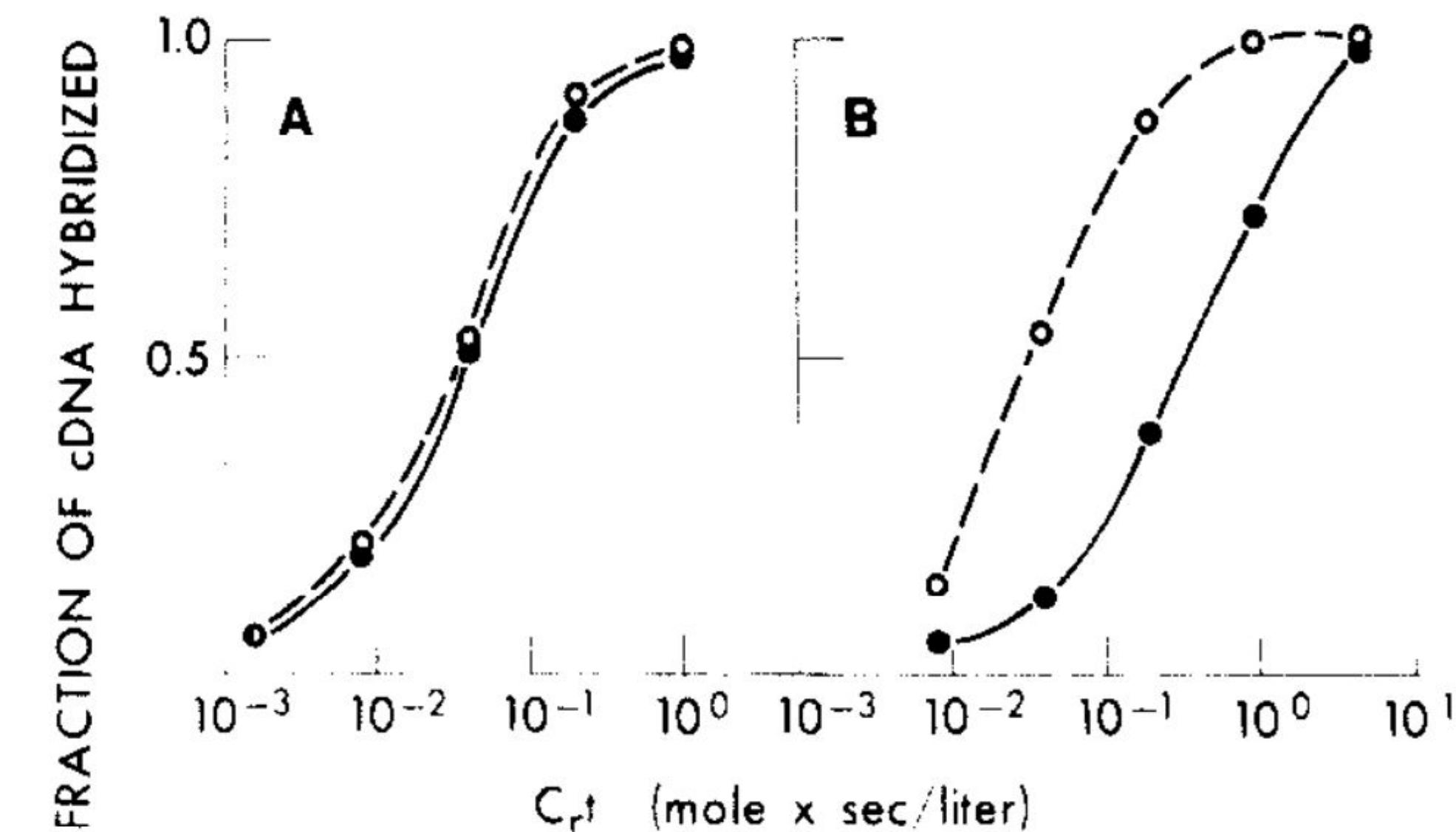
SRC: a proto-oncogene

Detection and Enumeration of Transformation-Defective Strains of Avian Sarcoma Virus with Molecular Hybridization

DOMINIQUE STEHELIN,¹ DONALD J. FUJITA, THOMAS PADGETT,
HAROLD E. VARMUS, AND J. MICHAEL BISHOP²

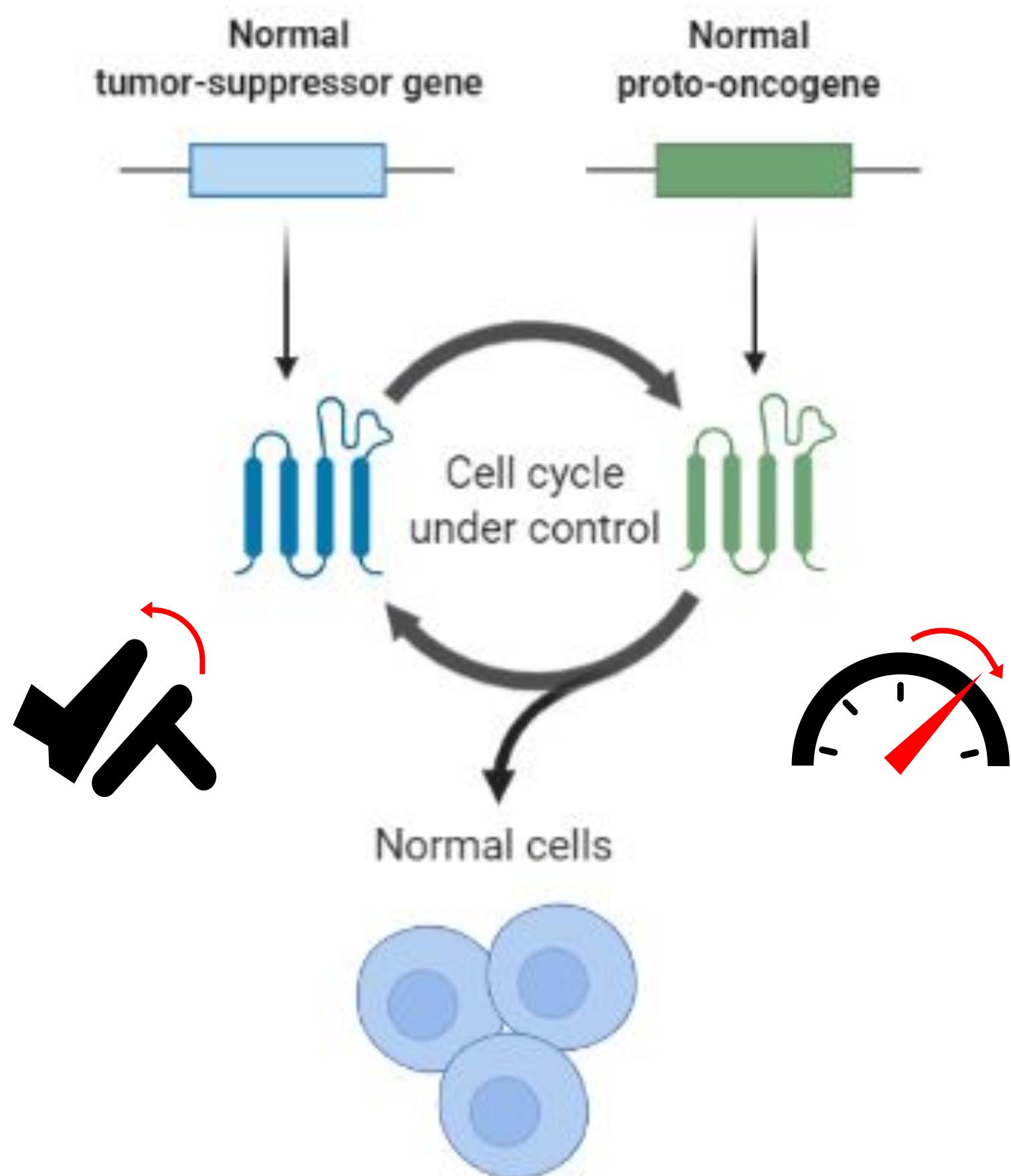
Department of Microbiology, University of California, San Francisco, California 94143

Accepted September 27, 1976

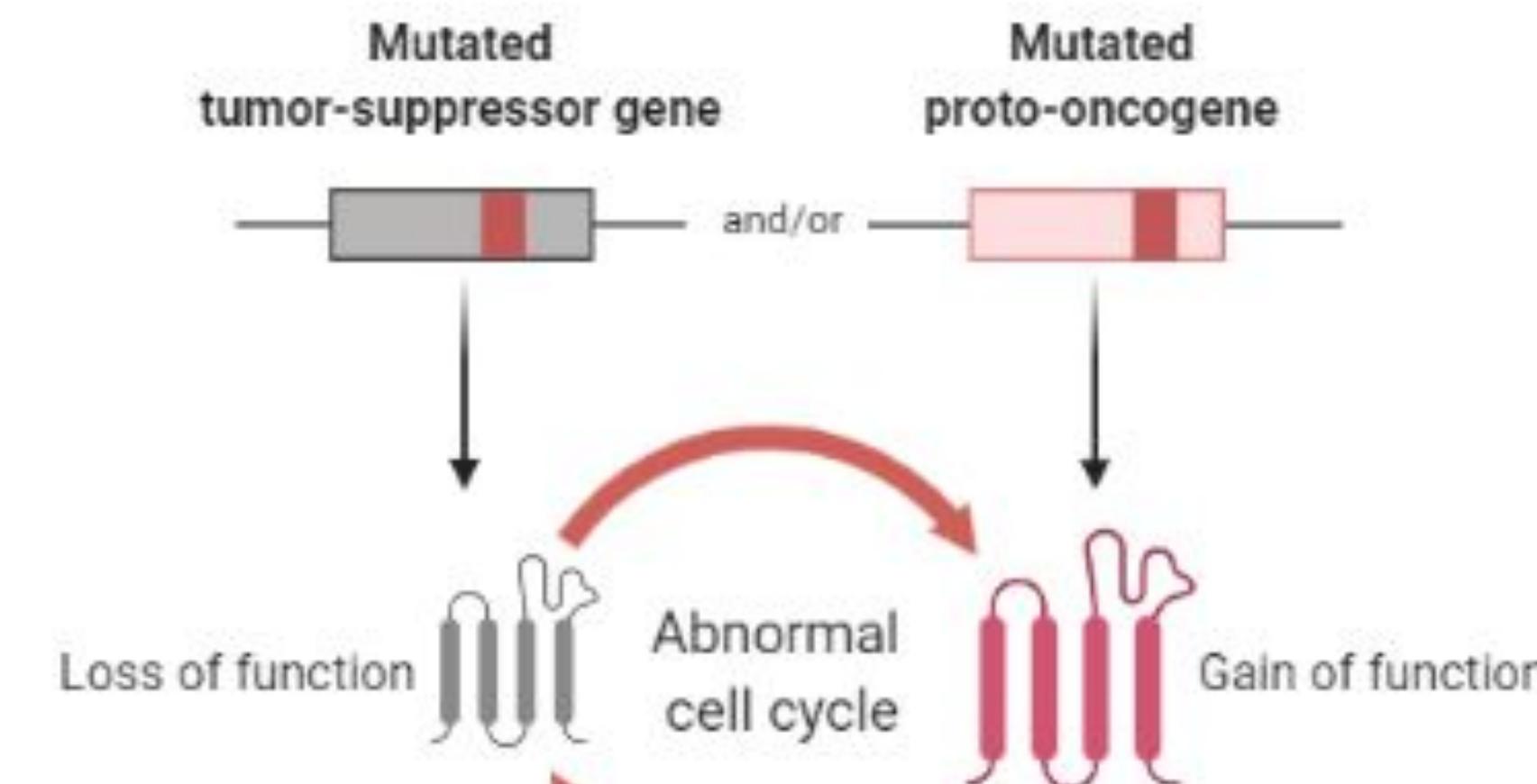


Oncogenes and Tumor Suppressor Genes

Normal Cell Division



Malignant Cell Division

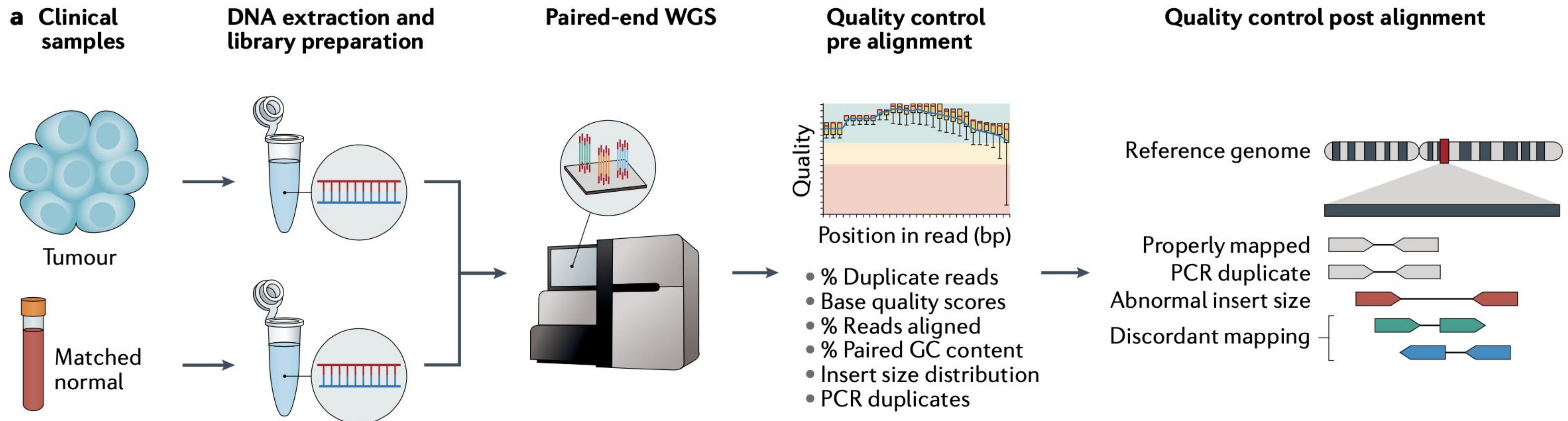


- Double hit
- Missense, splice site, nonsense mutations
- Copy number loss (Deletions)
- e.g. *TP53*, *VHL*

- Single hit
- Missense mutations
- Copy number gain (amplifications)
- e.g. *KRAS*, *MYC*

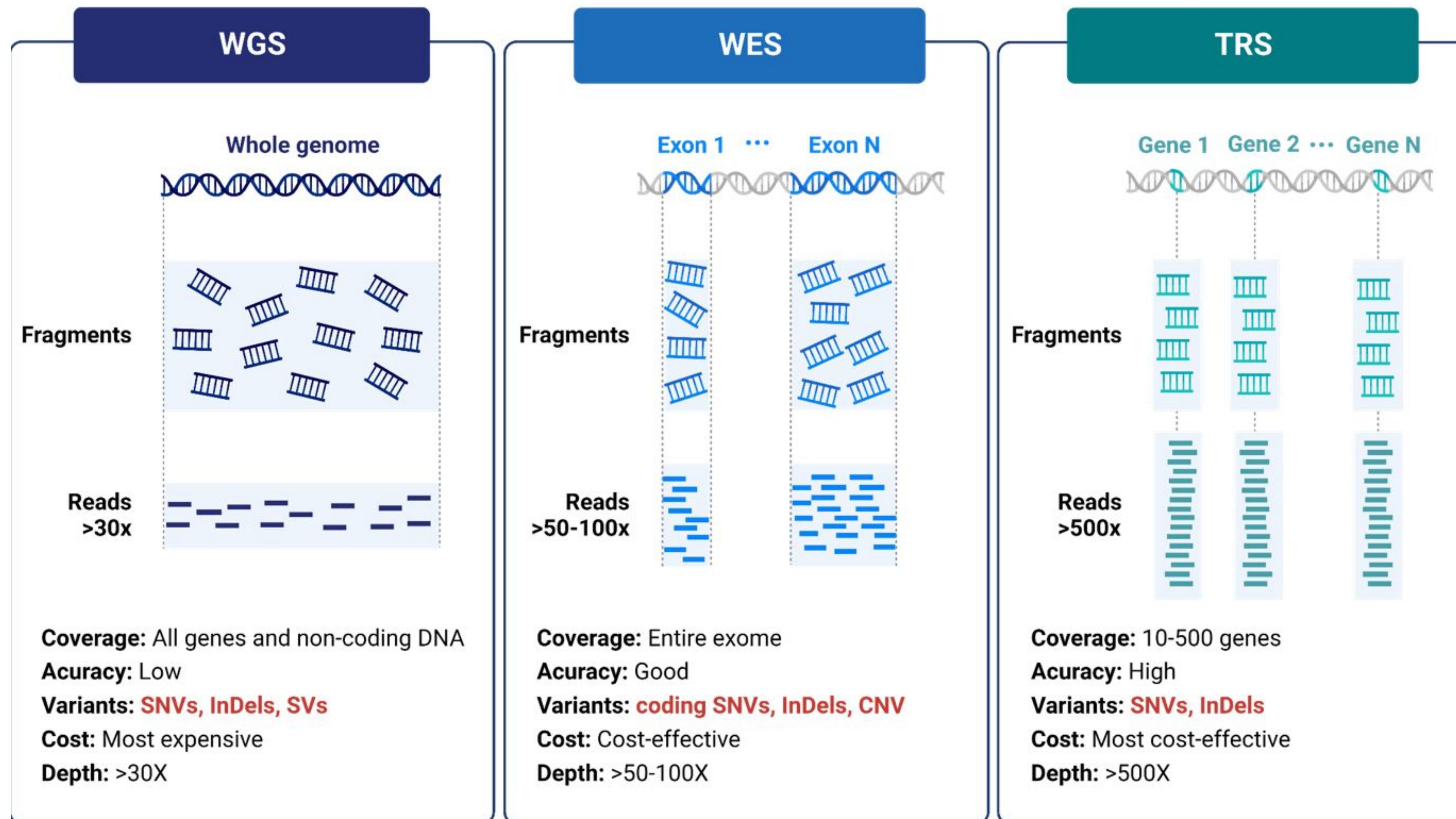
Identification of cancer drivers

Identification of genetic alterations in cancer genomes

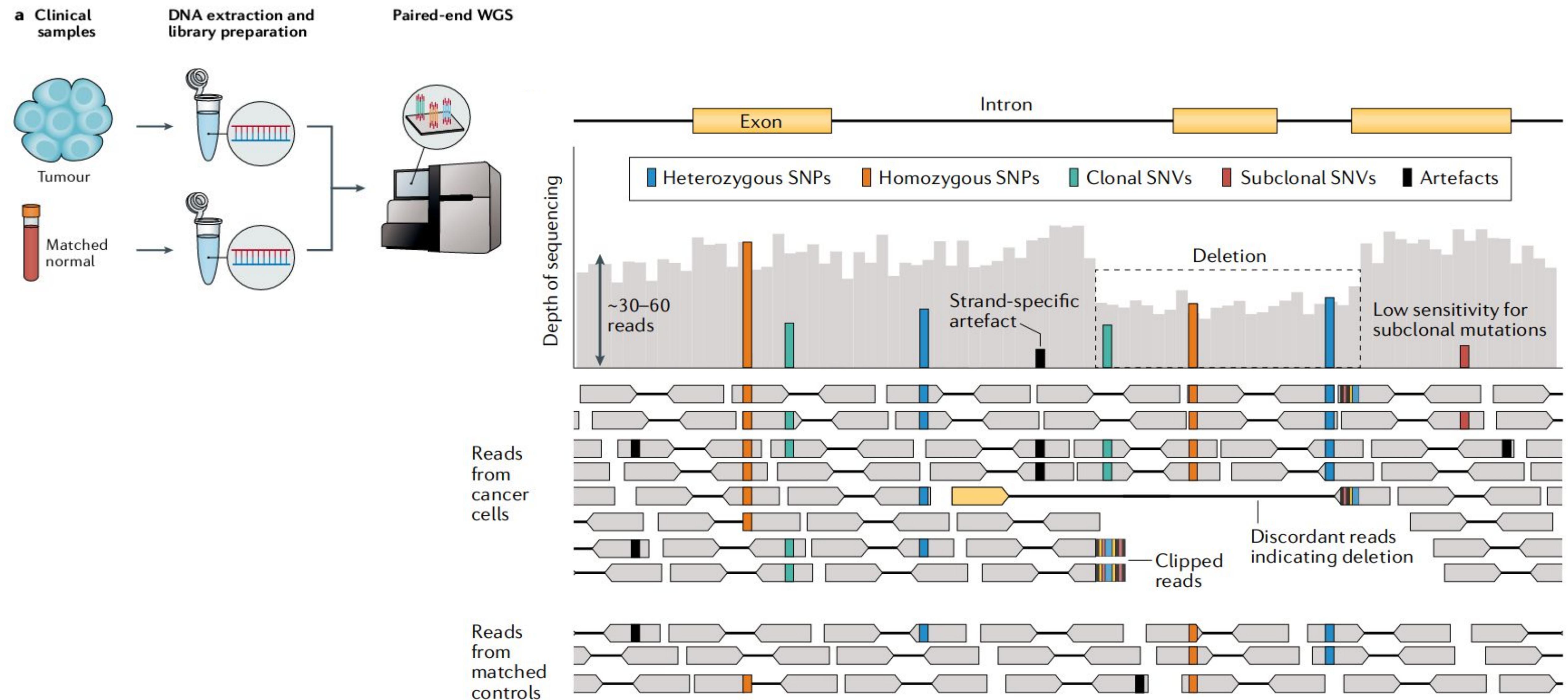


Identification of genetic alterations in cancer genomes

NGS: high-throughput sequencing

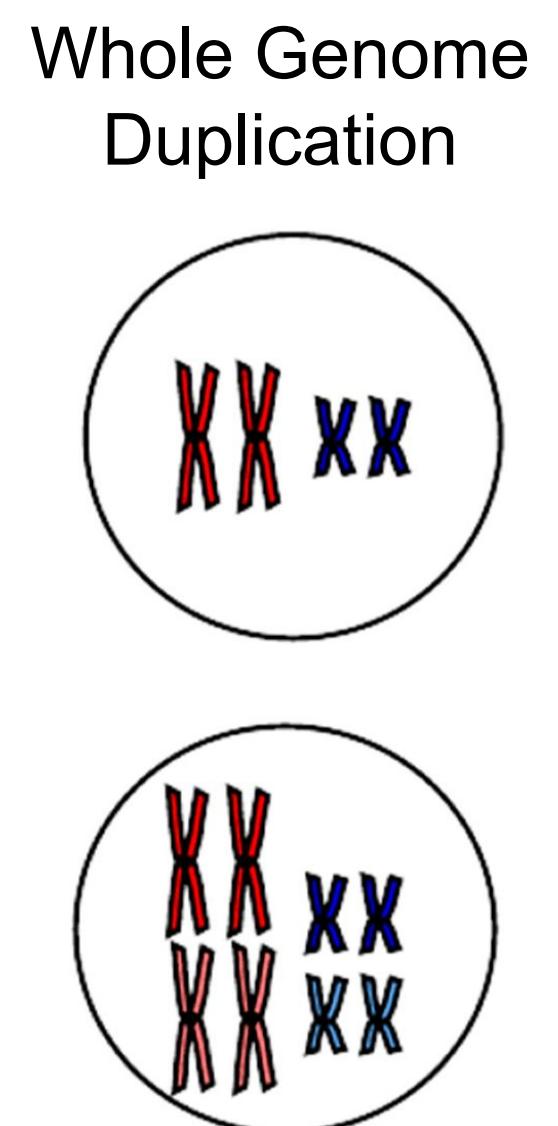
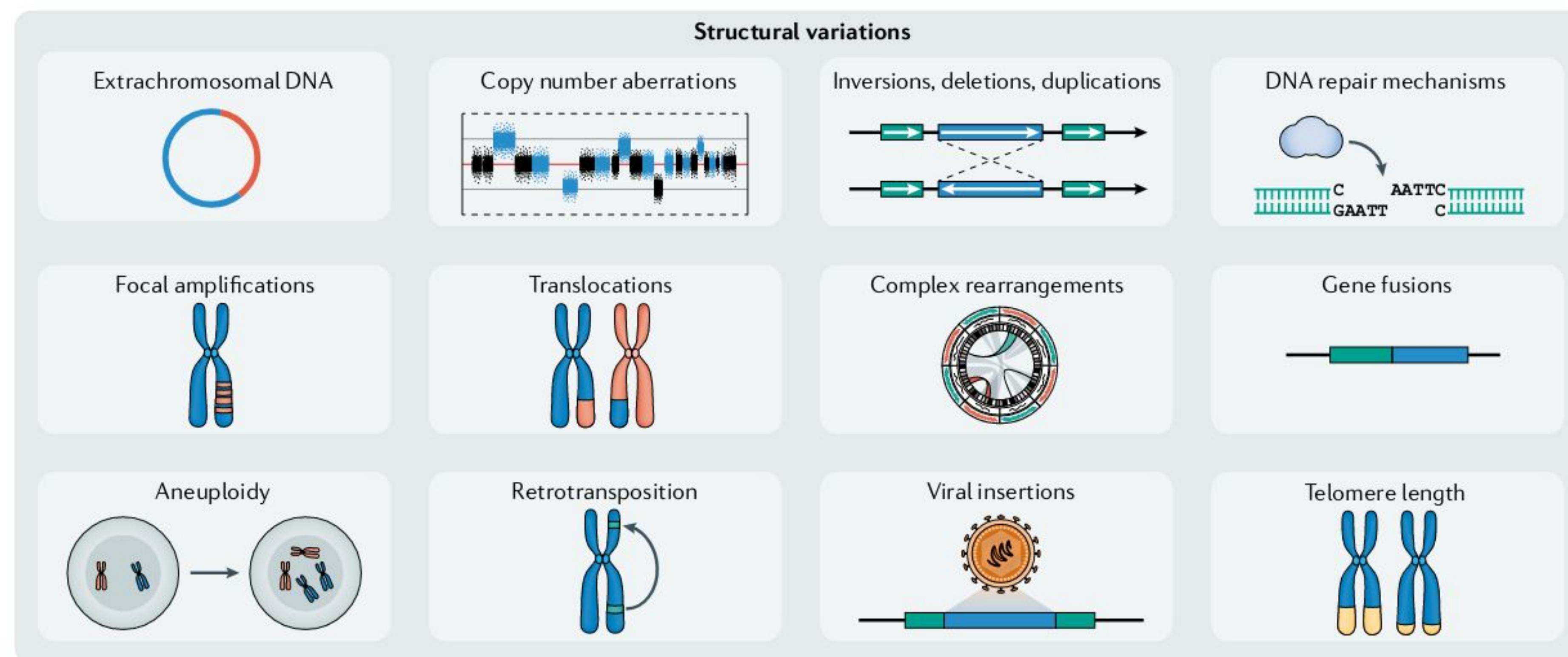
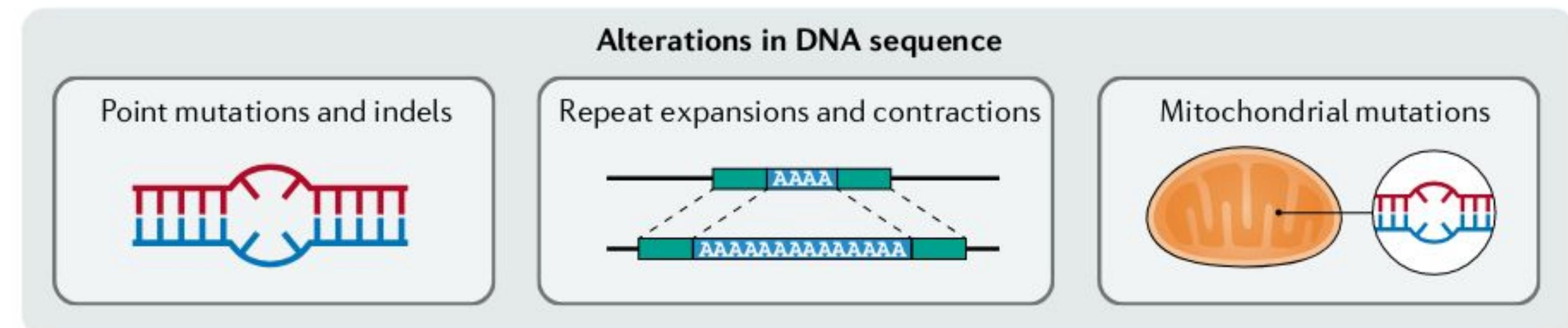
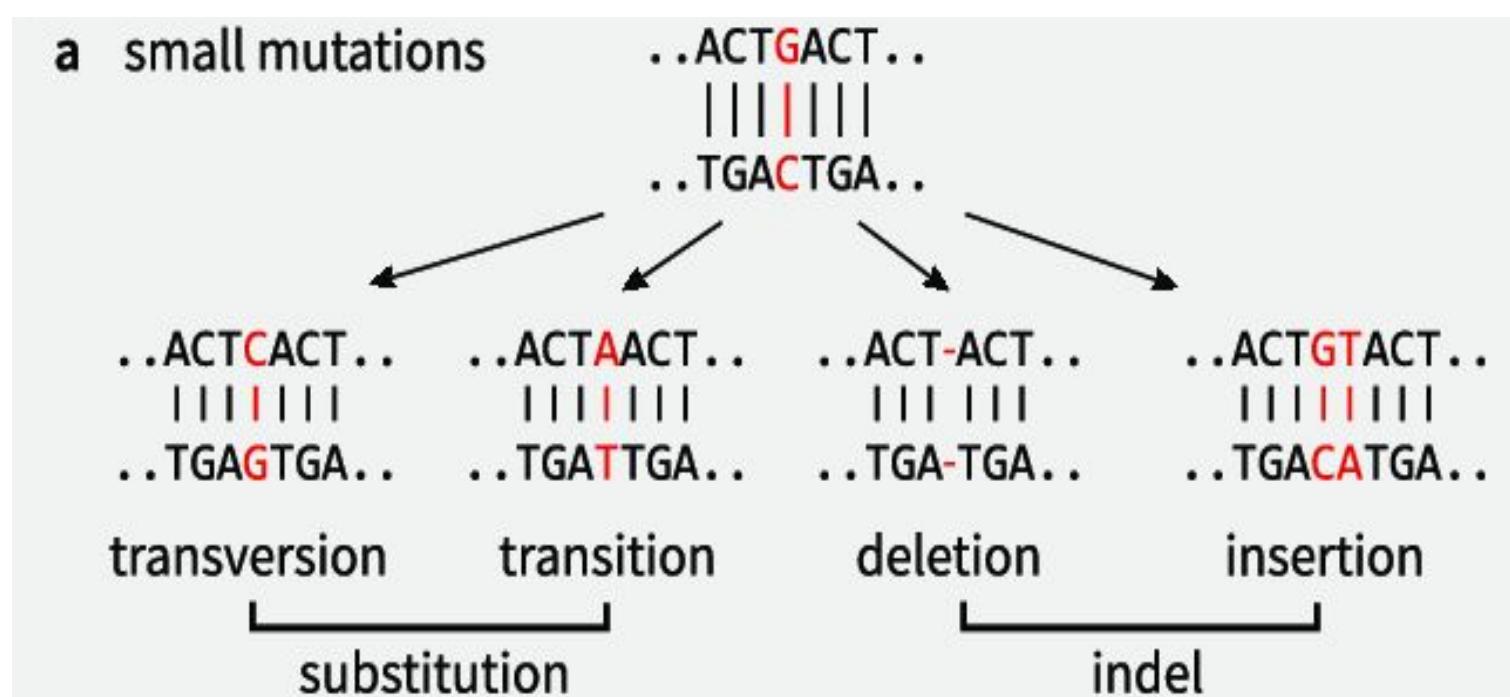


Identification of genetic alterations in cancer genomes

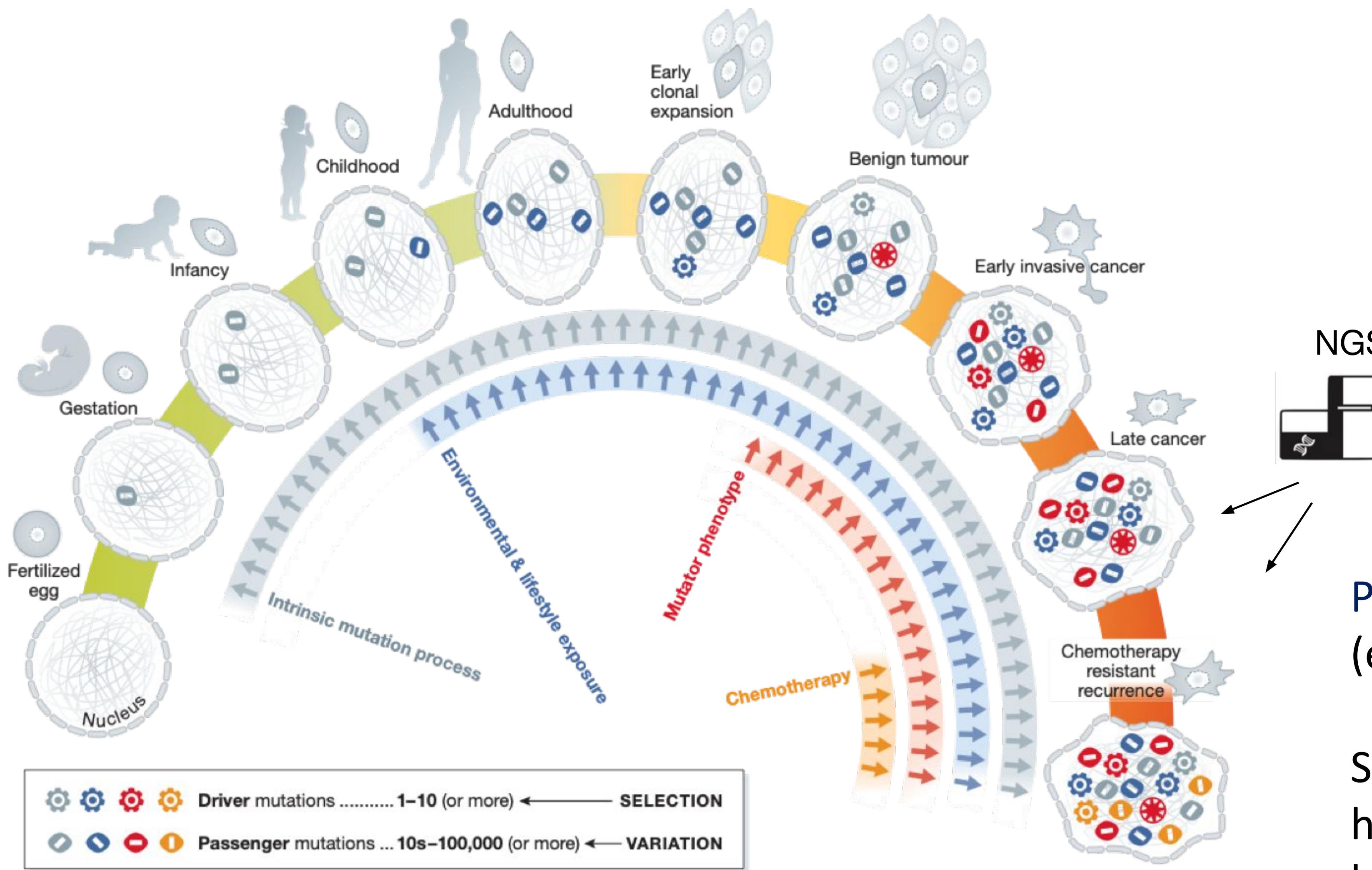


Identification of genetic alterations in cancer genomes

Point mutations

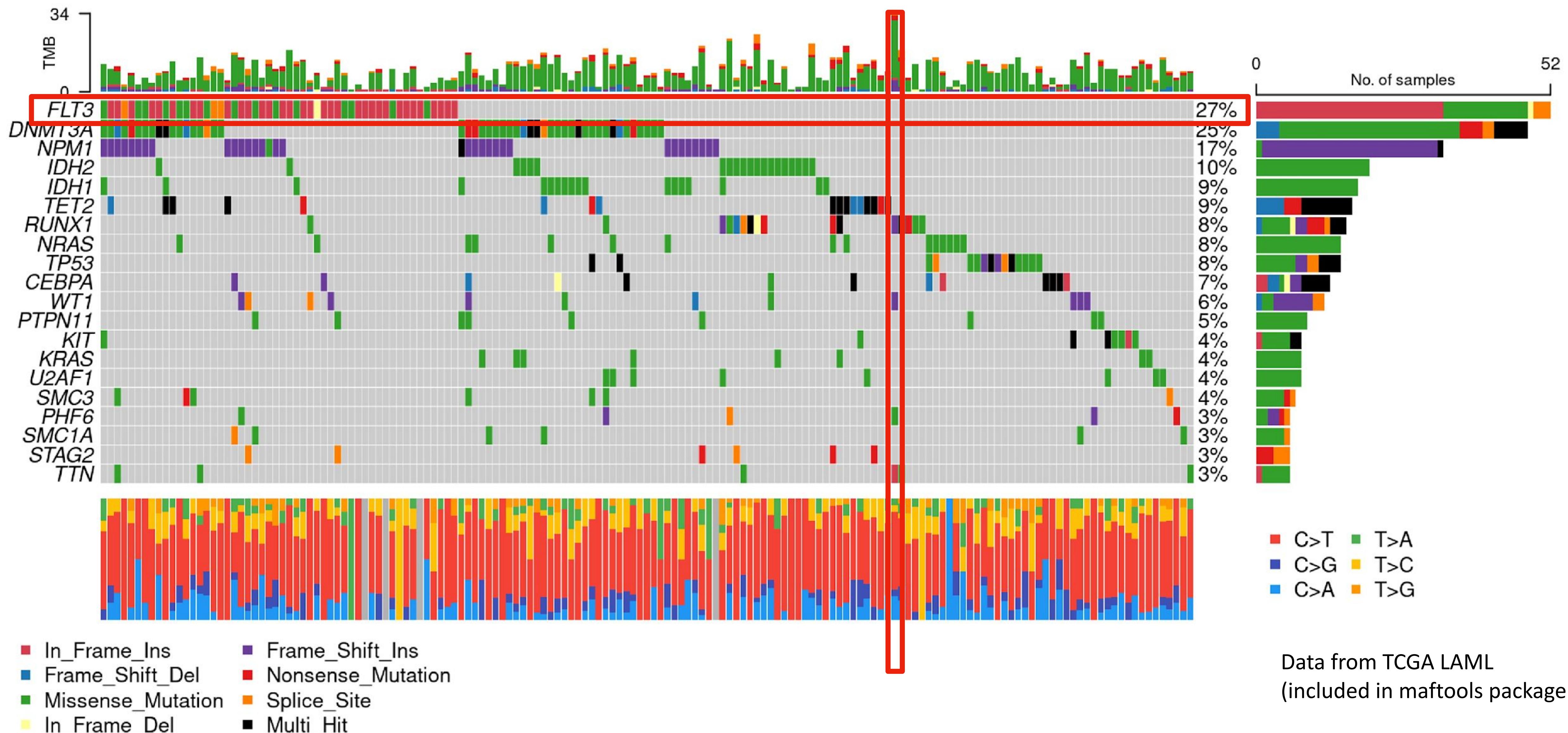


Separating drivers from passengers

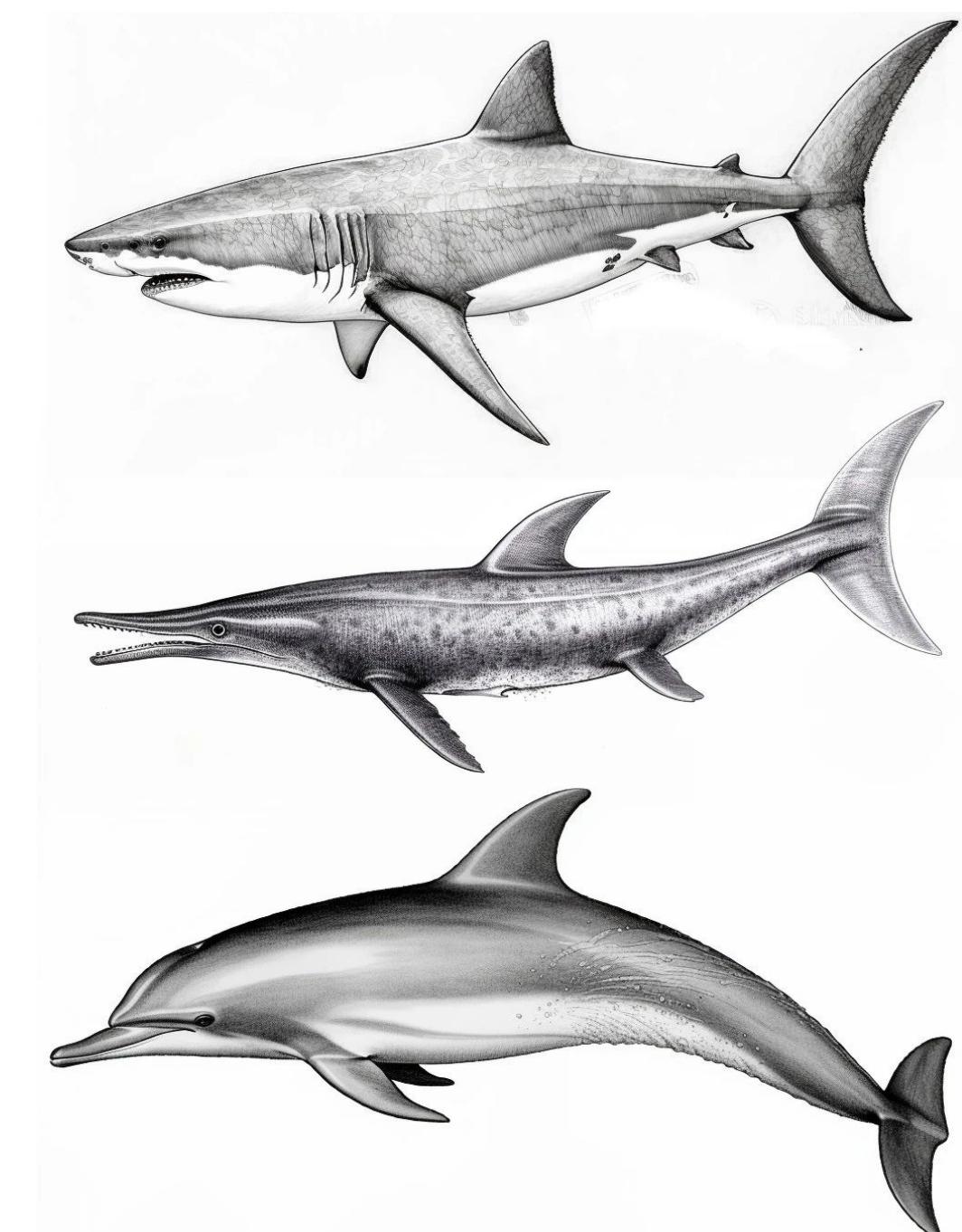
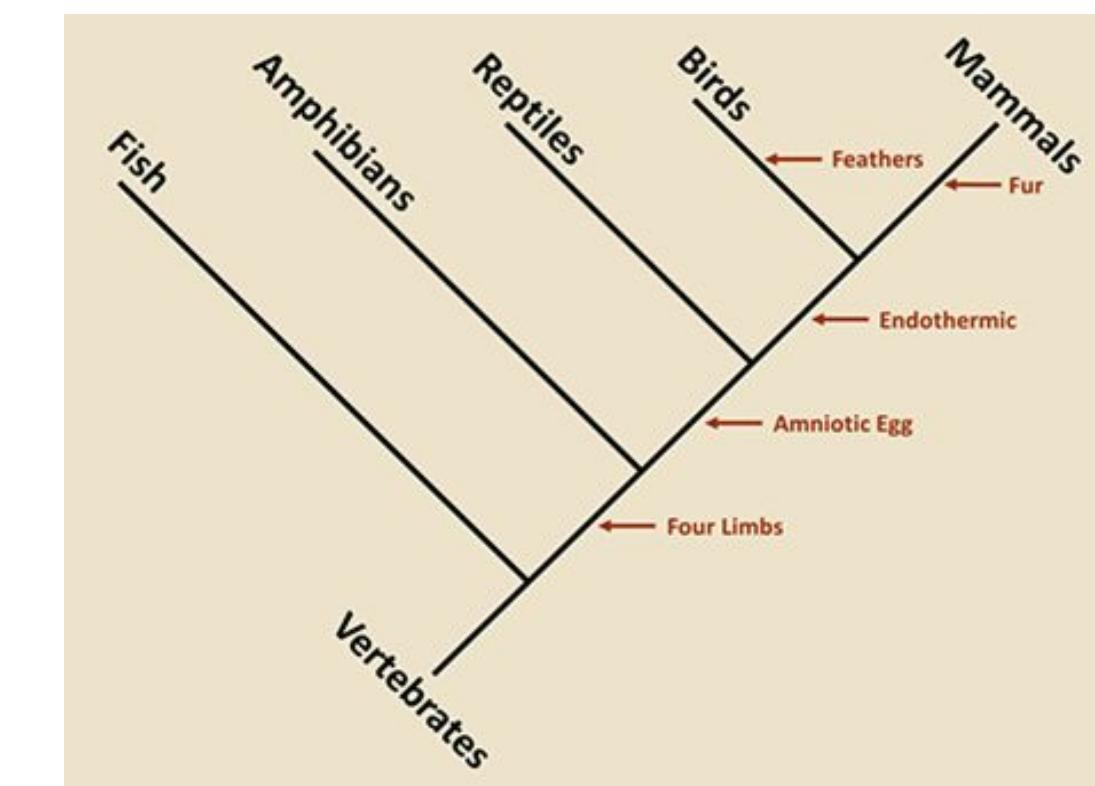
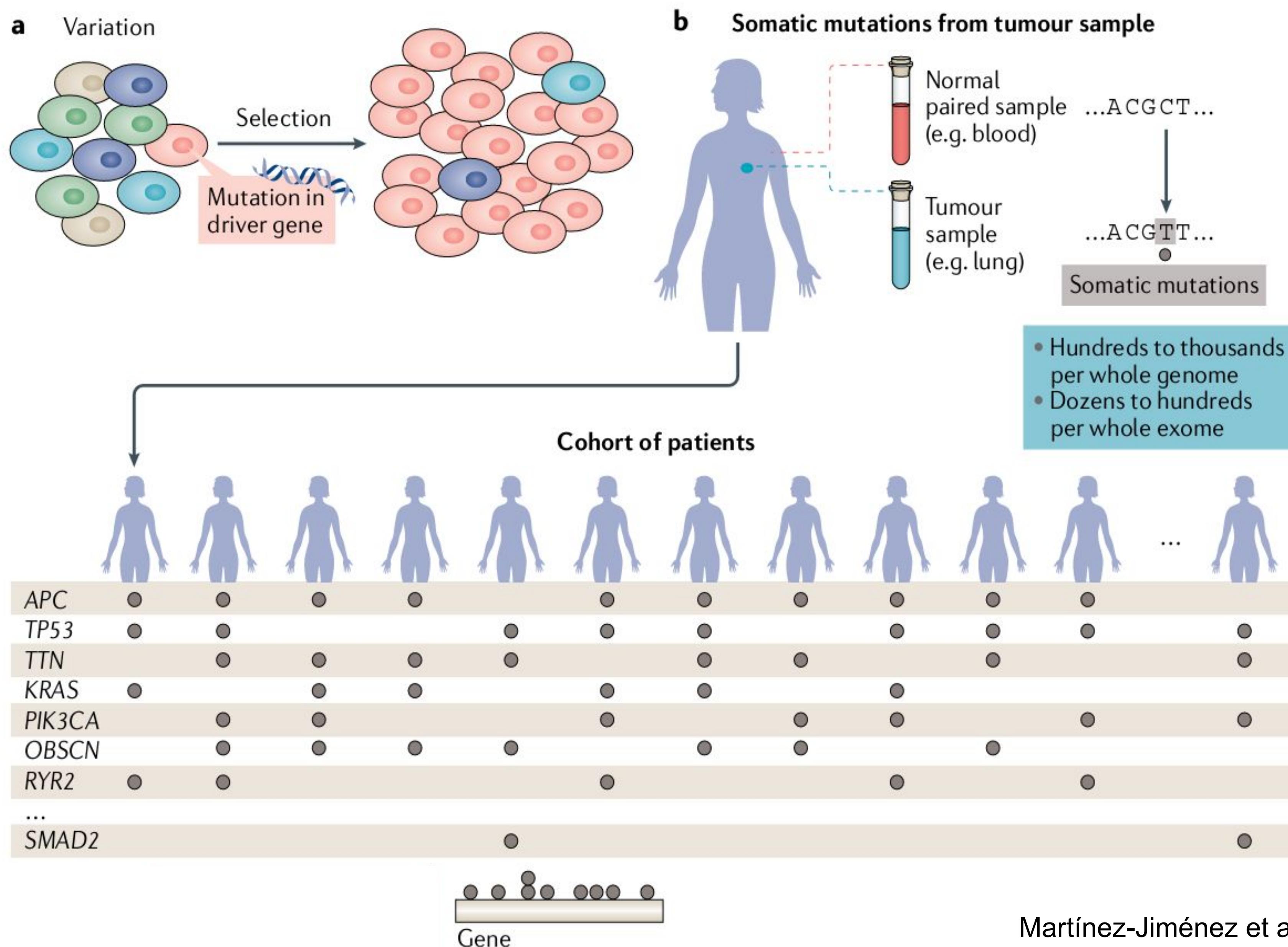


Oncoplots: Visualising mutations in a cohort

Altered in 159 (82.38%) of 193 samples.

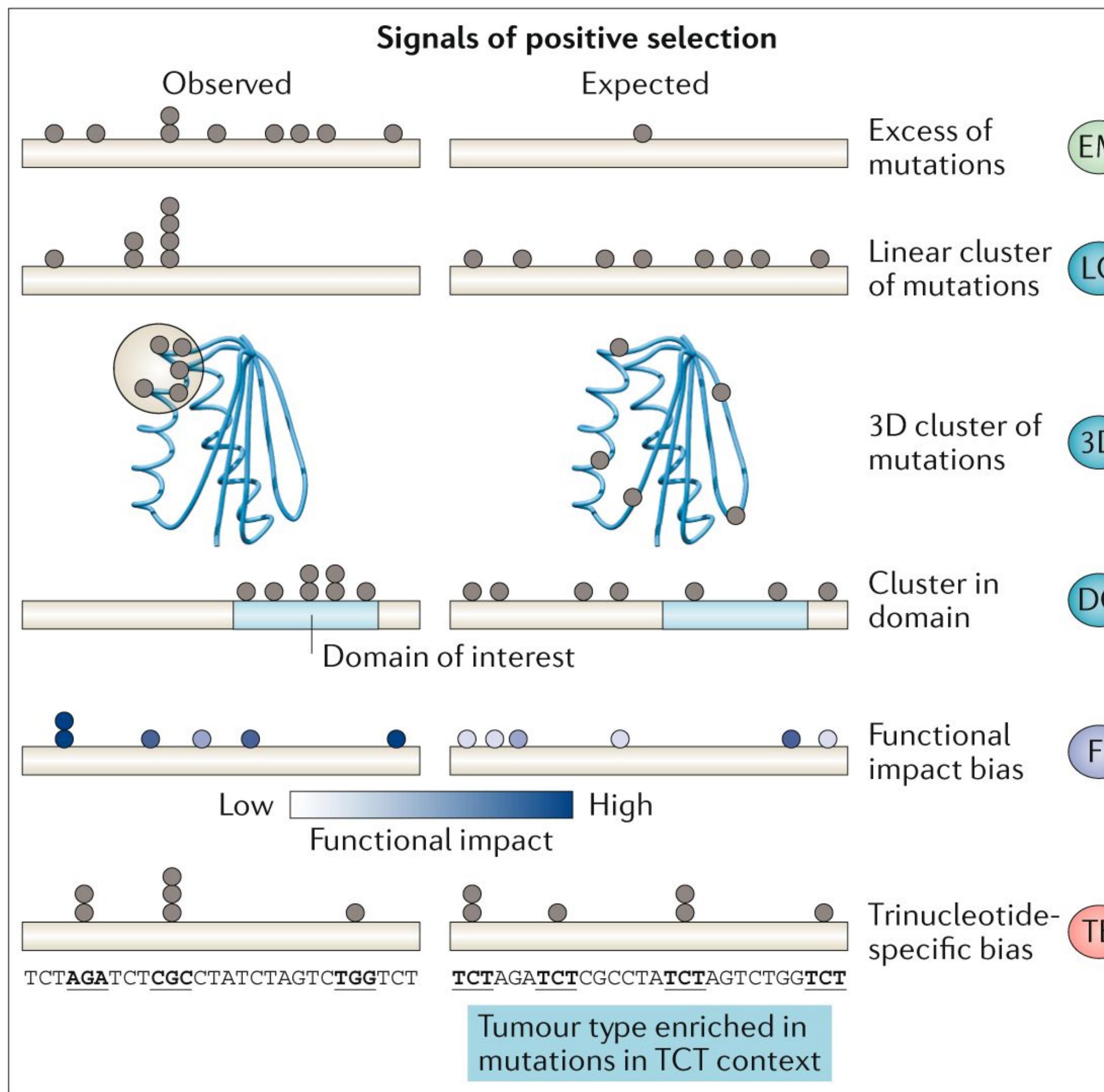


Identification of cancer driver genes: Cohort-based (recurrence) approaches



Signals of positive selection to identify cancer driver genes

Detection of positively selected genes

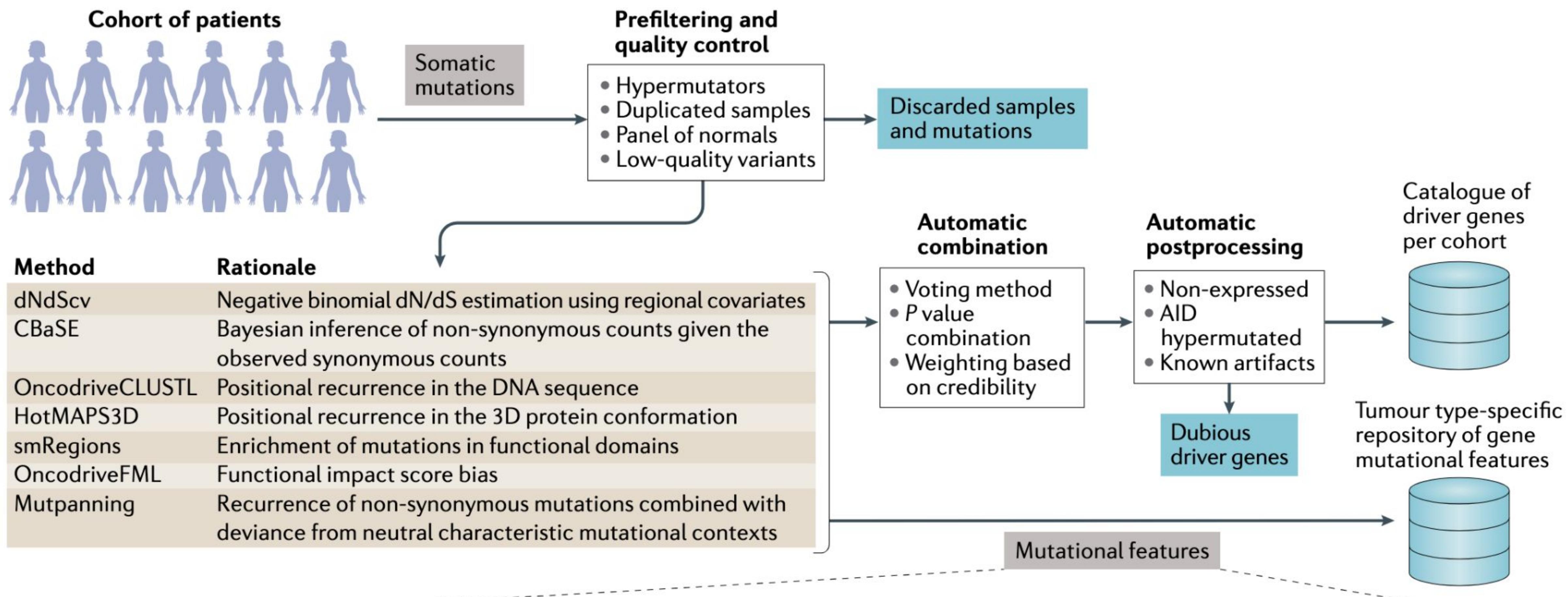


Cohort-specific catalogue of driver genes and their signals

APC	EM	FI	TB			
TP53	EM	LC	3D	DC	FI	TB
TTN						
KRAS	EM	LC	3D	FI		
PIK3CA	EM	LC	3D	FI	TB	
OBSCN						
RYR2						
...						
SMAD2	EM	LC	DC			

Consensus Driver Identification Pipeline: IntOGen

Schematic representation of the Integrative OncoGenomics (IntOGen) pipeline



<https://www.intogen.org/>



e.g. Mutation distribution of BRAF in breast cancer

[Search example](#) | [Show more examples](#)

Release 2023-05-31

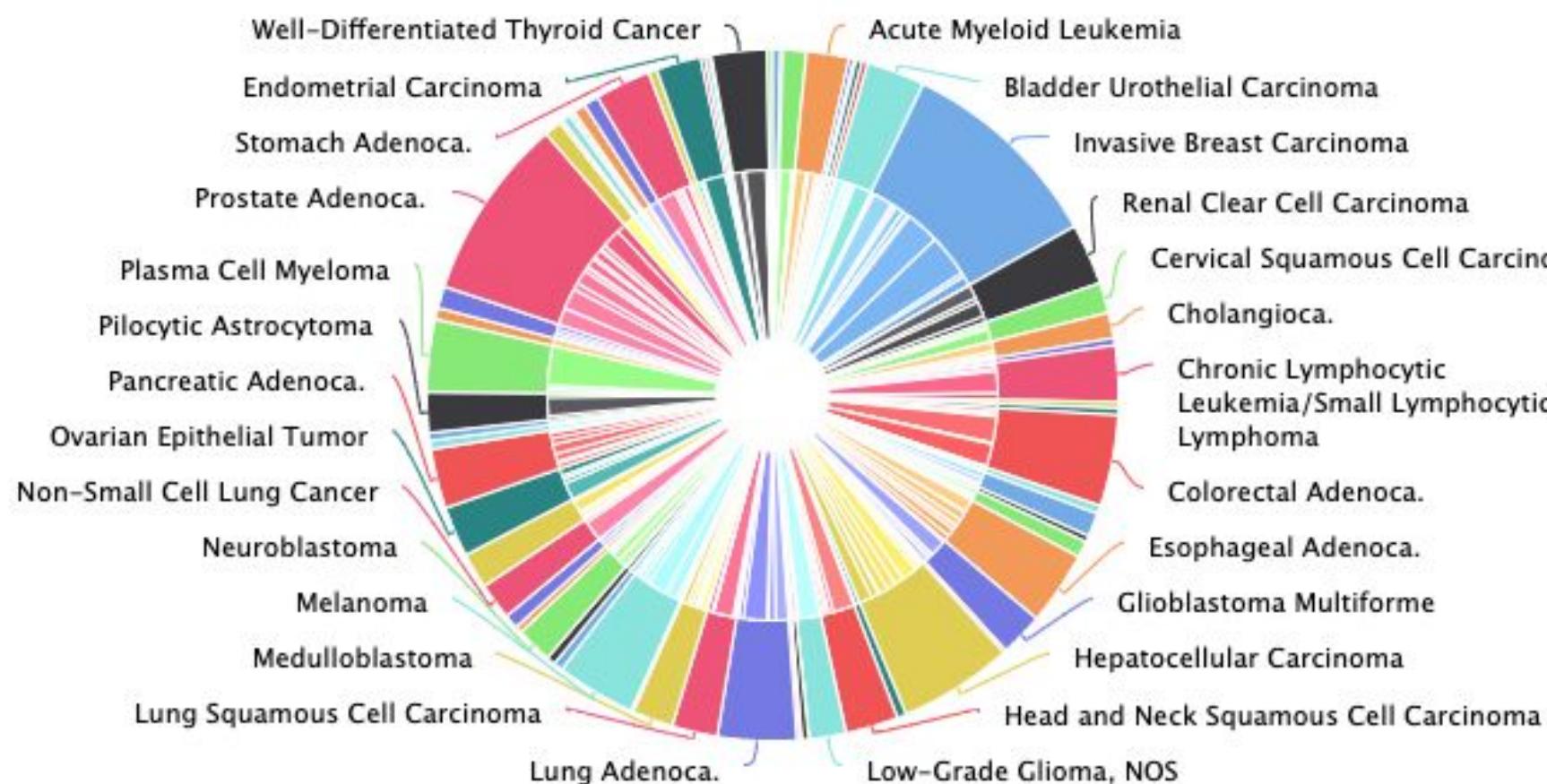
Plot

Table



IntOGen Samples

Cancer types and cohorts chart



Cancer types

73

Cohorts

266

Samples

33,019

Mutations

252,486,809

Driver genes

619

Challenges in Driver Identification from recurrence: Background mutation rate

Driver identification: Like finding needles in a haystack
Key: proper modelling of mutational process
Background mutation rate: Null (“expected”) model

- Null model difficult for structural variants:
 - CNAs (Gistic)
 - Gene fusions
 - Epigenetic alterations
- Better understood for point mutations
 - SNPs
 - Indels

Background mutation rate depends on:

- The nucleotide sequence
- Active mutational processes
- Identity of the cell/tissue
- Chromatin and regional context
- Formation of non-B-DNA structures
- Level of gene expression

Background mutation rate: Sequence composition

Most SNP mutational processes can be described using the tri-nucleotide context

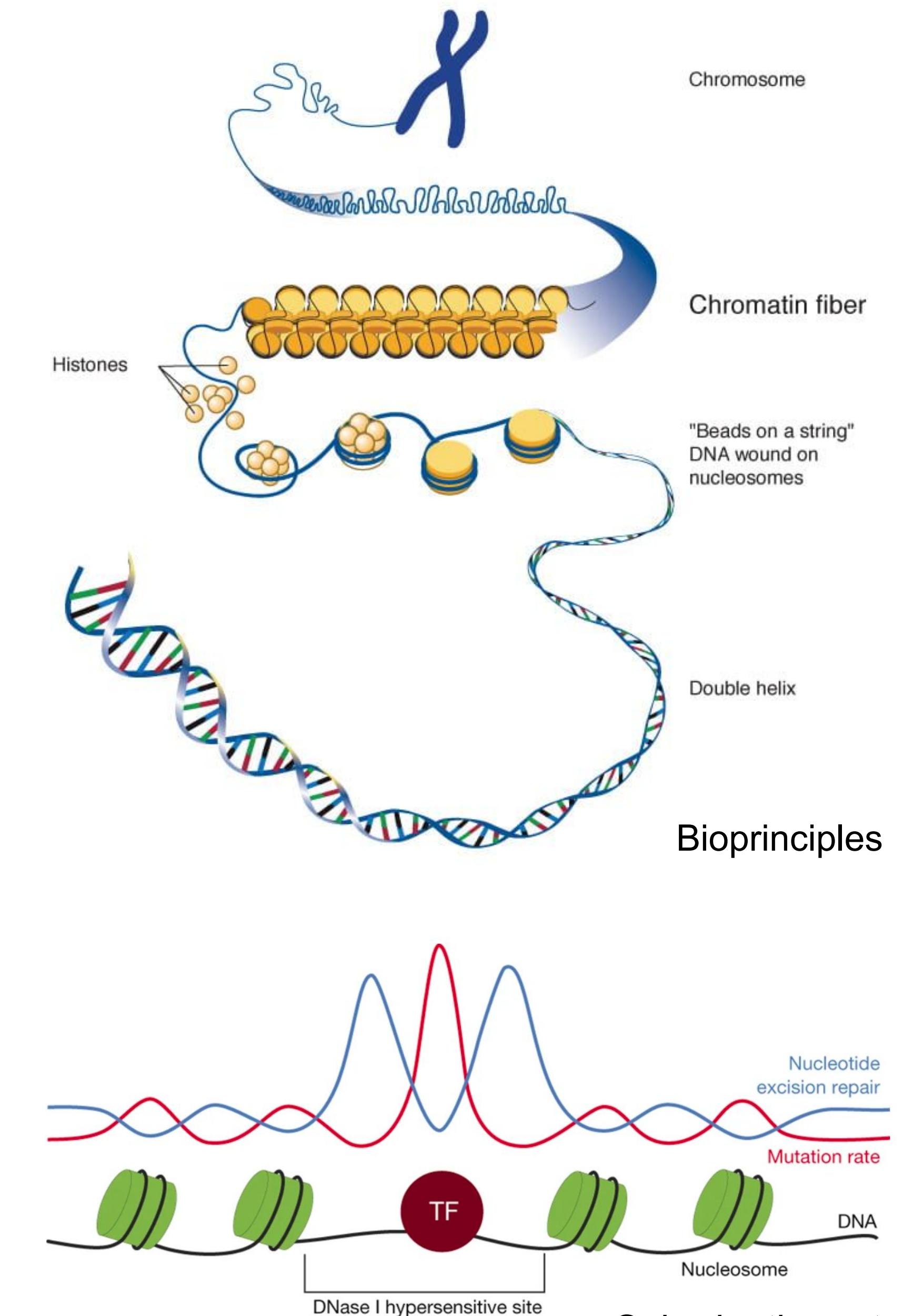
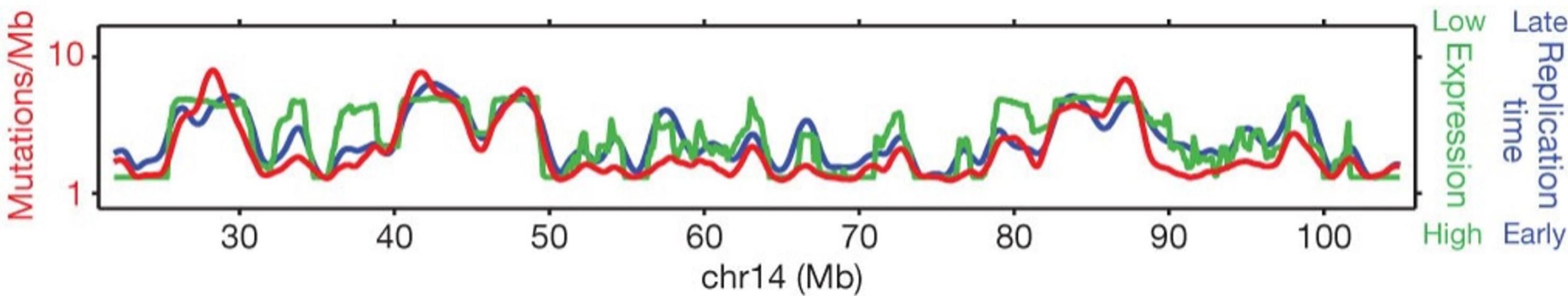
E.g. C>T at CpGs tend to have a rate per bp ~40x higher than others

Random mutations:



Background mutation rate: Regional variation

- Level of chromatin compaction
- Distribution of chromatin marks
- Occupancy by nucleosomes and other proteins
- Level of gene expression
- Replication timing



dN/dS

Concept Review: Types of point mutations

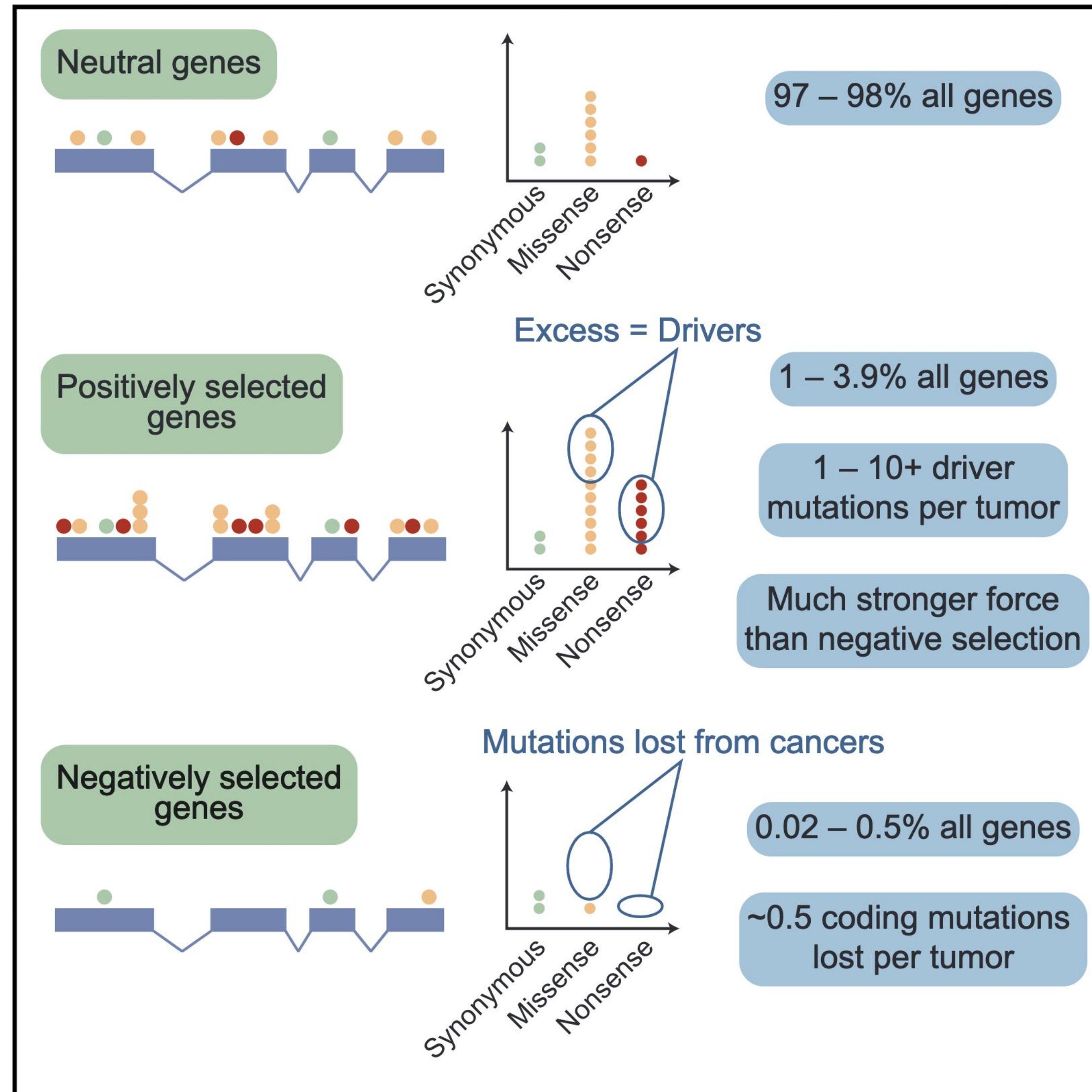
		Second letter						
		U	C	A	G			
First letter	U	UUU UUC UUA UUG	UCU UCC UCA UCG	UAU UAC UAA UAG	Tyr Stop Stop	UGU UGC UGA UGG	Cys Stop Trp	U C A G
	C	CUU CUC CUA CUG	CCU CCC CCA CCG	CAU CAC CAA CAG	His Pro Gin	CGU CGC CGA CGG	Arg	U C A G
	A	AUU AUC AUA AUG	ACU ACC ACA ACG	AAU AAC AAA AAG	Asn Thr Lys	AGU AGC AGA AGG	Ser Arg	U C A G
	G	GUU GUC GUA GUG	GCU GCC GCA GCG	GAU GAC GAA GAG	Asp Glu	GGU GGC GGA GGG	Gly	U C A G
Third letter								

Synonymous changes: e.g. TAT > TAC (Tyr → Tyr)

Non-synonymous changes:

- Missense changes: e.g. TAT > TGT (Tyr → Cys)
- Nonsense changes: e.g. TAT > TAA (Tyr → Stop*)
- Splice-site changes
- Others: Indels (in-frame, out-of-frame), stop-lost, etc

The idea behind dN/dS



Synonymous changes accumulate neutrally (generally)

- scale with background mutation rates
- comparing with them helps control for gene length and mutation rate differences

Calculating dN/dS

To formalize the observed versus expected test, estimate the coefficient selection **W=dN/dS** where:

dN = number of **non-synonymous** changes per non-synonymous site

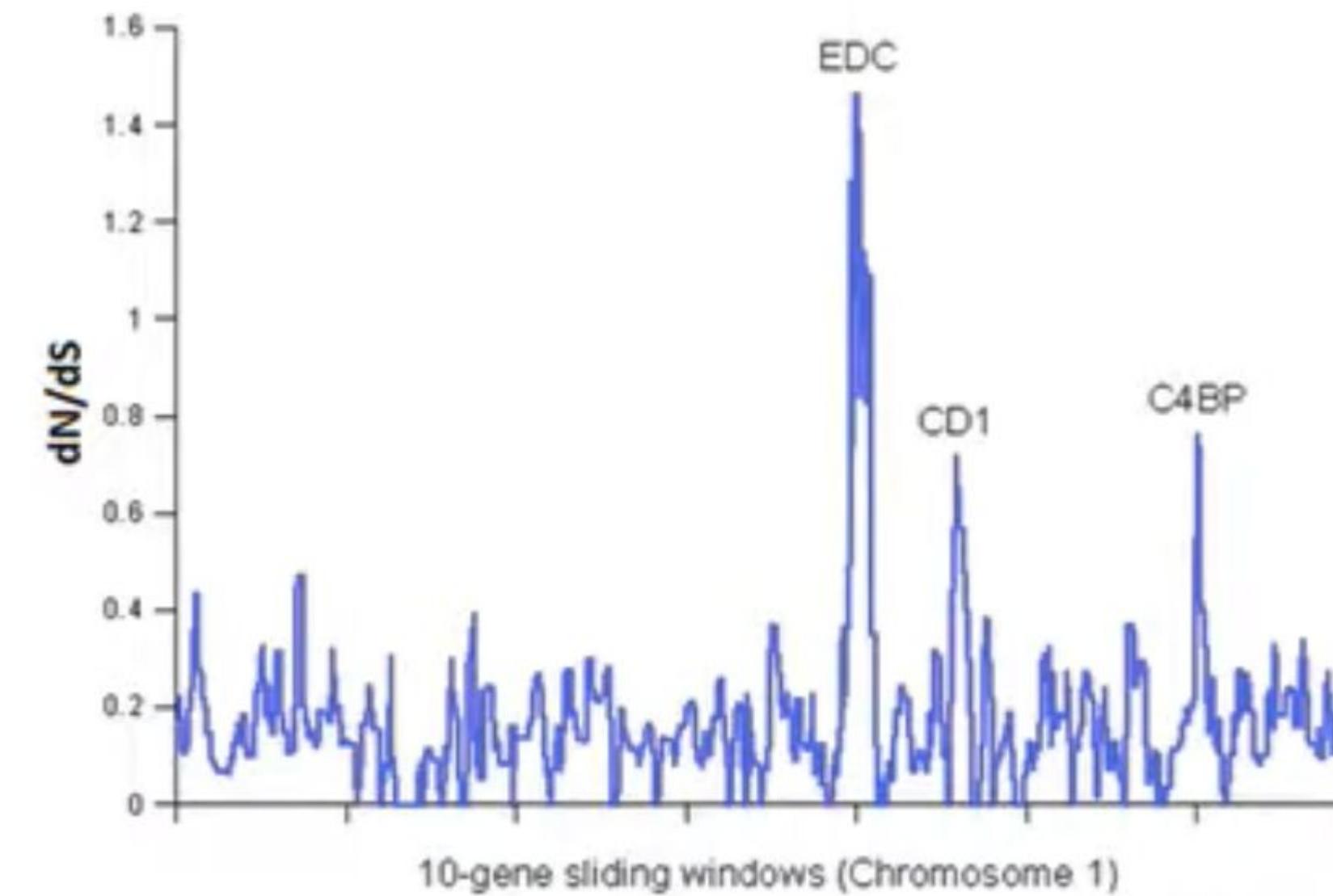
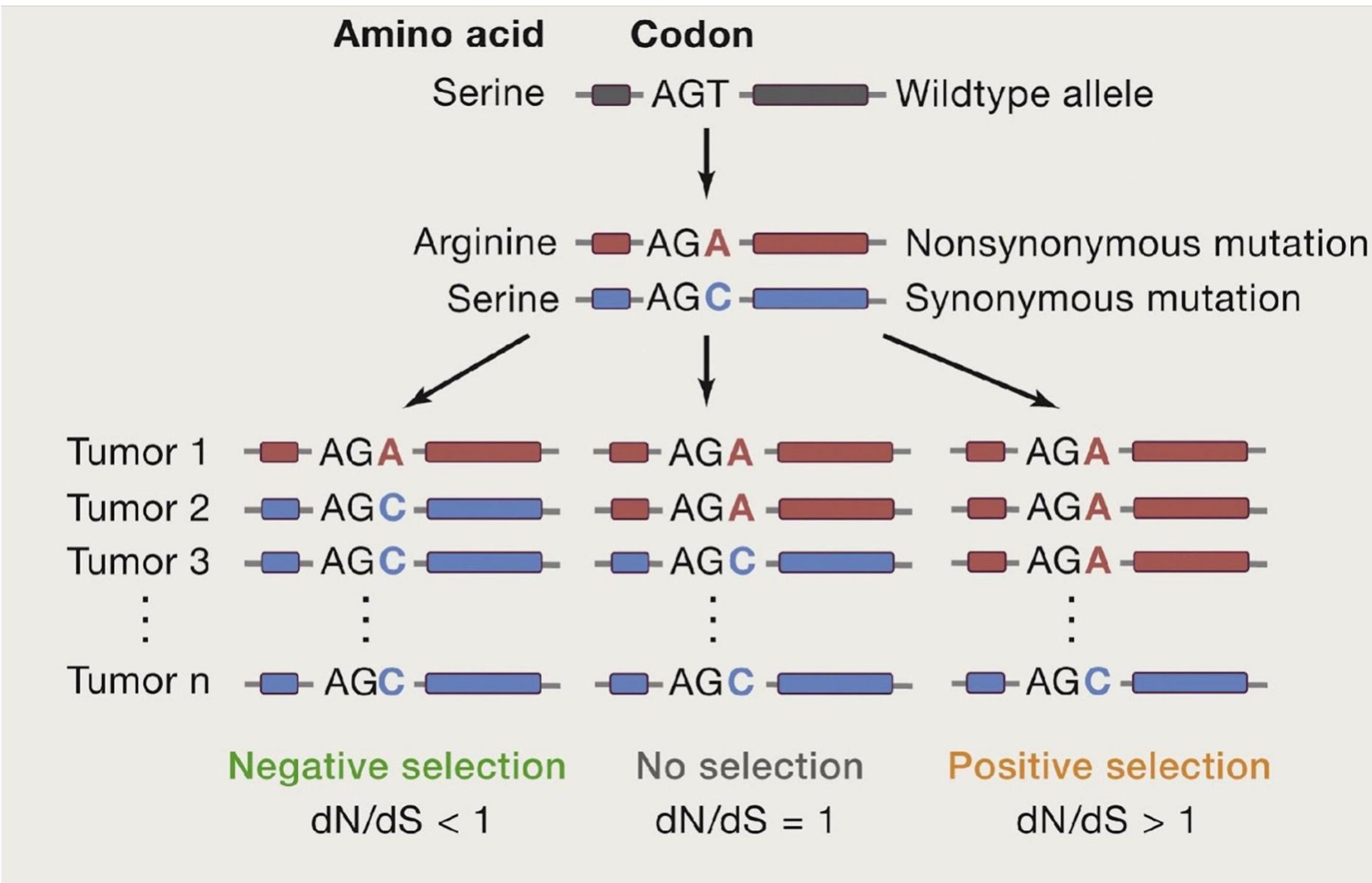
dS = number of **synonymous** changes per synonymous site

		ACT	CCG	GGG	CCC		Thr	Pro	Gly	Pro
		*		*	*					
		ACG	CCG	GGC	CTC		Thr	Pro	Gly	Leu
	site #	123	456	789	111	012				
syn		001	001	001	001		Total synonymous sites = 4			
non		110	110	110	110		Total nonsynonymous sites = 8			
							Total synonymous changes=2			
							Total nonsynonymous changes=1			
							$dN = 1/8 = 0.125$	$dS = 2/4 = 0.5$		
									$dN/dS = 0.125/0.5 = 0.25$	

<https://www.youtube.com/watch?v=ZlcHujBbQ2Q>

Interpreting dN/dS

Human vs chimp dN/dS



- dN/dS of 1 \rightarrow All observed non-syn are expected
- dN/dS of 2 \rightarrow 50% of non-syn are selected
- dN/dS of 10 \rightarrow 90% of non-syn are selected
- dN/dS of 100 \rightarrow 99% of non-syn are selected

w = dN/dS (coefficient of selection)

(w-1)/w = fraction mutations selected

Refining dN/dS for somatic evolution: dNdScv (R package)

dndslc - Traditional dN/dS calculation

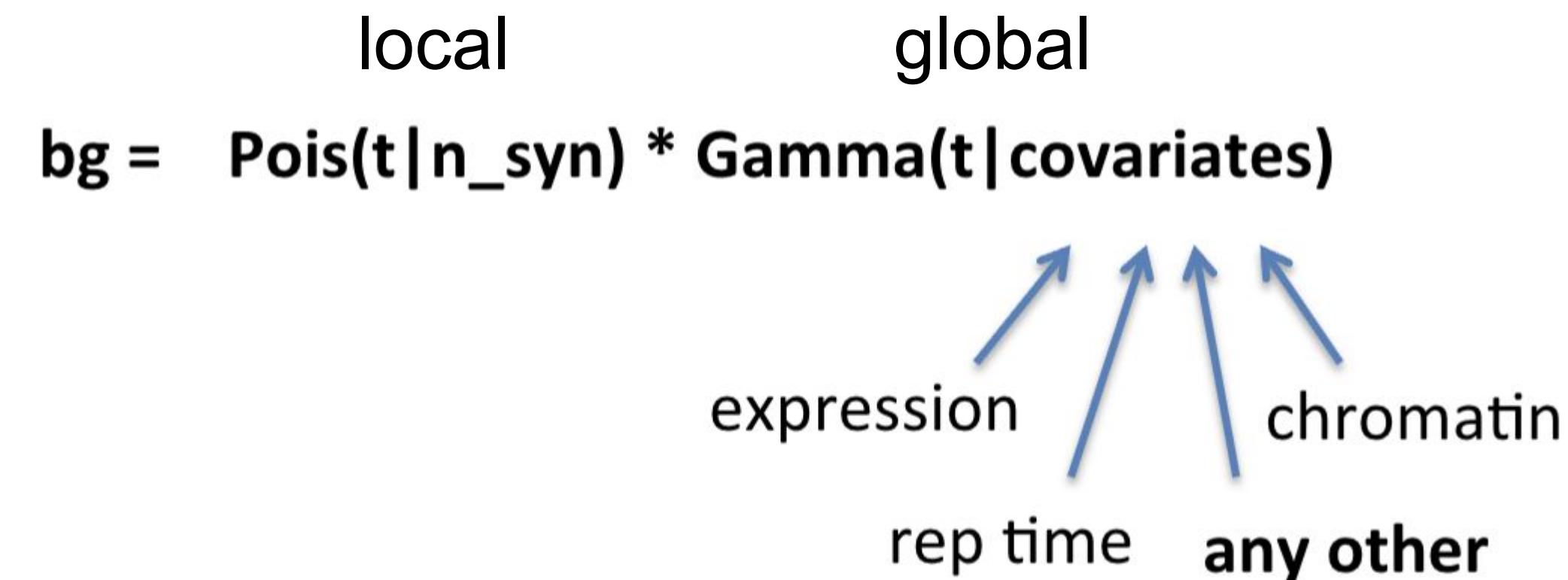
Assumes all mutations are equally likely

Underpowered for small cohorts or small genomic regions

dndscv - Refined dN/dS calculation

Accounts for:

- Context-dependent mutational processes (6*16*2 rate parameters)
- Mutation rate variation along the genome (eg between genes)
- Separately models missense, nonsense, splice and indel mutations



dndslc and dndscv converge in large datasets with sufficient number of synonymous mutations per gene!

Evaluation of statistical significance:
Likelihood ratio tests with Benjamini Hochberg for multiple hypothesis testing

Mutation Hotspots

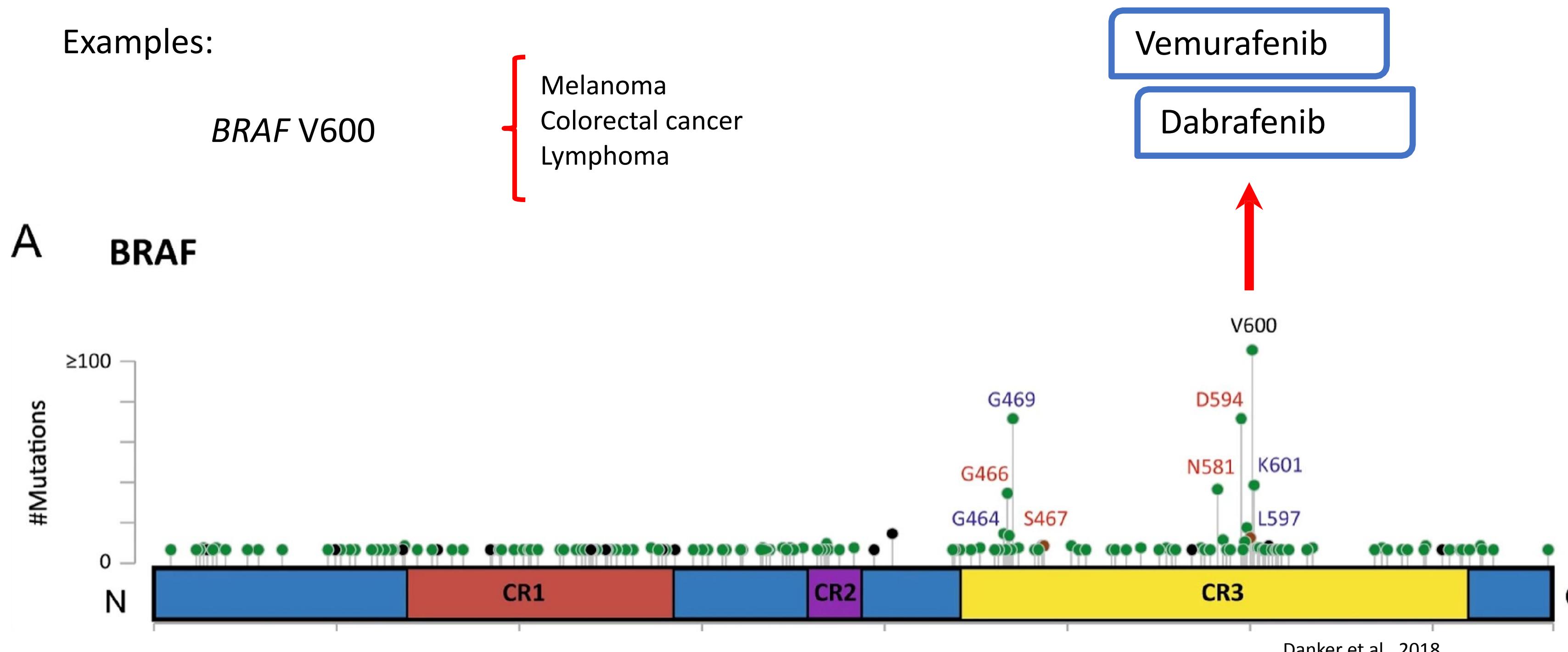
Resolution matters!

Same gene can have different regions under different selection pressures:
constraint (negative selection), neutral evolution (no selection), rapid adaptation (positive selection)

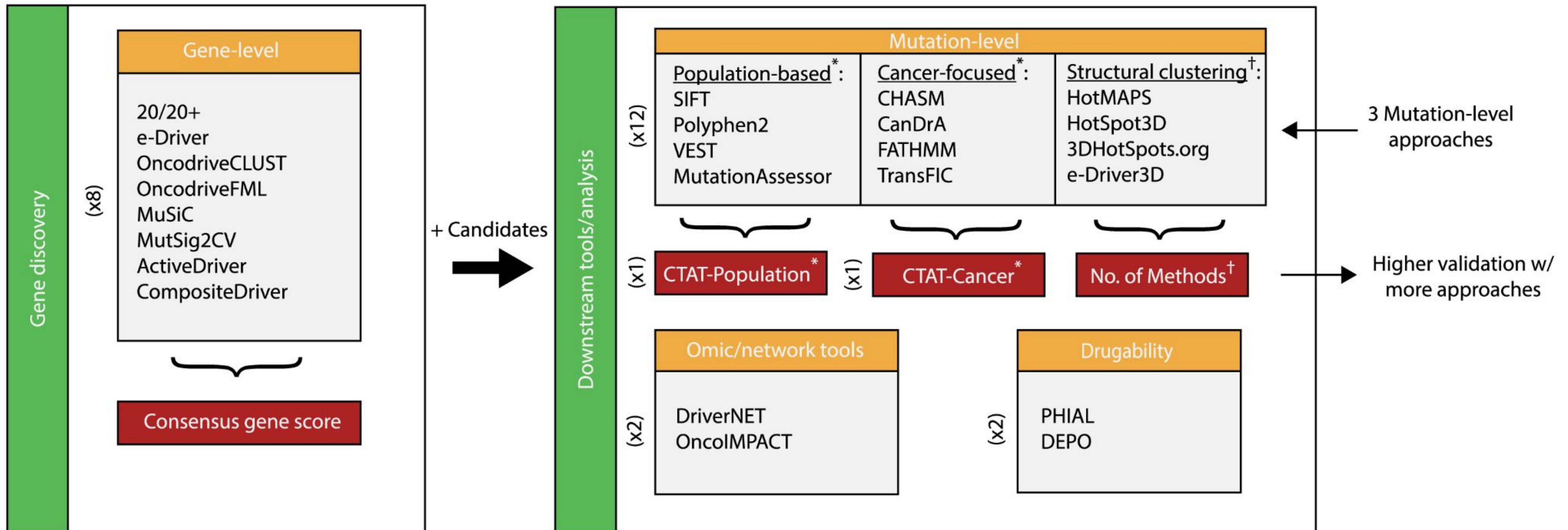
dN/dS at gene level will average out signals!

dNdScv estimates selection coefficients at the levels of genome, gene, domain, codons/aa sites

Lollipop plots:
Visualising mutations in a gene



Identification of cancer driver genes: Bailey et al.



Data Curation

Somatic mutations were filtered to avoid sequencing artifacts and hypermutated samples that would lower the statistical power of this study.

Tool Deployment

Multiple tools using distinct genetic signals were used to statistically identify cancer drivers. Results were reported in a standardized format.

Outlier Adjustment

The mutational heterogeneity in cancer leads to models which occasionally provide unreliable results. Outlier results received less weight.

Manual Curation

Manual evaluation of genes by cancer experts to eliminate false positives and rescue prominent known cancer driver genes.

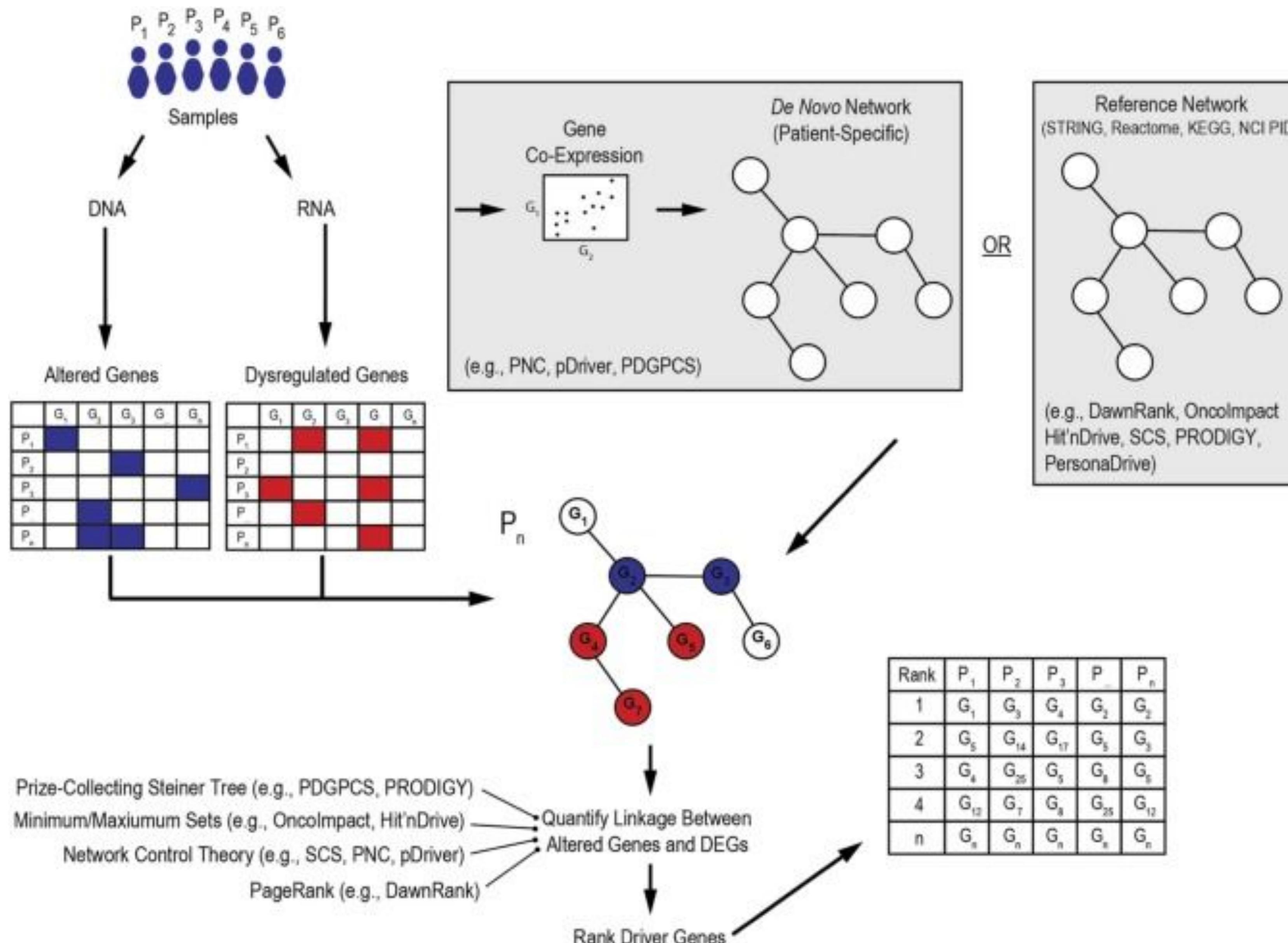
Downstream Tools/Analysis

Downstream analysis of gene expression, clonality, protein structure, and clinically-related features.

Functional Validation

Predicted mutations were confirmed in an independent dataset of experimentally validated mutations using two cell line models.

Identification of cancer driver genes: Network-based approaches



Other kinds of drivers

Structural drivers

- Copy number gains / loses
- Gene fusions
- Rearrangements

Challenges

- Modelling the background
- Involve multiple chromosomes
- Require long read sequencing

Non-coding somatic drivers

- lncRNAs
- UTRs/promoters
- tRNAs
- small RNAs
- micro RNAs

Example:

TERT promoter

Challenges

- Sequencing and mapping artefacts
- Incomplete annotation of regulatory regions
- Unknown functional effects

Time for practicals!

Driver identification: dN/dS (dndscv)

Visualisation: Oncoplots and Lollipop plots (maf-tools)

Clinical relevance of cancer drivers

Why identify drivers?

- Biomarkers of prognosis
- Biomarkers of response to treatments
- Targeted cancer therapies

Bottleneck: Requires identification of actionable genetic alterations

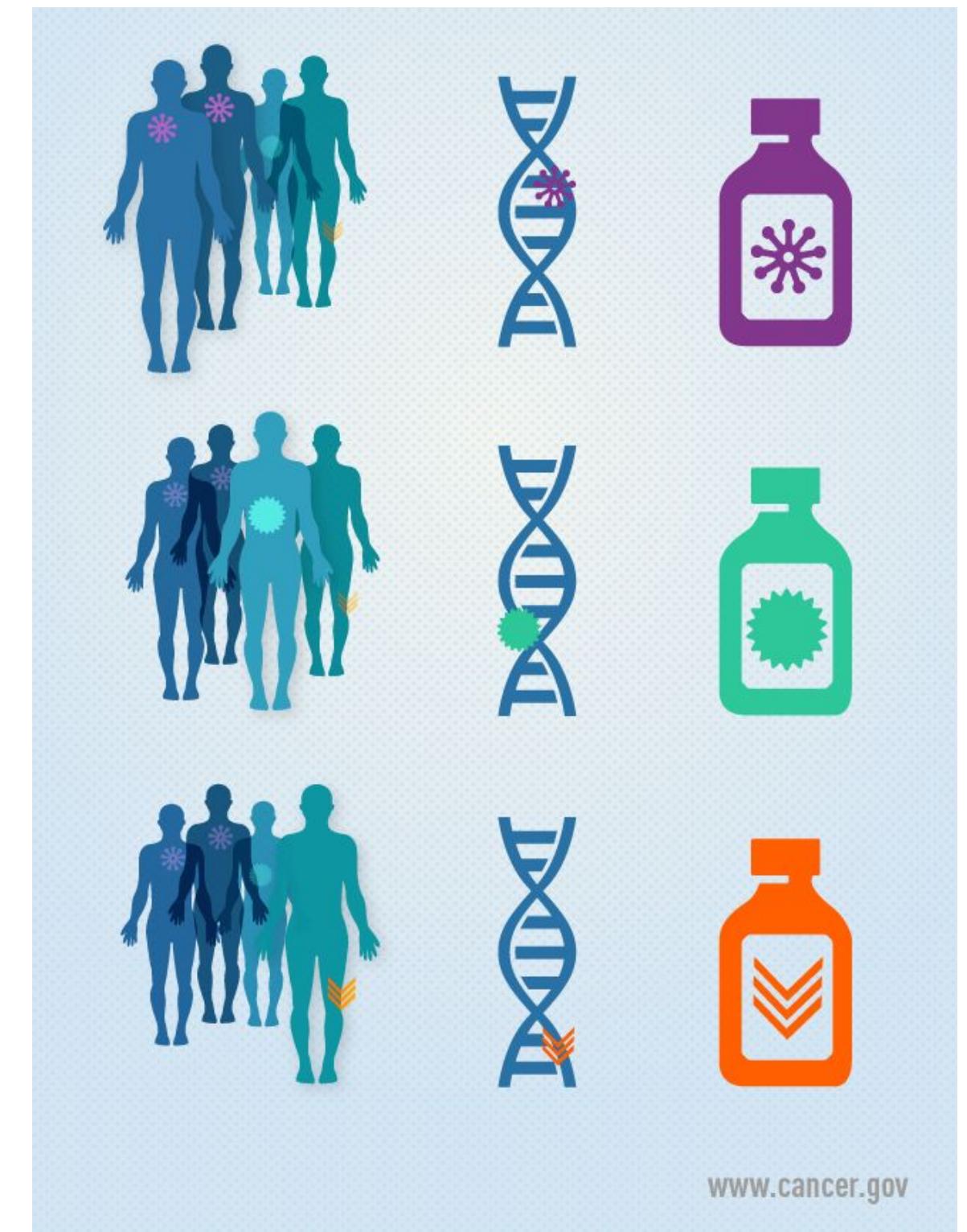
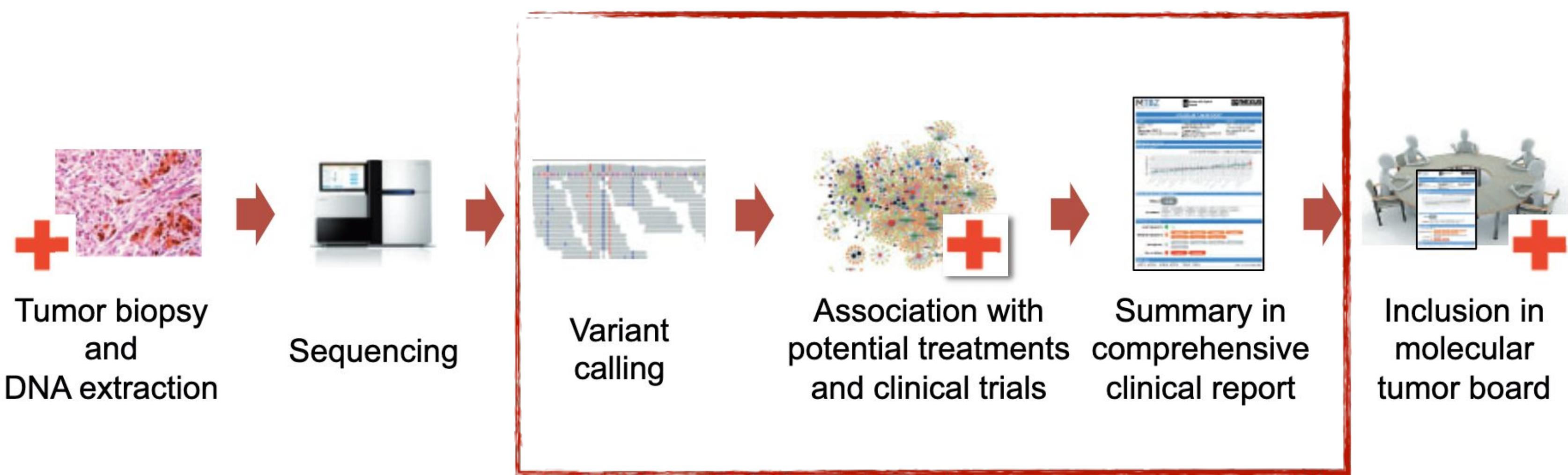
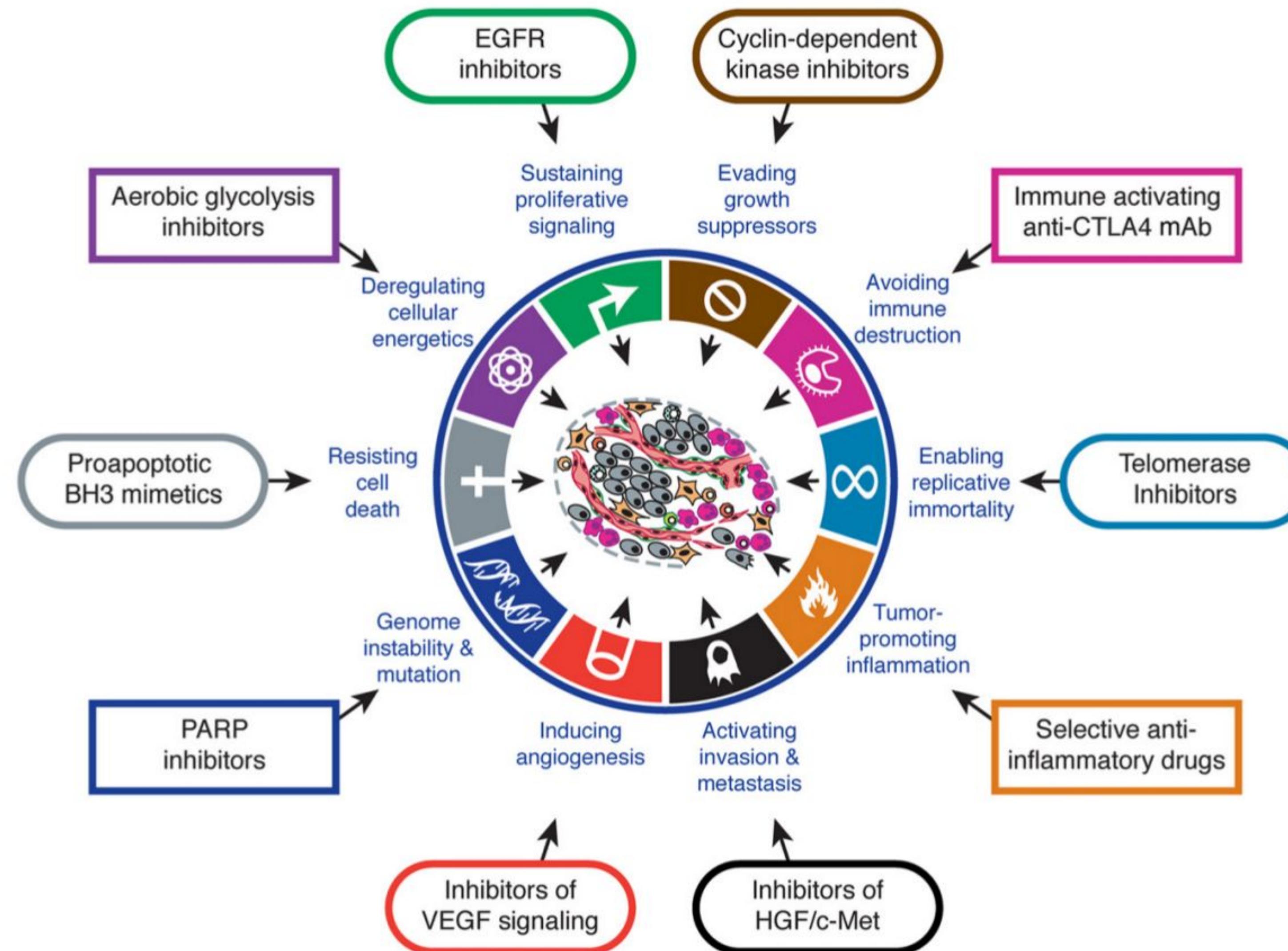


Image adapted from <https://www.cancer.gov/>

Therapeutic targeting of the hallmarks of cancer



Reporting summary

MUTATION SUMMARY

Mutational burden: 833 SNVs (292 non-synonymous), 17 CNVs (affecting 2028 genes)

High mutational burden indicates potential benefit of anti-CTLA4 therapy

HLA Type: A*03:02, A*03:02, B*07:01, B*08:01, C*07:02, C*07:02

Somatic mutations in melanoma-typical genes (based on mycancergenome.org and customer requests):

BRAF: p.Val600Glu CDKN2a/b: Copy number loss x1

Melanoma-typical genes without somatic mutations (based on mycancergenome.org and customer requests):

CTNNB1, CDK4/6, EGFR, GNA11, GNAQ, KIT, MEK1, MET, NF1, NRAS, PIK3CA, PTEN

CANCER TYPE-SPECIFIC THERAPIES

Gene	Treatment	Mutation	Variant Frequency / Copy number	Pathway	Confidence	References
BRAF	Dabrafenib, Vemurafenib	c.1799T>A p.Val600Glu	38.1%	MAPK signaling	I.A*	2,3,4,5

NON CANCER TYPE-SPECIFIC THERAPIES

Gene	Treatment	Mutation	Variant Frequency / Copy number	Pathway	Confidence	References
BRCA2	Olaparib	Copy number loss	x1 (hemizygous deletion)	DNA repair	I.C	7,12,13,14, 26, 27
SRC	Dasatinib, Bosutinib, Ponatinib	Copy number gain	x3	Growth signal transduction	I.B	18,19, 20,25

INVESTIGATIONAL THERAPIES

Gene	Treatment	Mutation	Variant Frequency / Copy number	Pathway	Confidence	References
CDKN2A	PD0332991 (Palbociclib)	Copy number loss	x1 (heterozygous deletion)	Cell cycle	I.B*	6,7,8, 9,10,11

THERAPIES WITH POTENTIAL LACK OF BENEFIT

Therapy	Responsible Mutation	Description	References
Sunitinib, Sorafenib	FLT3 Copy loss	Inhibitor treatment not advisable for copy loss. In addition, absence of cKIT mutation	21,22,23
Regorafenib	FRK Copy loss	Inhibitor treatment not advisable for copy loss.	24

COMPREHENSIVE LIST OF MUTATIONS

See Appendix for comprehensive overview and details of all 833 SNVs and 17 CNVs.

Association of genomic alterations with potential treatments and clinical trials

DGIdb

The previous version of DGIdb can be found at old.dgide.org until June 1st, 2024.

[About](#) [Publications](#) [Types/Directionalities](#) [Interaction Score](#) [FAQ](#) [Known Data Clients](#) [Contact](#) [Current Contributors](#) [Acknowledgements](#)

About

Mining the Druggable Genome for Personalized Medicine

Integration of the Drug-Gene Interaction Database (DGIdb 4.0) with open crowdsource efforts. Freshour S*, Kiwala S*, Cotto KC*, Coffman AC, McMichael JF, Song J, Griffith M, Griffith OL, Wagner AH. Nucleic Acids Research. 2020 Nov 25; doi: <https://doi.org/10.1093/nar/gkaa1084>. PMID: 33237278

In the era of clinical sequencing and personalized medicine, investigators are frequently presented with lists of mutated or otherwise altered genes implicated in disease for a specific patient or cohort. Numerous resources exist to help form hypotheses about how such genomic events might be targeted therapeutically. However, utilizing these resources typically involves tedious manual review of literature, clinical trial records, and knowledgebases. Few currently exist which collect and curate these resources and provide a simple interface for searching lists of genes against the existing compendia of known or potential drug-gene interactions. The drug-gene interaction database (DGIdb) attempts to address this challenge. Using a combination of expert curation and text-mining, drug-gene interactions have been mined from DrugBank, PharmGKB, ChEMBL, Drug Target Commons, and others. Genes have also been categorized as potentially druggable according to membership in selected pathways, molecular functions and gene families from the

Oncokb™ Levels of Evidence Actionable Genes Oncology Therapies Cancer Genes API / License About News FAQ 🔍 ⓘ

Welcome to Oncokb™

MSK's Precision Oncology Knowledge Base

An FDA-Recognized Human Genetic Variant Database*

858

Genes

7729

Alterations

137

Cancer Types

138

Drugs

Search Gene / Alteration / Cancer Type / Drug / Genomic Variant ⓘ

Therapeutic Levels

Diagnostic Levels

Prognostic Levels

FDA Levels

① Level 1
FDA-approved drugs
53 Genes

② Level 2
Standard care
26 Genes

③ Level 3
Clinical evidence
33 Genes

④ Level 4
Biological evidence
27 Genes

Level R1/R2
Resistance
11 Genes

CANCER GENOME INTERPRETER HOME ANALYSIS ABOUT FAQ CONTACT Login

[Cancer Biomarkers](#) [Validated Oncogenic Mutations](#) [Cancer Genes](#) [Cancer Bioactivities](#)

Cancer Biomarkers database

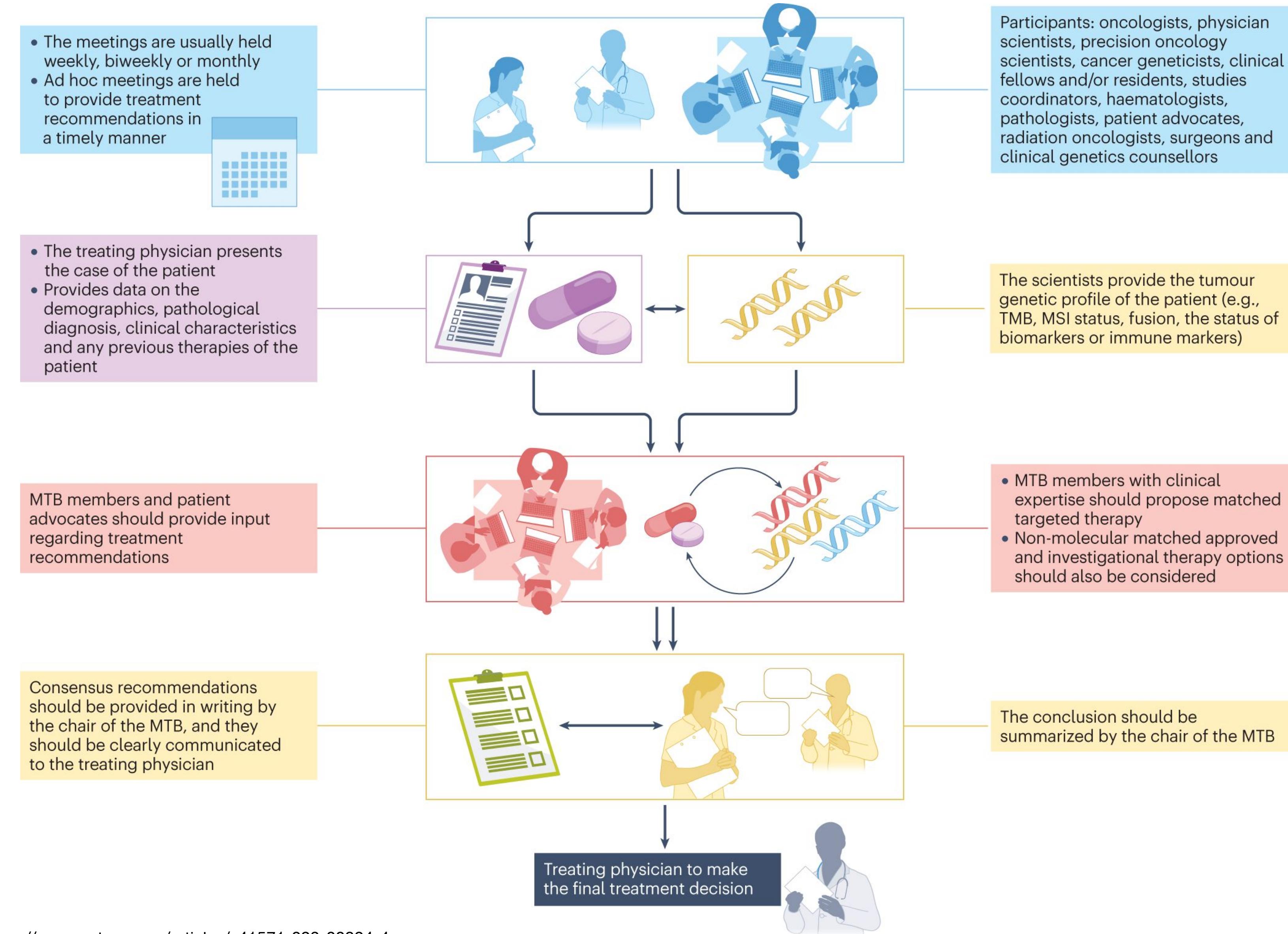
Last update: 2022/10/17

The Cancer Biomarkers database is curated and maintained by [several clinical and scientific experts](#) in the field of precision oncology supported by the European Union's Horizon 2020 funded [project](#). This database is currently being integrated with knowledge databases of other institutions in a [collaborative effort](#) of the [Global Alliance for Genomics and Health](#). The contribution of the community is encouraged and proposals of edition or comments about the information contained in this database can be given by contacting us [here](#) or by using the feedback icon located at the left of each entry of the table. The database follows the data model originally described by [Dienstmann et al.](#) This table provides a summary of the content of the database that can be interactively browsed. Additional information, including the genomic coordinates of the variants, can be accessed via the download feature. This database is licensed under a [Creative Commons Public Domain Dedication \(CC0 1.0 Universal\)](#). When referring to this database, please cite: Cancer Genome Interpreter Annotates The Biological And Clinical Relevance Of Tumor Alterations; doi: <https://doi.org/10.1101/140475>.

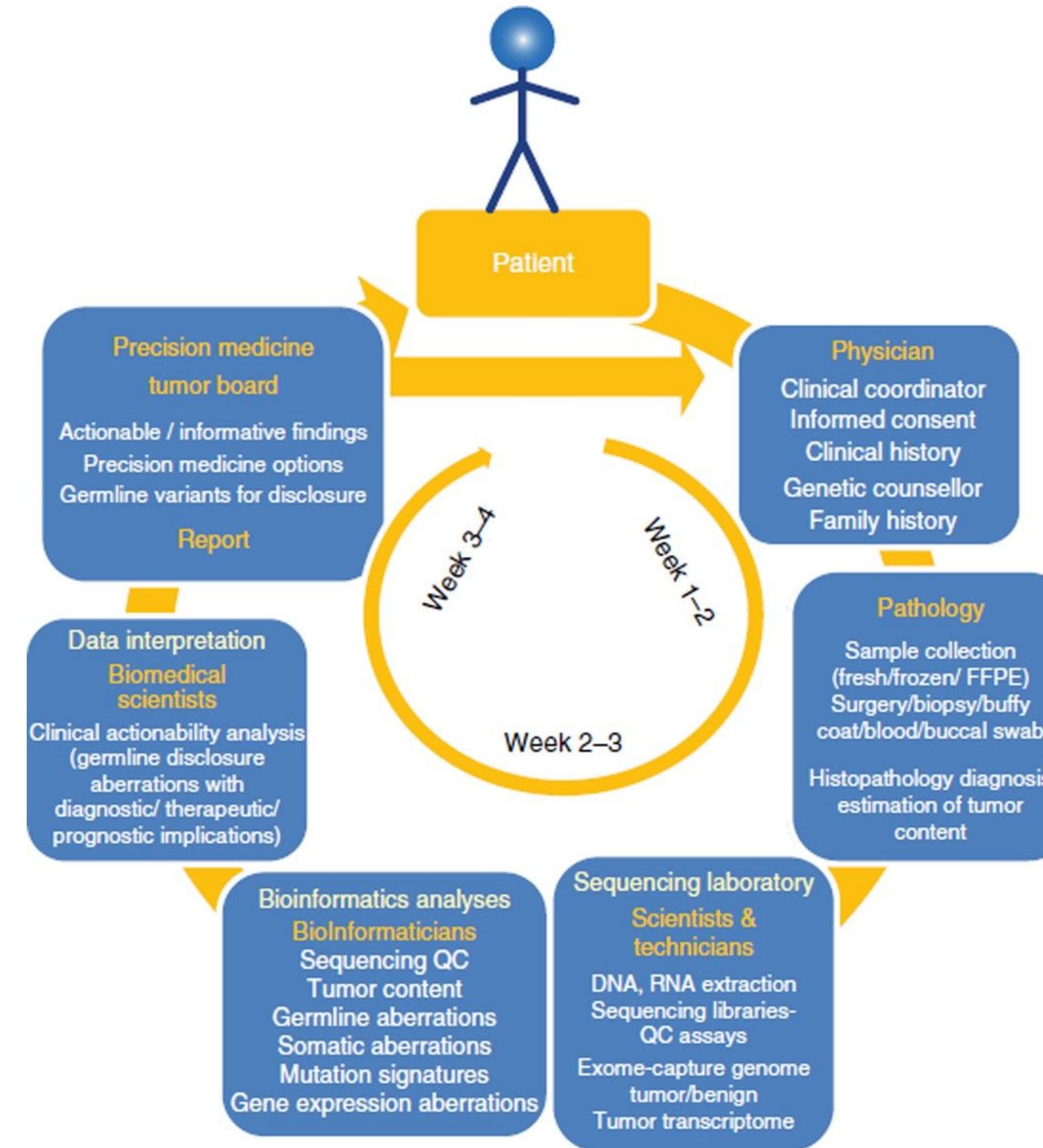
Download Feedback

Biomarker ⓘ	Drug ⓘ	Effect ⓘ	Evidence ⓘ	Source ⓘ
ABL1 (E255V,Y253H,F359V)	Ponatinib (BCR-ABL inhibitor 3rd gen&PA)	Responsive	NCCN guidelines	
ABL1 (F359V,F359C,F359I,Y)	Dasatinib (BCR-ABL inhibitor 2nd gen)	Responsive	NCCN guidelines	
ABL1 (I242T,M244V,K247R)	Imatinib (BCR-ABL inhibitor 1st gen&KIT)	Resistant	European LeukemiaNet	
ABL1 (T315A,F317L,F317V,I)	Bosutinib (BCR-ABL inhibitor 3rd gen)	Responsive	NCCN guidelines	
ABL1 (T315A,F317L,F317V,I)	Nilotinib (BCR-ABL inhibitor 2nd gen)	Responsive	NCCN guidelines	
ABL1 (T315I)	Asciminib (Kinase inhibitor)	Responsive	FDA guidelines	
ABL1 (T315I)	Nilotinib (BCR-ABL inhibitor 2nd gen)	Resistant	NCCN guidelines	
ABL1 (T315I)	Bosutinib (BCR-ABL inhibitor 3rd gen)	Resistant	NCCN guidelines	
ABL1 (T315I)	Axitinib (VEGFR inhibitor)	Responsive	Pre-clinical	

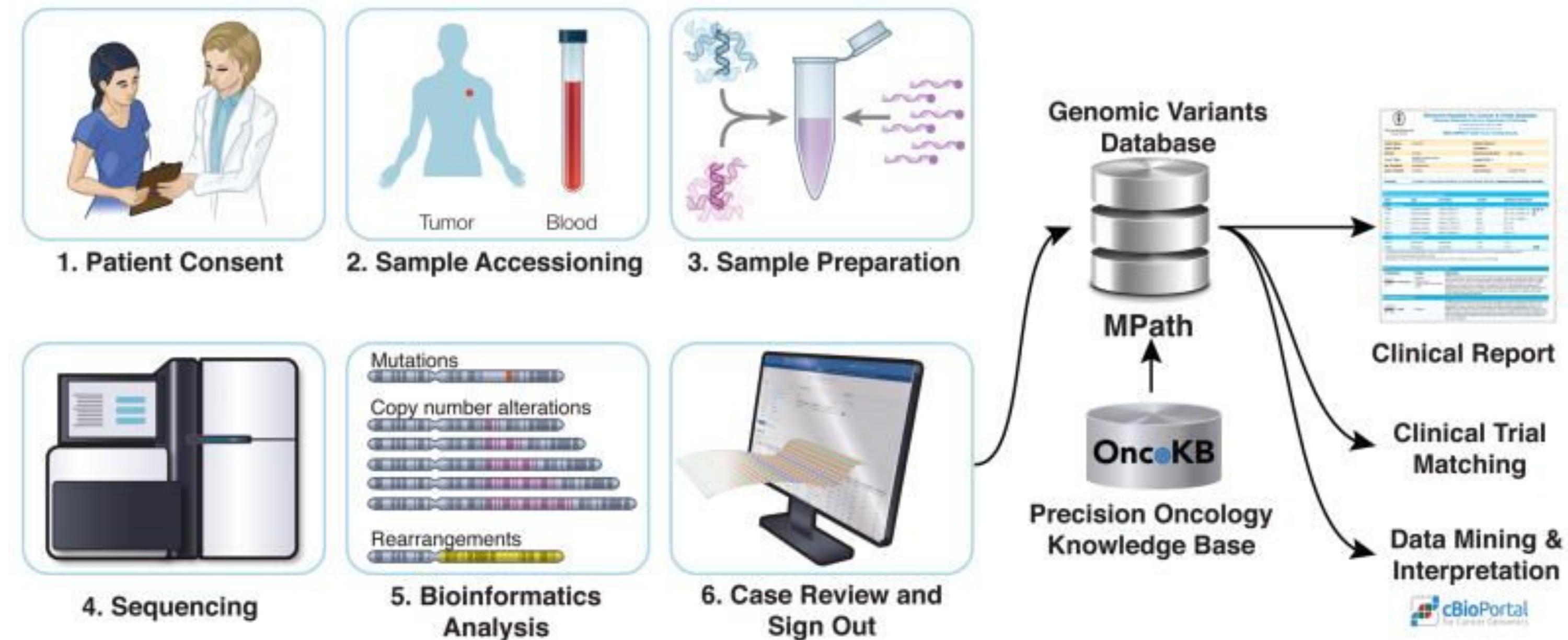
Molecular tumour board (MTB)



Workflow of integrative clinical sequencing for precision oncology



MSK-IMPACT: Integrated Mutation Profiling of Actionable Cancer Targets

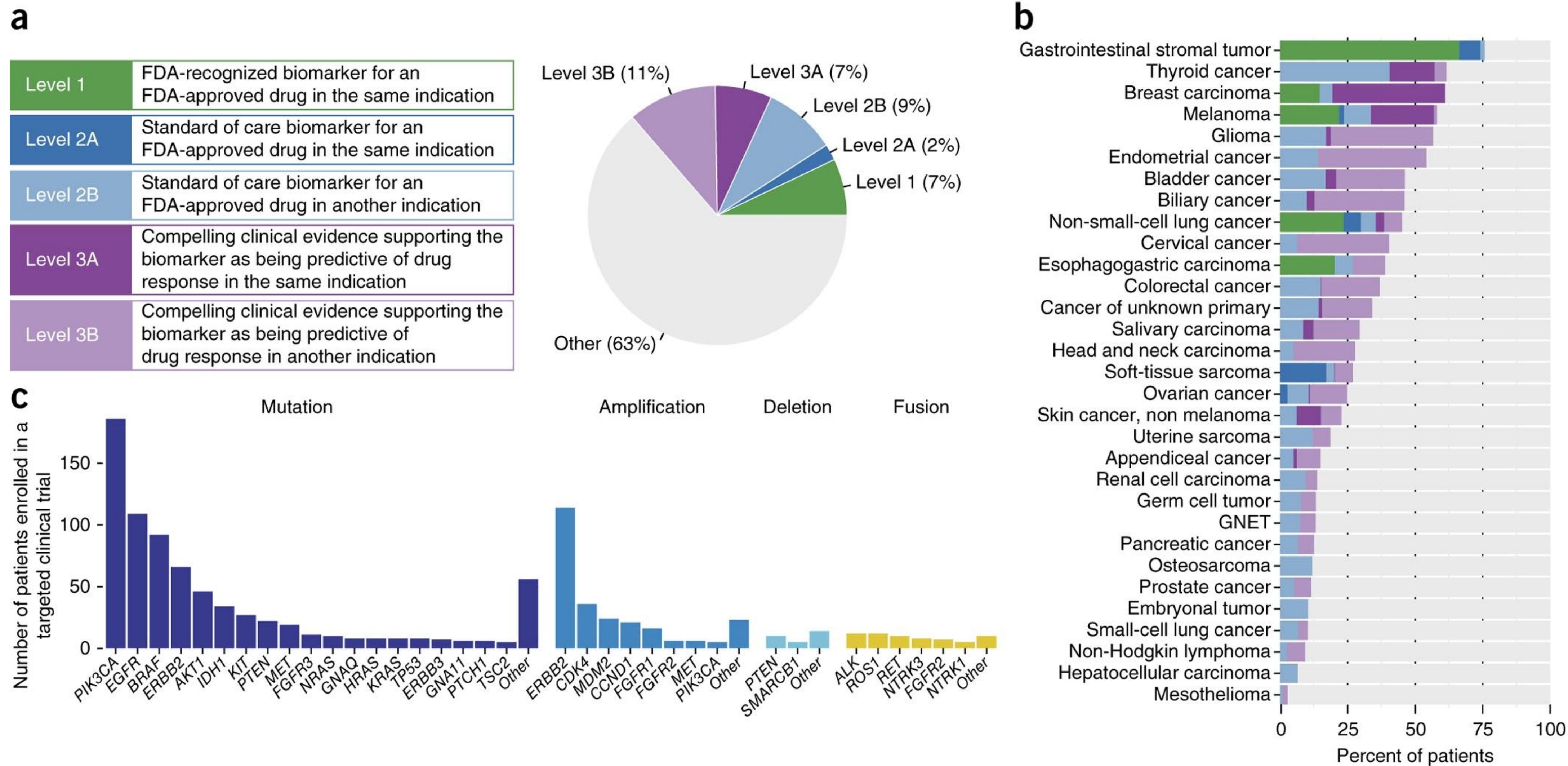


January 2014 and May 2016: **12,670 advanced tumours from 11,369 patients**

Gene panel covering: ~400 cancer related genes

median turnaround time of <21 days

Clinical actionability of somatic alterations revealed by MSK-IMPACT



36.7% of patients (n=3,792) harbored at least one actionable alteration

Mutations identified using WGS in genes indicated for testing in the NGTDC

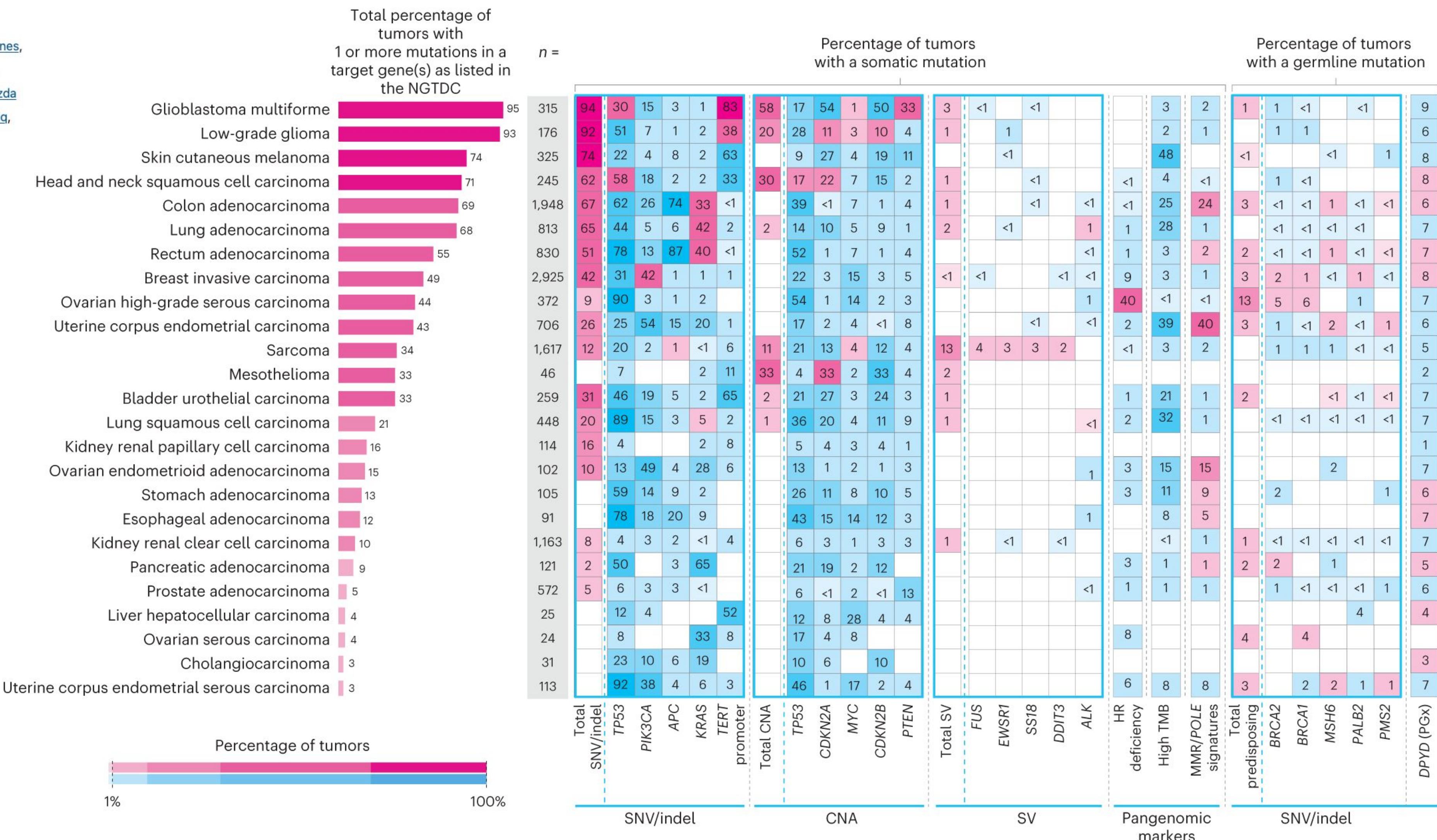
Insights for precision oncology from the integration of genomic and clinical data of 13,880 tumors from the 100,000 Genomes Cancer Programme

[Alona Sosinsky](#), [John Ambrose](#), [William Cross](#), [Clare Turnbull](#), [Shirley Henderson](#), [Louise Jones](#),
[Angela Hamblin](#), [Prabhu Arumugam](#), [Georgia Chan](#), [Daniel Chubb](#), [Boris Noyvert](#), [Jonathan](#)

Mitchell, Susan Walker, Katy Bowman, Dorota Pasko, Marianna Buongermino Pereira, Nadezda Volkova, Antonio Rueda-Martin, Daniel Perez-Gil, Javier Lopez, John Pullinger, Afshan Siddiqi,

Tala Zainy, Tasnim Choudhury ... Nirupa Murugesu  + Show author

[Nature Medicine](#) 30, 279–289 (2024) | [Cite this article](#)



“We note that the clinical actionability of these mutations will be dependent on the individual case and clinical circumstances, such as the stage of the tumor and associated comorbidities of the participant. This highlights the need for clinical interpretation and discussion where clinically appropriate within a GTAB.”

NGTDC: National Genomic Test Directory for Cancer

GTAB: Genomic Tumour Advisory Board

Timing driver alterations during clonal evolution

Dynamic adaptive landscapes redefine driver identities over time

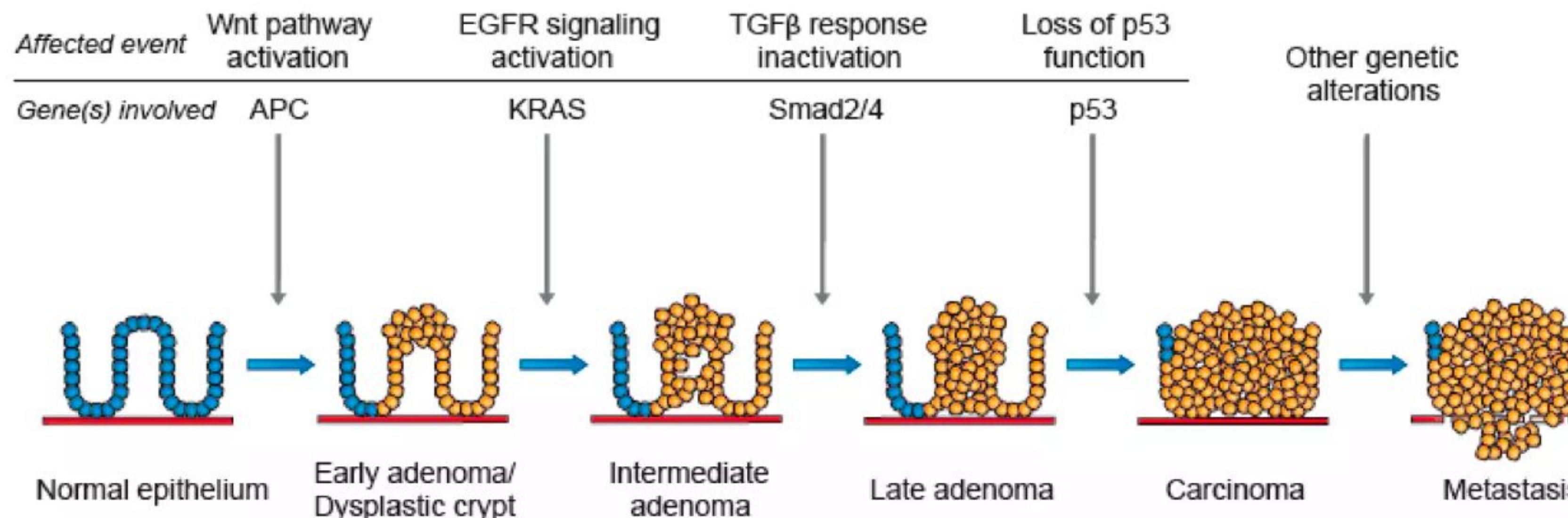
Adaptive landscapes are not static.

Former passengers may become drivers and vice-versa under changing contexts such as:

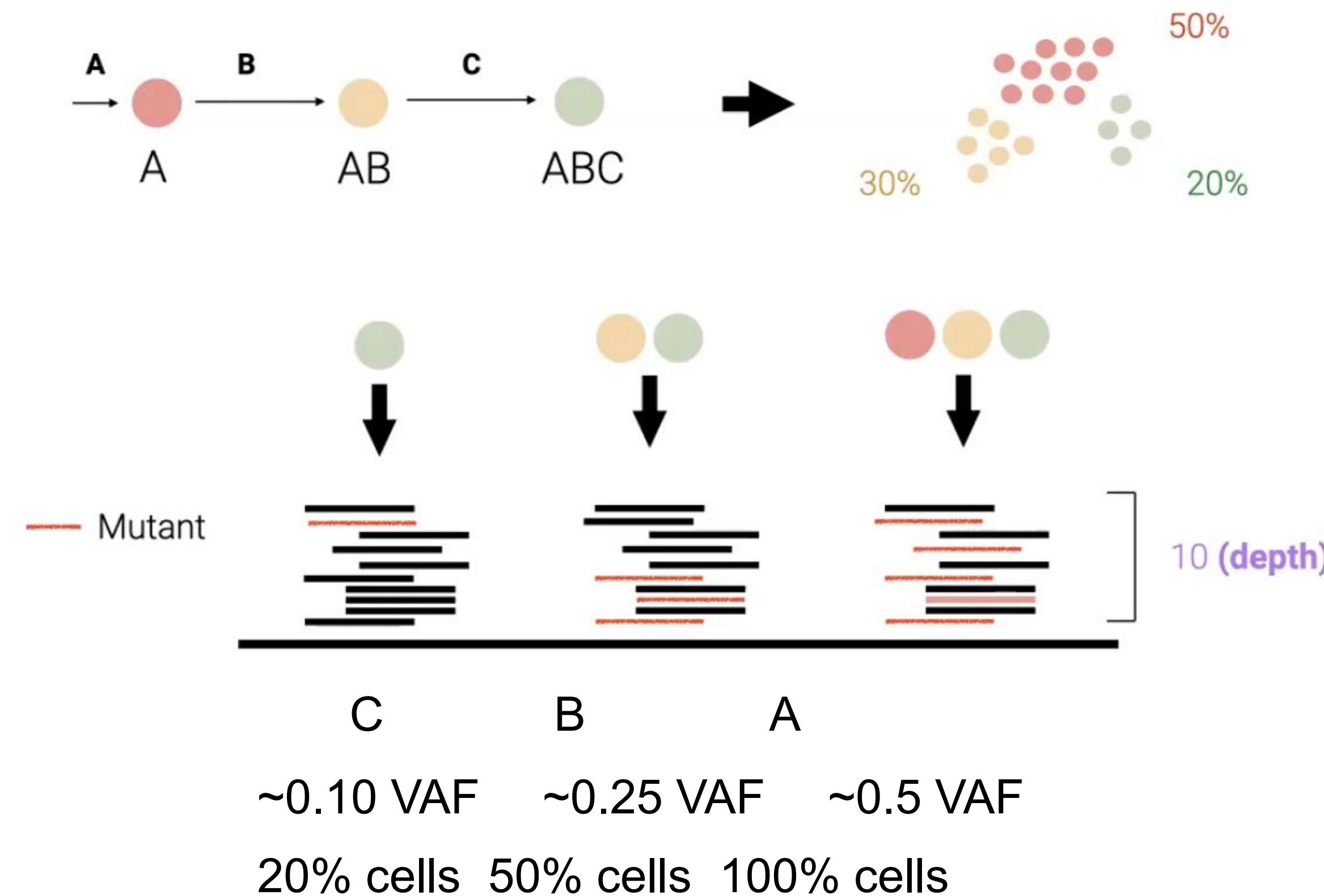
- changing microenvironments (oxygen, nutrients, pH, immune infiltration, etc.)
- metastasis and dissemination to new niches
- genomic background
- therapeutic interventions
- host-level factors (inflammation, hormones, aging, metabolic state, etc.)

You might want to know when the driver mutations occurred during clonal evolution

- longitudinal or multi-site data
- time estimates from end point data

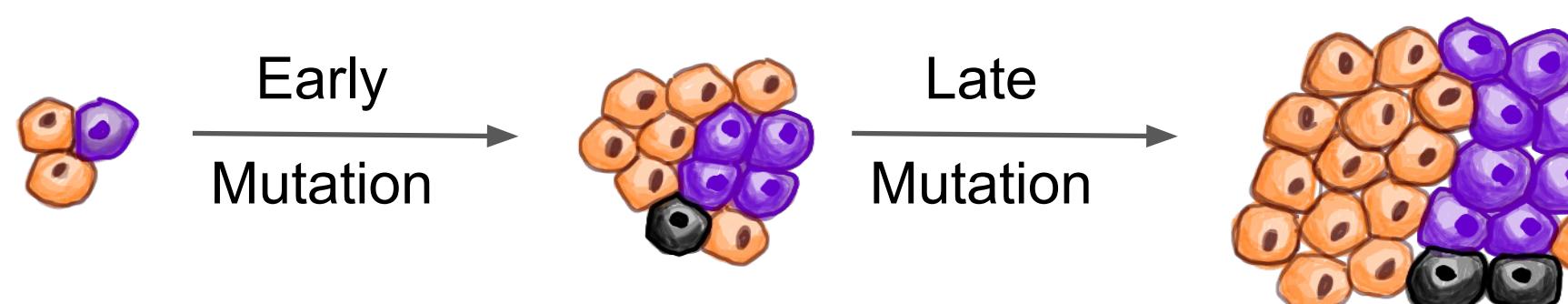


Timing point mutations



Variant Allele Frequency (VAF) = # reads with variant / read depth

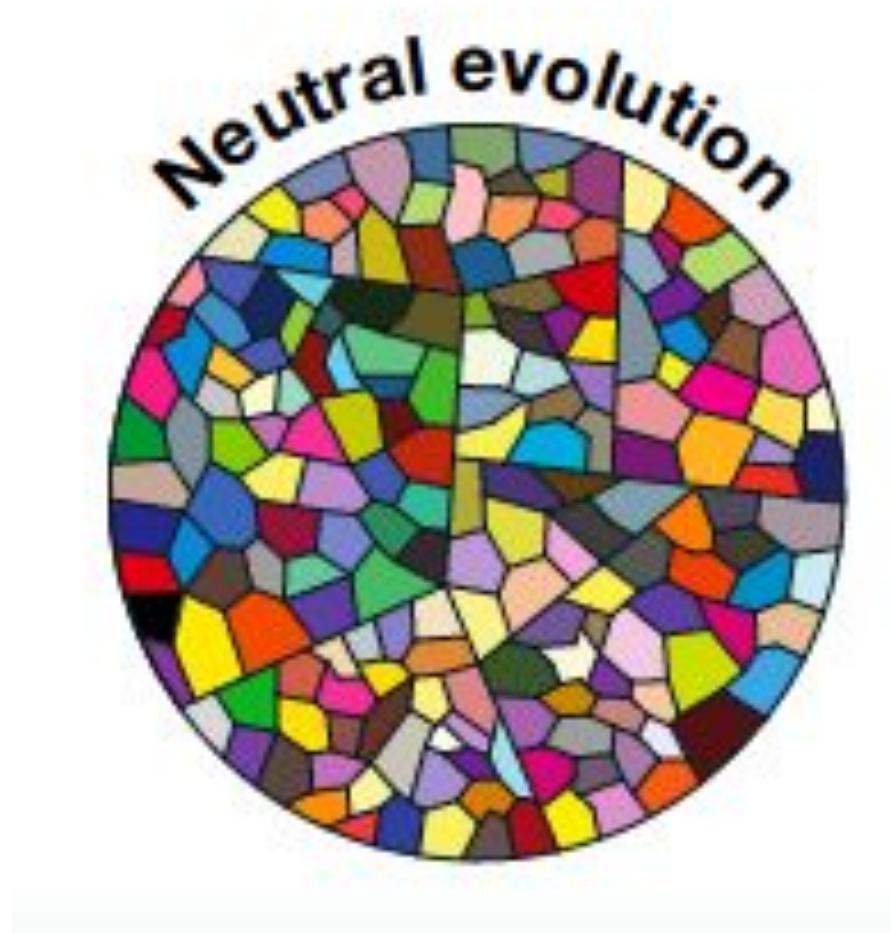
Cancer Cell Fraction (CCF) = VAF corrected for tumor purity and copy number



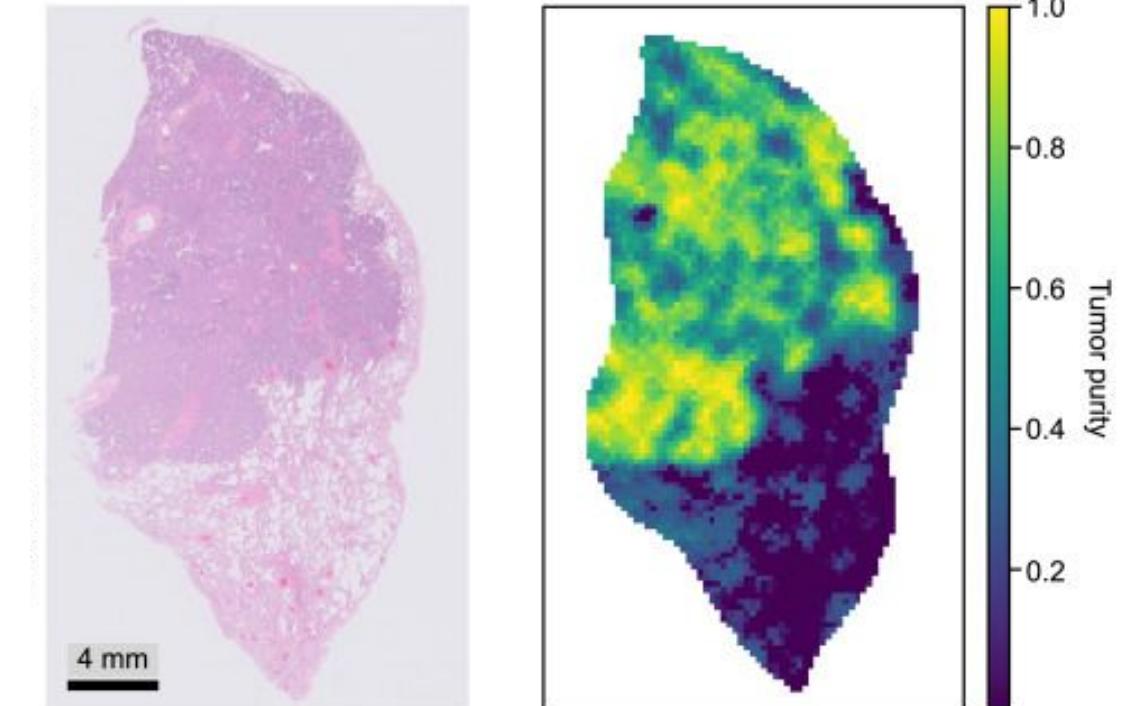
High CCF variants → **Clonal** or early
Low CCF variants → **Subclonal** or late

Challenges in timing point mutations

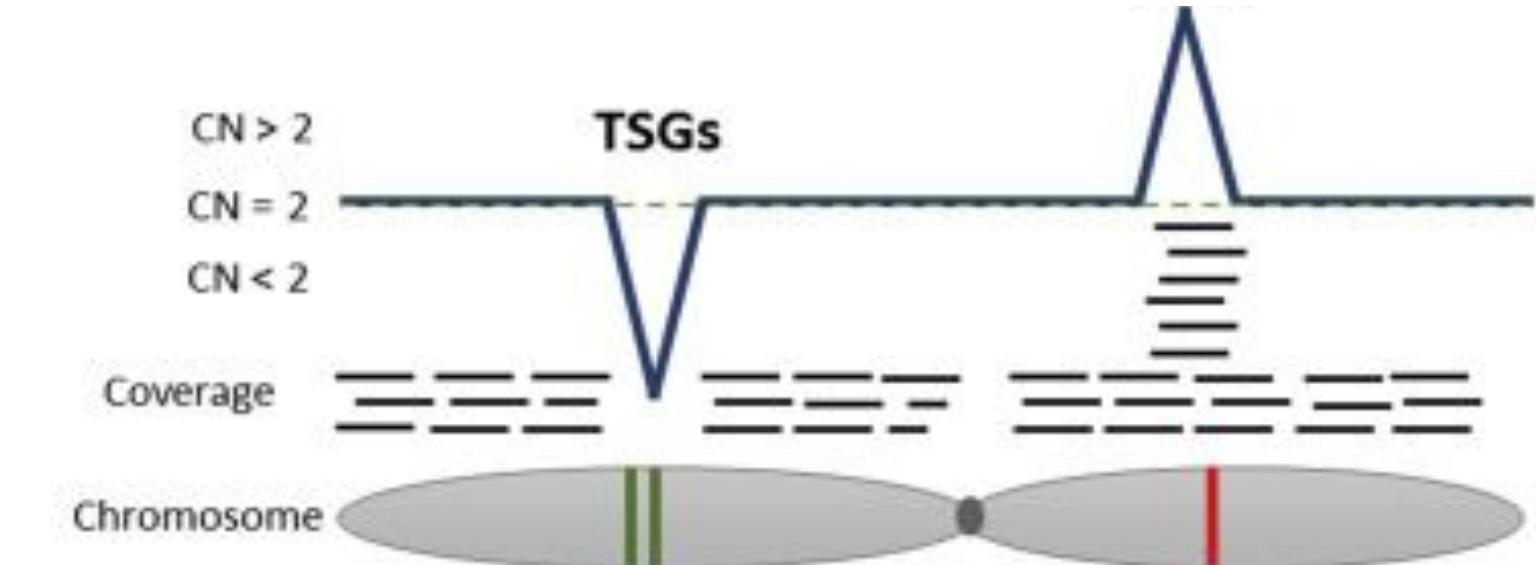
- Depends on purity and copy number estimates
- Assumes a neutral evolution model



Tumor purity
= # cancer cells sequences / # all cells sequenced

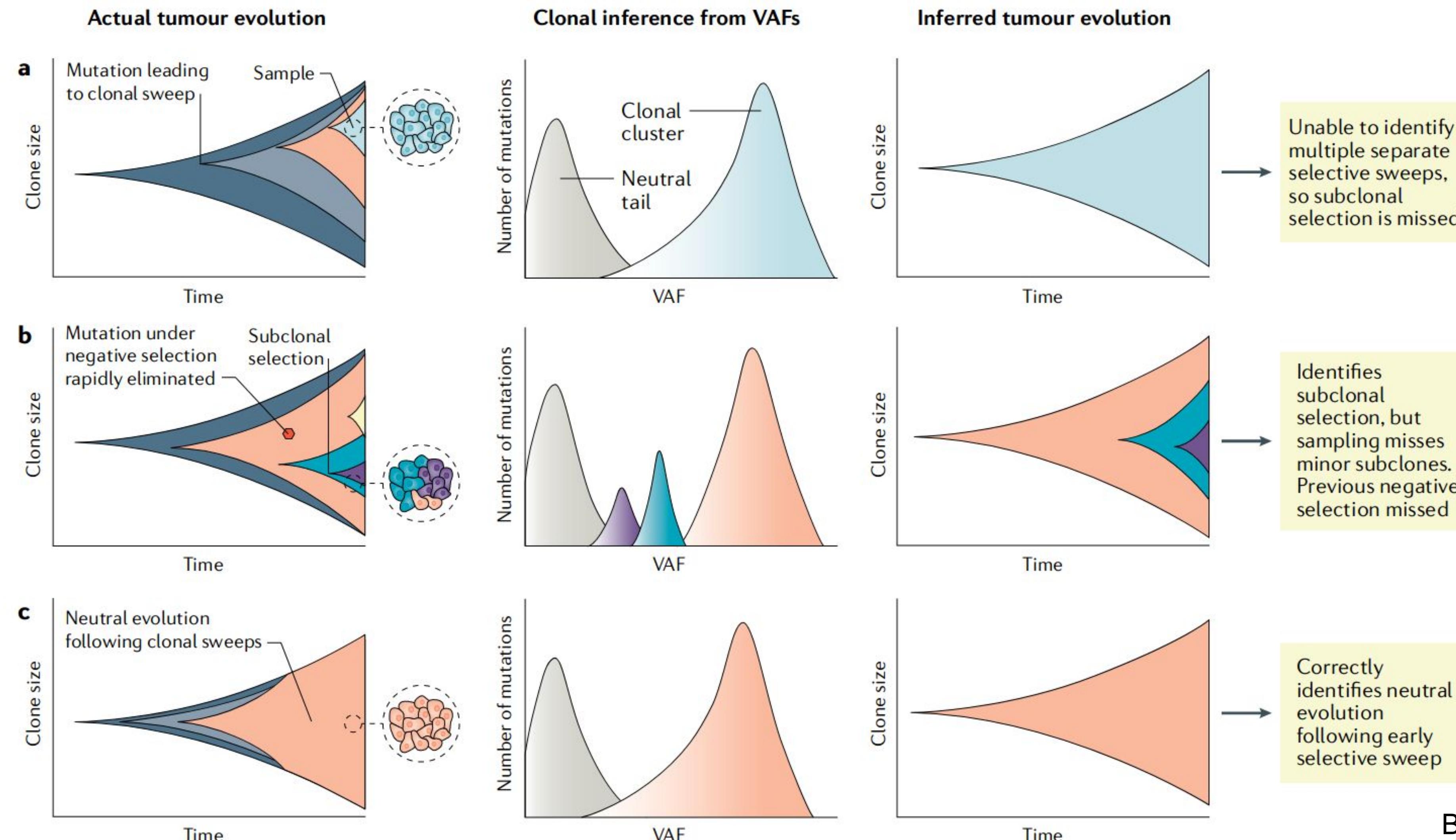


Copy number gains and losses



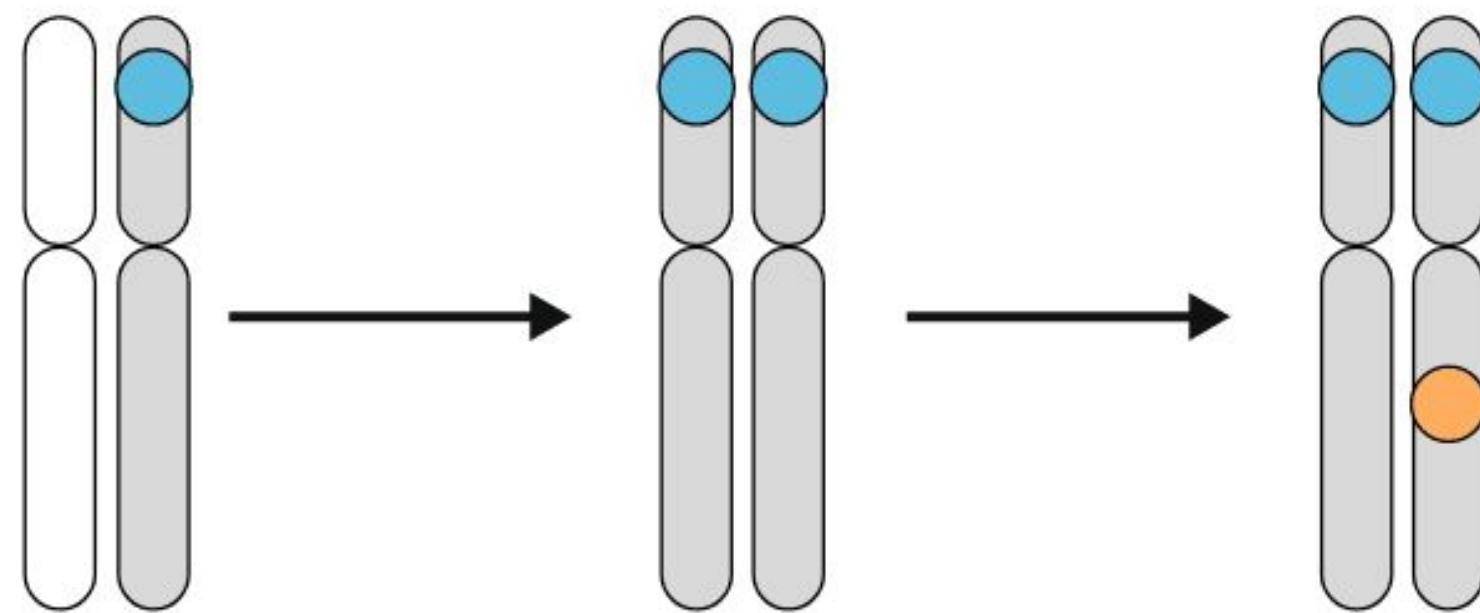
Challenges in timing point mutations

- Temporal resolution is limited
- Clonal heterogeneity and sampling bias

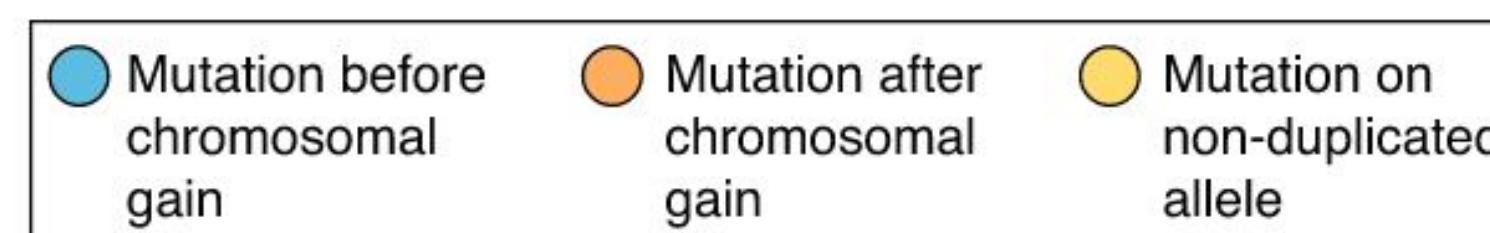
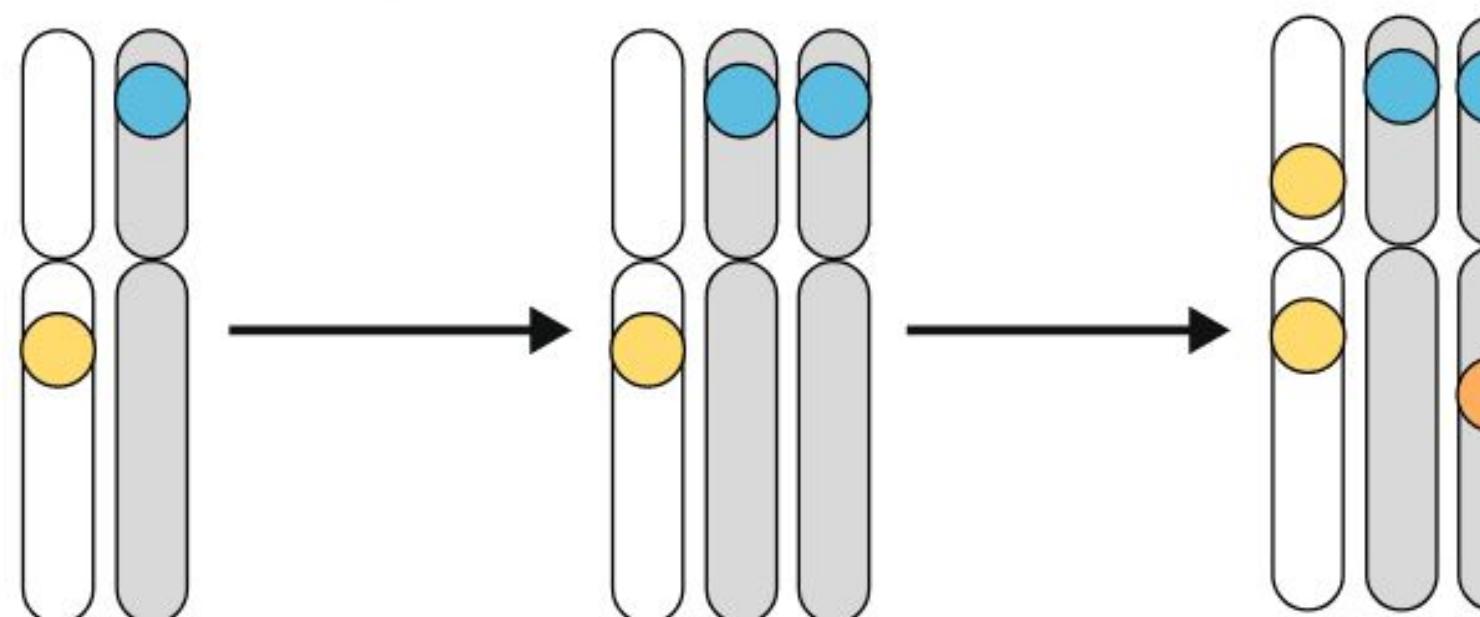


Timing copy number alterations

a Copy neutral loss-of-heterozygosity

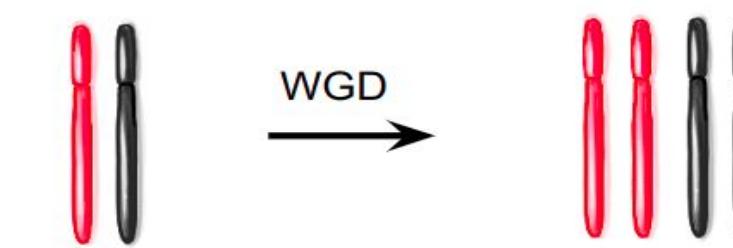


b Gain of single allele



No deletion

Two types of alleles (balanced)



Post WGD deletion

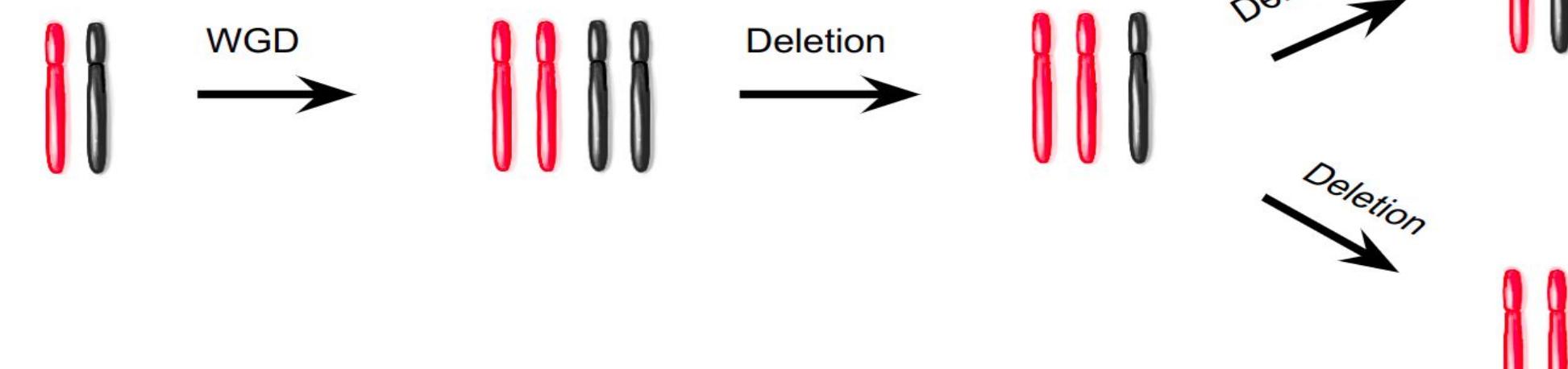
= Deletion resulting in two types of alleles (imbalanced)

Only possibility



Pre WGD deletion

= Deletion resulting in one type of allele



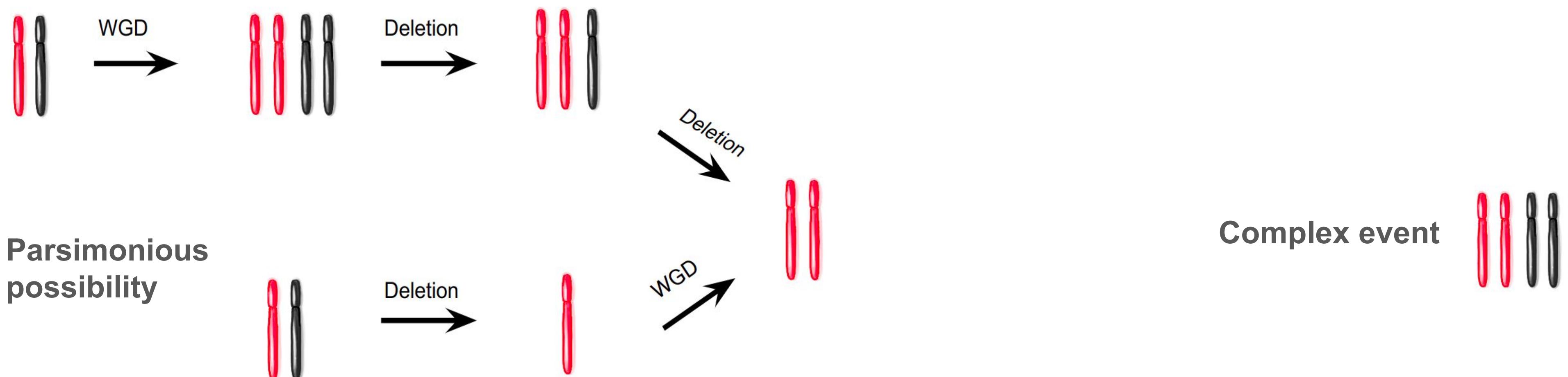
Parsimonious possibility

Zack et al., 2013



Challenges in timing copy number alterations

- Relative estimates (chronologically time point mutations first for absolute estimates)
- Assumes parsimony
 - Cannot time non-parsimonious true events
 - Cannot time complex events with multiple equally likely possibilities



WHEN? WHERE? WHO?

When? WHERE? ANY QUESTIONS?

HOW? Why? WHEN? What?

How? Why? WHEN? What?

WHAT? WHEN? What?

What? WHEN? What?

WHERE? When? What?

Where? When? What?

WHO? When? Why?

Who? When? Why?

WHEN? When? Why?

When? Why?

WHAT? When? Why?

What? When? Why?

WHERE? When? Why?

Where? When? Why?

WHO? When? Why?

Who? When? Why?