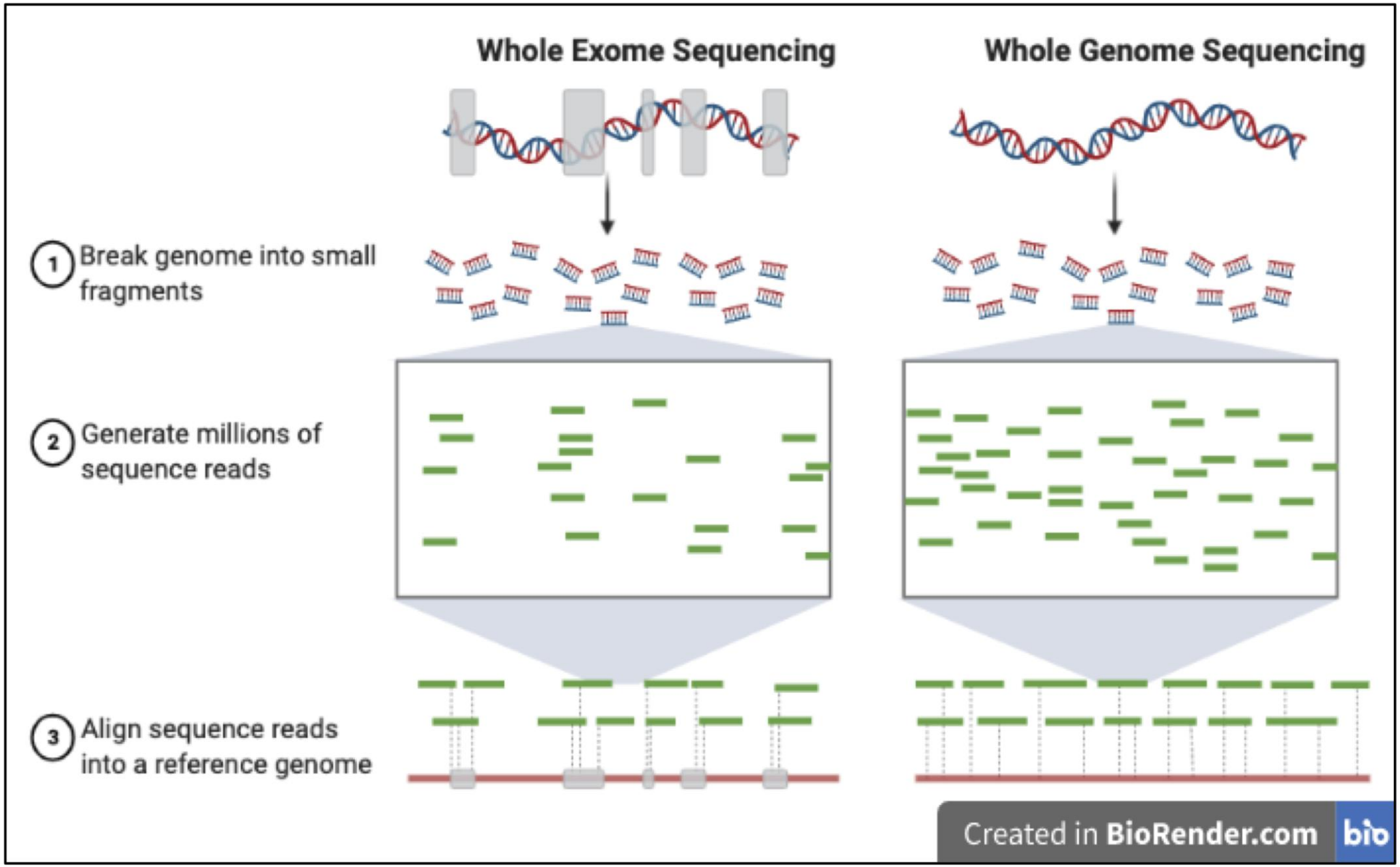
scDNA-seq

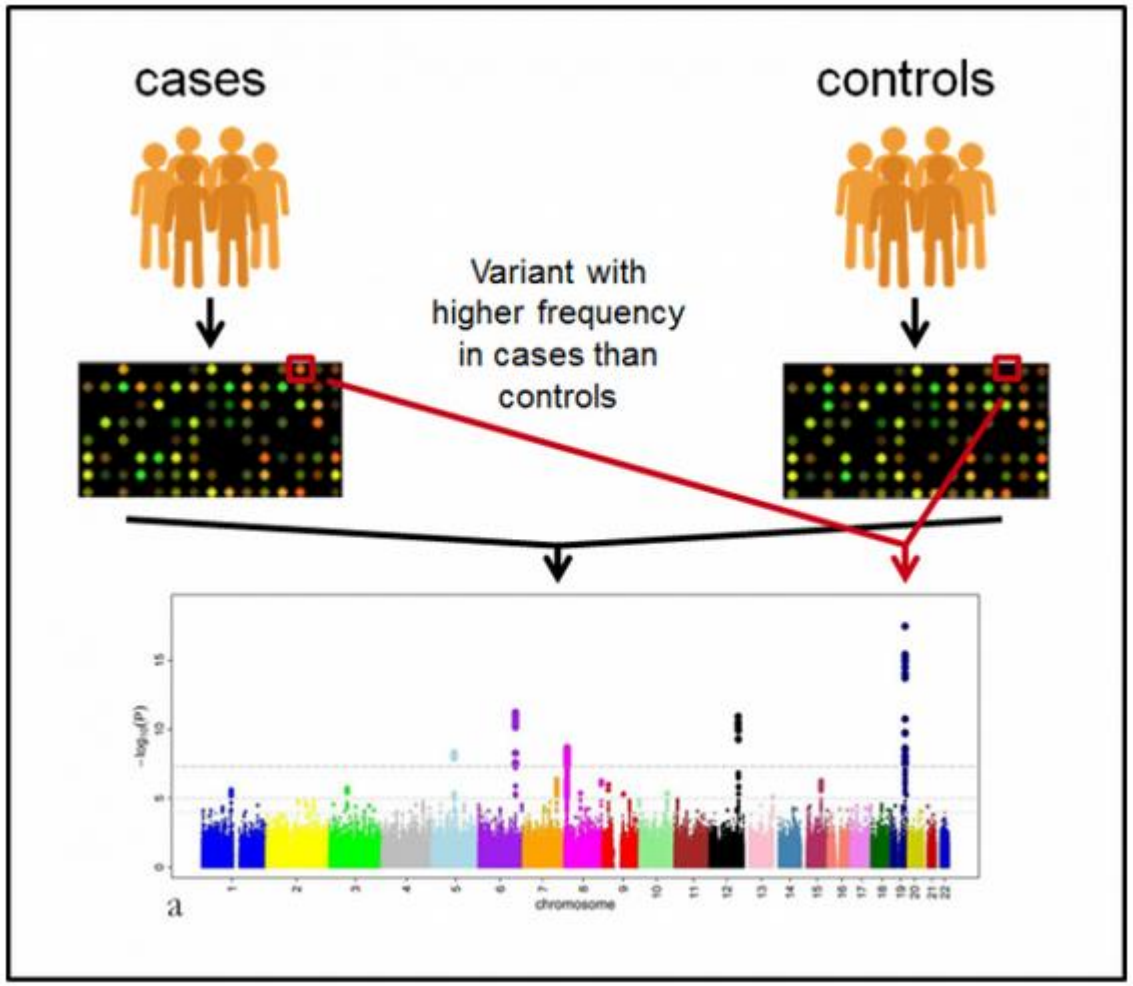
- Bulk DNA-seq


Whole Exome Sequencing
Whole Genome Sequencing
1 Break genome into small fragments
2 Generate millions of sequence reads
3 Align sequence reads into a reference genome
Created in BioRender.com
bio

- Bulk DNA-seq

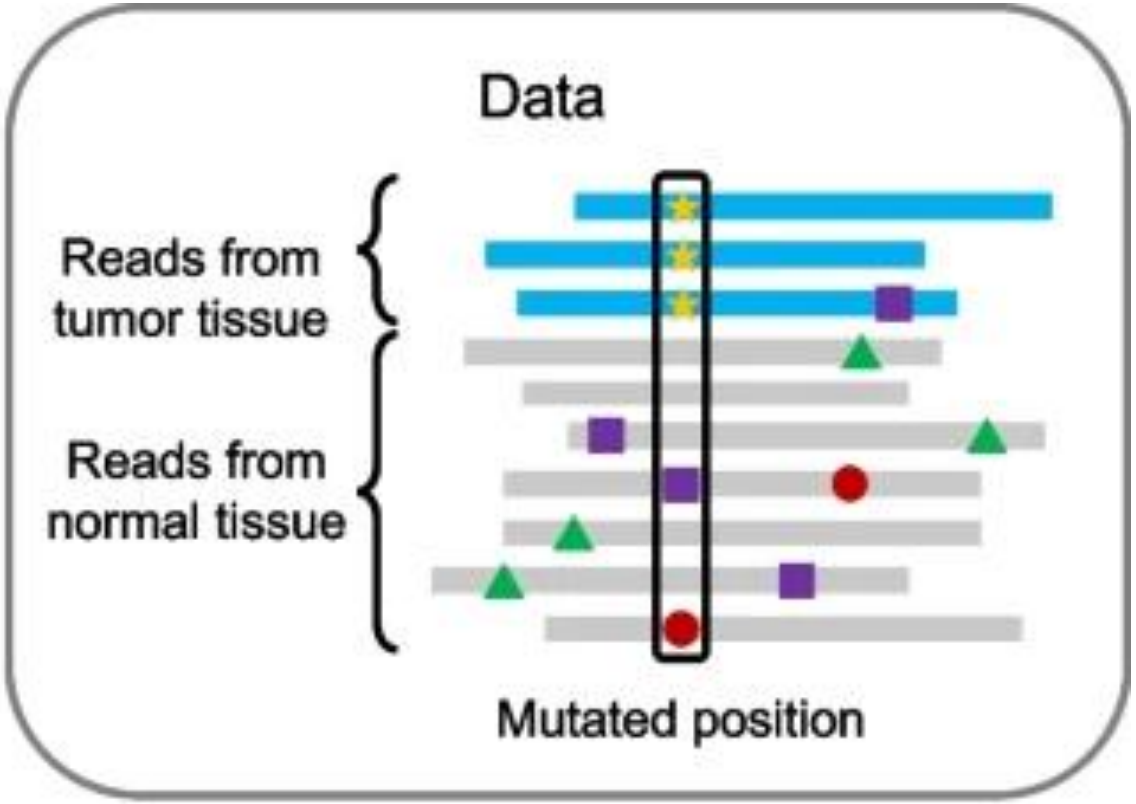
CCGTTAGAGTTACAATTCGA
CCGTTAGAGTAACAATTCGA
CCGTTAGAGTTACAATTCGA
CCGTTAGAGTTACAATTCGA
CCGTTAGAGTAACAATTCGA
CCGTTAGAGTAACAATTCGA
CCGTTAGAGTTACAATTCGA
CCGTTAGAGTTACAATTCGA
CCGTTAGAGTTACAATTCGA

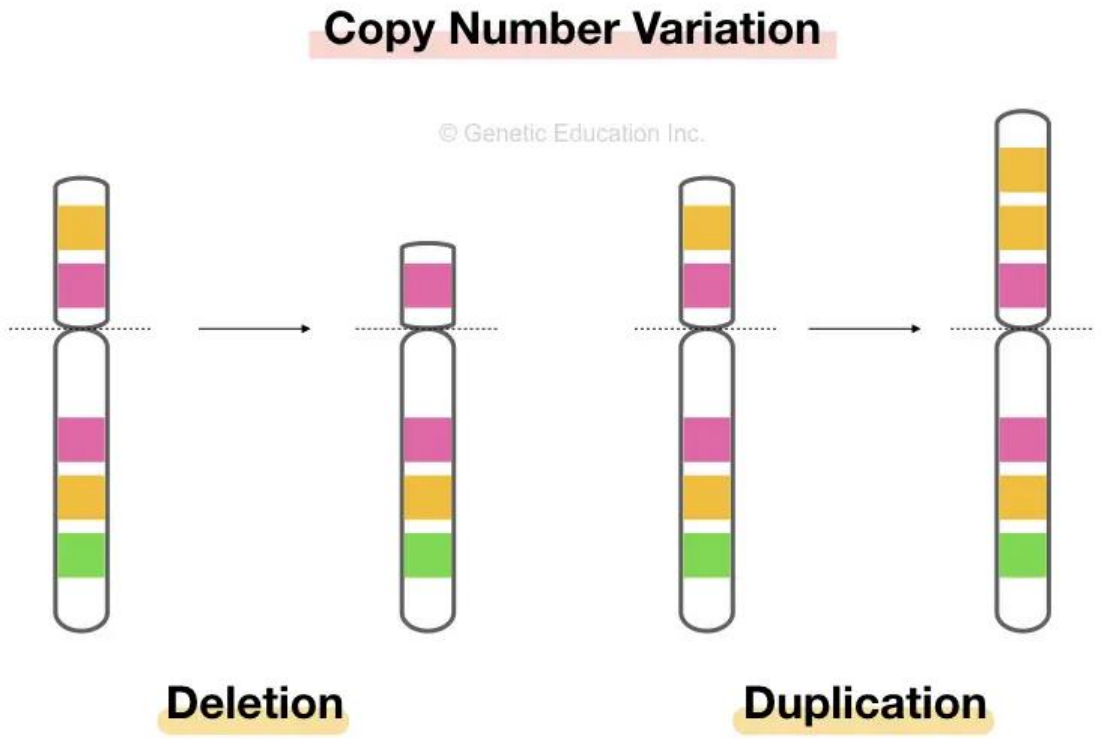
SNP calling → population heterogeneity



GWAS

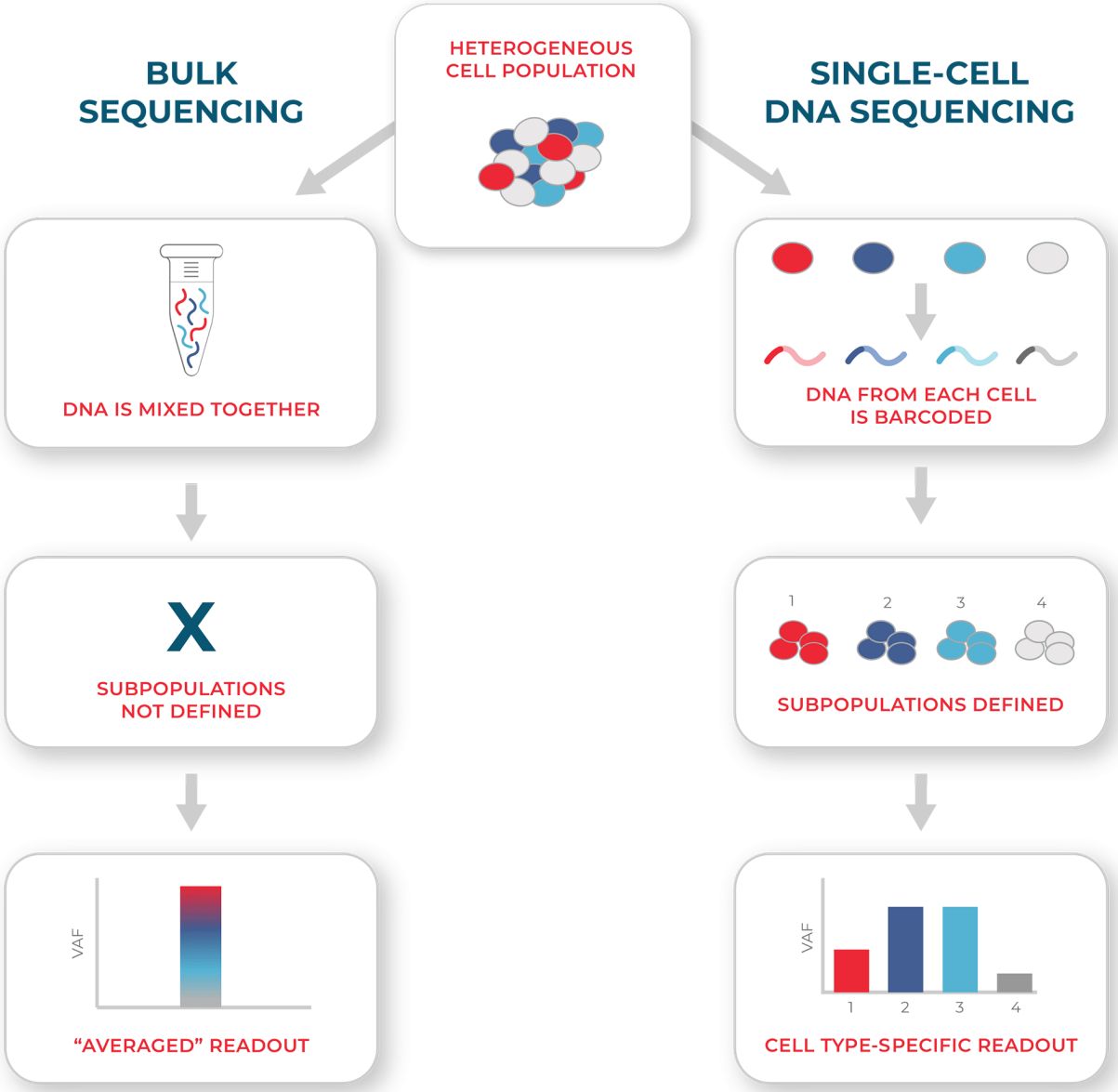
- Bulk DNA-seq

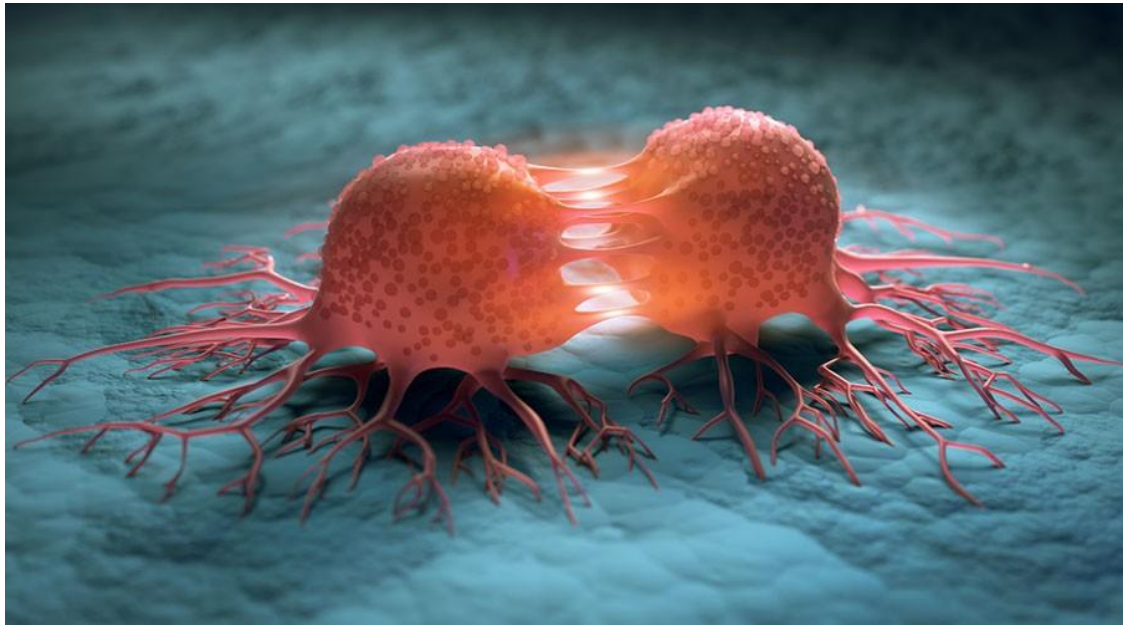




-Germline mutation (vs Reference)
-Somatic mutation (vs individual)

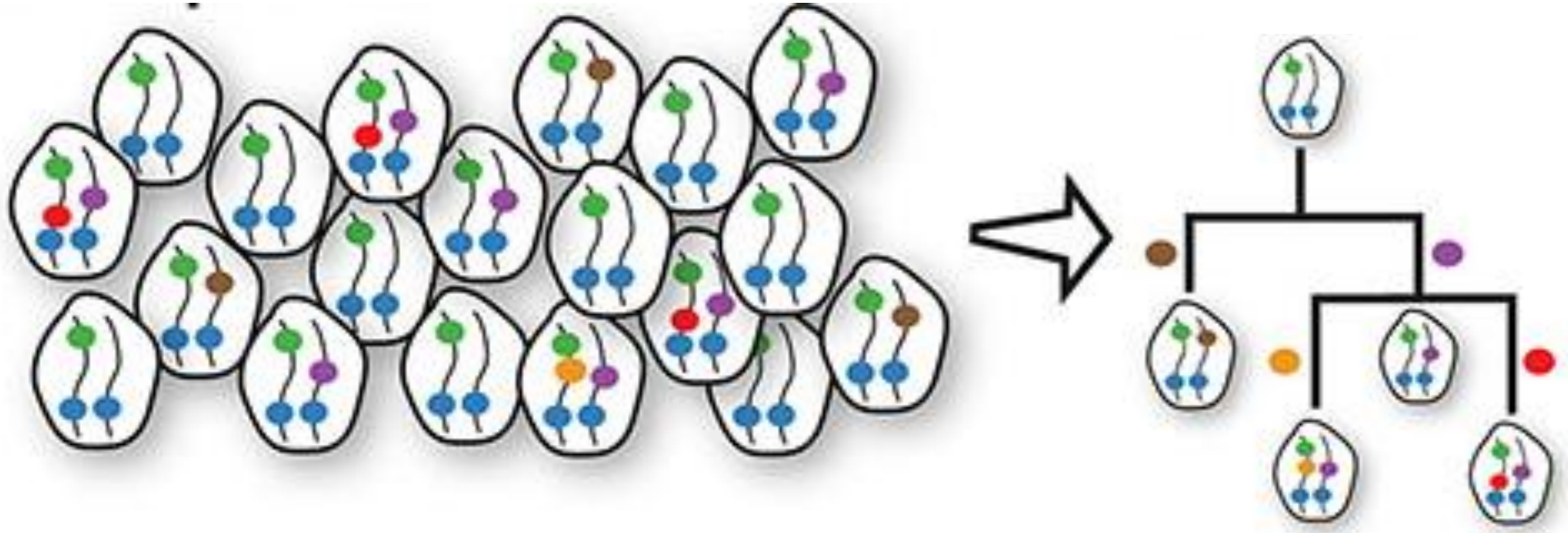
- scDNA-seq



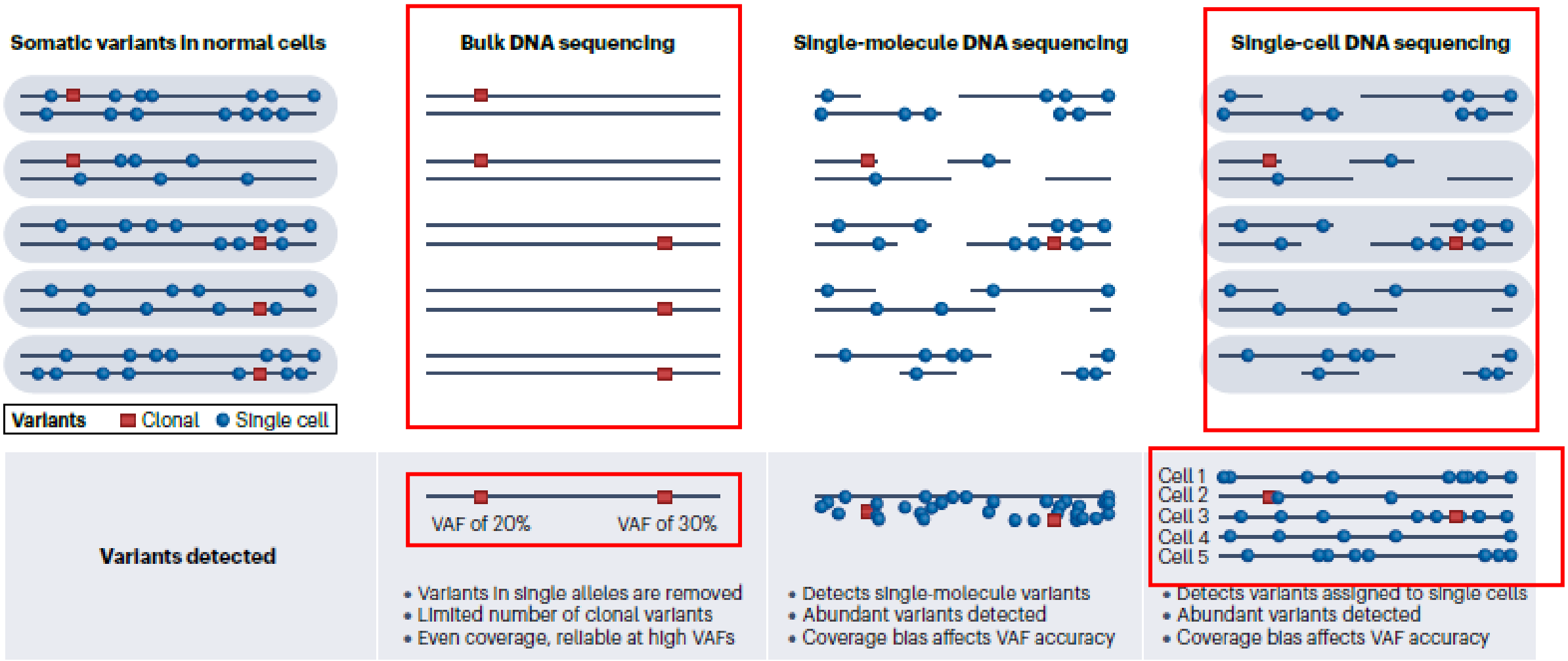


Especially, cancer

- scDNA-seq

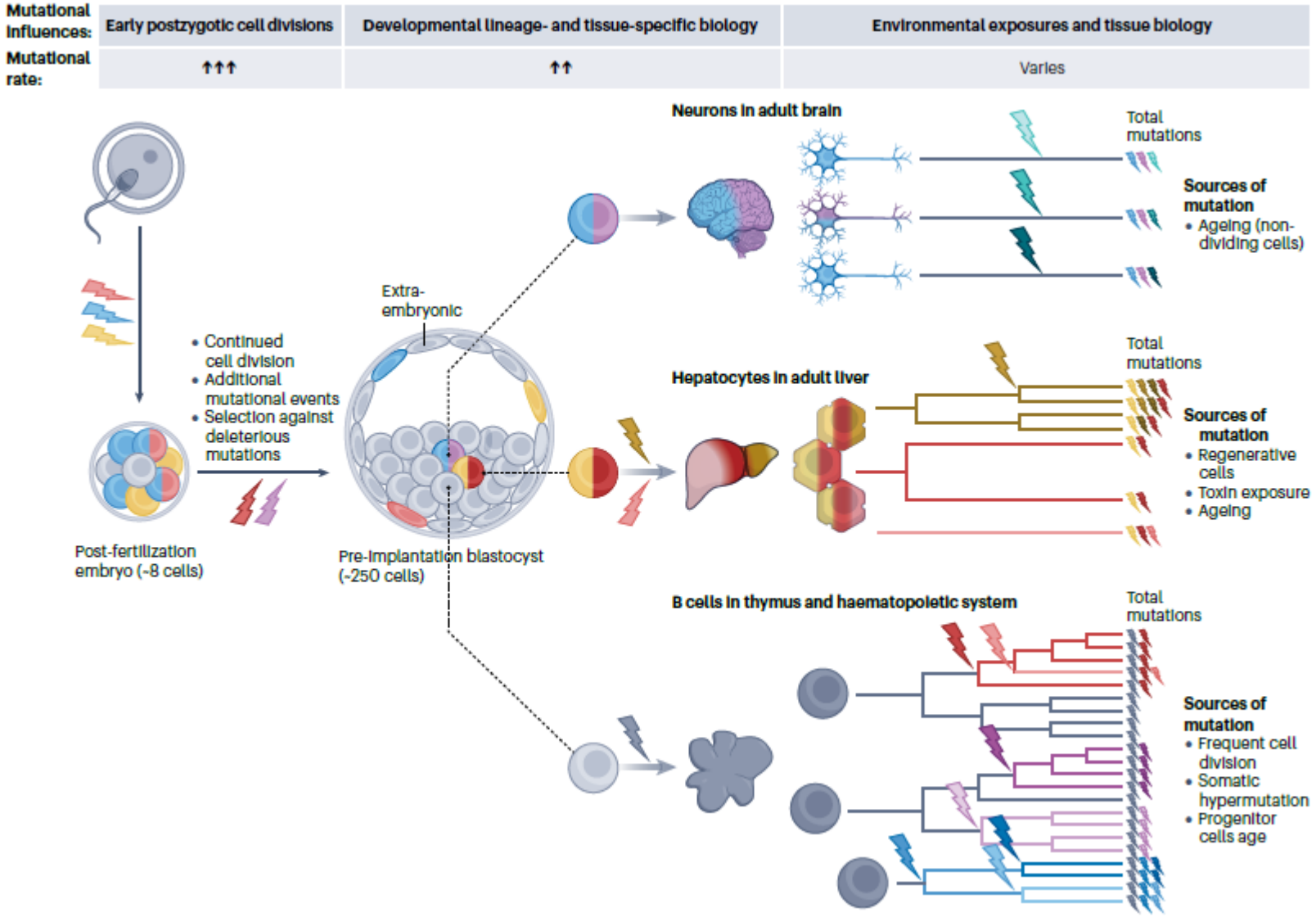


- scDNA-seq



Advances in single-cell DNA sequencing enable insights into human somatic mosaicism

- scDNA-seq



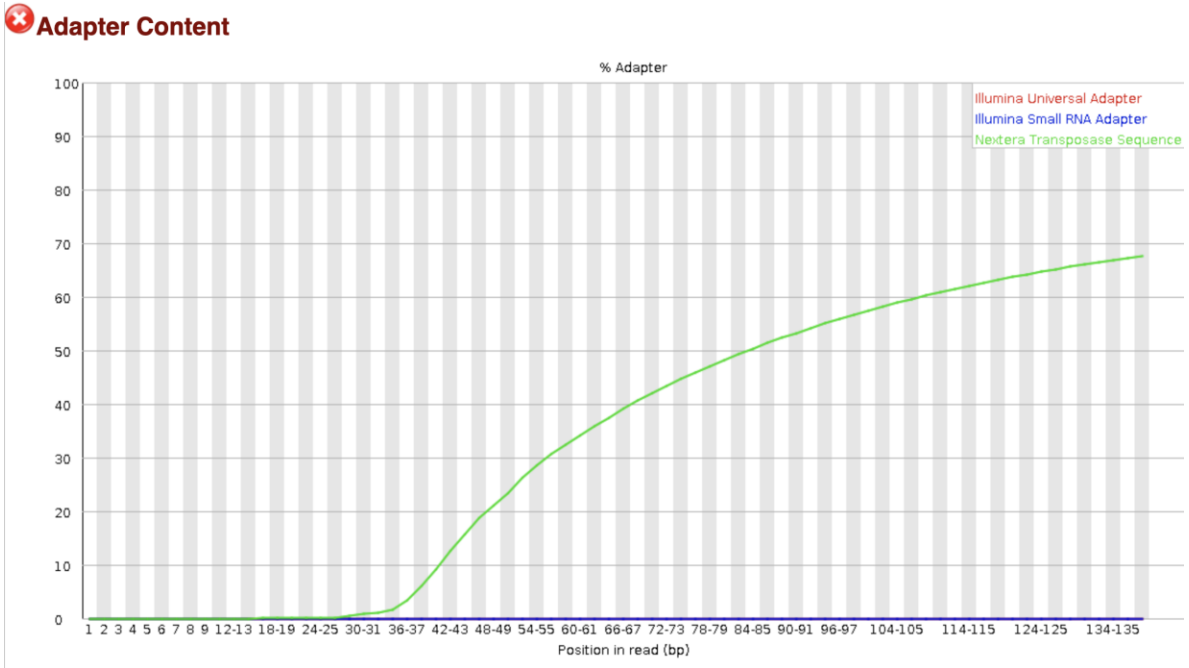
-Can detect which cell has a mutation
→ Lineage tracing → when does mutation occur

+ aging → more mutation!

- Preprocessing

-QC: FastQC or MultiQC
-Adapter trimming: Trimmomatic, fastp
-Barcode demultiplexing (if applicable): cellranger-dna
→ Alignment + CNV
-SMART-seq2: each fastq → each cell → go to alignment

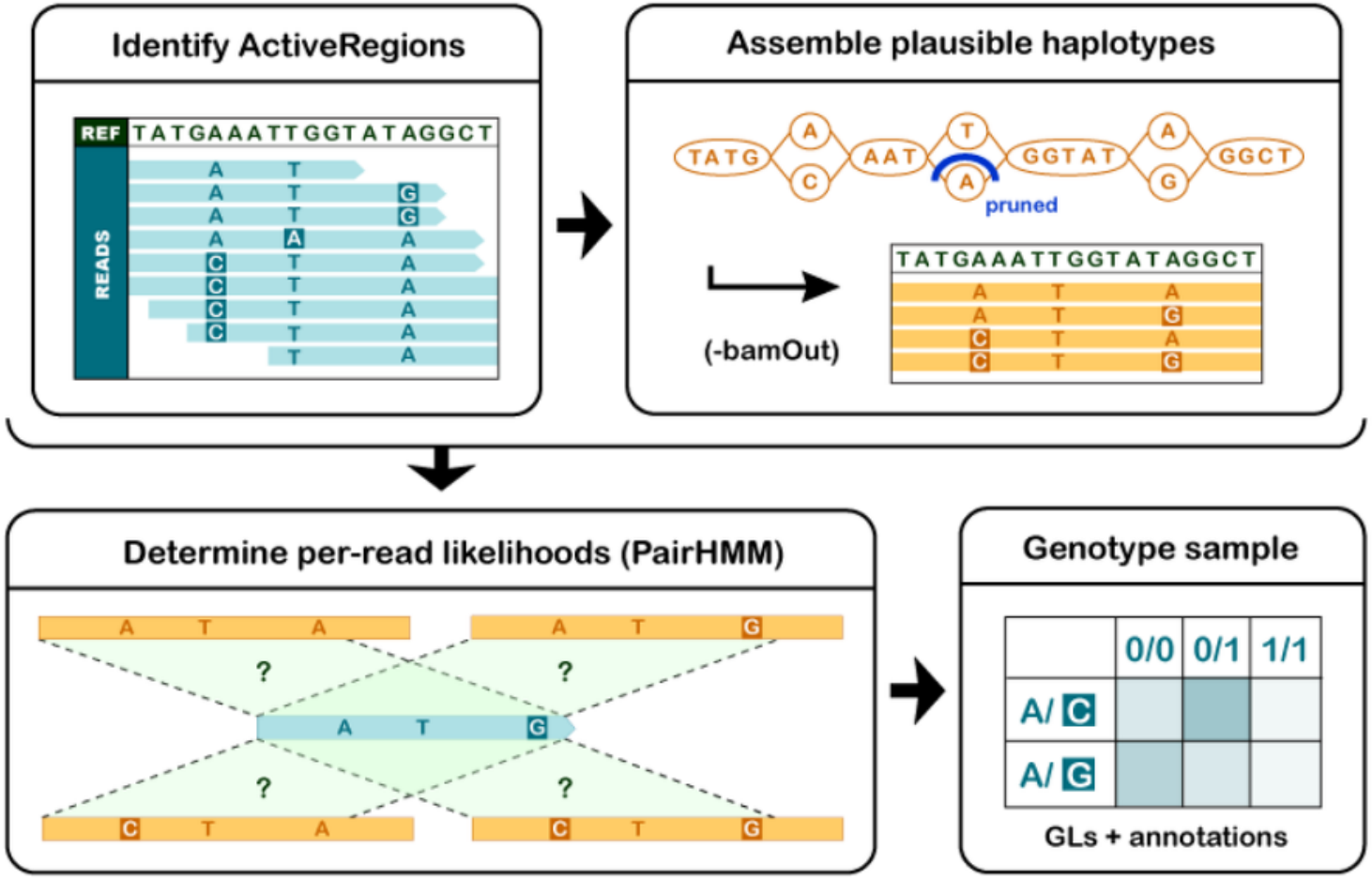
-Alignment: BWA-MEM, Bowtie2 (support: Samtools)
-Remove low-quality cells: read depth, duplicates …



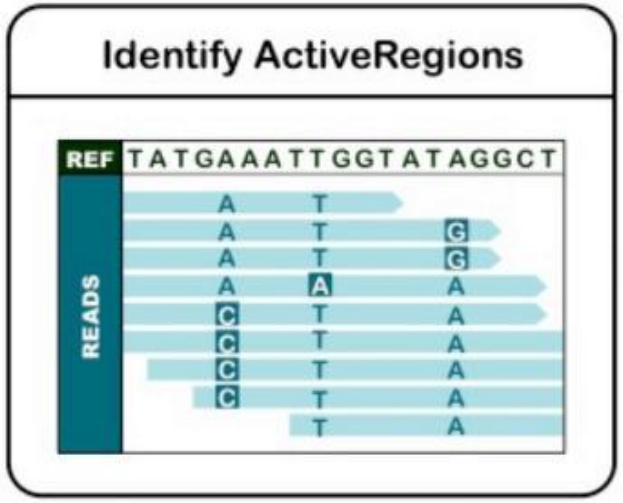
- Mutation calling

GATK, Mutect2, VarScan, FreeBayes
→ GATK:variant calling + variant annotation

-Variant calling: Haplotypecaller



- Haplotypecaller

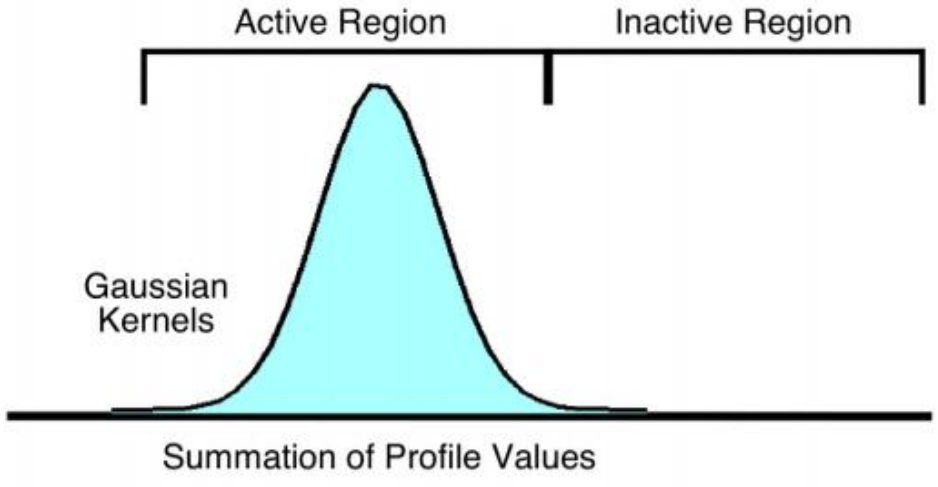


- Sliding window along the reference
- Count mismatches, indels and soft clips

➢ **Measure of entropy**

-sliding → count mismatch →entropy

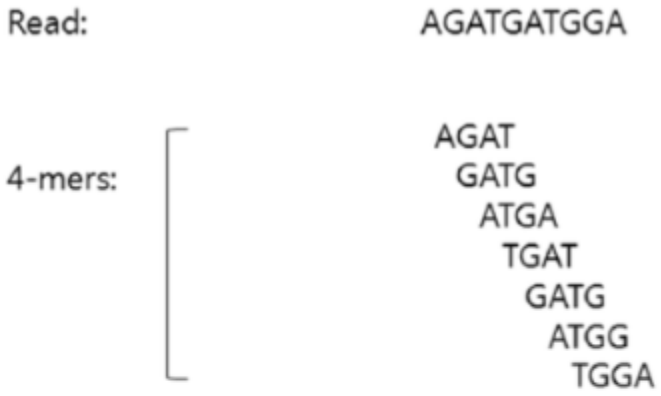
Entropy values → Gaussian kernels → above threshold → “active region”

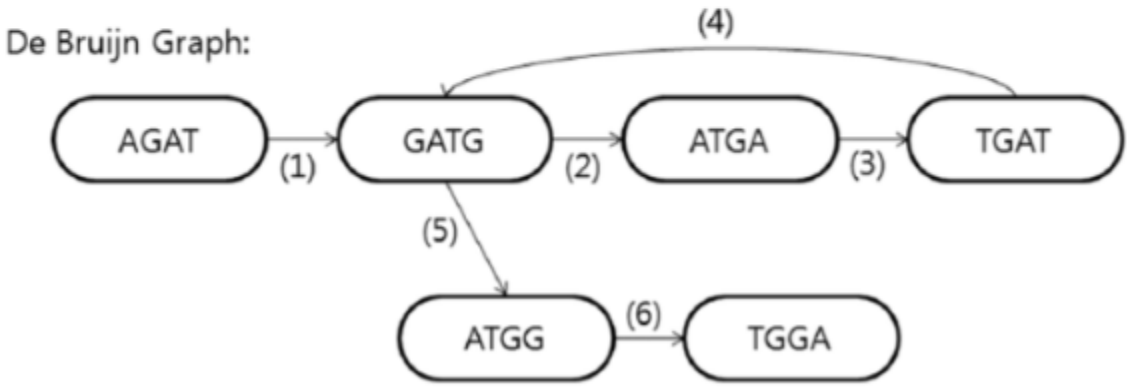


Over threshold:
Trigger “ActiveRegion” to be processed

- Haplotypecaller

-Active region sequence → De Brujin graph → extract possible haplotype (contig)
-Smith-Waterman algorithm: match between haplotype and reference





-Active seq → 4-mer sliding
→ Graph → if there is a overlap (same 4-mer again) → ex: loop 4
→ Final contig: AGATGGA

- Haplotypecaller

-PairHMM algorithm → reads ~ Haplotype likelihood martix

Haplotypes

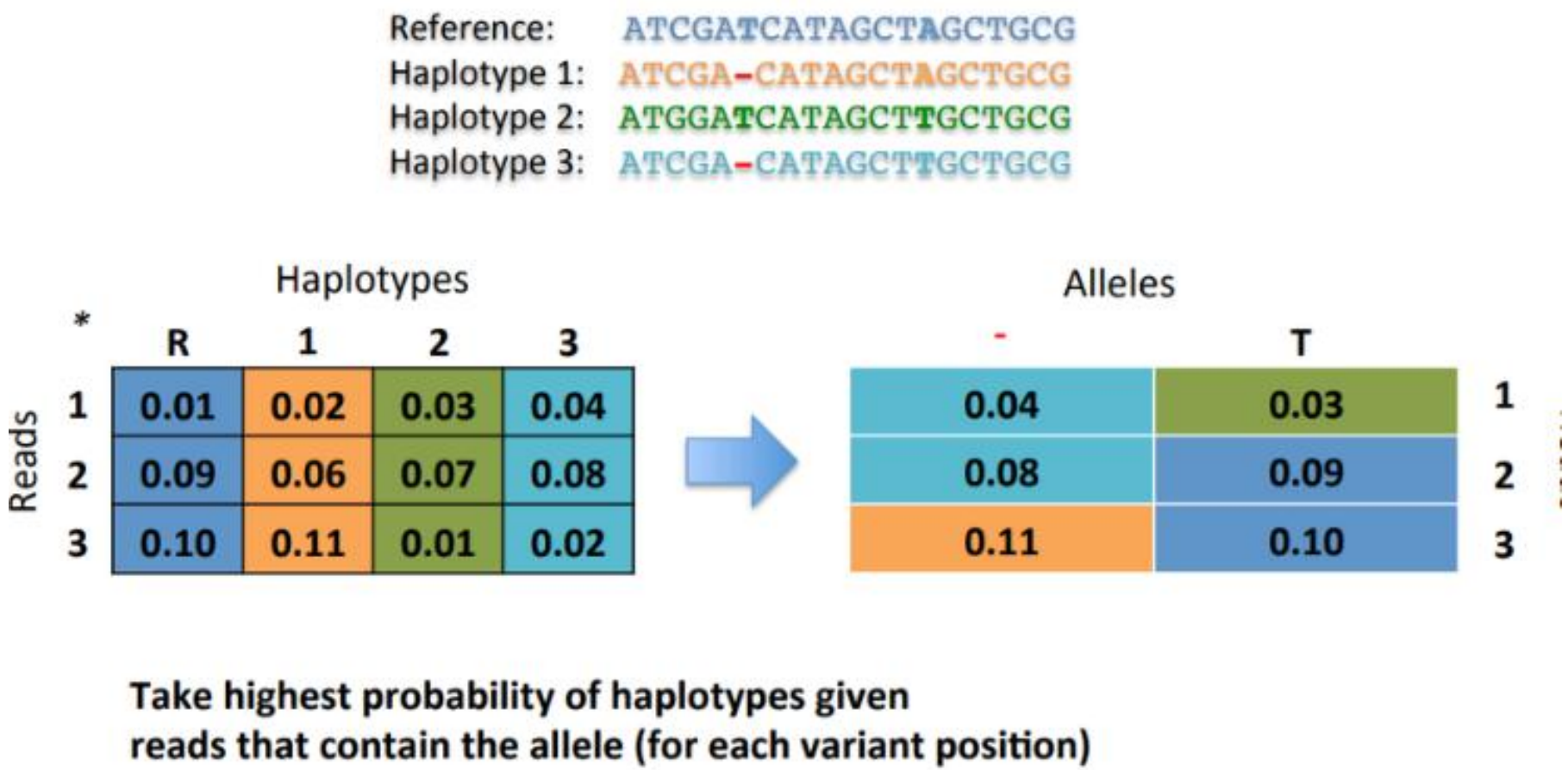
Reads

$$\begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & & & A_{2n} \\ \vdots & & & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nn} \end{bmatrix}$$

A_{ij} = probability of haplotype vs read

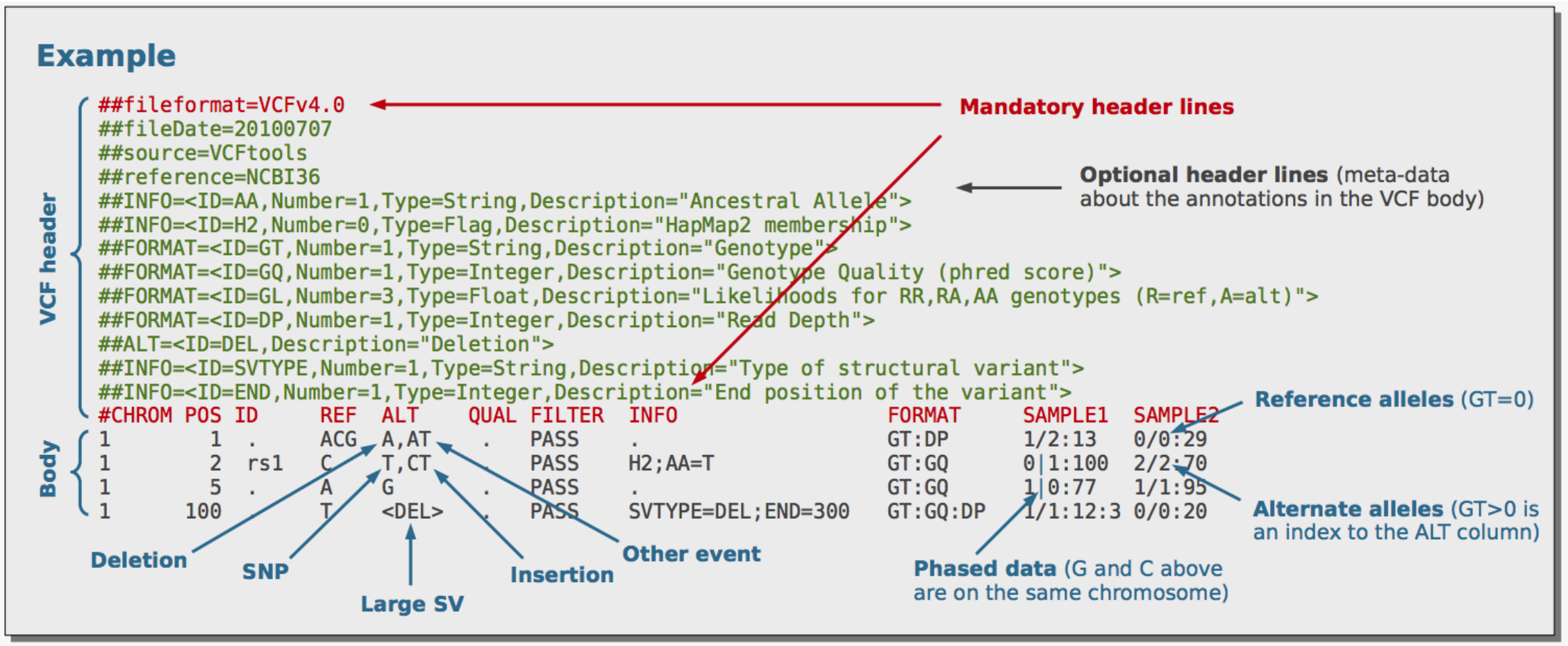
-> likelihoods of the haplotypes given the reads
-> store in matrix

- Haplotypecaller



-Bayesian statistics → determine genotype for diploid

- VCF file


Example
VCF header
Body
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1 SAMPLE2
1 1 . ACG A,AT . PASS . GT:DP 1/2:13 0/0:29
1 2 rs1 C T,CT . PASS H2;AA=T GT:GQ 0|1:100 2/2:70
1 5 . A G . PASS . GT:GQ 1|0:77 1/1:95
1 100 . T <DEL> . PASS SVTYPE=DEL;END=300 GT:GQ:DP 1/1:12:3 0/0:20
Mandatory header lines
Optional header lines (meta-data about the annotations in the VCF body)
Reference alleles (GT=0)
Alternate alleles (GT>0 is an index to the ALT column)
Deletion
SNP
Large SV
Insertion
Other event
Phased data (G and C above are on the same chromosome)

- Variant annotation

coding : protein coding region
synonymous : no codon alteration
nonsynonymous : codon alteration

nonsense : STOP 코돈으로 변화
missense : 단백질에서 아미노산의 변화를 만드는 코돈의 변화

frameshift : indel SNP → codon frame change

UTR : Untraslated region (UTR-3, UTR-5)

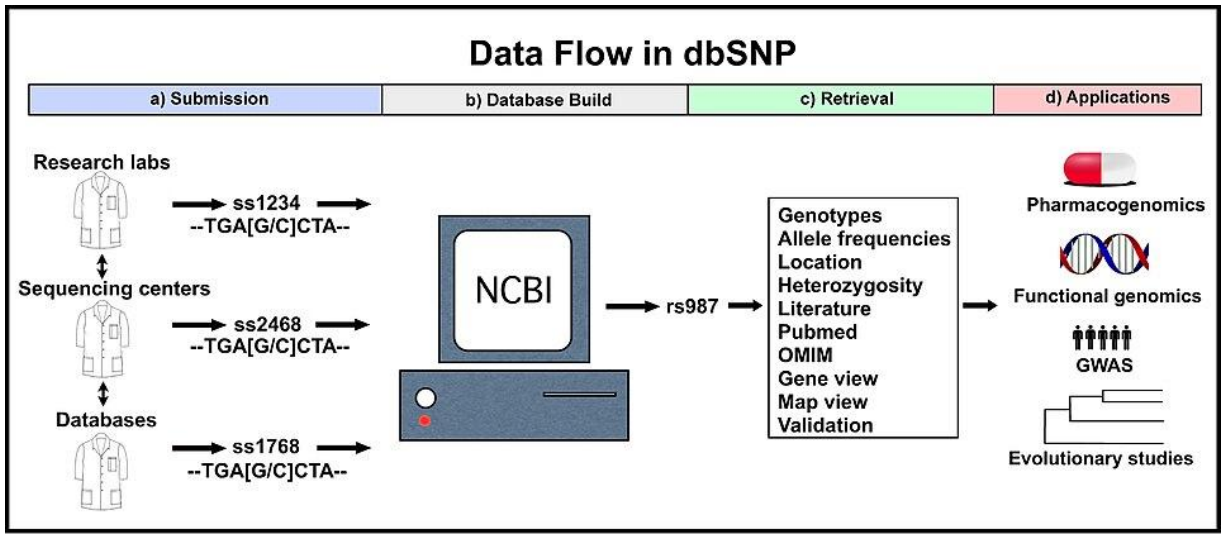
splice-site : splicing site (splice-3 : 3' acceptor dinucleotide, splice-5 : 5' donor dinucleotide)

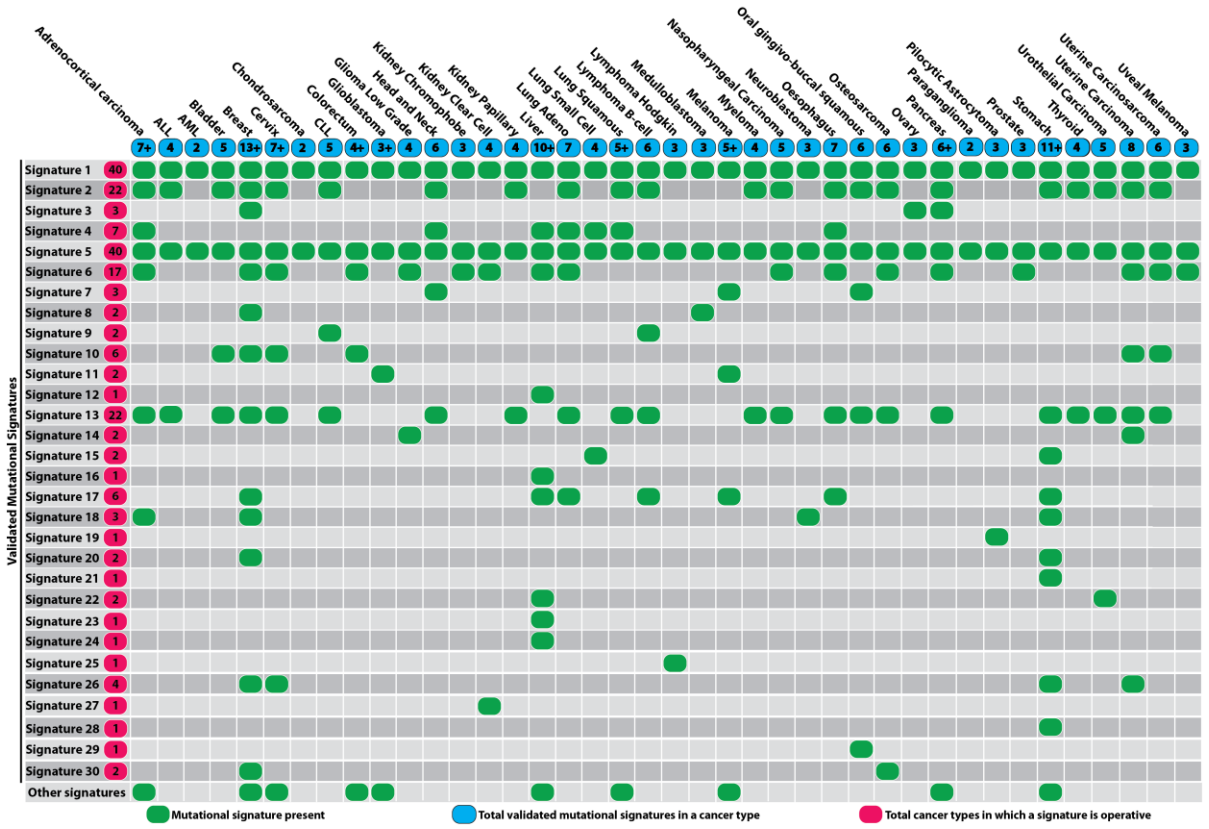
- Variant annotation

-Database

dbSNP, gnomAD: mutation DB

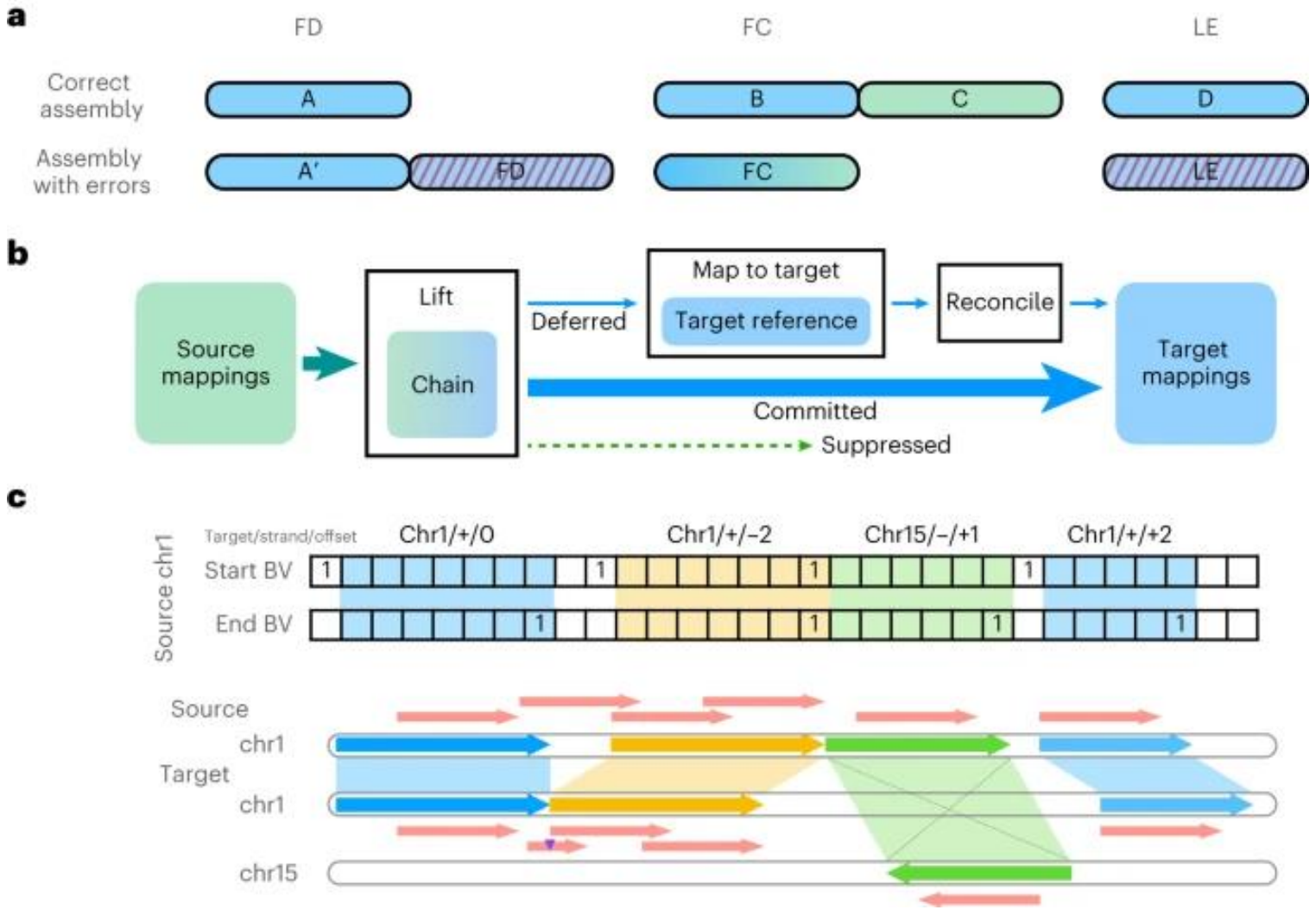
COSMIC (Catalogue Of Somatic Mutations In Cancer): cancer associated mutation DB





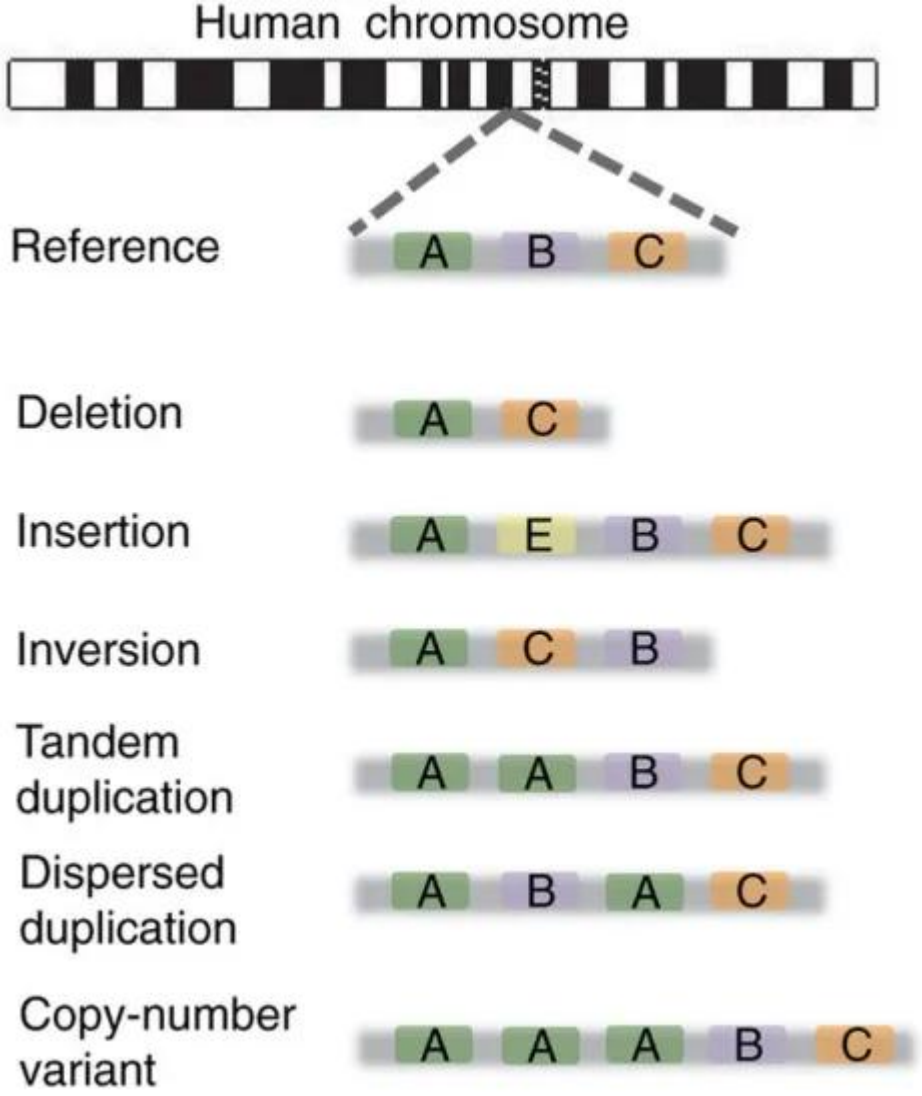
- Lift-over

-Old genome build (ex: GRCh37, hg19 or mm9)
→ Current version: BRCh38.p14, GRCm39 (20250810)
→ Genomic coordinate should be matched for compatibility



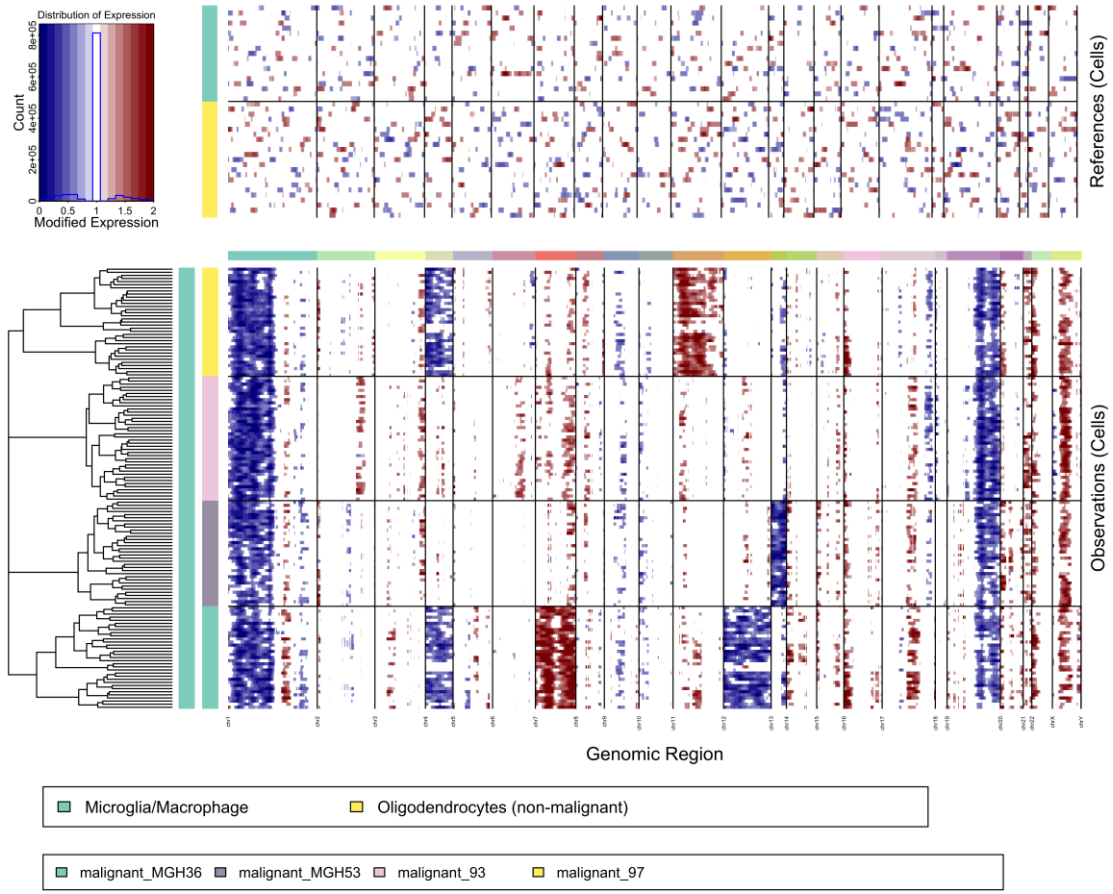
Improved sequence mapping using a complete reference genome and lift-over

- Structural variation

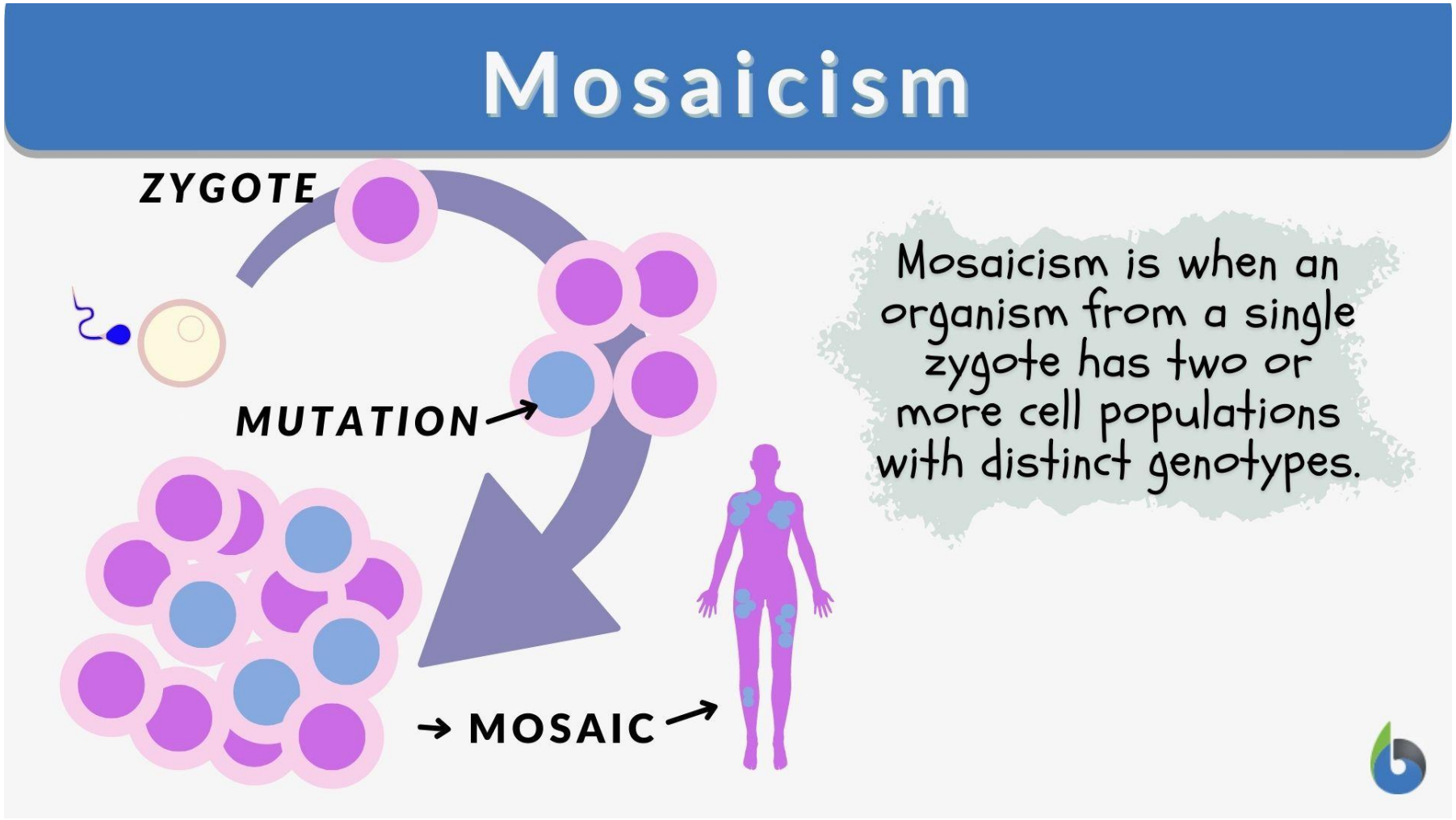


-Require Strand-seq, longread-seq
cf: MosaiCatcher v2

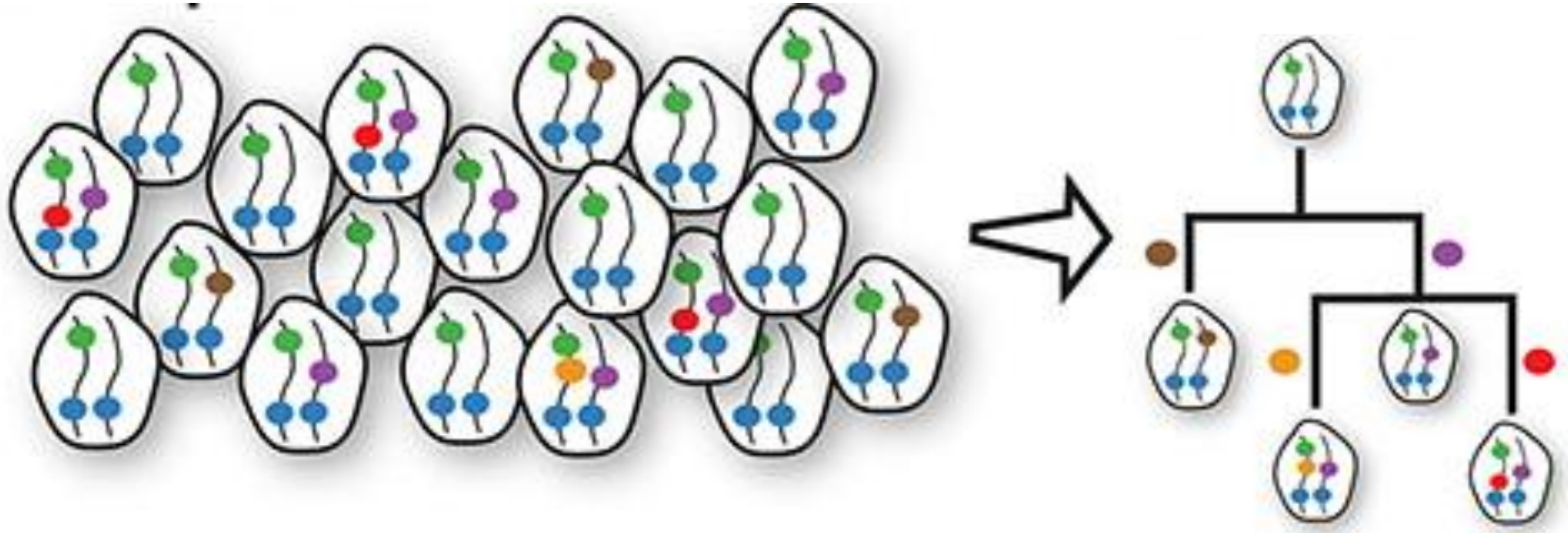
-CNV
-cellranger-dna
-inferCNV (also compatible with scRNA-seq)
→ Need to assign Normal cells



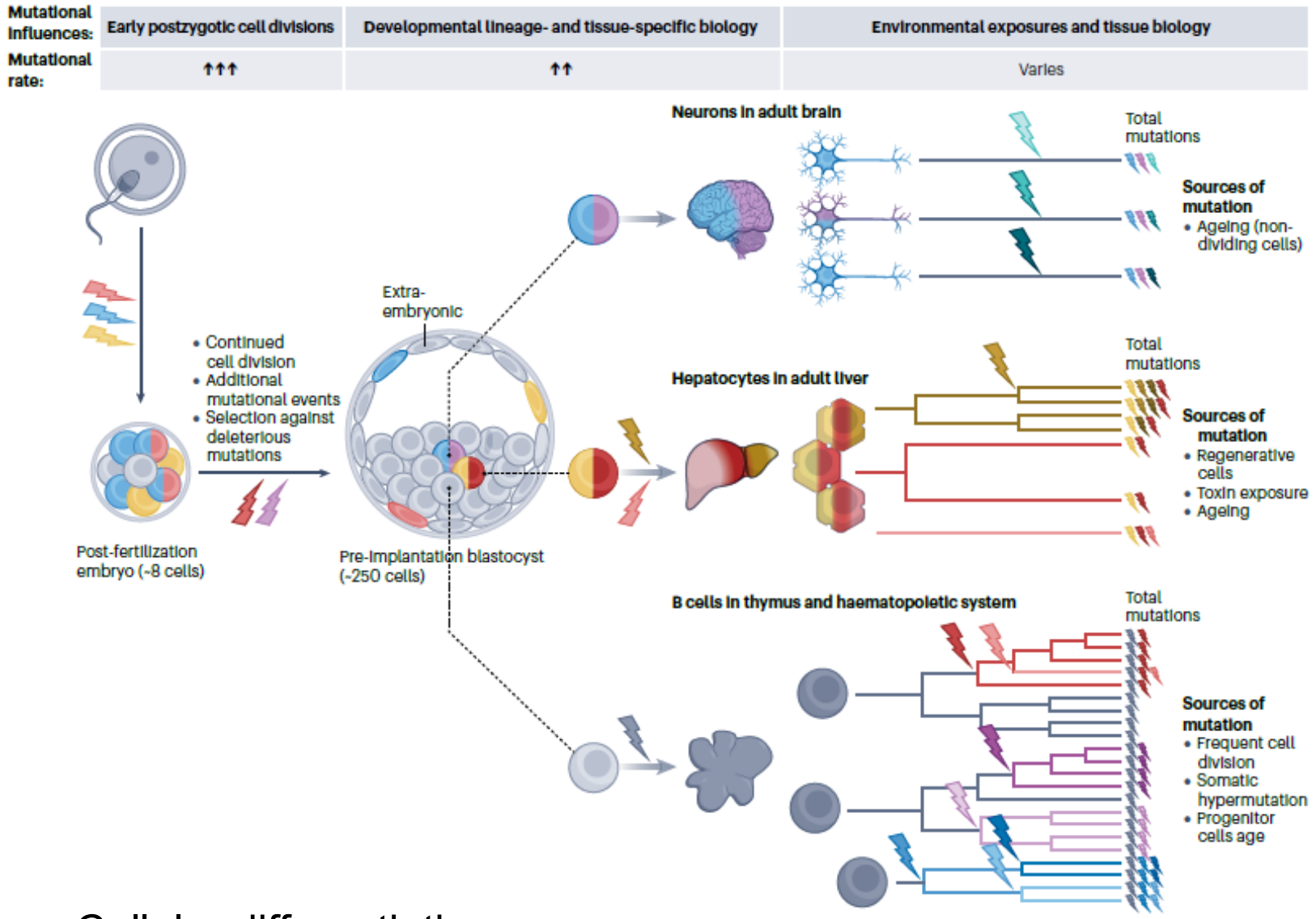
- Cellular mosaicism



- Phylogeny algorithm



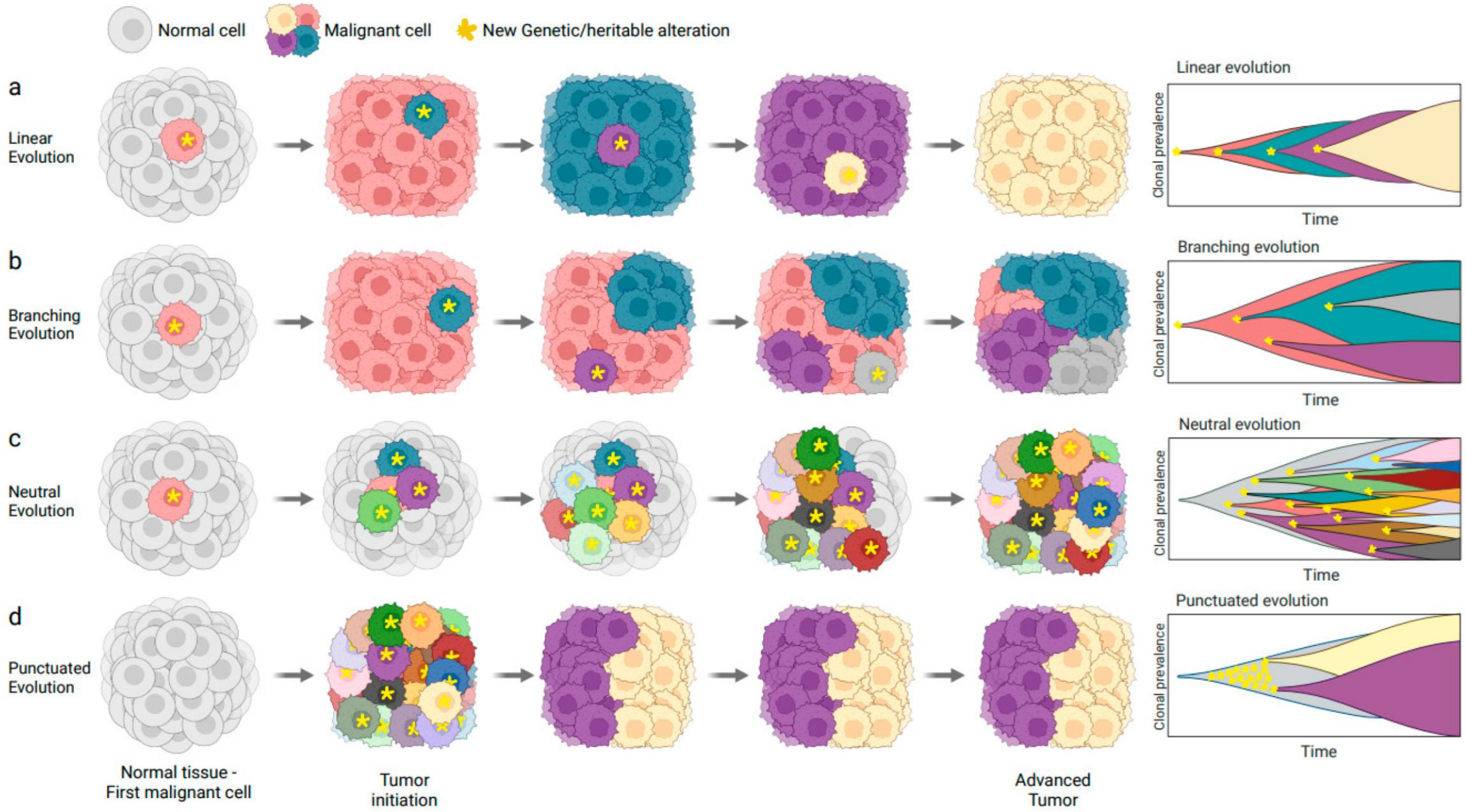
- Phylogeny algorithm



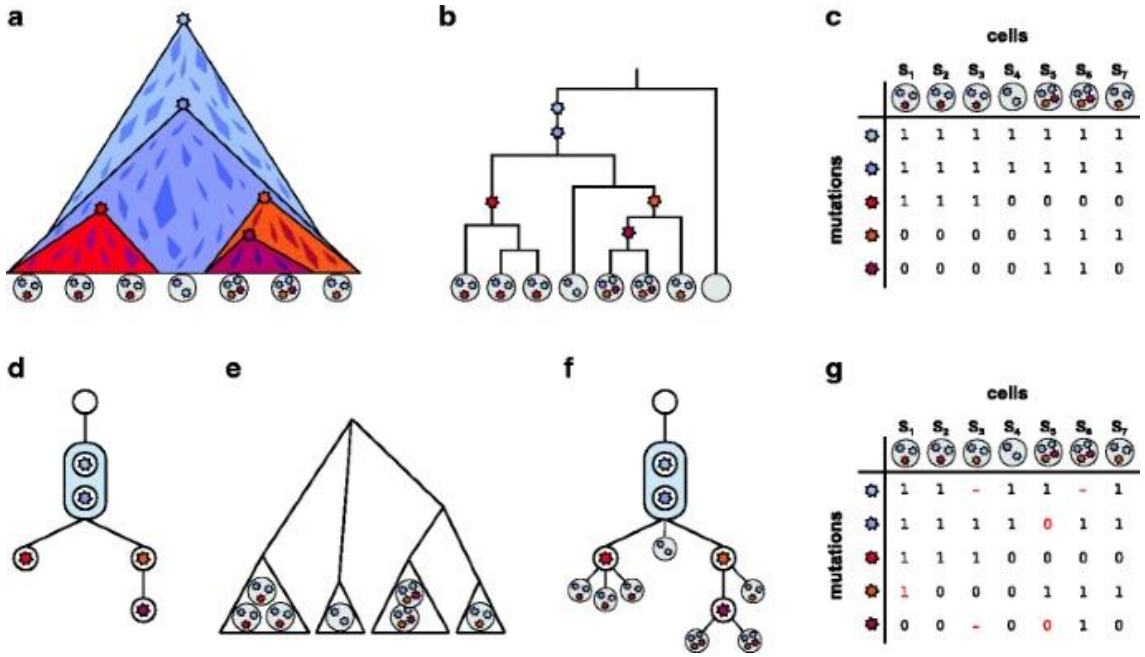
Cellular differentiation
Aging

- Phylogeny algorithm

Cancer evolution



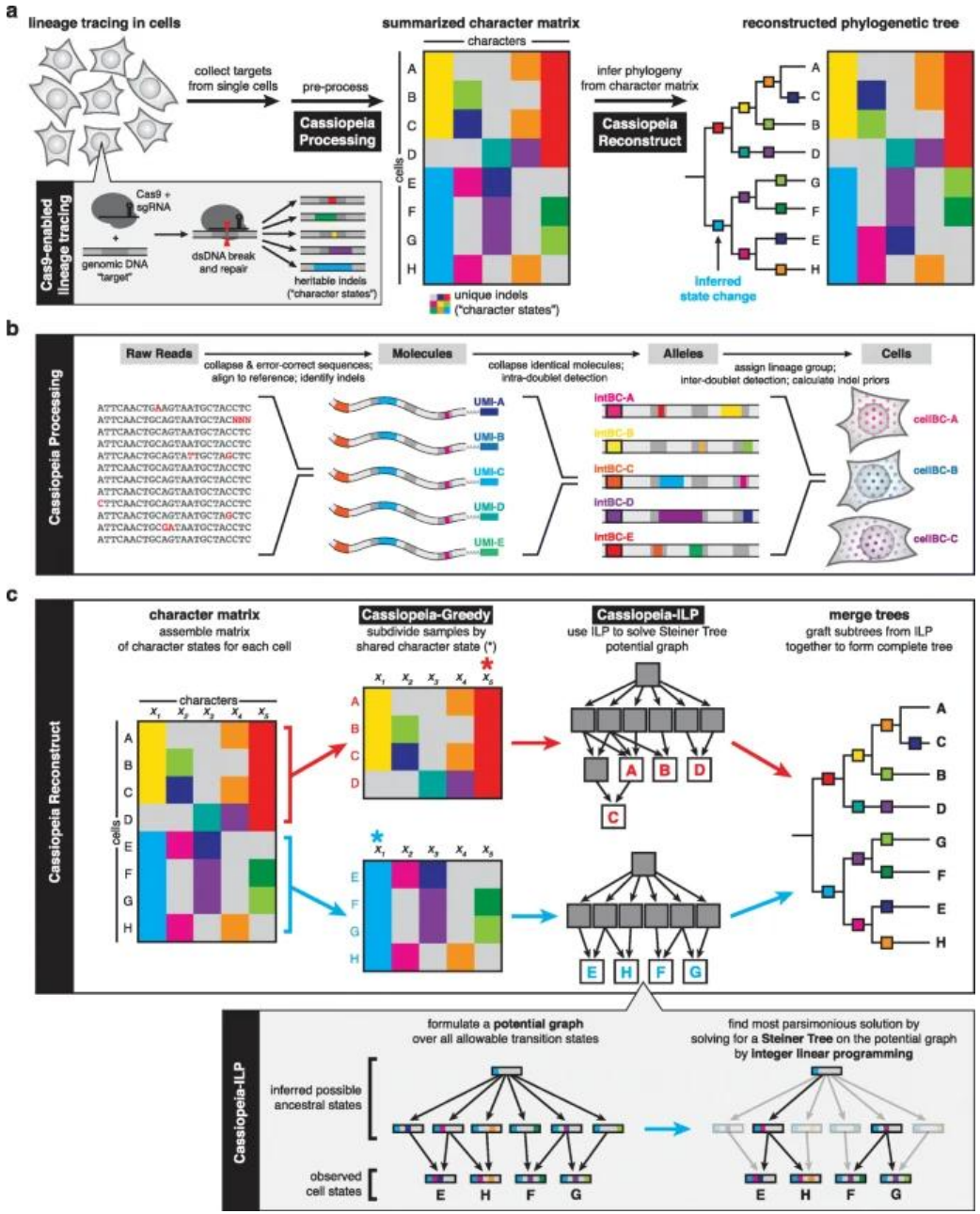
- Phylogeny algorithm

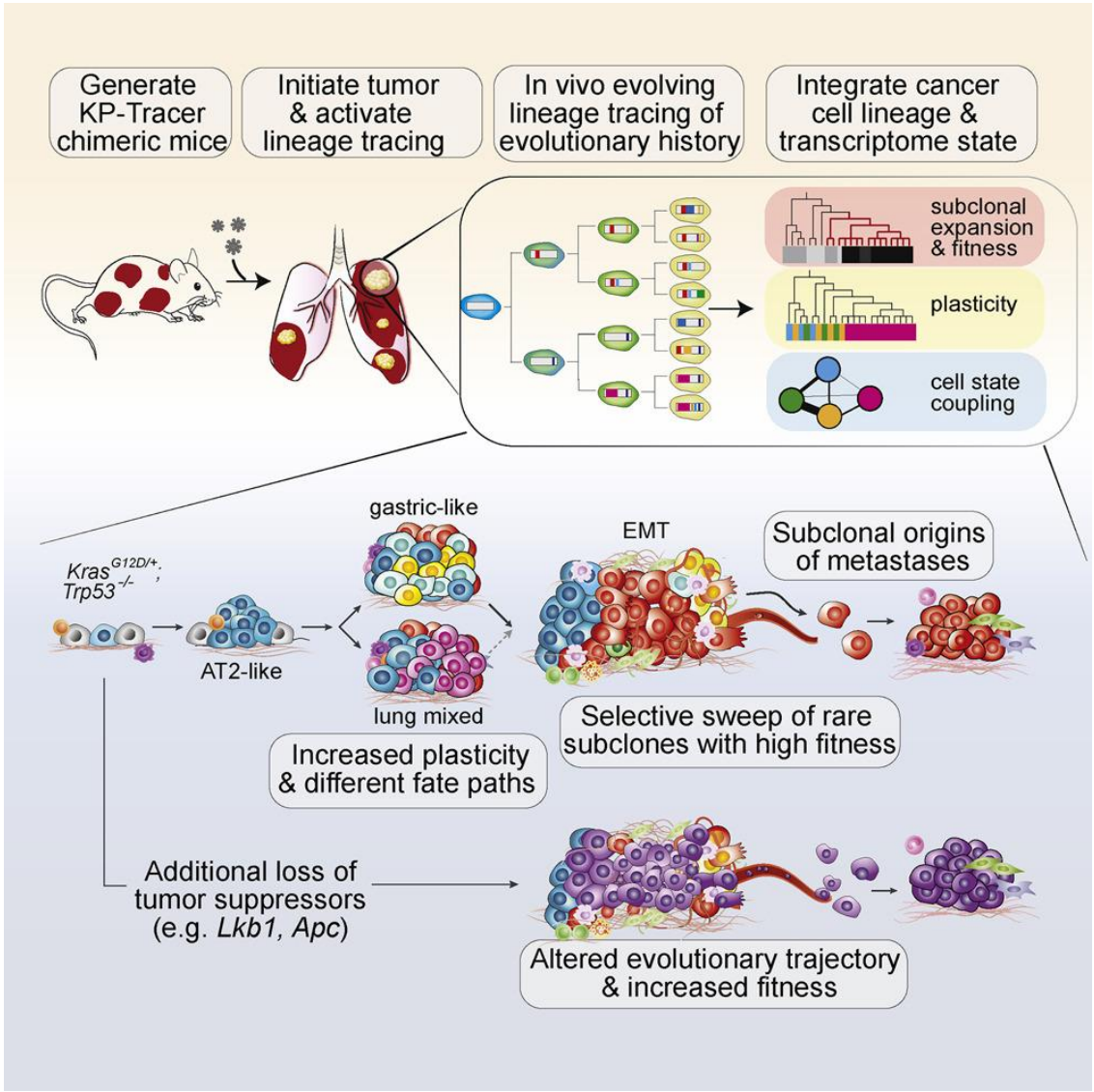


SCITE, 2016
Or CellPhy, 2022

→ Mutation profile → matrix
→ Tree generation
(based on root cell or the least mutated cell)

- Phylogeny algorithm



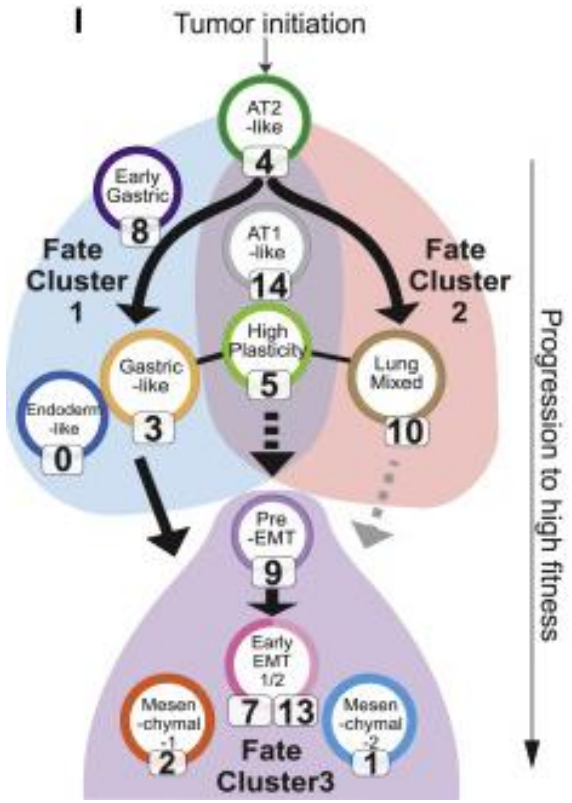


Lineage-tracing by barcode
Cf: Casiopea

- Genotyping & Phenotyping at single-cell level

-Typical 10x platform: 3 or 5’ bias
→ Poor mutation calling

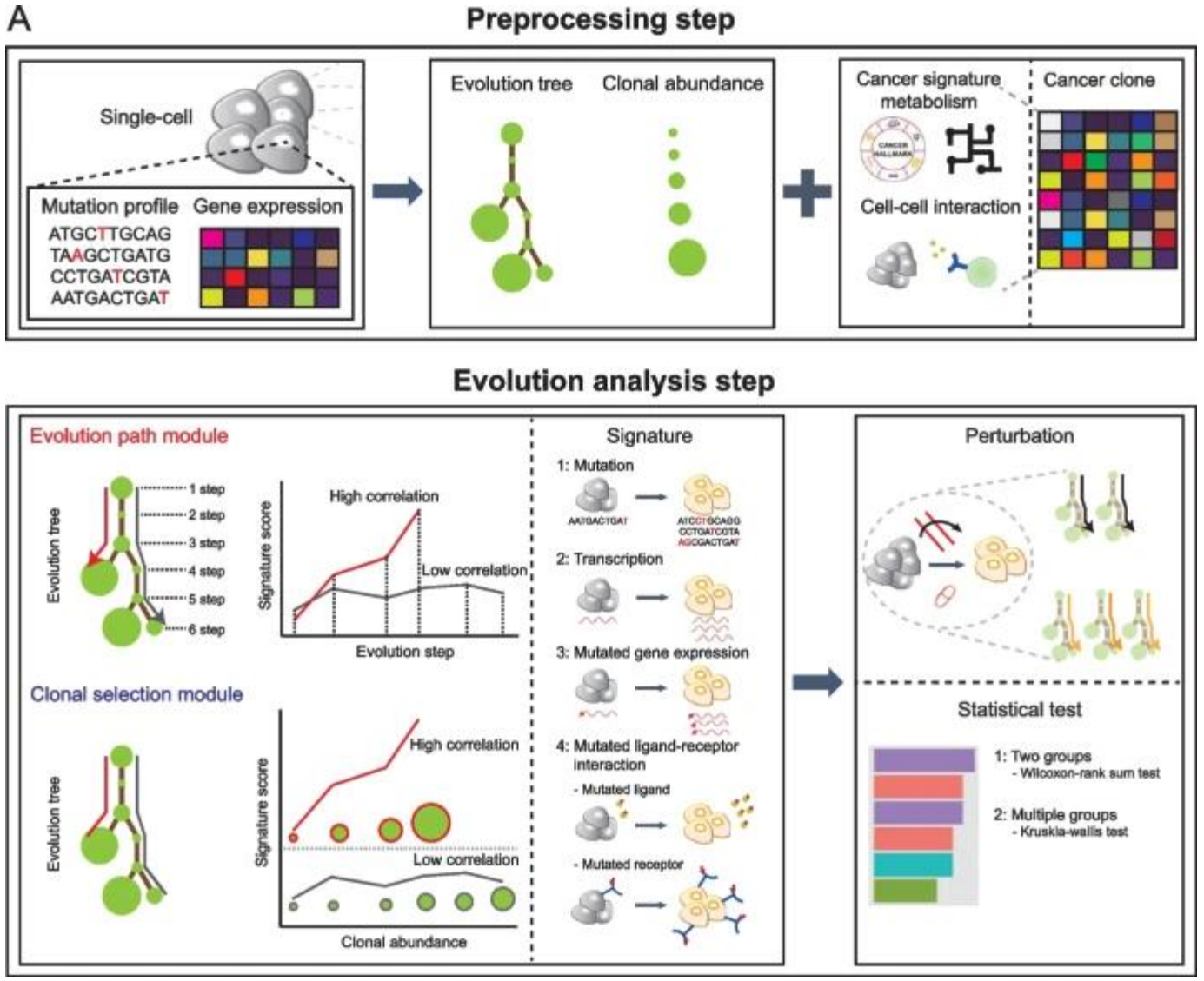
-Full-length sequencing: SMART-seq2 or Longread sequencing
→ Although it is limited to expressed genes but still it is “functional” mutation



Genetic – Transcriptomic interaction
→ Which clone: which phenotype
→ What kind of evolution occurs

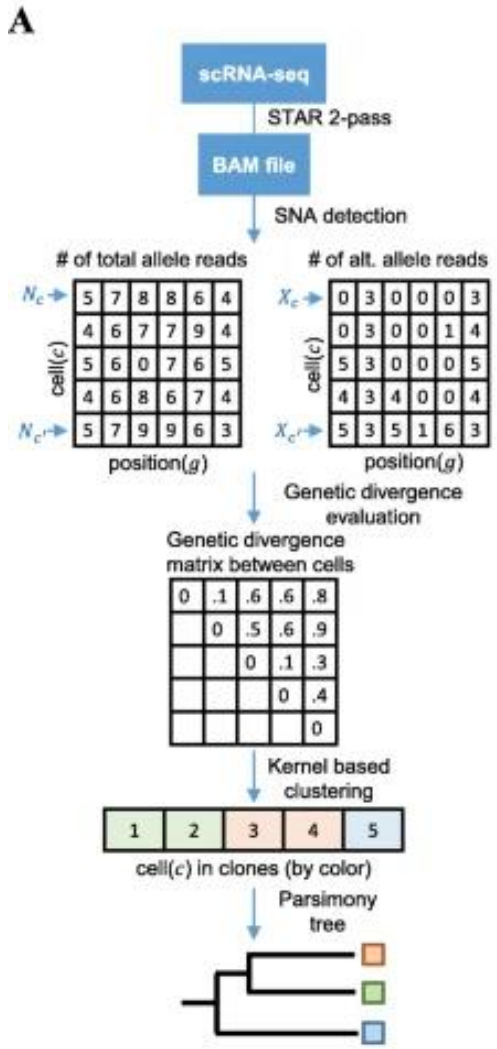
- Genotyping & Phenotyping at single-cell level

Canvolution: Joint analysis of mutational and transcriptional landscapes in human cancer reveals key perturbations during cancer evolution



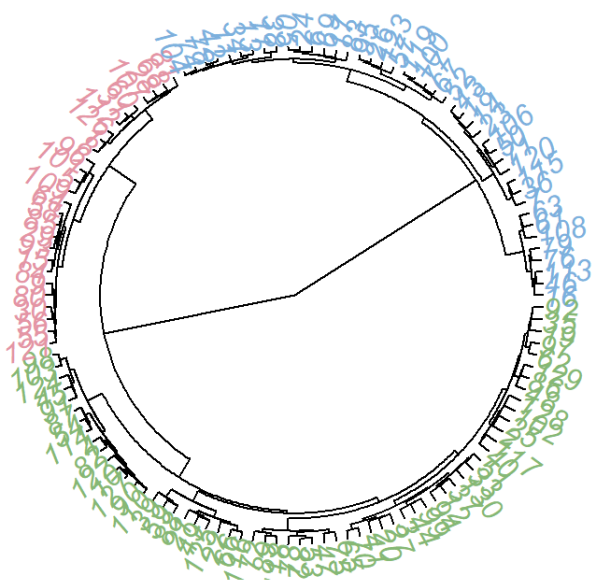
- Clonotyping by scRNA-seq

DENDRO



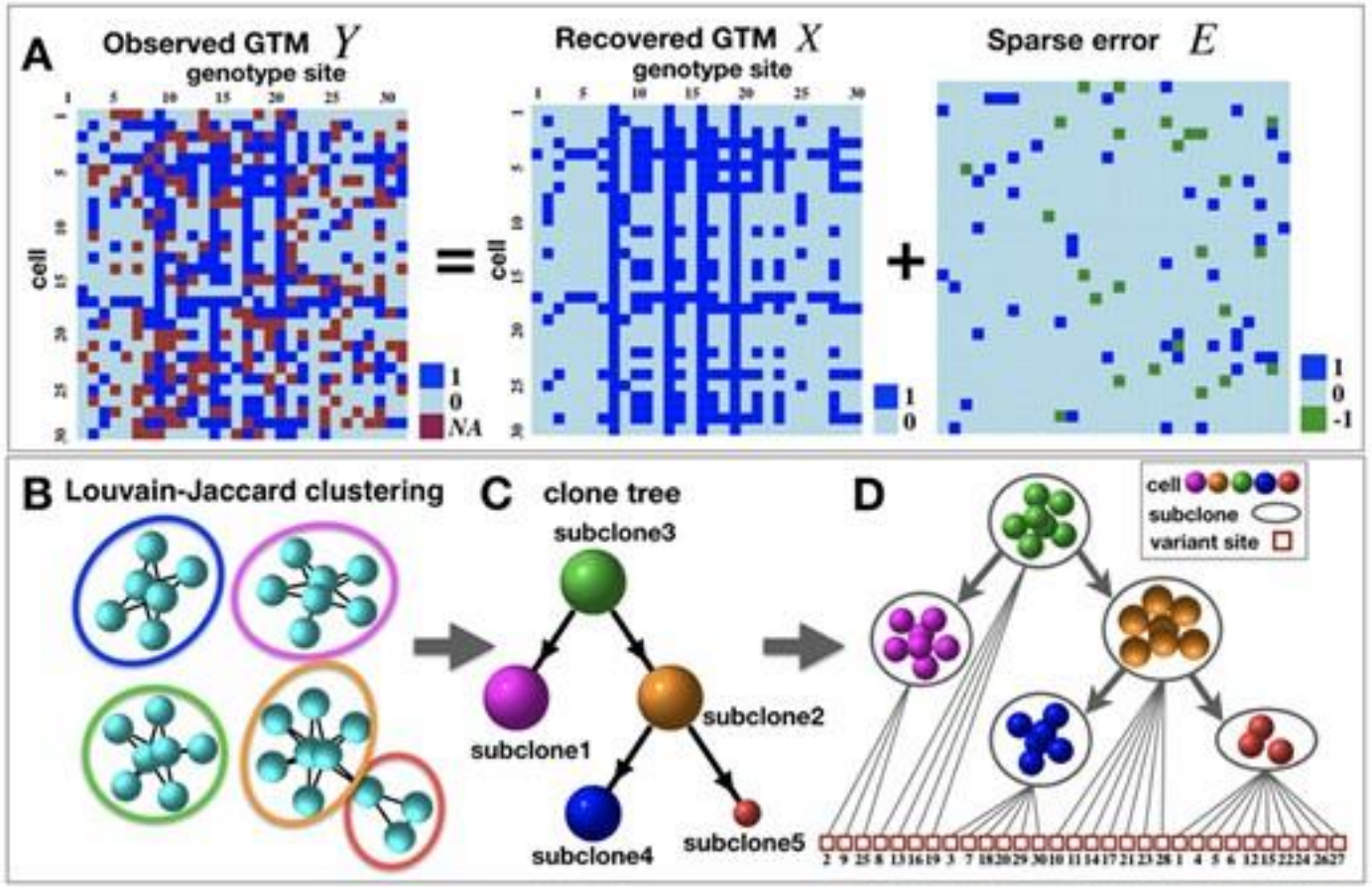
Designed for scRNA-seq
-Total read vs Alt read
→Distance measurement
→clustering: clonotyping
→ Parsimony Tree

DENDRO Result



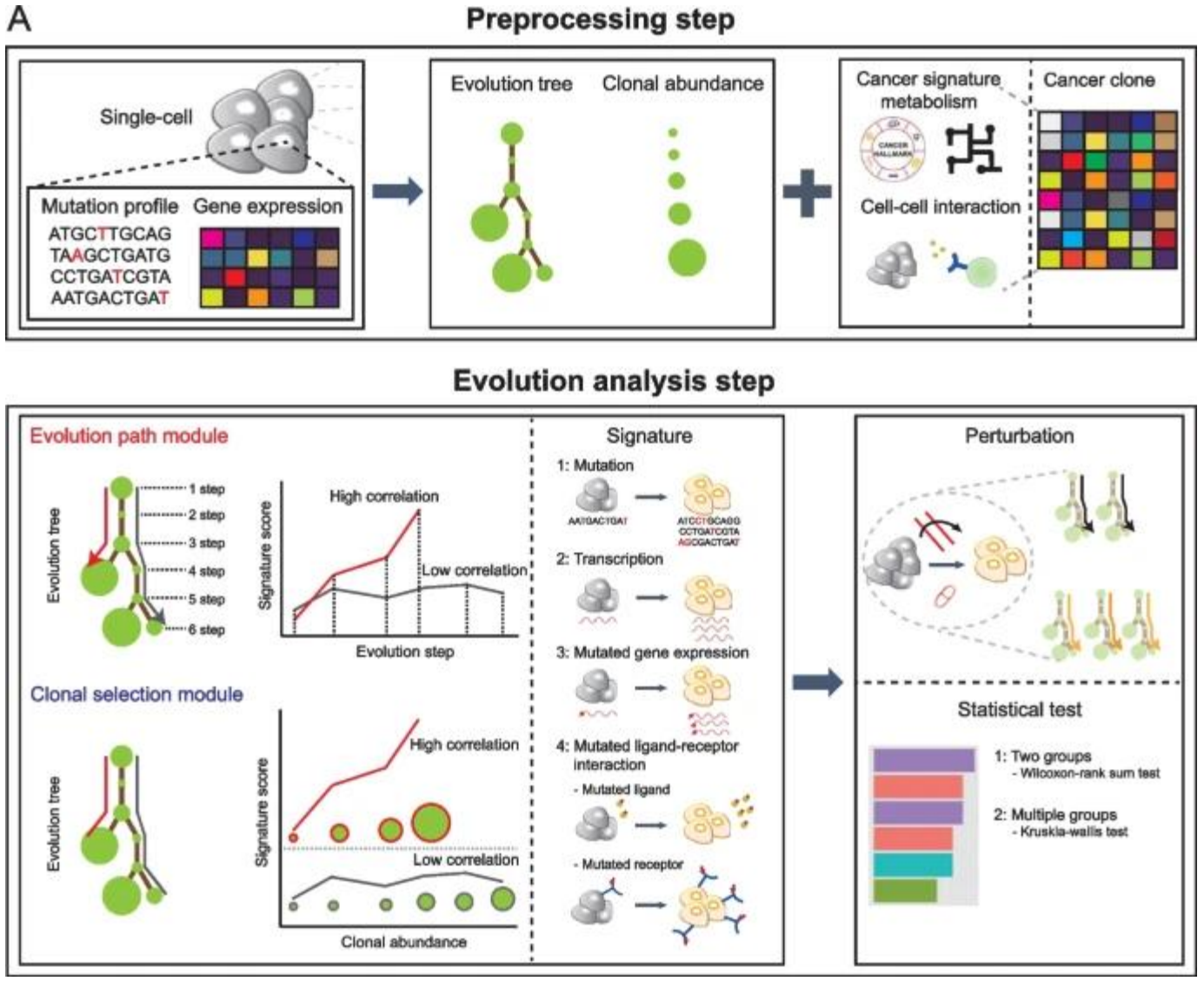
- Clonotyping by scRNA-seq

Robustclone
-Order of clonotype



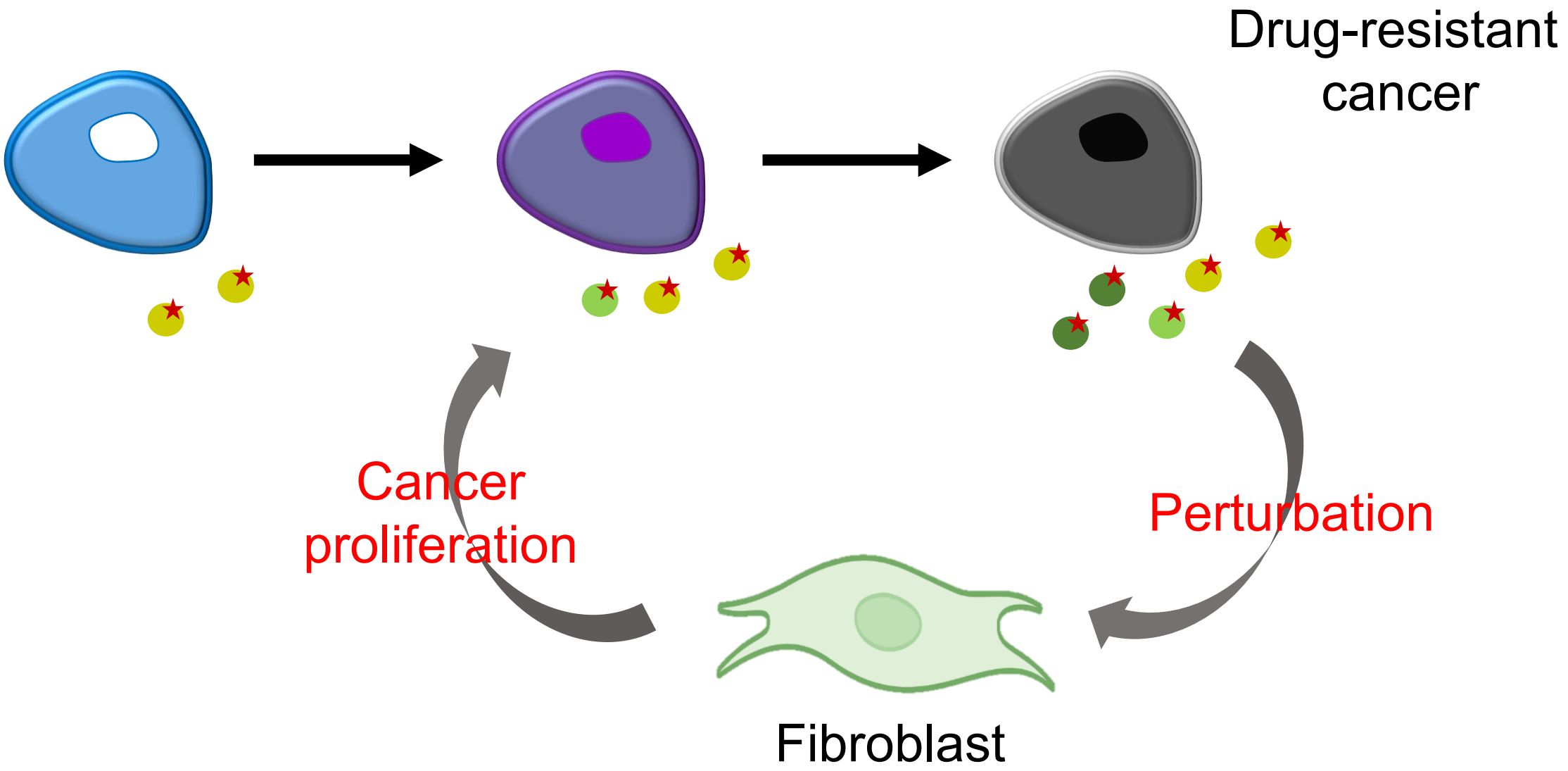
- Genotyping & Phenotyping at single-cell level

Canvolution: Joint analysis of mutational and transcriptional landscapes in human cancer reveals key perturbations during cancer evolution



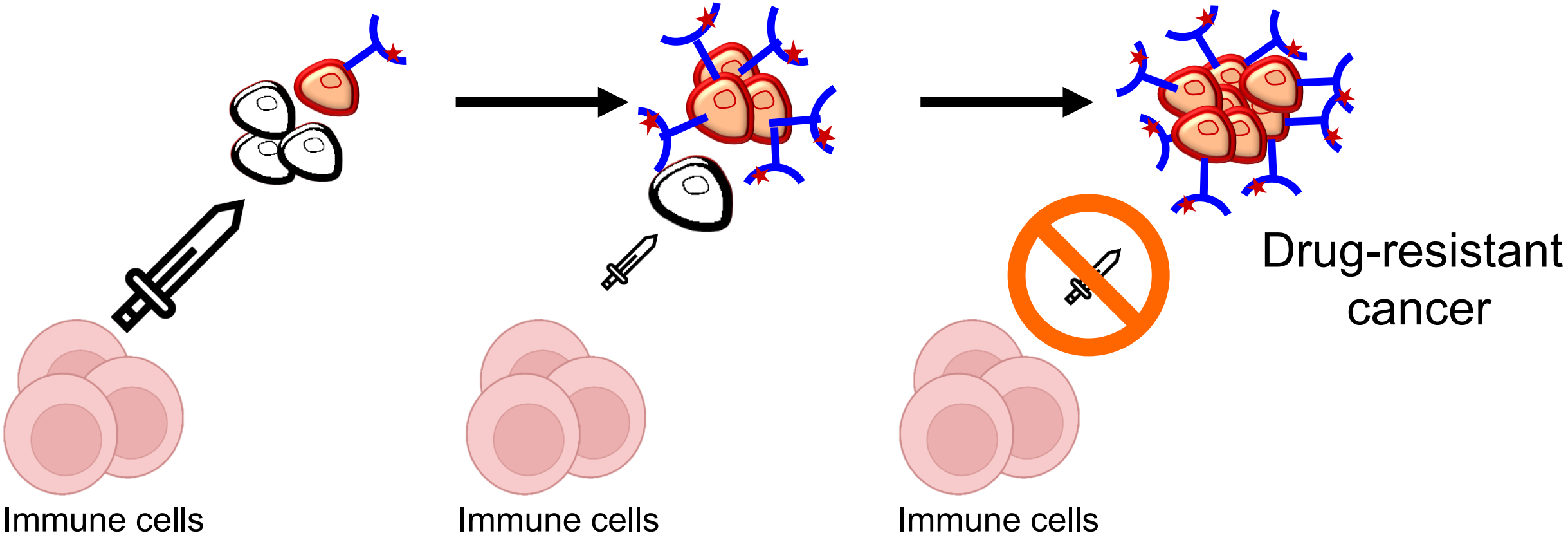
- Canovolution

- What phenotype is arising during cancer evolution?



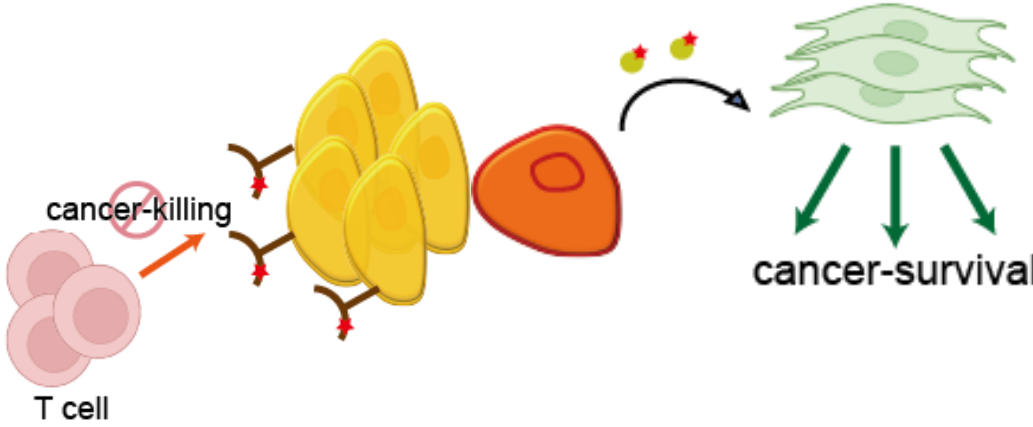
- Canvolution

- What phenotype is abundant during cancer evolution?



- Mutation in TGFBR2 & TNFR
- Block cancer-killing mechanisms → higher survival

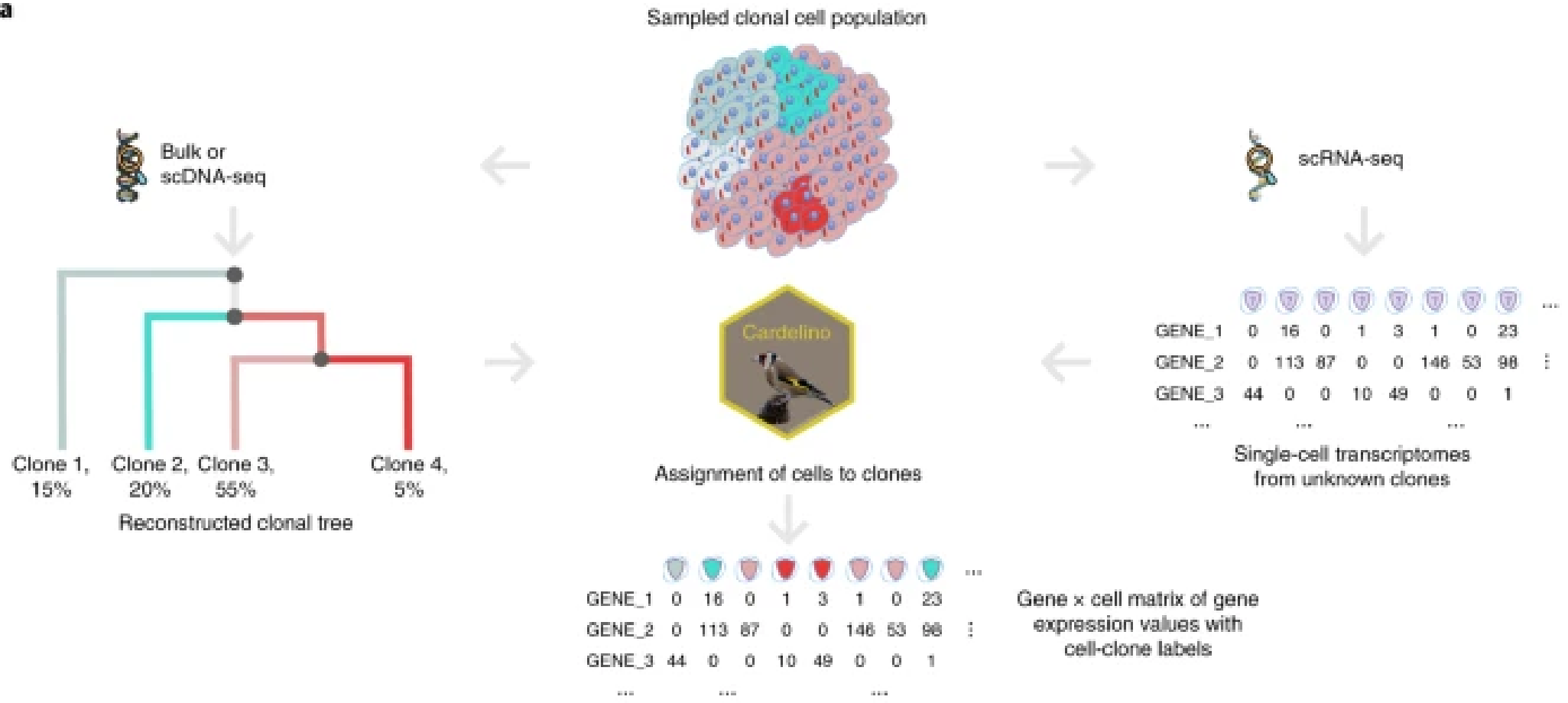
- Canvolution



Darwin theory → competition

Cooperation by distinctive roles from each other

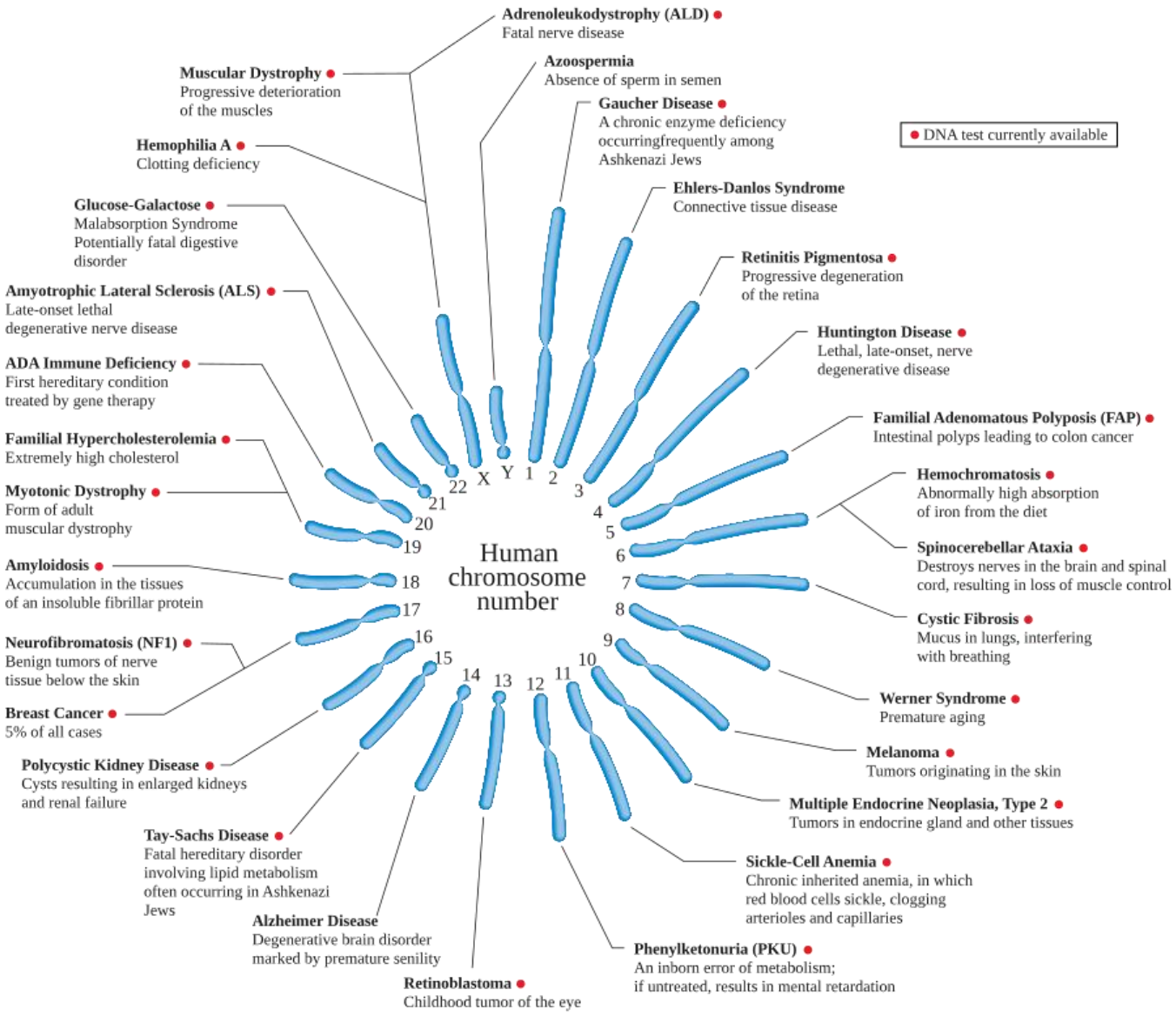
- Cardelino



-Obtaining both genomic and transcriptomic data from the same cell is quite challenging
→ Genomic data (different sample) → merge with scRNA-seq

Limitation: mutational profile should be similar across sampling site
Low sensitivity to detect the clonotype

- Genetic association



-Genetic diseases

→ Genetic mutations can affect disease

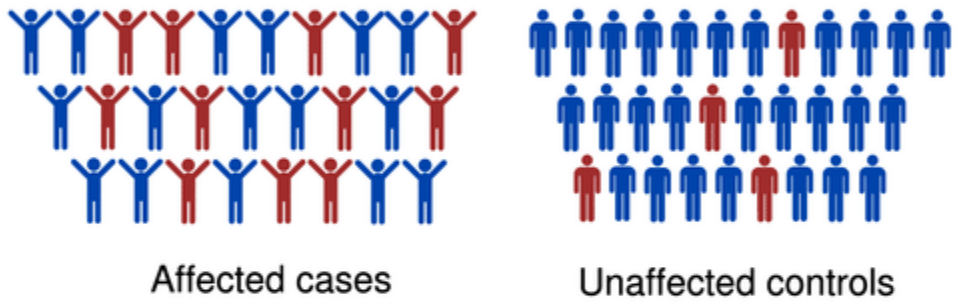
- Genetic association

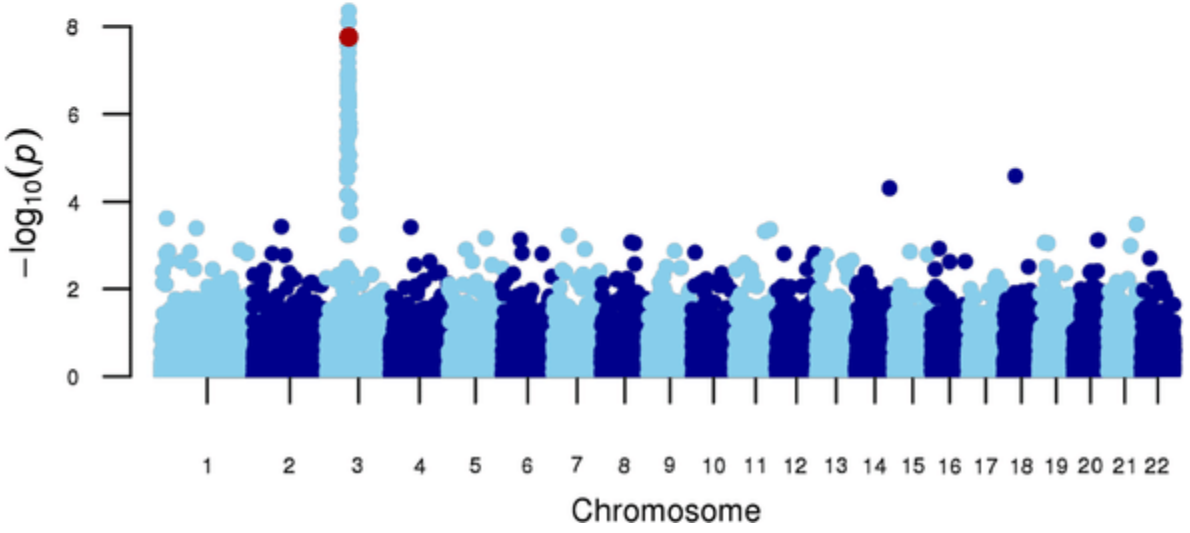
Sequence Variation

ATGCCAGTGTTTCAAGATGCTTGGCCAGCTGGACGAGGGCGATGAC
ATGCCAGTGTTTCAAGATG**T**TTGGCCAGCTGGACGAGGGCGATGAC

Disease

GWAS

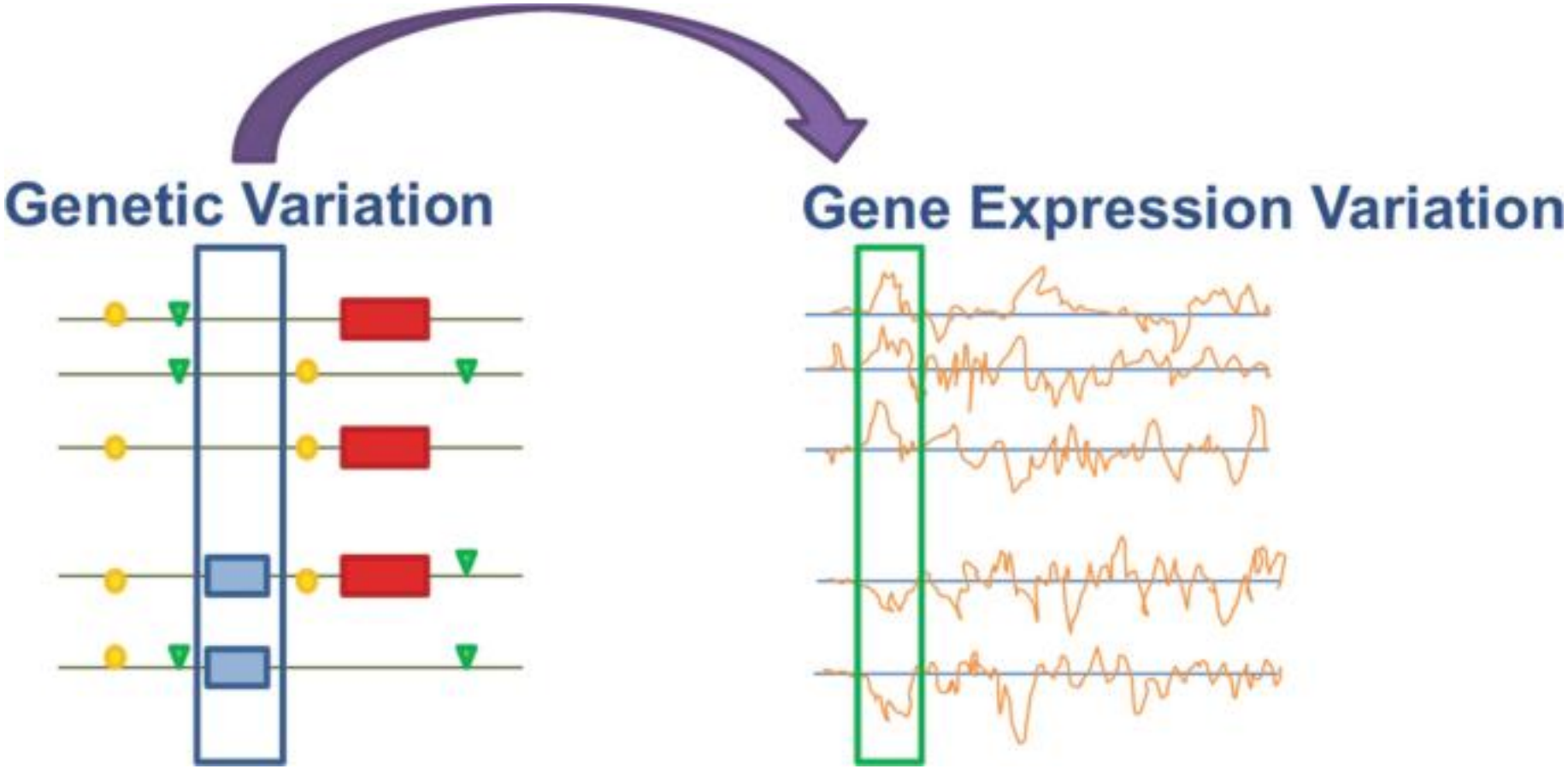




-GWAS
Genome-wide association study

→ All the genomic region
→ Associated with disease

- Single-cell eQTL



- Single-cell eQTL

-Cell-type-level eQTL → A better understanding of how genetic variants affect gene expression

Single-cell eQTLGen
Consortium