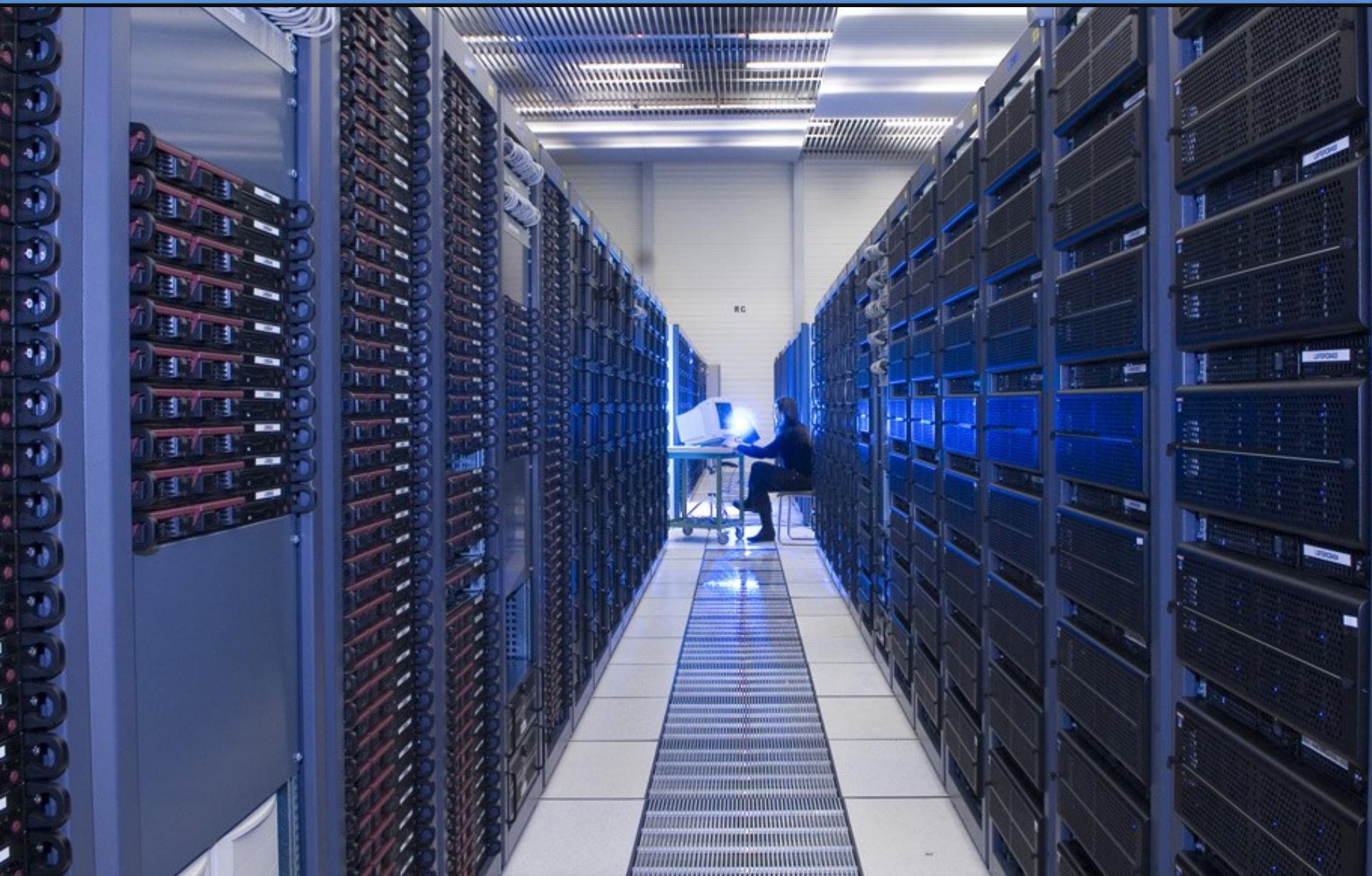


Cancer Sequencing



What is Cancer?

Definitions

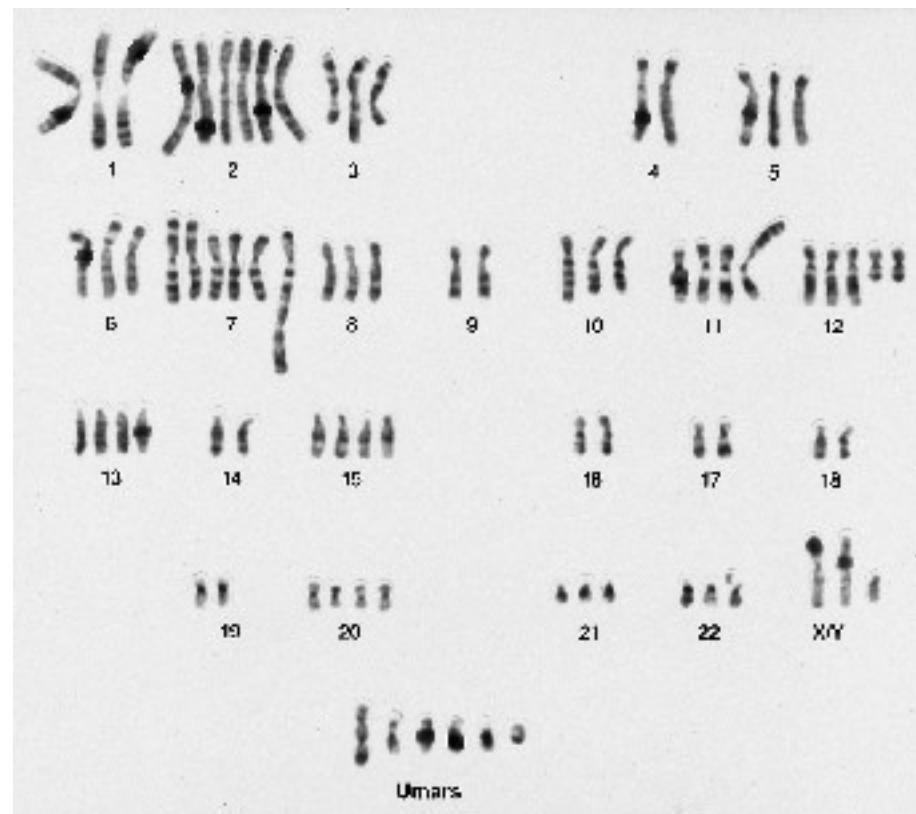
- A class of diseases characterized by malignant growth of a group of cells
 - Growth is uncontrolled
 - Invasive and Damaging
 - Often able to metastasize
- An instance of such a disease (a malignant tumor)
- A disease of the genome



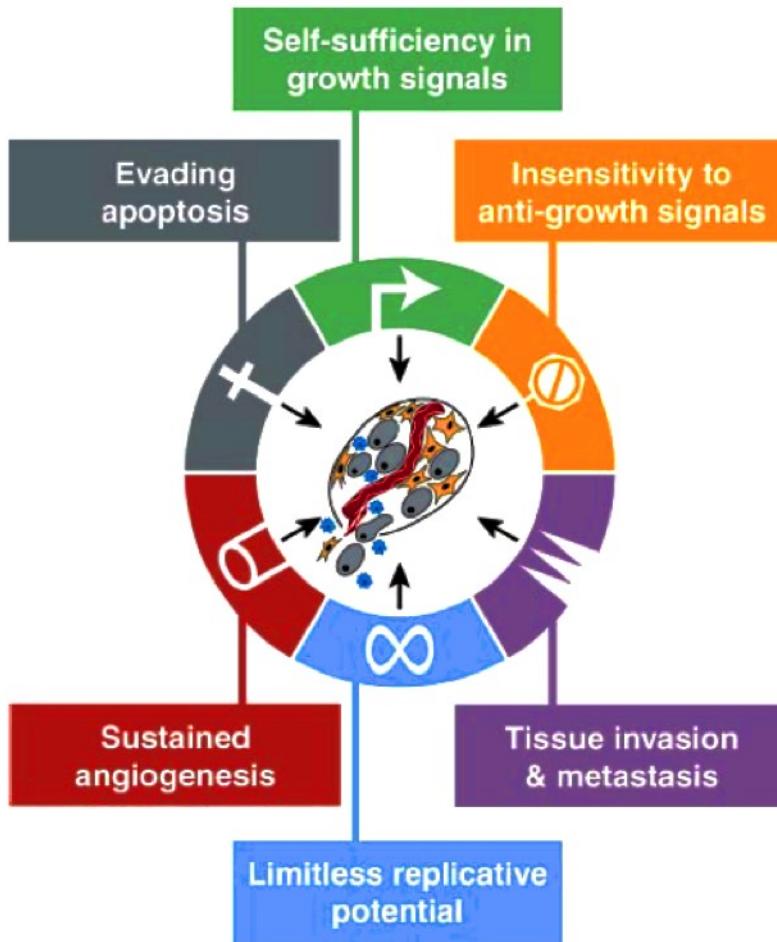
What is Cancer?

Definitions

- A class of diseases characterized by malignant growth of a group of cells
 - Growth is uncontrolled
 - Invasive and Damaging
 - Often able to metastasize
- An instance of such a disease (a malignant tumor)
- A disease of the genome



Fundamental Changes in Cancer Cell Physiology



Exploitation of natural pathways for cellular growth

- Growth Signals (e.g. TGF family)
- Angiogenesis
- Tissue Invasion & Metastasis

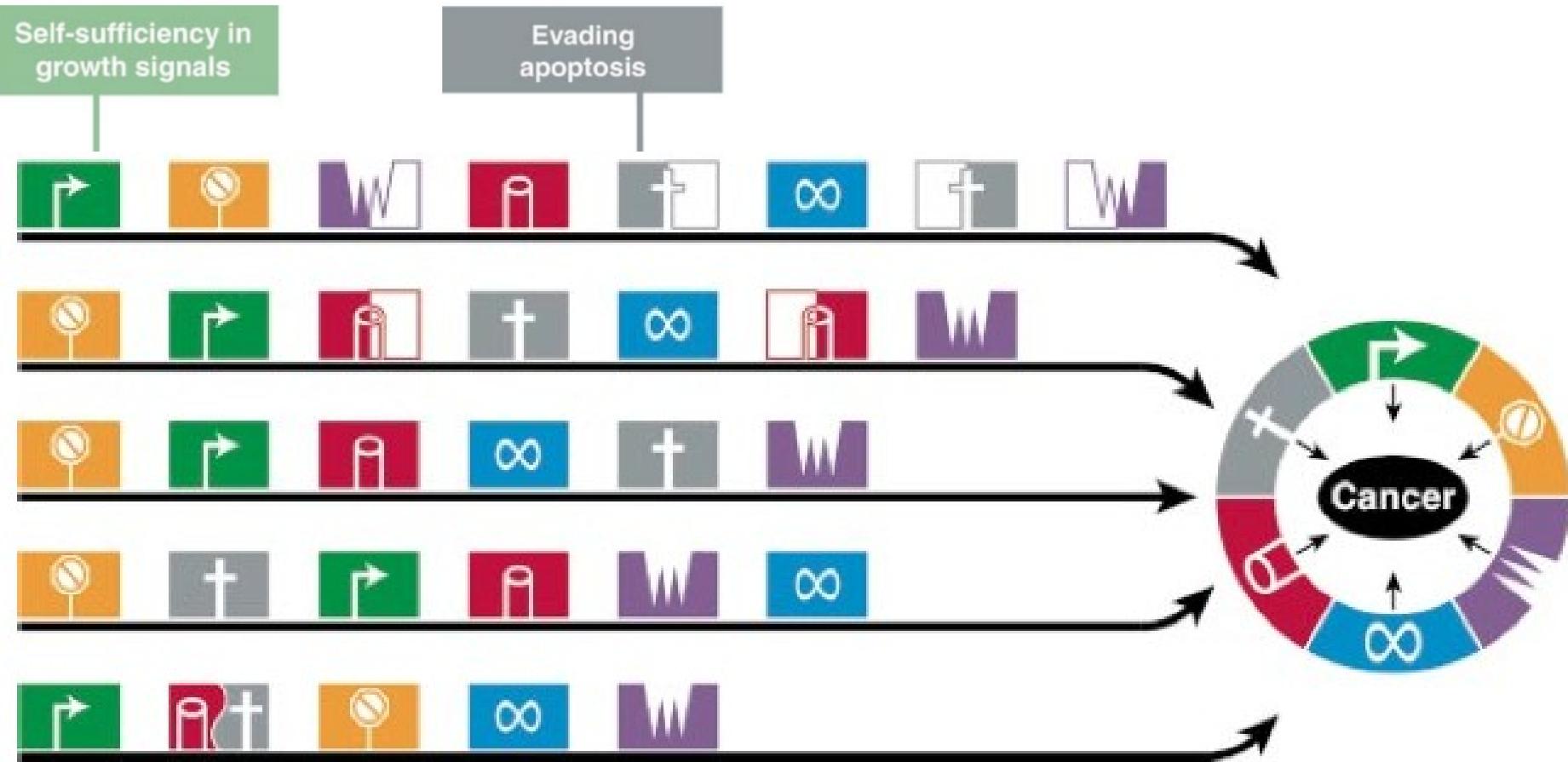
Evasion of anti-cancer control mechanisms

- Apoptosis (e.g. p53)
- Antigrowth signals (e.g. pRb)
- Cell Senescence

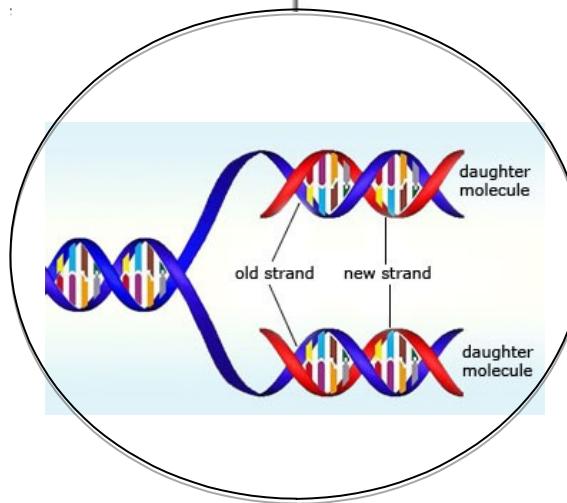
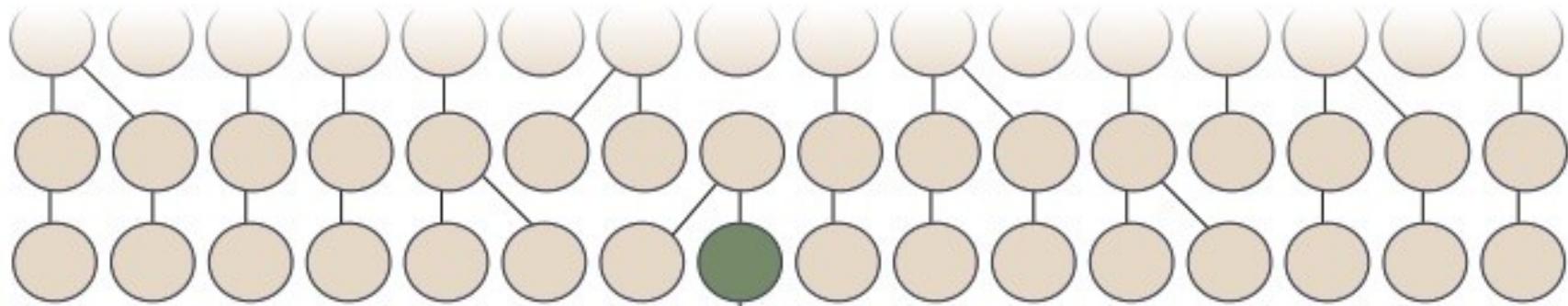
Acceleration of Cellular Evolution Via Genome Instability

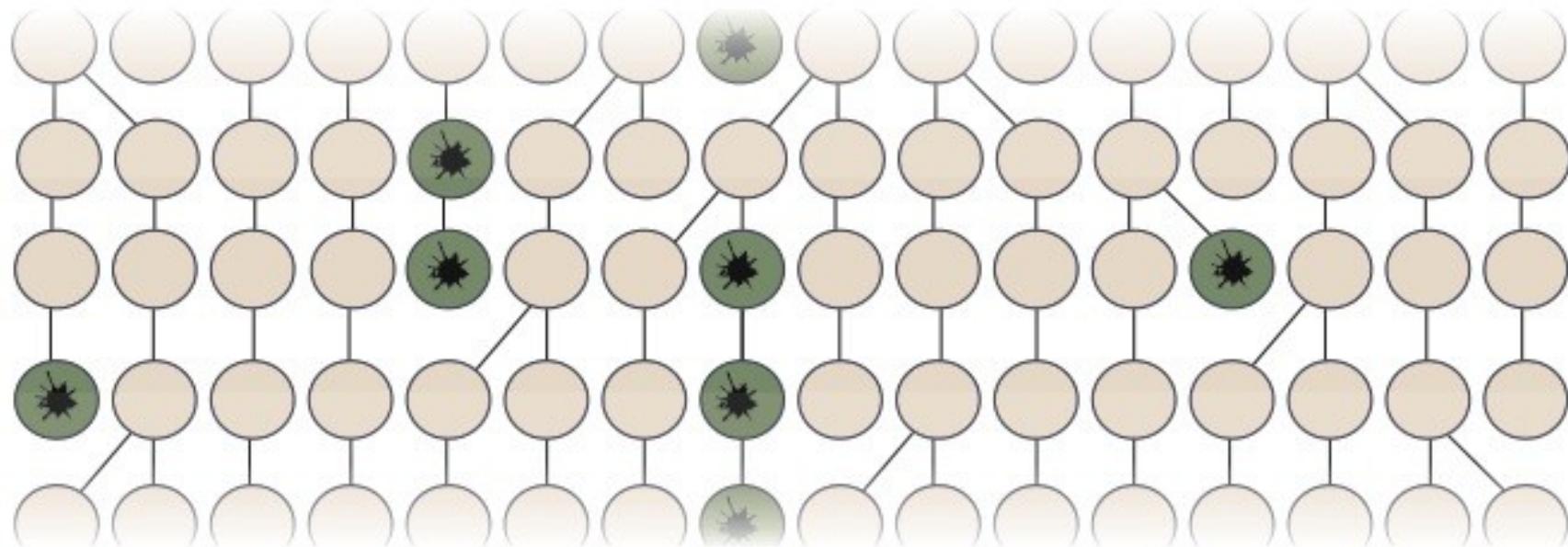
- DNA Repair
- DNA Polymerase

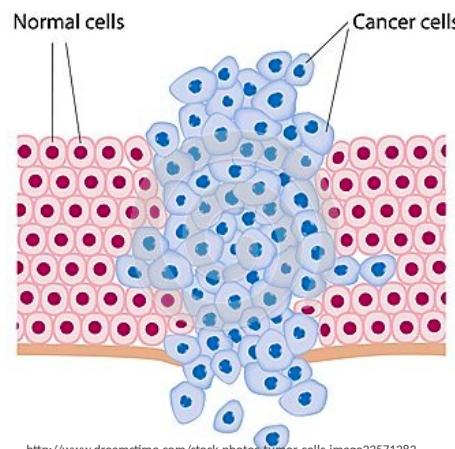
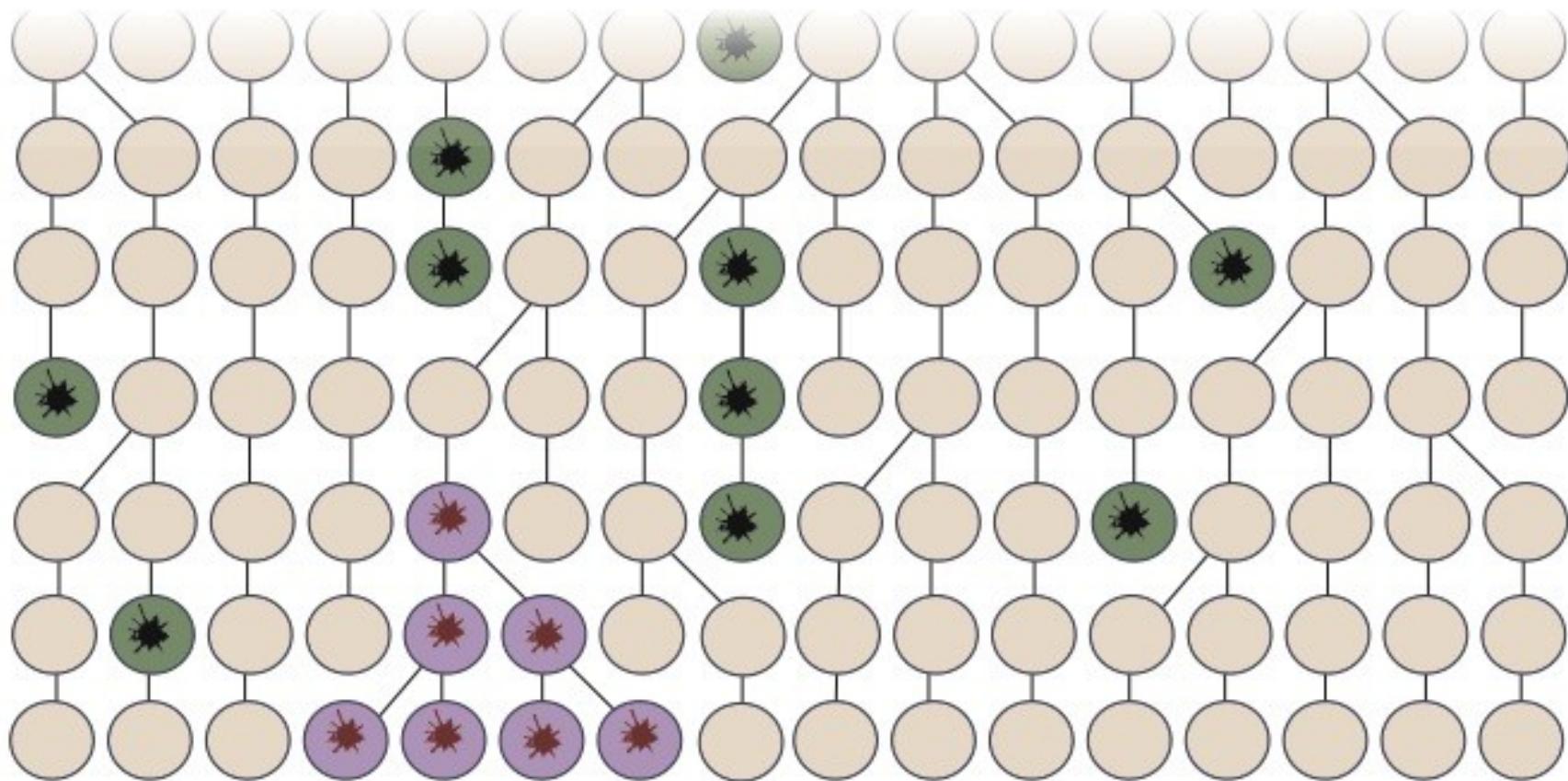
Many Paths Lead to Cancer Self-Sufficiency

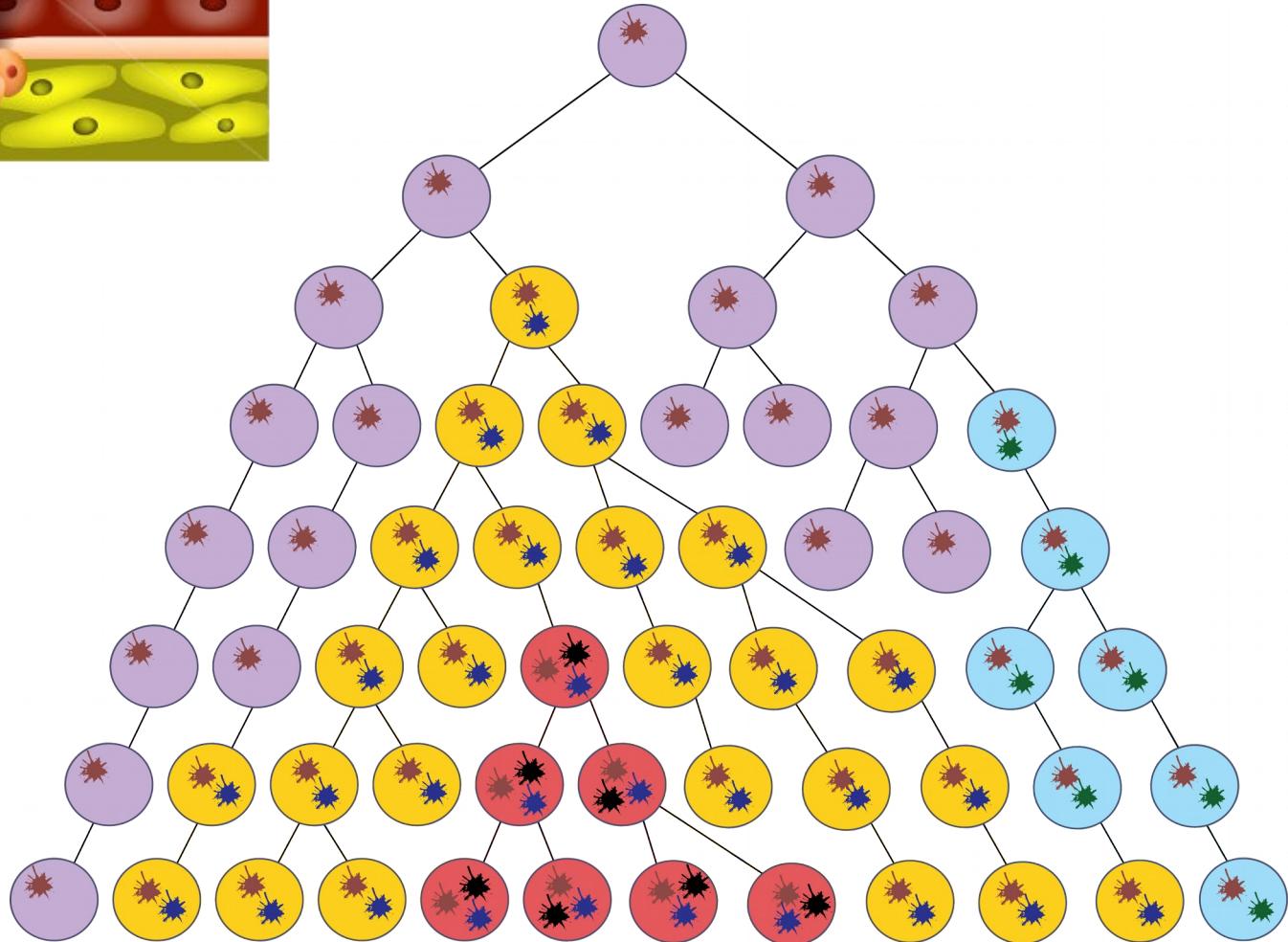
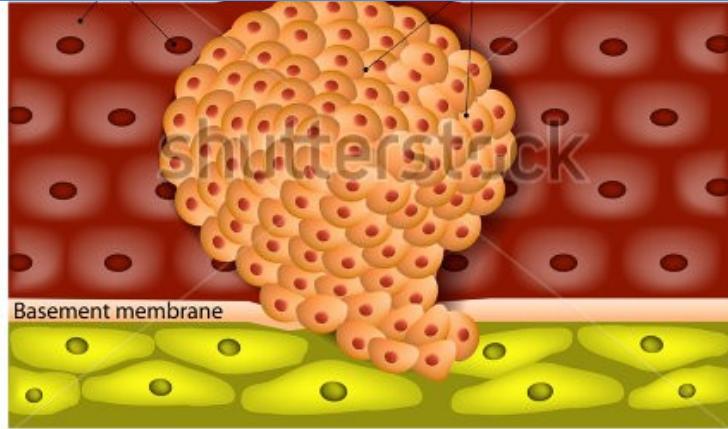


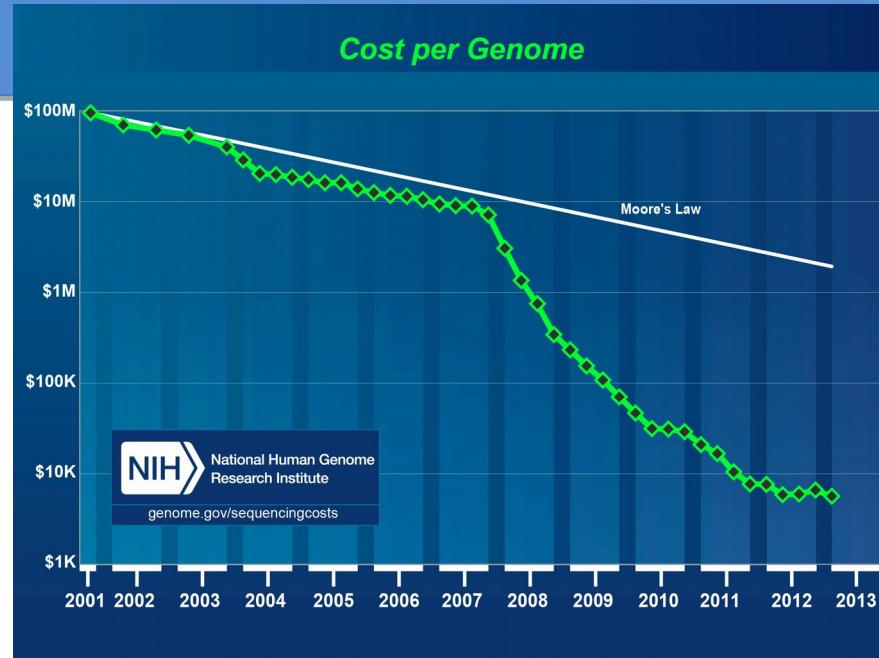
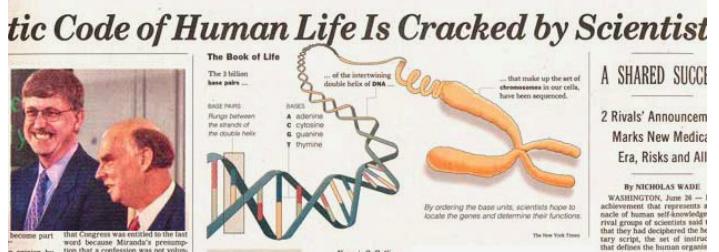






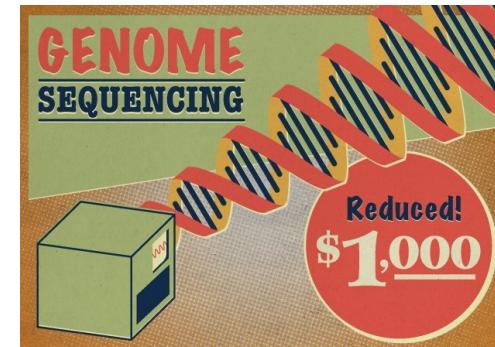






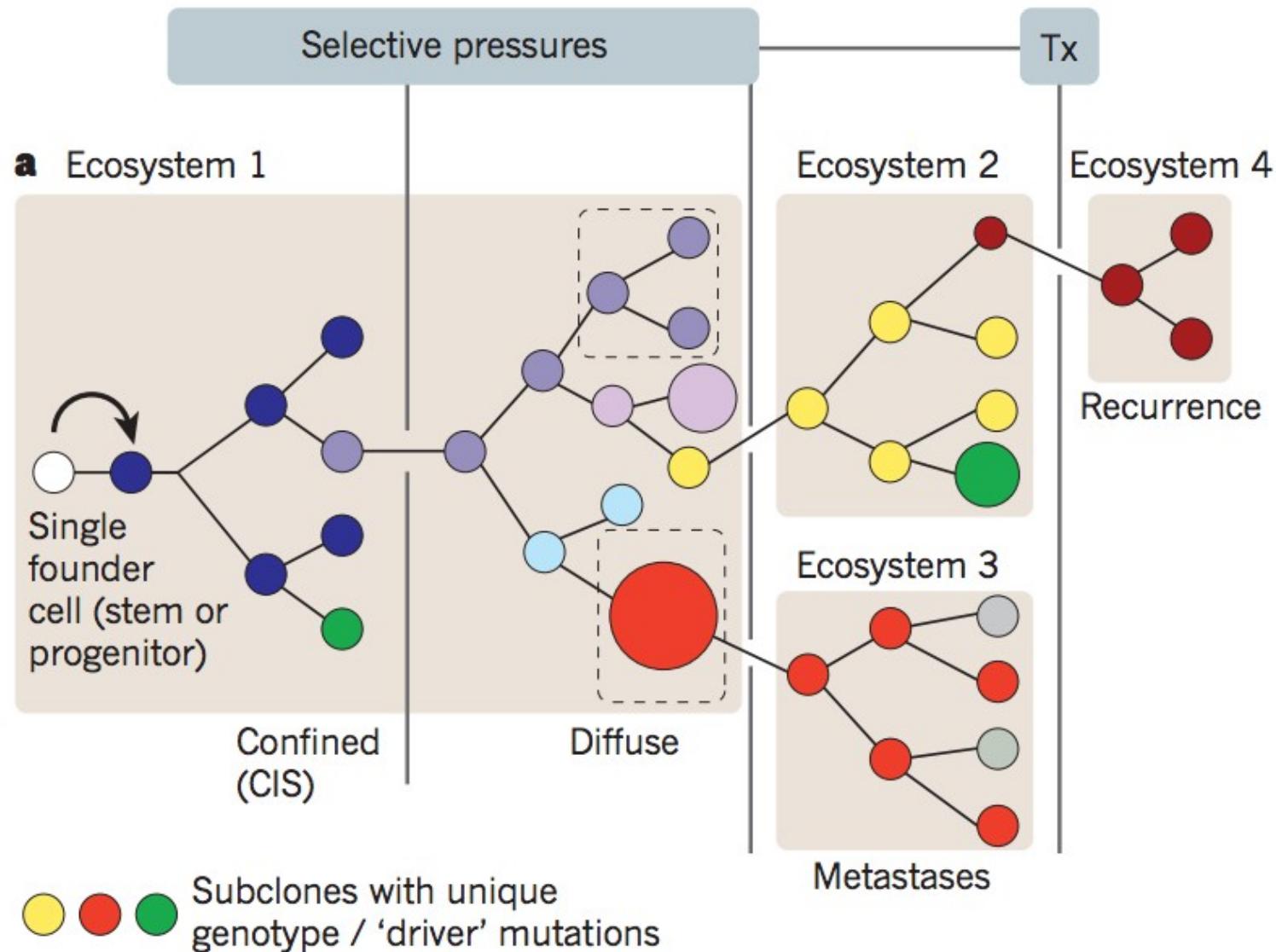
2000: initial draft

2003: Complete human reference genome
(\$3 billion)



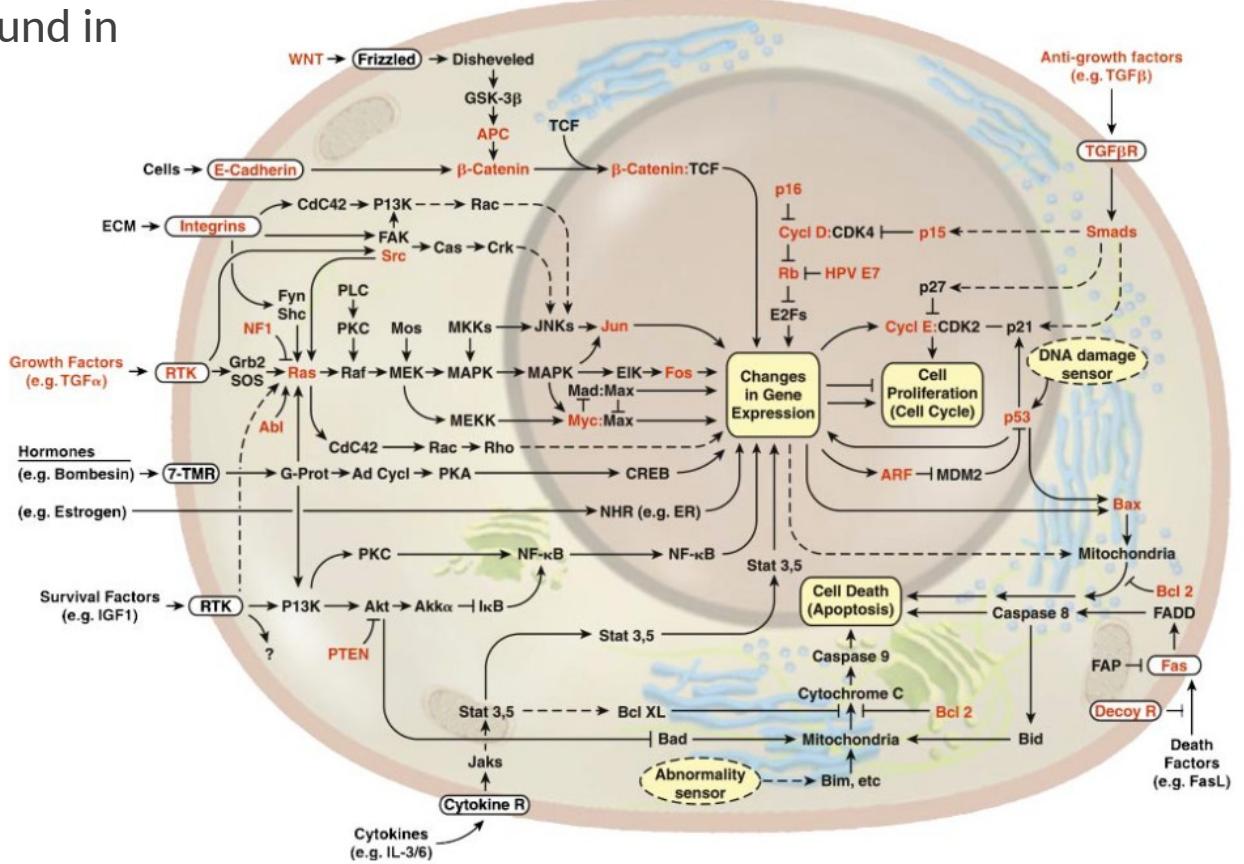
2014

Cancer Heterogeneity



Why Sequence Cancer Genomes?

- Better understand cancer biology
 - Pathway information
 - Types of mutations found in different cancers



Why Sequence Cancer Genomes?

- Better understand cancer biology
 - Pathway information
 - Types of mutations found in different cancers
- Cancer Diagnosis
 - Genetic signatures of cancer types will inform diagnosis
 - Non-invasive means of detecting or confirming presence of cancer
- Improve cancer therapies
 - Targeted treatment of cancer subtypes



Catalogue Of Somatic Mutations In Cancer

Samples	544809
Mutations	141212
Papers	10383
Whole Genomes	29

COSMIC Database, v48, July 2010
<http://www.sanger.ac.uk/genetics/CGP/cosmic/>

Why Sequence Cancer Genomes?

- Better understand cancer biology
 - Pathway information
 - Types of mutations found in different cancers
- Cancer Diagnosis
 - Genetic signatures of cancer types will inform diagnosis
 - Non-invasive means of detecting or confirming presence of cancer
- Improve cancer therapies
 - Targeted treatment of cancer subtypes



Catalogue Of Somatic Mutations In Cancer

Samples	1058292
Mutations	2710449
Papers	20247
Whole Genomes	15047

COSMIC Database, v71, Oct 2014
<http://www.sanger.ac.uk/genetics/CGP/cosmic/>

Why Sequence Cancer Genomes?

- Better understand cancer biology
 - Pathway information
 - Types of mutations found in different cancers
- Cancer Diagnosis
 - Genetic signatures of cancer types will inform diagnosis
 - Non-invasive means of detecting or confirming presence of cancer
- Improve cancer therapies
 - Targeted treatment of cancer subtypes

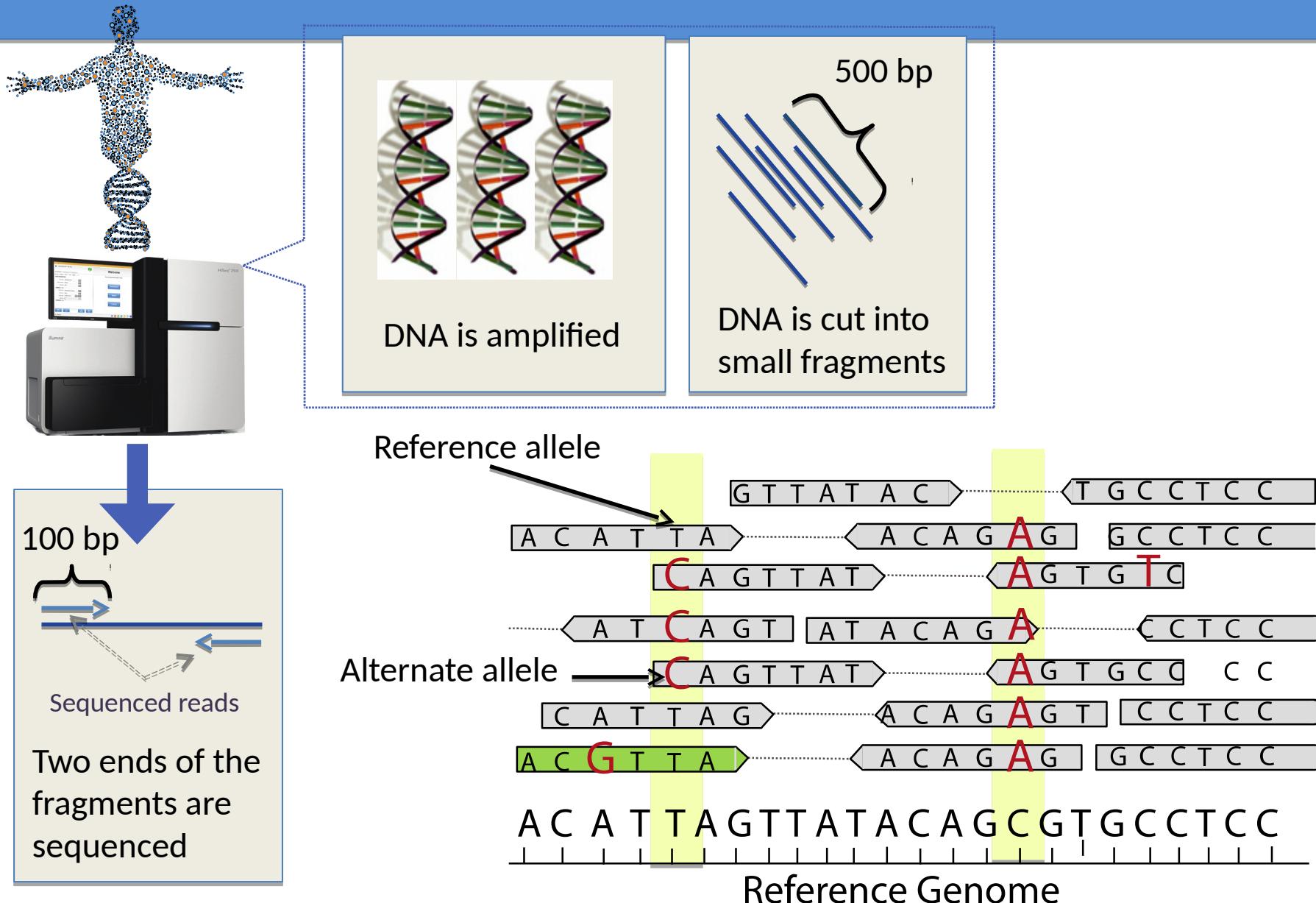


Catalogue Of Somatic Mutations In Cancer

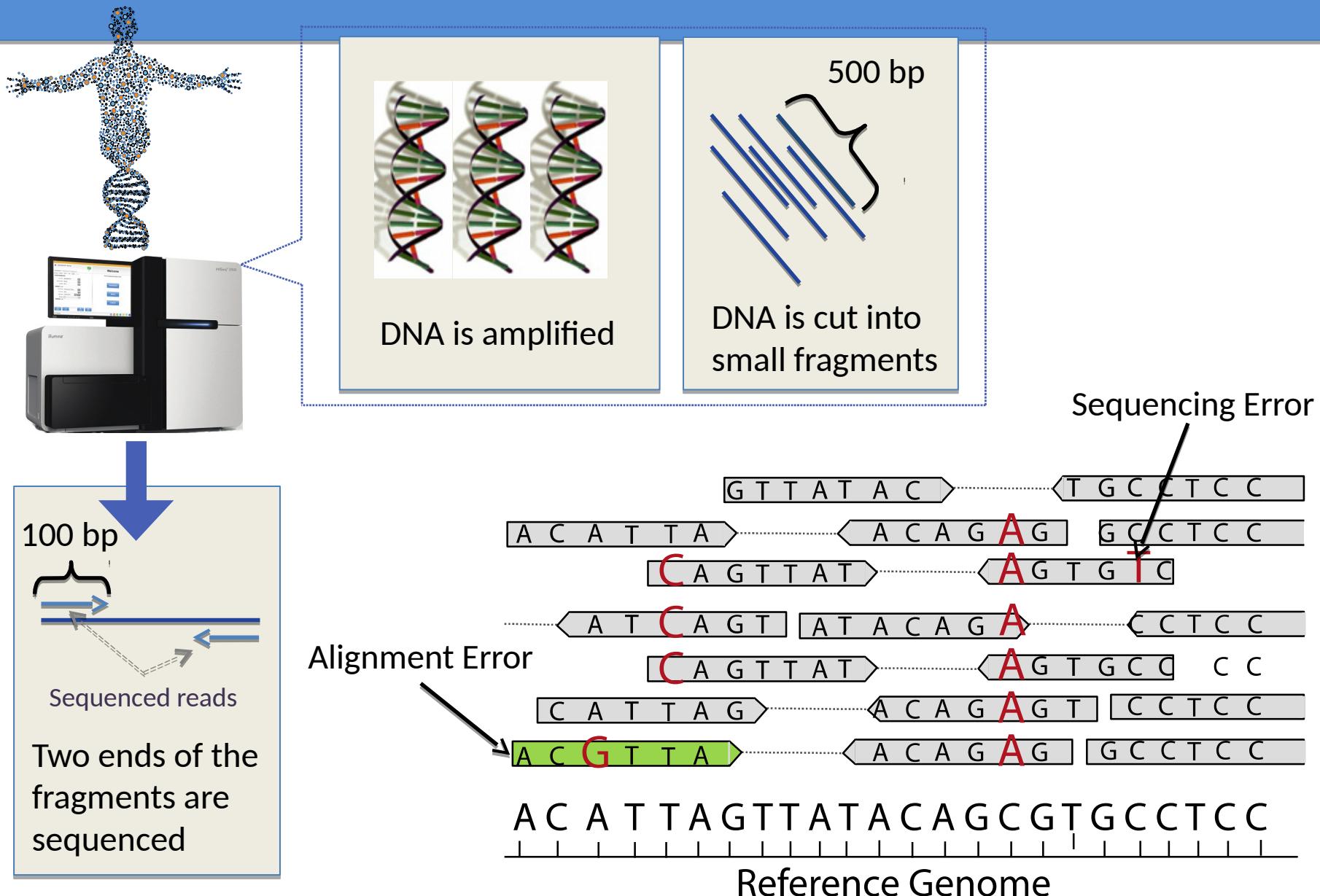
Samples	1,144,255
Mutations	3,480,051
Papers	22,276
Whole Genomes	22,690

COSMIC Database, v74, Sept 2015
<http://www.sanger.ac.uk/genetics/CGP/cosmic/>

Genome Background



Genome Background

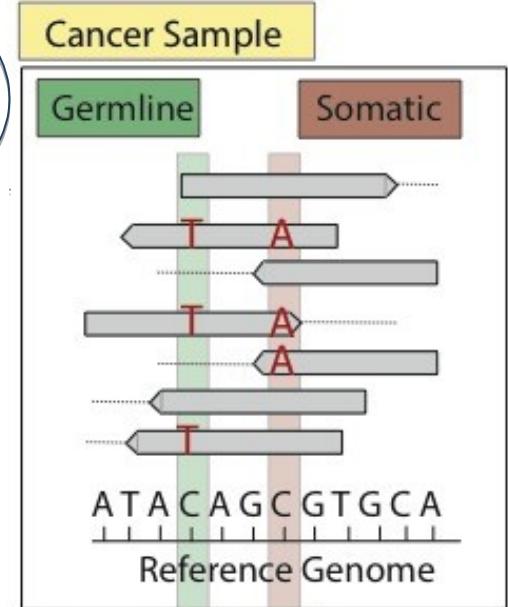
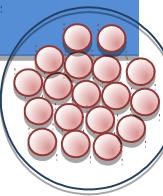
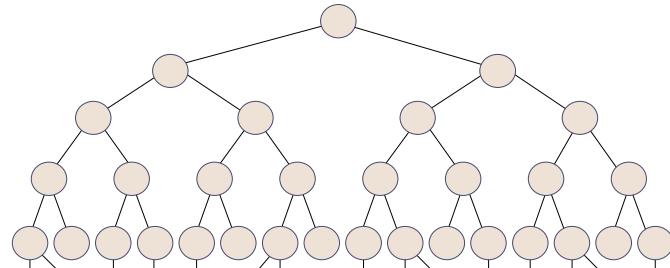


Genome Background

Types of SNVs in a cancer sample:

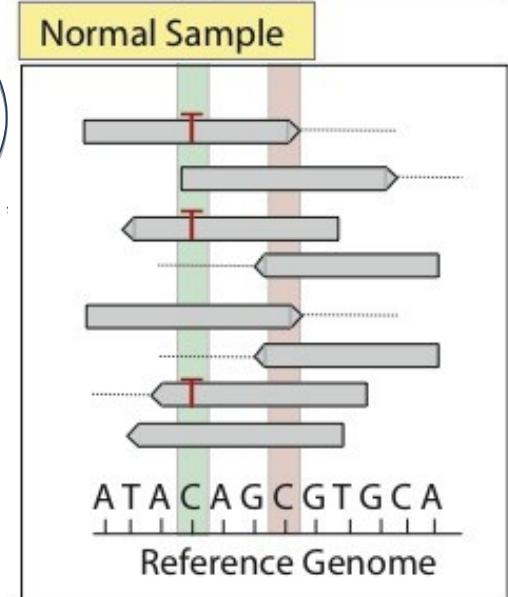
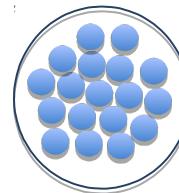
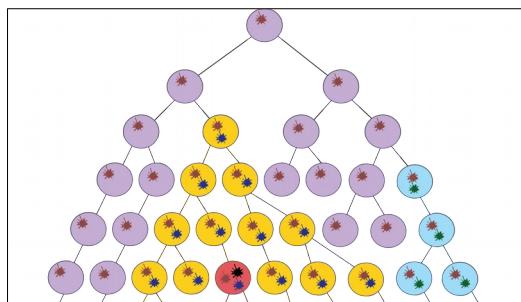
1. Germline (SNPs)

- Inherited
- All cells have it



2. Somatic

- Acquired during cancer progression
- Not present in normal cells



Considerations for Cancer Sequencing

- Factors that effect mutation signal
 - Limited genetic material (lower depth)
 - Mixture of tumor and normal tissue
 - Cancer Heterogeneity
- Factors that introduce noise
 - Formalin-fixed and Paraffin-embedded samples
 - Increased number of mutations and unusual genomic rearrangements
- General Consideration
 - Each individual has many unique mutations that could be confused with cancer causing mutations

Human Genome Variation

SNP

TGCT**T**GAGA
TGCCGAGA

Novel Sequence

TGCT**TCG**GAGA
TGC - - - GAGA

Inversion



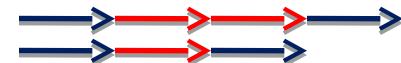
Mobile Element or
Pseudogene Insertion



Translocation



Tandem Duplication



Microdeletion

TGC - - AGA
TGCCGAGA

Transposition



Large Deletion



Novel Sequence
at Breakpoint

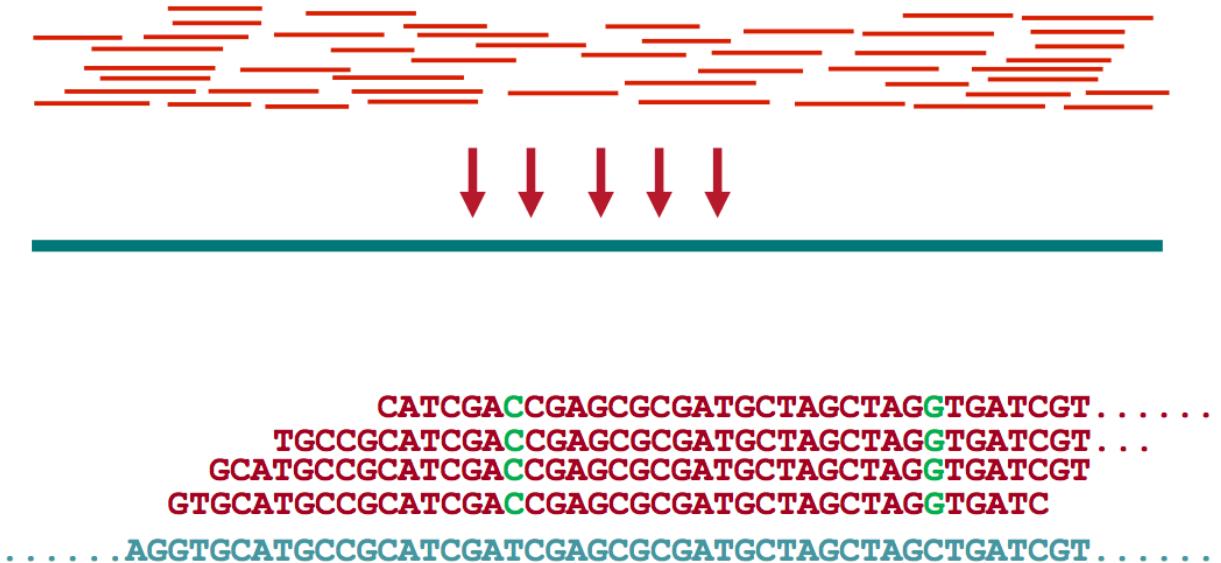


Variant Types

Variant Types
Single Nucleotide Variants(SNVs)
Small Insertion / Deletion (indels)
Copy Number Variants (CNVs)
Structural Variants (SVs)
Novel Sequence

SNV Calling

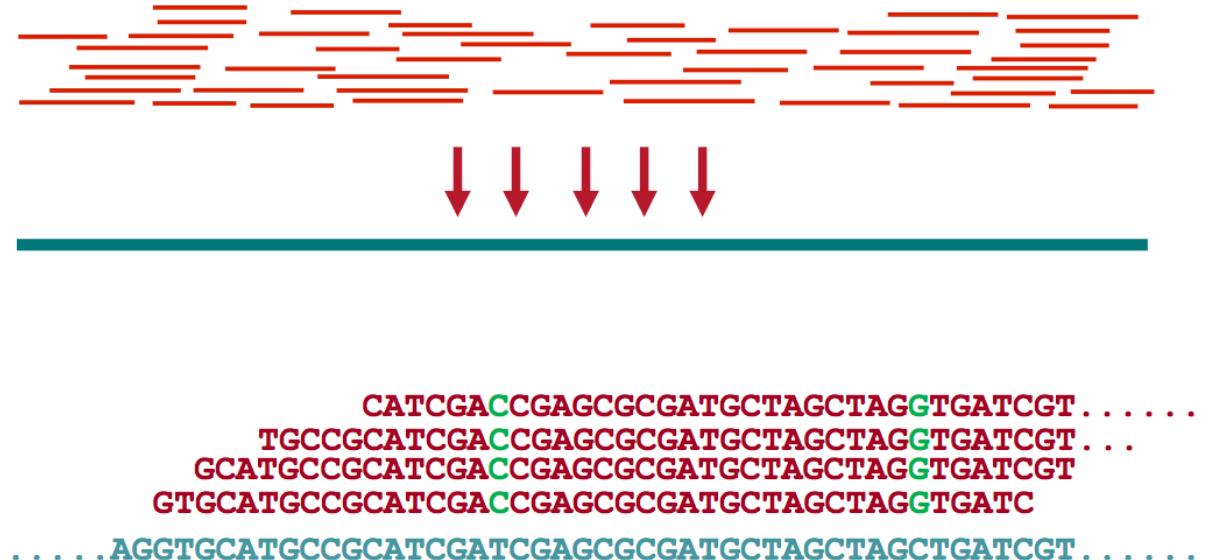
Variant Types
Single Nucleotide Variants(SNVs)
Small Insertion / Deletion (indels)
Copy Number Variants (CNVs)
Structural Variants (SVs)
Novel Sequence



- A bayesian approach is the most general and common method of calling SNVs
 - MAQ, SOAPsnp, Genome Analyis ToolKit (GATK), SAMtools

SNV Calling

Variant Types
Single Nucleotide Variants(SNVs)
Small Insertion / Deletion (indels)
Copy Number Variants (CNVs)
Structural Variants (SVs)
Novel Sequence



Prior of the genotype Likelihood of the genotype

$$\Pr\{G|D\} = \frac{\Pr\{G\} \Pr\{D|G\}}{\sum_i \Pr\{G_i\} \Pr\{D|G_i\}}, \text{ [Bayes' rule]}$$

$$\Pr\{D|G\} = \prod_j \left(\frac{\Pr\{D_j|H_1\}}{2} + \frac{\Pr\{D_j|H_2\}}{2} \right) \text{ where } G = H_1 H_2$$

Diploid assumption

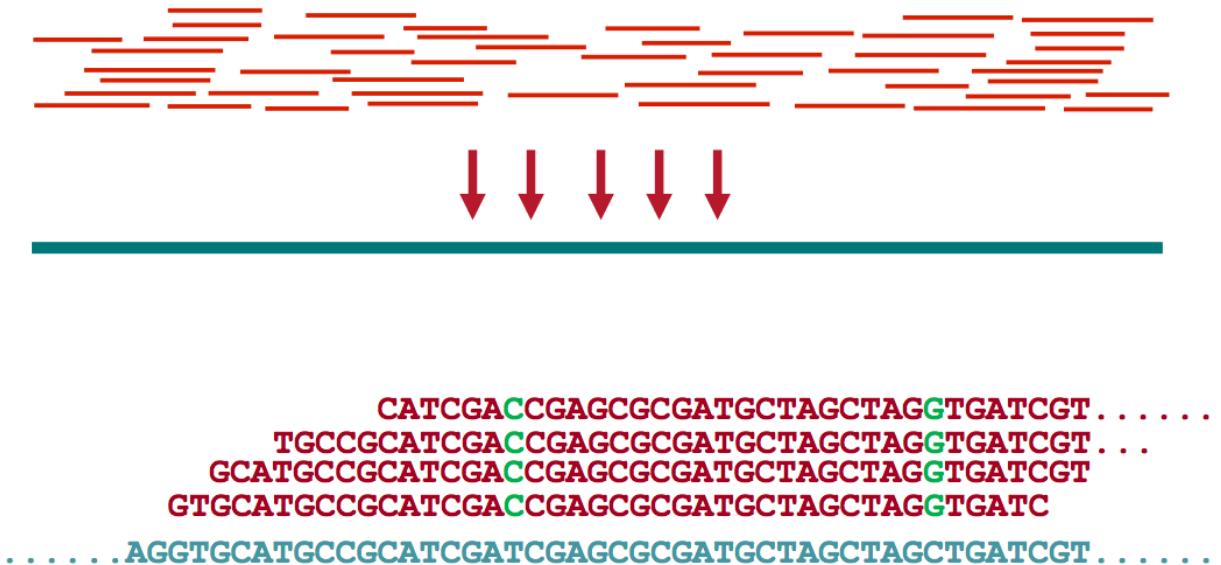
Bayesian model $\Pr\{D|H\}$ is the haploid likelihood function

$$\Pr\{D_j|H\} = \Pr\{D_j|b\}, \text{ [single base pileup]}$$

$$\Pr\{D_j|b\} = \begin{cases} 1 - \epsilon_j & D_j = b, \\ \epsilon_j & \text{otherwise.} \end{cases}$$

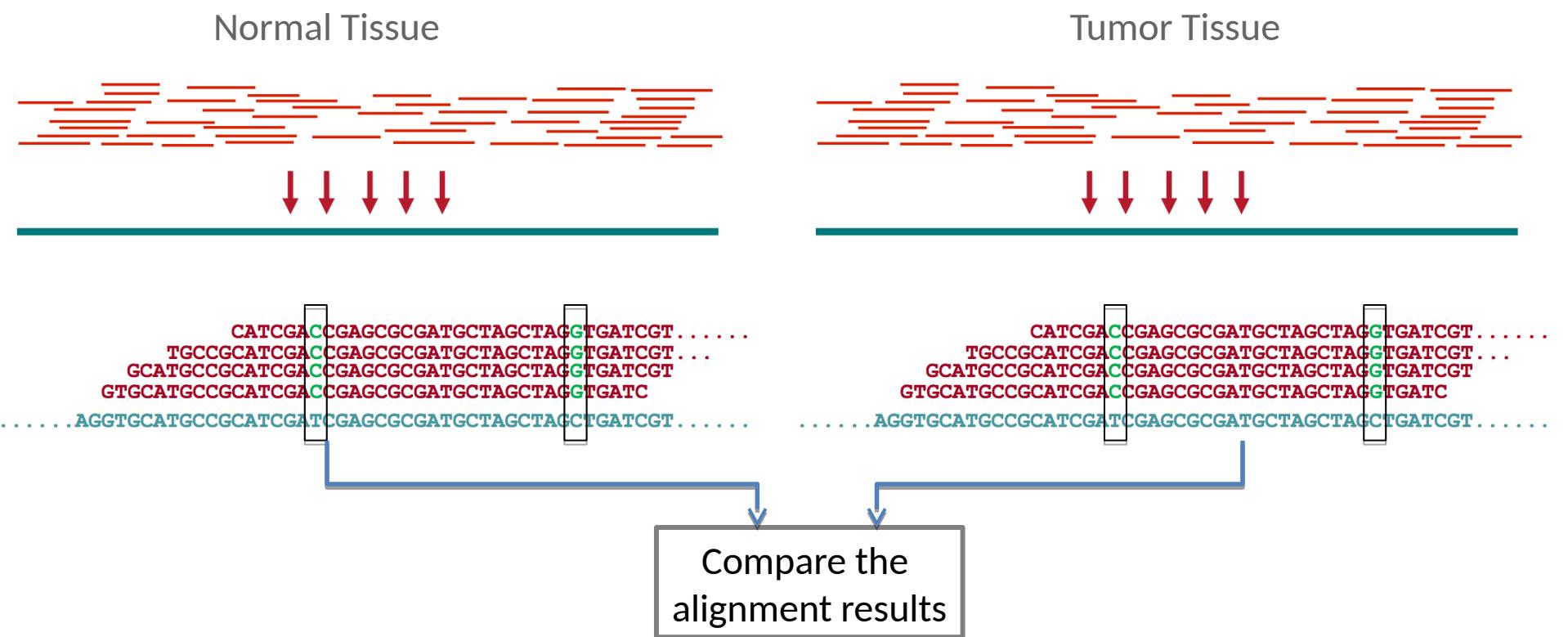
SNV Calling

Variant Types
Single Nucleotide Variants(SNVs)
Small Insertion / Deletion (indels)
Copy Number Variants (CNVs)
Structural Variants (SVs)
Novel Sequence



- A given human genome (germline) differs from the reference genome at millions of positions.
- A cancer genome differs from the healthy genome of its host by tens of thousands of positions at most, which is several orders of magnitude fewer differences than germline versus reference
- How do we distinguish germline mutations from somatic mutations?

Somatic SNV calling



- Most naïve: use a standard SNV caller on both datasets. If there is a mutation found in the tumor sample but not the normal, it is somatic!

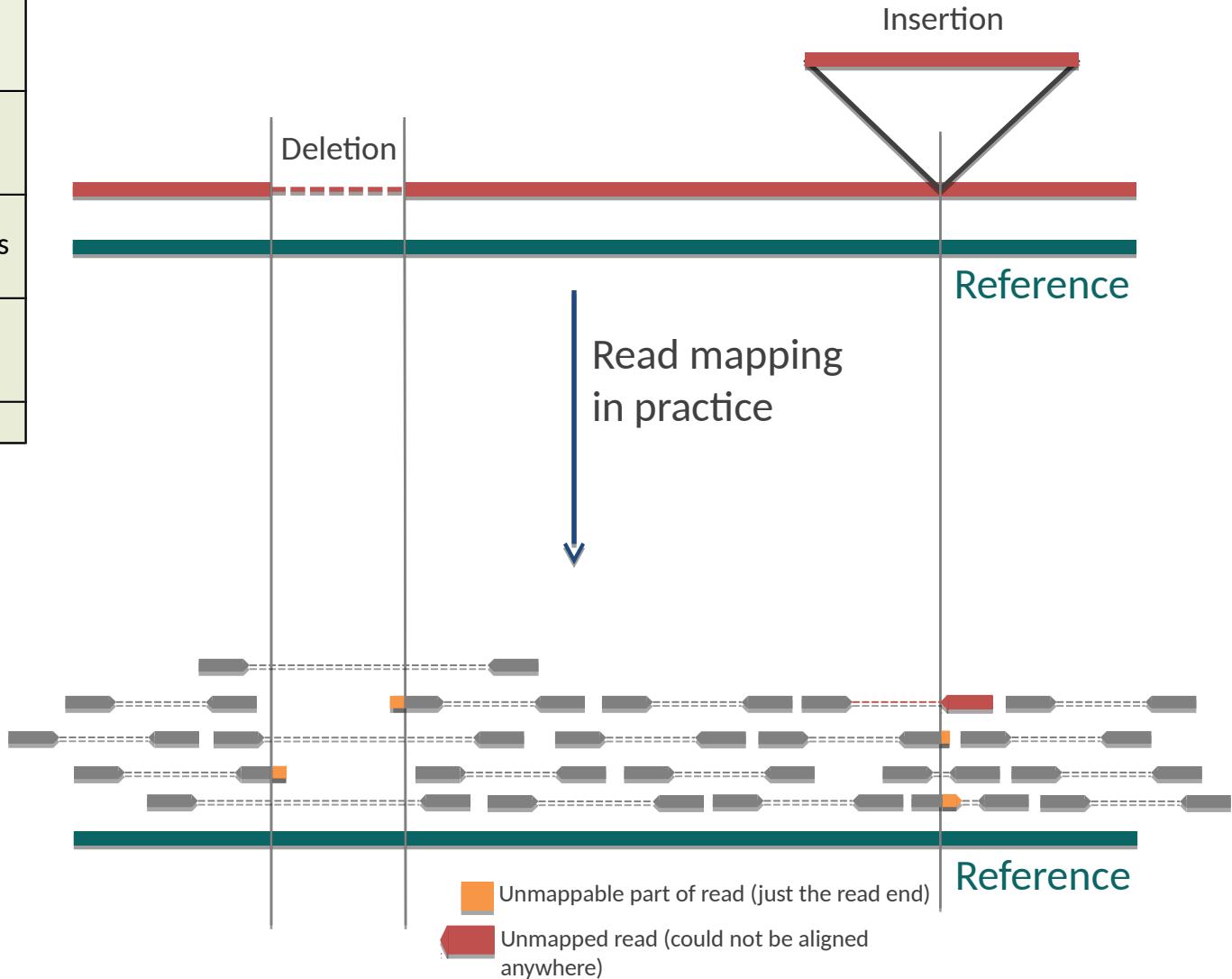
Short Indel Calling

Variant Types
Single Nucleotide Variants(SNVs)
Short Insertion / Deletion (indels)
Copy Number Variants (CNVs)
Structural Variants (SVs)
Novel Sequence



Short Indel Calling

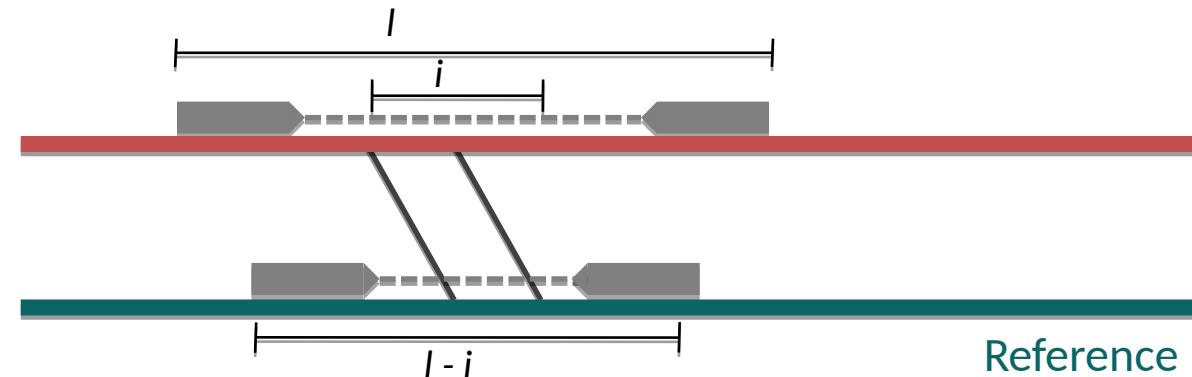
Variant Types
Single Nucleotide Variants(SNVs)
Short Insertion / Deletion (indels)
Copy Number Variants (CNVs)
Structural Variants (SVs)
Novel Sequence



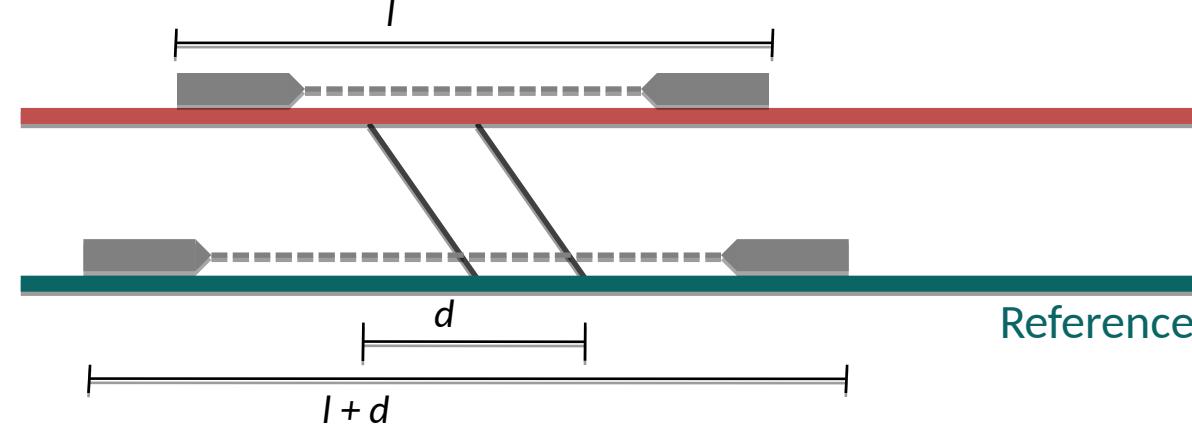
Short Indel Calling - Discordant Reads Pairs

Variant Types
Single Nucleotide Variants(SNVs)
Short Insertion / Deletion (indels)
Copy Number Variants (CNVs)
Structural Variants (SVs)
Novel Sequence

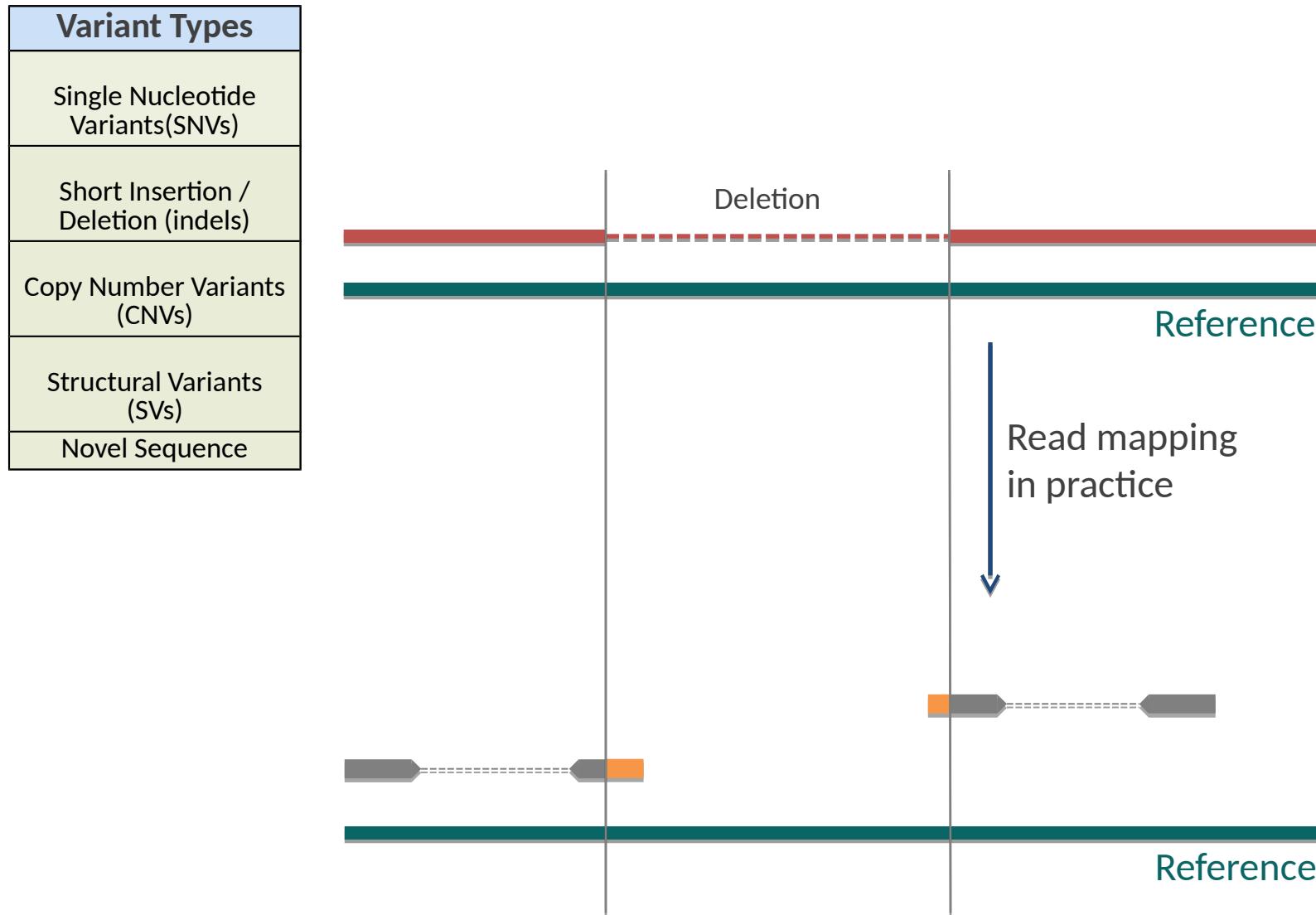
I) Insertion



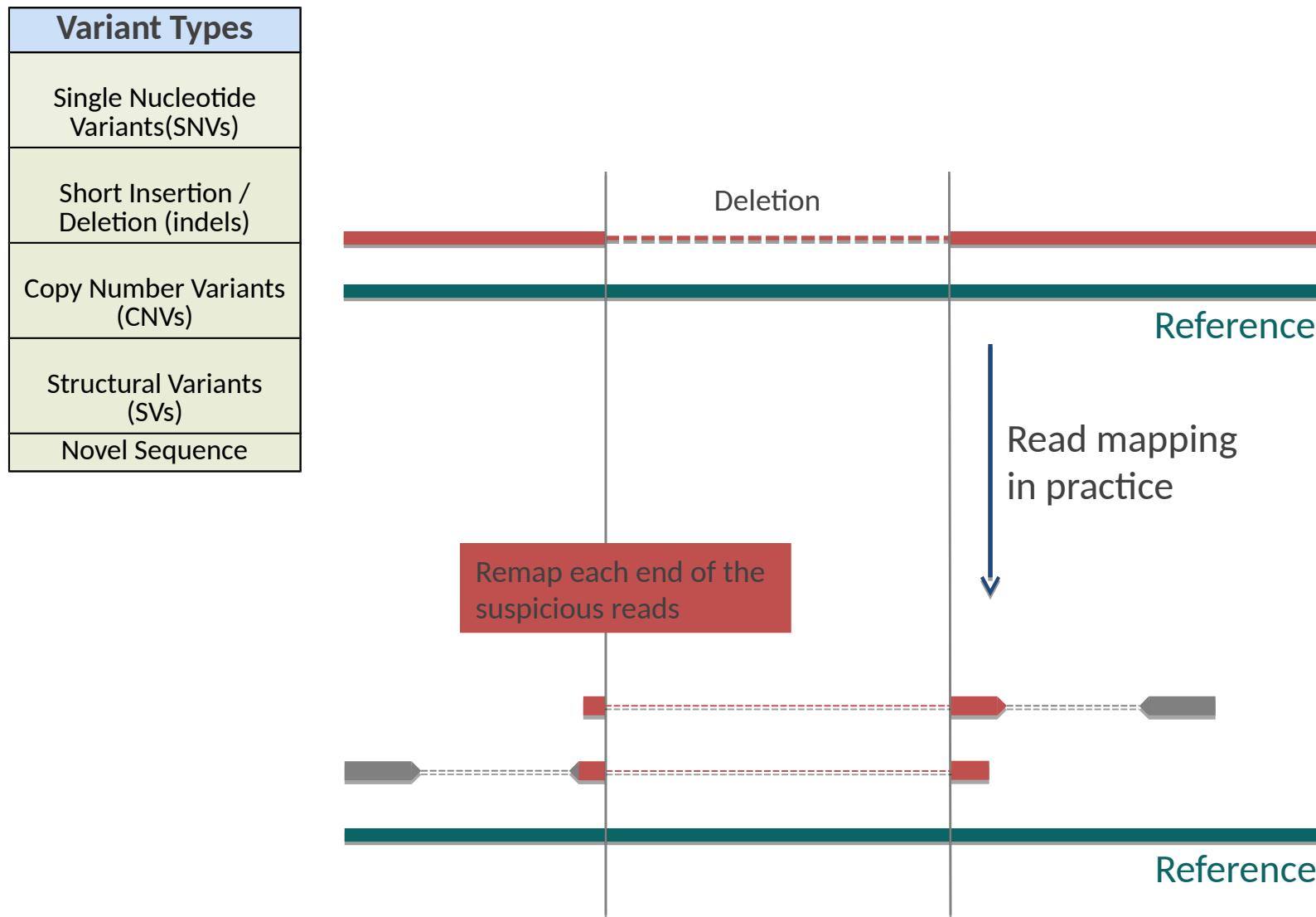
II) Deletion



Short Indel Calling - Split Read Mapping



Short Indel Calling - Split Read Mapping



Copy Number Variants

Variant Types
Single Nucleotide Variants(SNVs)
Short Insertion / Deletion (indels)
Copy Number Variants (CNVs)
Structural Variants (SVs)
Novel Sequence

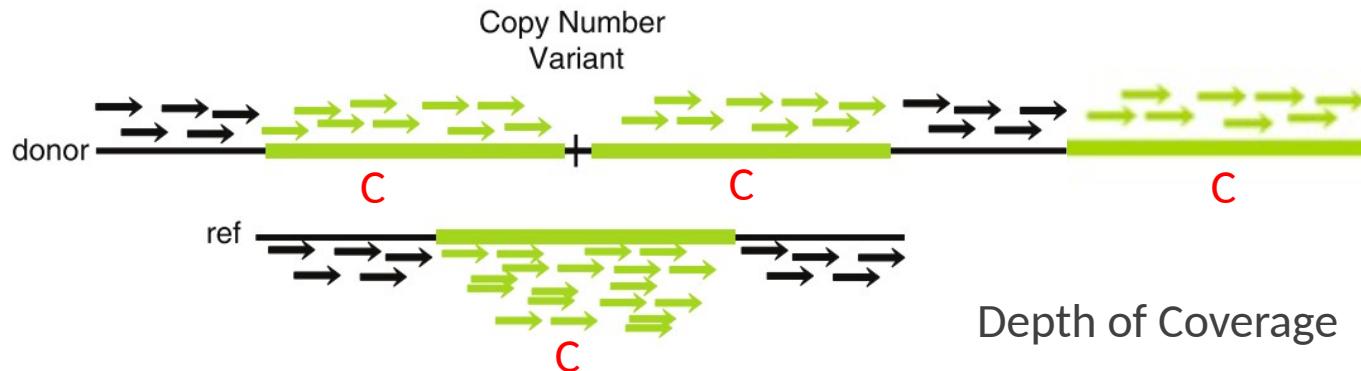


Ref:



Copy Number Variants

Variant Types
Single Nucleotide Variants(SNVs)
Short Insertion / Deletion (indels)
Copy Number Variants (CNVs)
Structural Variants (SVs)
Novel Sequence

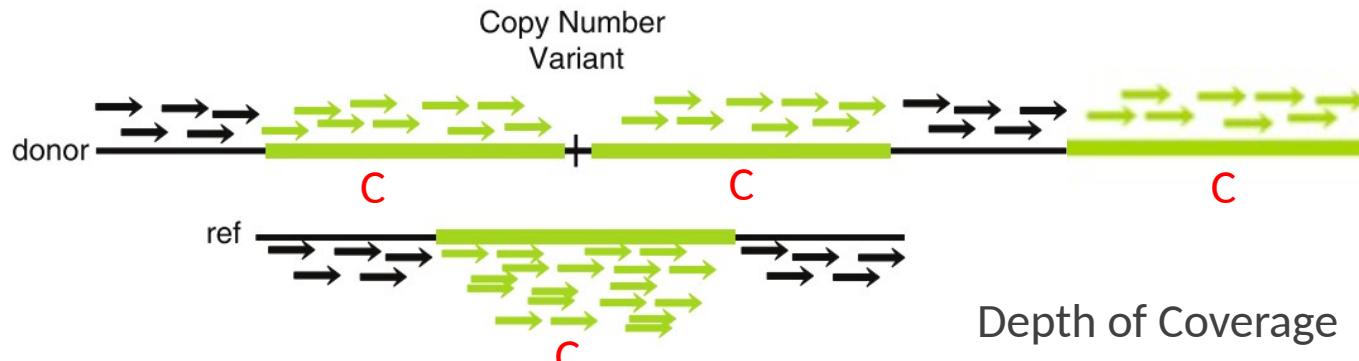


Modified from Dalca and Brudno. 2010. Genome variation discovery with high-throughput sequencing data. *Briefings in bioinformatics* 11, no. 1: 3-14



Copy Number Variants

Variant Types
Single Nucleotide Variants(SNVs)
Small Insertion / Deletion (indels)
Copy Number Variants (CNVs)
Structural Variants (SVs)
Novel Sequence



- Problems with DOC
 - Very sensitive to stochastic variance in coverage
 - Sensitive to bias coverage (e.g. GC content).
 - Impossible to determine non-reference locations of CNVs
- Graph methods using paired-end reads help overcome some of these problems



Variant Types

Variant Types
Single Nucleotide Variants(SNVs)
Short Insertion / Deletion (indels)
Copy Number Variants (CNVs)
Structural Variants (SVs)
Novel Sequence



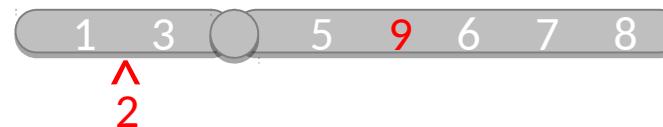
Structural Rearrangement



Translocation



Inversion



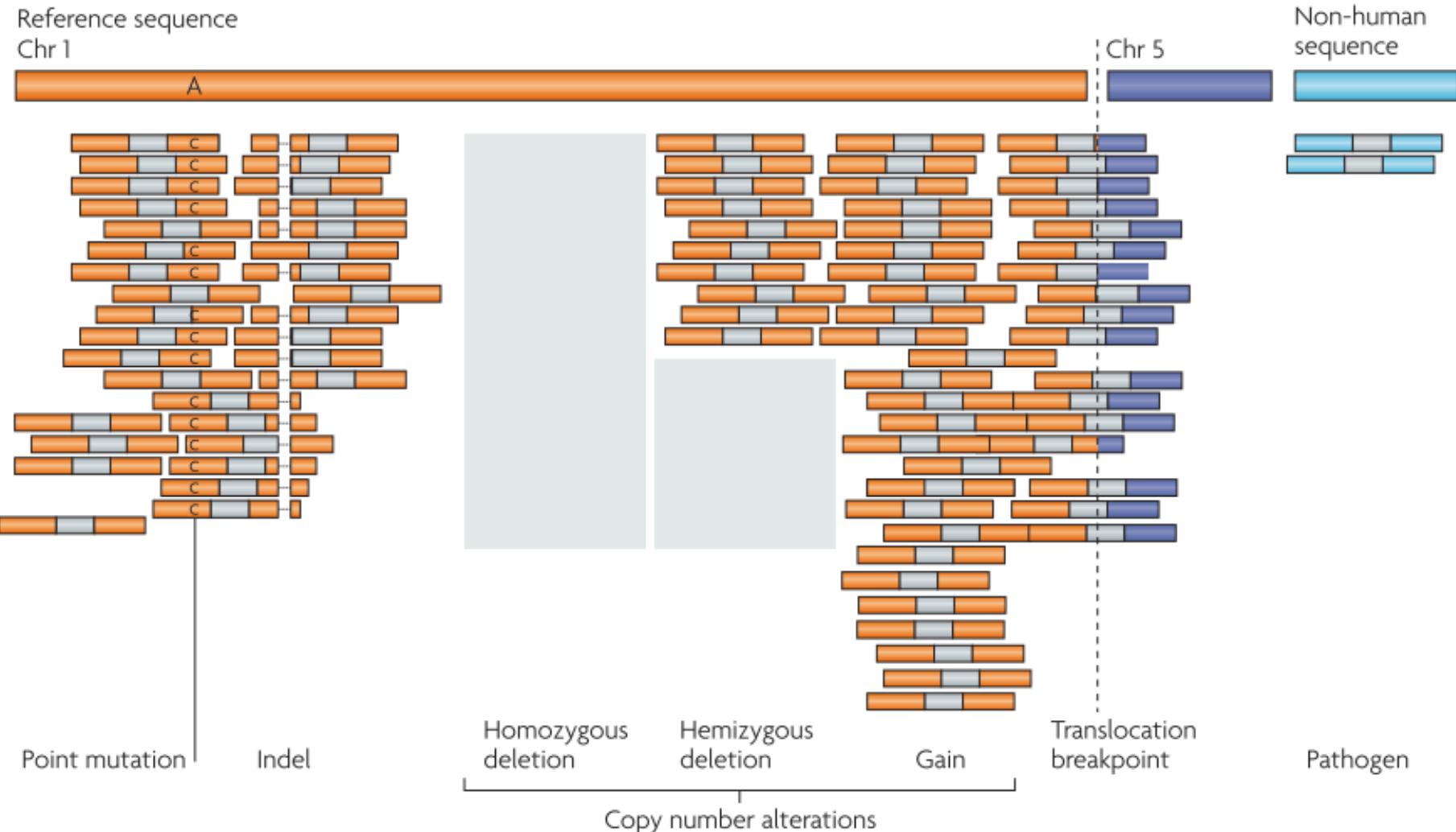
Large Insertion / Deletion



Ref:



Summary of Variant Types



Tracking the Evolution of Cancer

Genome evolution during progression to breast cancer

Daniel E. Newburger, Dorna Kashef-Haghghi, Ziming Weng, et al.

Genome Res. 2013 23: 1097-1108 originally published online April 8, 2013

Access the most recent version at doi:[10.1101/gr.151670.112](https://doi.org/10.1101/gr.151670.112)

Cell



The Life History of 21 Breast Cancers

Serena Nik-Zainal,^{1,19} Peter Van Loo,^{1,2,3,19} David C. Wedge,^{1,19} Ludmil B. Alexandrov,¹ Christopher D. Greenman,^{1,4,5} King Wai Lau,¹ Keiran Raine,¹ David Jones,¹ John Marshall,¹ Lucy A. Stebbings,¹ Andrew Menzies,¹ Sandra Martin,¹ Jonathan Hinton,¹ Andrew Menzies,¹ Lucy A. Stebbings,¹ Catherine Leroy,¹ Mingming Jia,¹ Richard Rance,¹ Laura J. Mudie,¹ Stephen J. Gamble,¹ Philip J. Stephens,¹ Stuart McLaren,¹ Patrick S. Tarpey,¹ Eli Papenmannuil,¹ Helen R. Davies,¹ Ignacio Varela,¹ David J. McBride,¹ Graham R. Bignell,¹ Kenric Leung,¹ Adam P. Butler,¹ Jon W. Teague,¹ Sandra Martin,¹ Göran Jönsson,⁶ Odette Mariani,⁷ Sandrine Boyault,⁸ Penelope Miron,⁹ Aquila Fatima,⁸ Anita Langerød,¹⁰ Samuel A.J.R. Aparicio,^{11,12} Andrew Tutt,⁶ Anieta M. Sieuwerts,¹³ Åke Borg,¹⁴ Gilles Thomas,⁸ Anne Vincent Salomon,⁷ Andrea L. Richardson,^{9,15} Anne-Lise Børresen-Dale,^{10,16} P. Andrew Futreal,¹ Michael R. Stratton,¹ Peter J. Campbell^{1,17,18,*} and Breast Cancer Working Group of the International Cancer Genome Consortium



Cell

Mutational Processes Molding the Genomes of 21 Breast Cancers

Serena Nik-Zainal,¹ Ludmil B. Alexandrov,¹ David C. Wedge,¹ Peter Van Loo,^{1,2,3} Christopher D. Greenman,^{1,4,5} Keiran Raine,¹ David Jones,¹ Jonathan Hinton,¹ John Marshall,¹ Lucy A. Stebbings,¹ Andrew Menzies,¹ Sandra Martin,¹ Kenric Leung,¹ Lina Chen,¹ Catherine Leroy,¹ Manasa Ramakrishna,¹ Richard Rance,¹ King Wai Lau,¹ Laura J. Mudie,¹ Ignacio Varela,¹ David J. McBride,¹ Graham R. Bignell,¹ Susanna L. Cooke,¹ Adam Shlien,¹ John Gamble,¹ Ian Whitmore,¹ Mark Maddison,¹ Patrick S. Tarpey,¹ Helen R. Davies,¹ Eli Papenmannuil,¹ Philip J. Stephens,¹ Stuart McLaren,¹ Adam P. Butler,¹ Jon W. Teague,¹ Göran Jönsson,¹³ Judy E. Garber,⁷ Daniel Silver,⁷ Penelope Miron,⁷ Aquila Fatima,⁷ Sandrine Boyault,⁸ Anita Langerød,⁹ Andrew Tutt,¹⁰ John W.M. Martens,¹¹ Samuel A.J.R. Aparicio,^{5,12} Åke Borg,¹³ Anne Vincent Salomon,⁷ Gilles Thomas,⁸ Anne-Lise Børresen-Dale,^{10,15} Andrea L. Richardson,¹⁶ Michael S. Neuberger,¹⁷ P. Andrew Futreal,¹ Peter J. Campbell,^{1,18,19} Michael R. Stratton^{1,*} and the Breast Cancer Working Group of the International Cancer Genome Consortium

REVIEWS

Evolution of the cancer genome

Lucy R. Yates¹ and Peter J. Campbell^{1,2}

NATURE REVIEWS | GENETICS

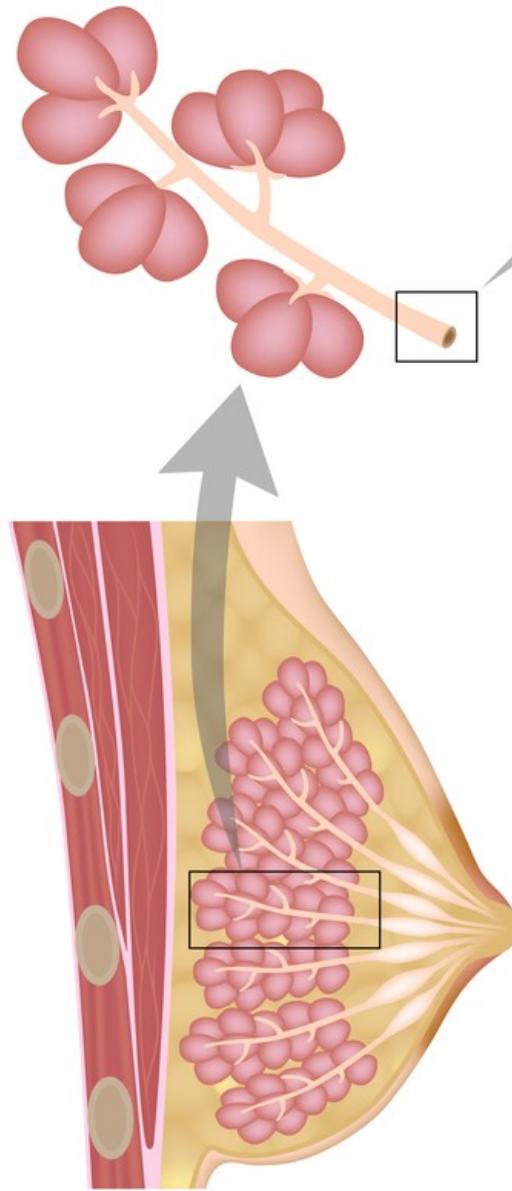
VOLUME 13 | NOVEMBER 2012

Clonal evolution in cancer

Mel Greaves¹ & Carlo C. Maley²

NATURE | VOL 481 | 19 JANUARY 2012

Mammary gland



Basement membrane

Normal duct cells

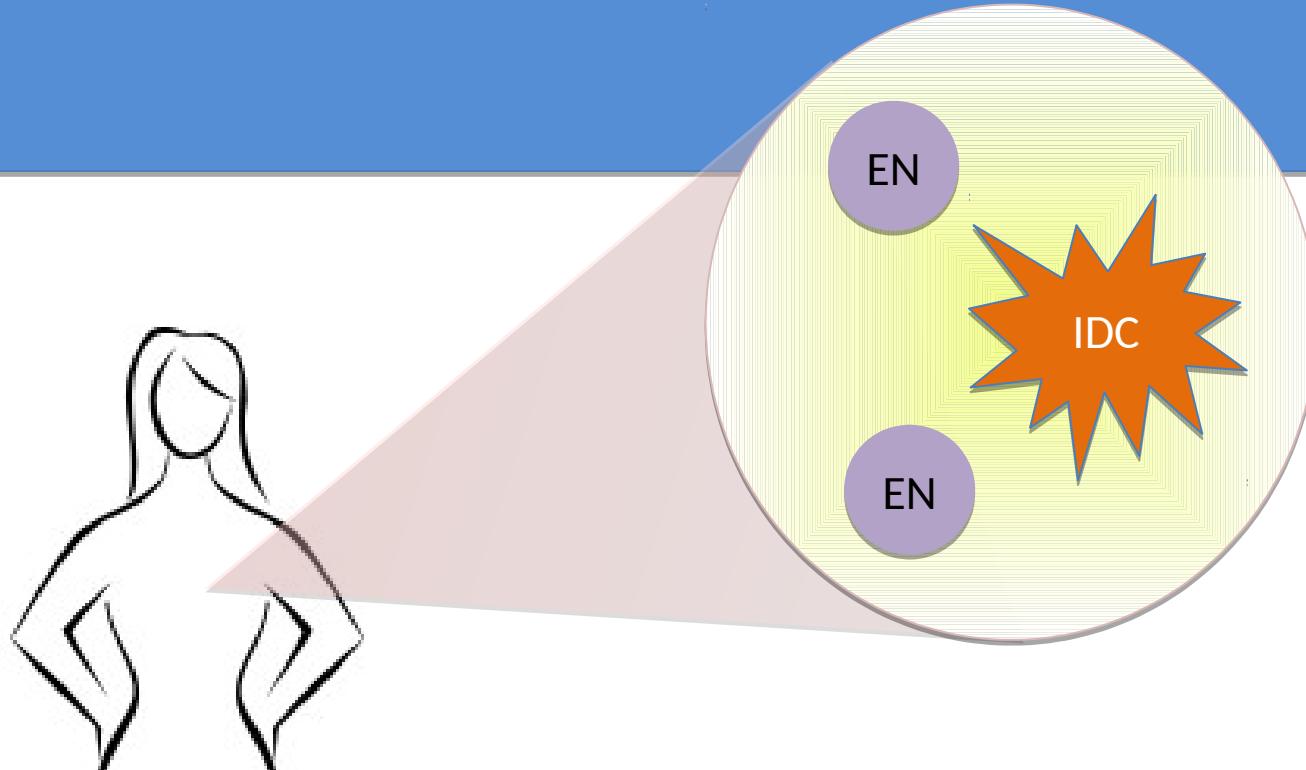
Normal duct

Cancer cells

Ductal carcinoma in situ (DCIS)

Invasive cancer cells

Invasive Ductal Carcinoma (IDC)



Normal

Early Neoplasia

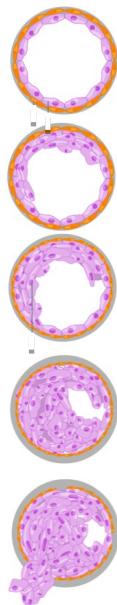
Ductal Carcinoma
In situ

Invasive Ductal
Carcinoma

Progression

Lesions

Patients



	P1	P2	P3	P4	P5	P6
Lymph	●			●	●	●
Normal	●	●	●			
EN	●	●	●	●	●	●
ENA	●				●	●
DCIS		●	●			●
IDC	●	●	●	●	●	●

Questions ^{>50x whole genome coverage per sample}

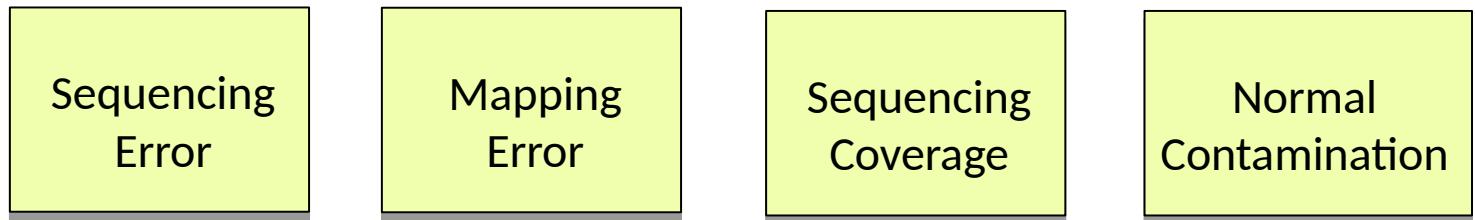
Types of Variants

- Somatic SNVs
- Aneuploidies

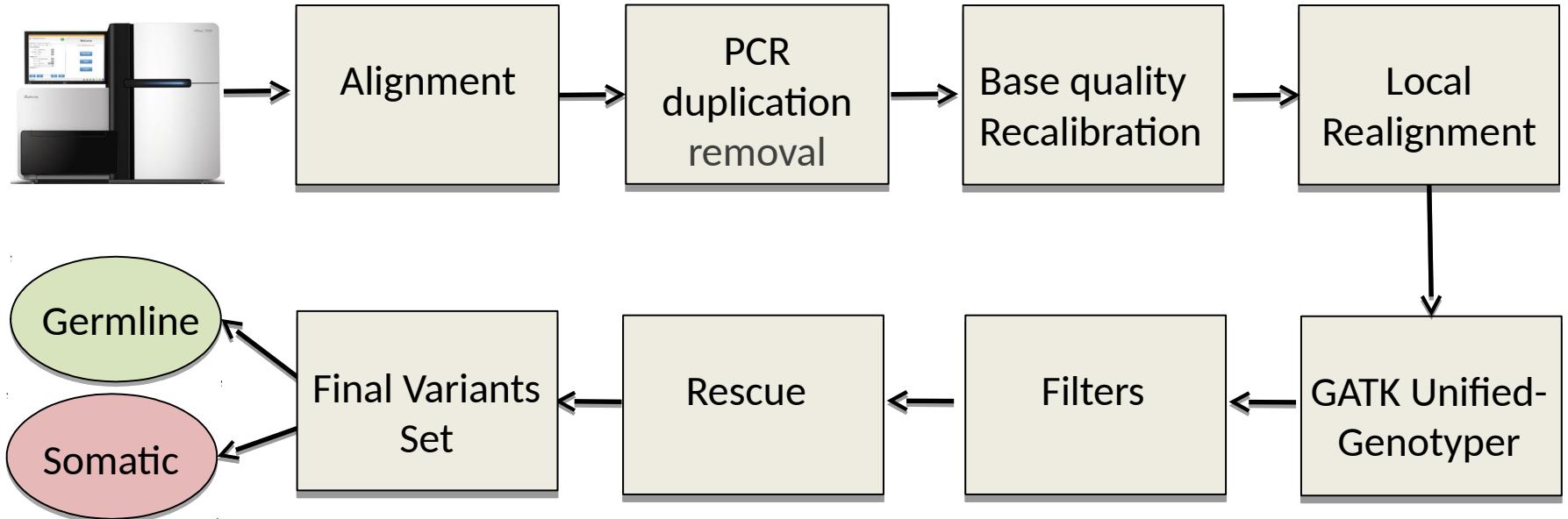
1. What is the **genetic** relationship of lesions to one another?
2. What are the **early** genomic events?

Variant calling process

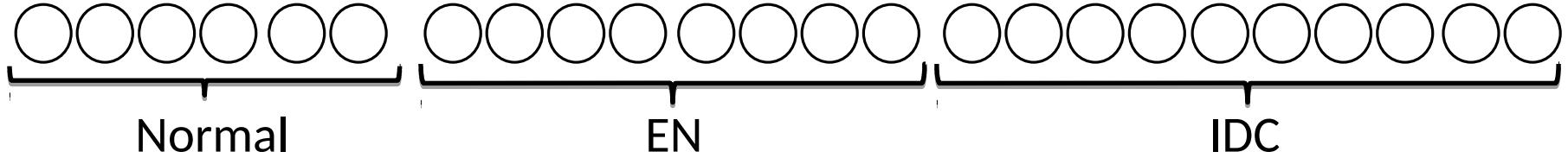
Challenges



Pipeline



Somatic SNVs as lineage markers



Somatic SNVs as lineage markers

EN
(300 SNVs)

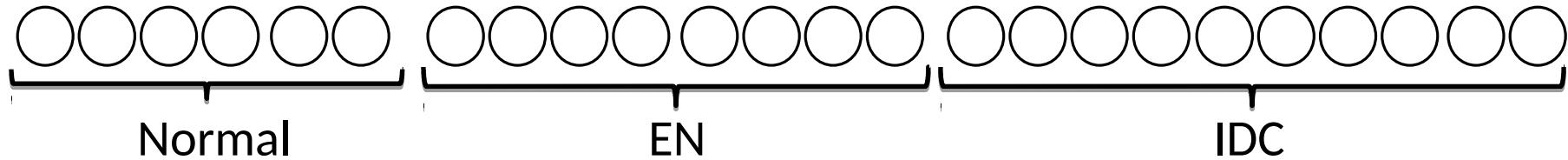
100
EN
only

200
shared

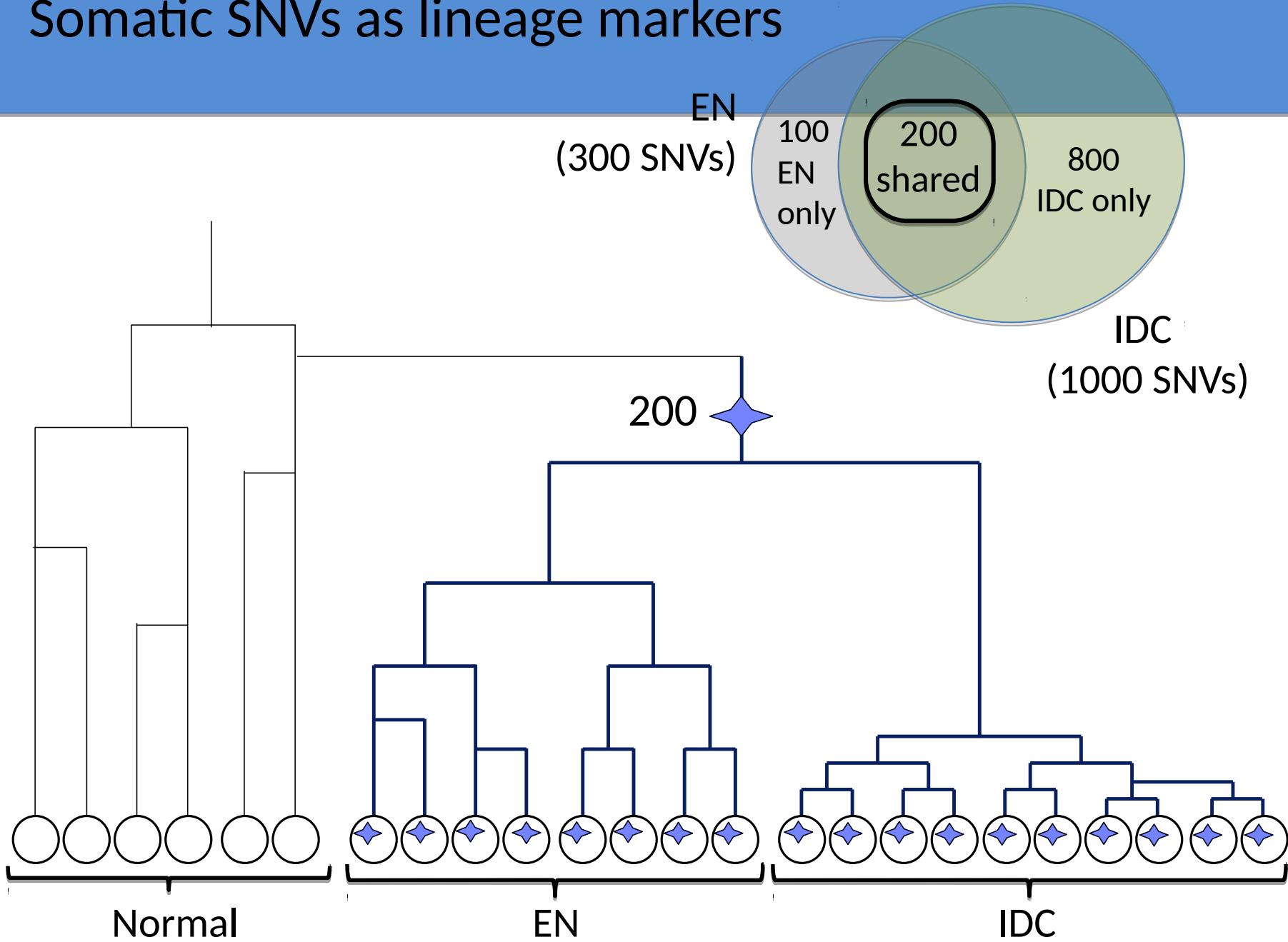
800
IDC only

IDC

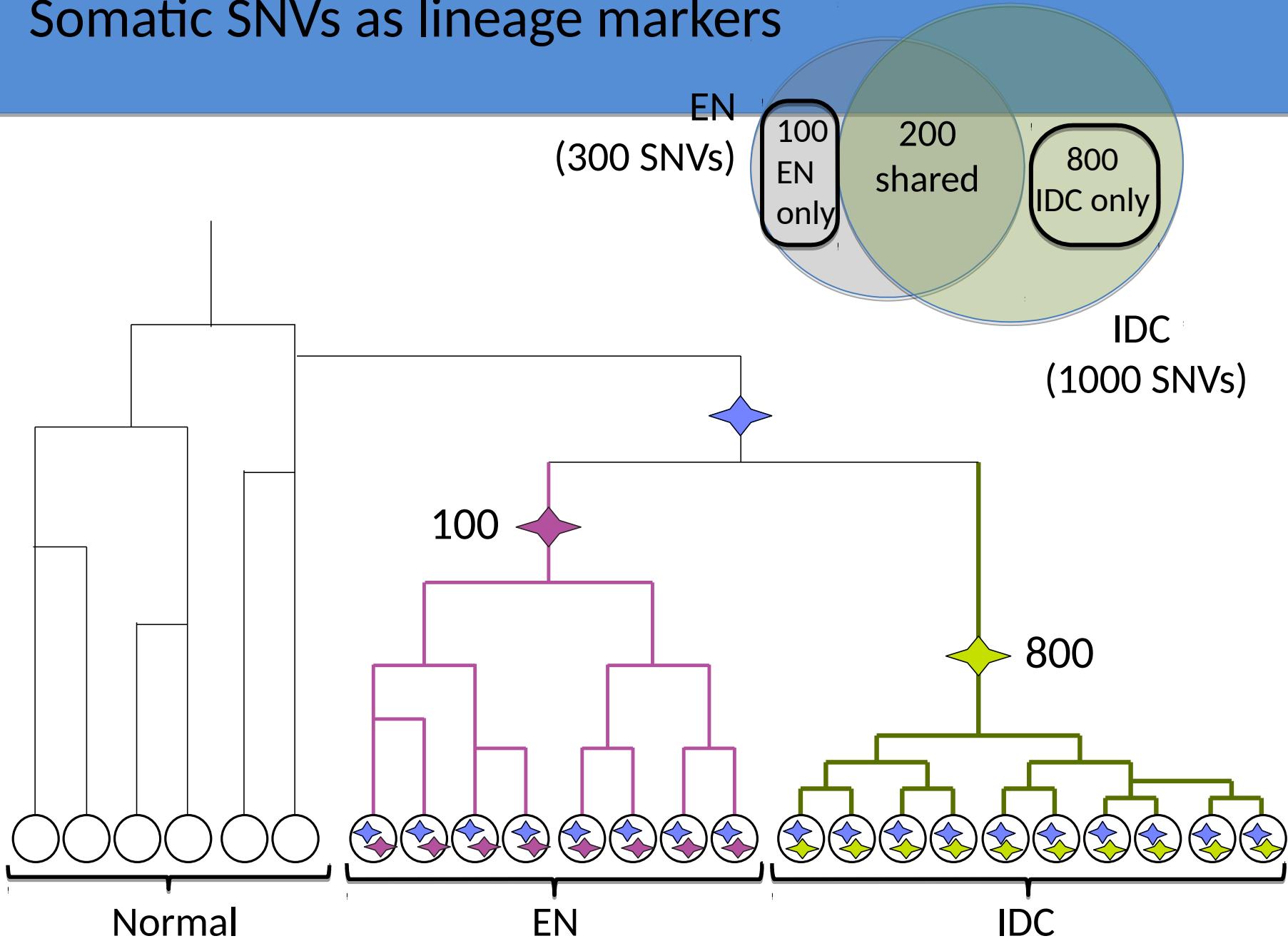
(1000 SNVs)



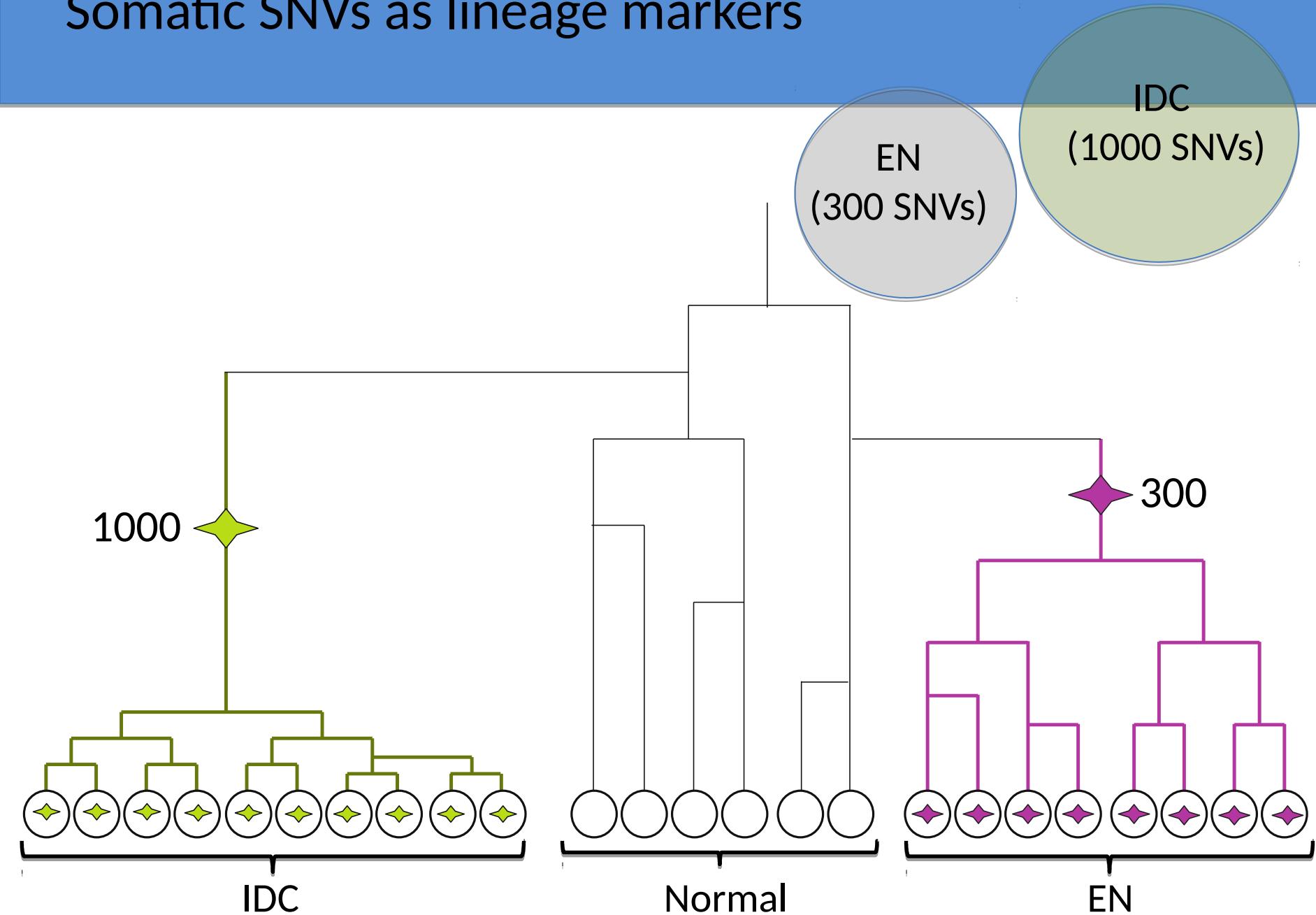
Somatic SNVs as lineage markers



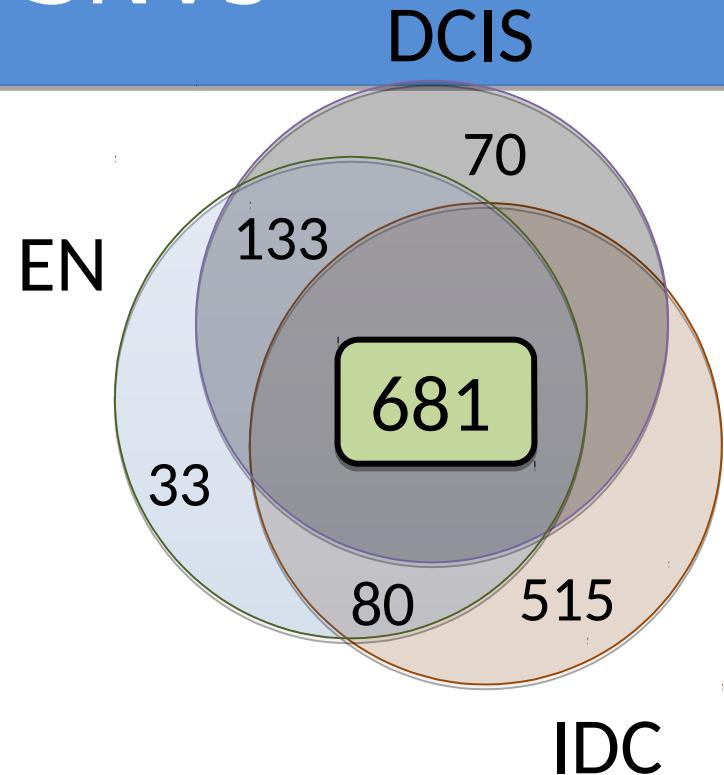
Somatic SNVs as lineage markers



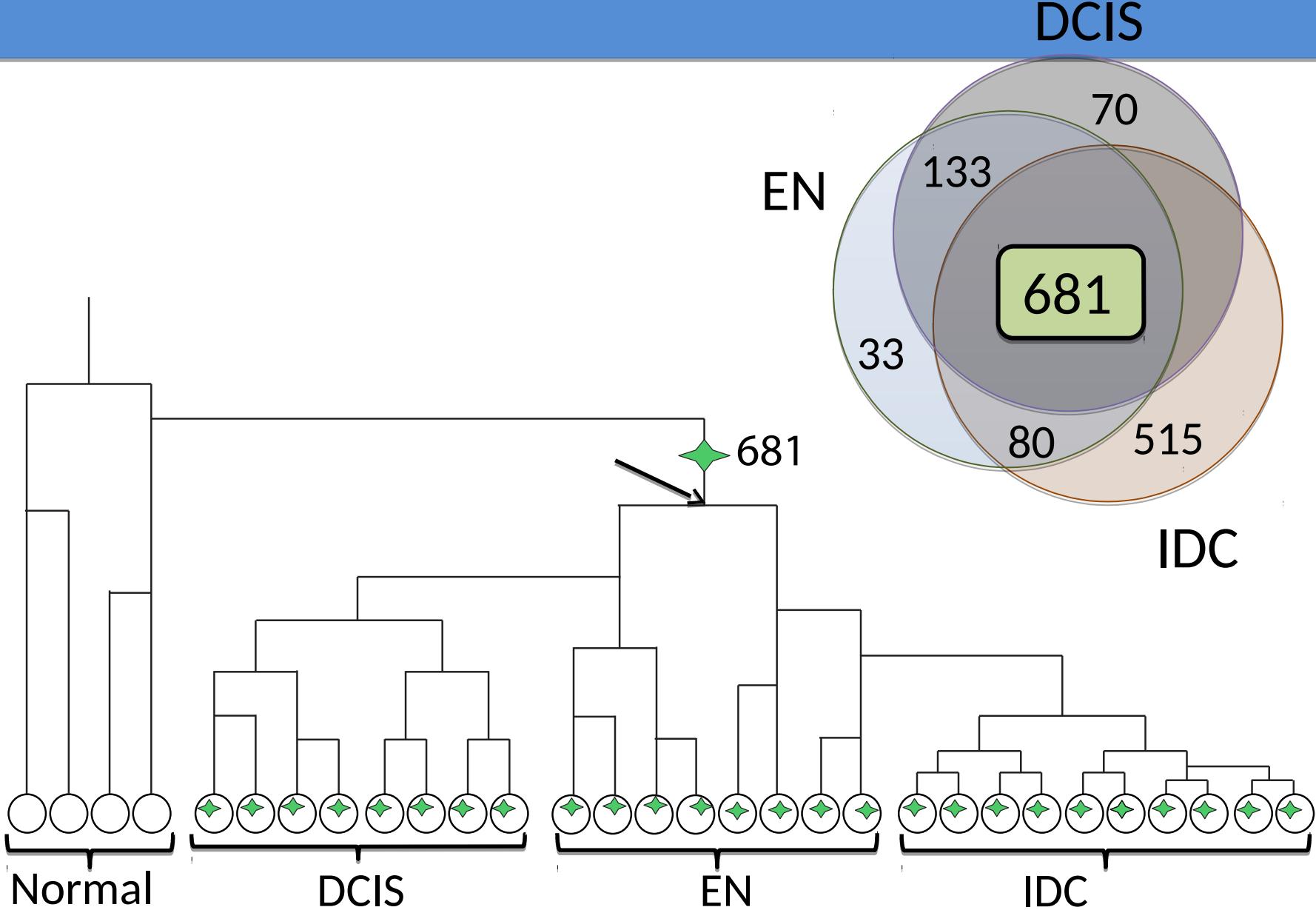
Somatic SNVs as lineage markers



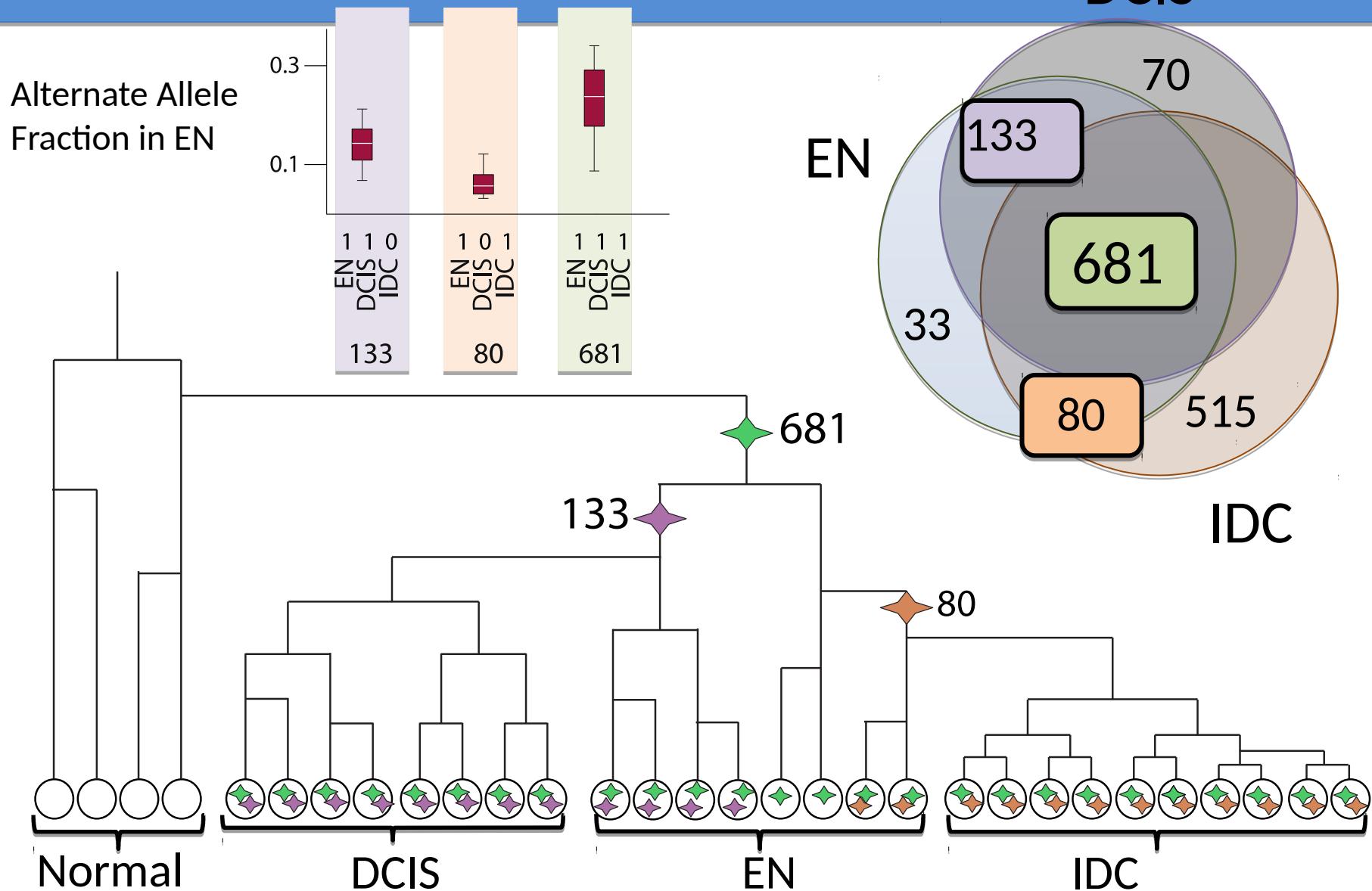
Patient 2 - SNVs



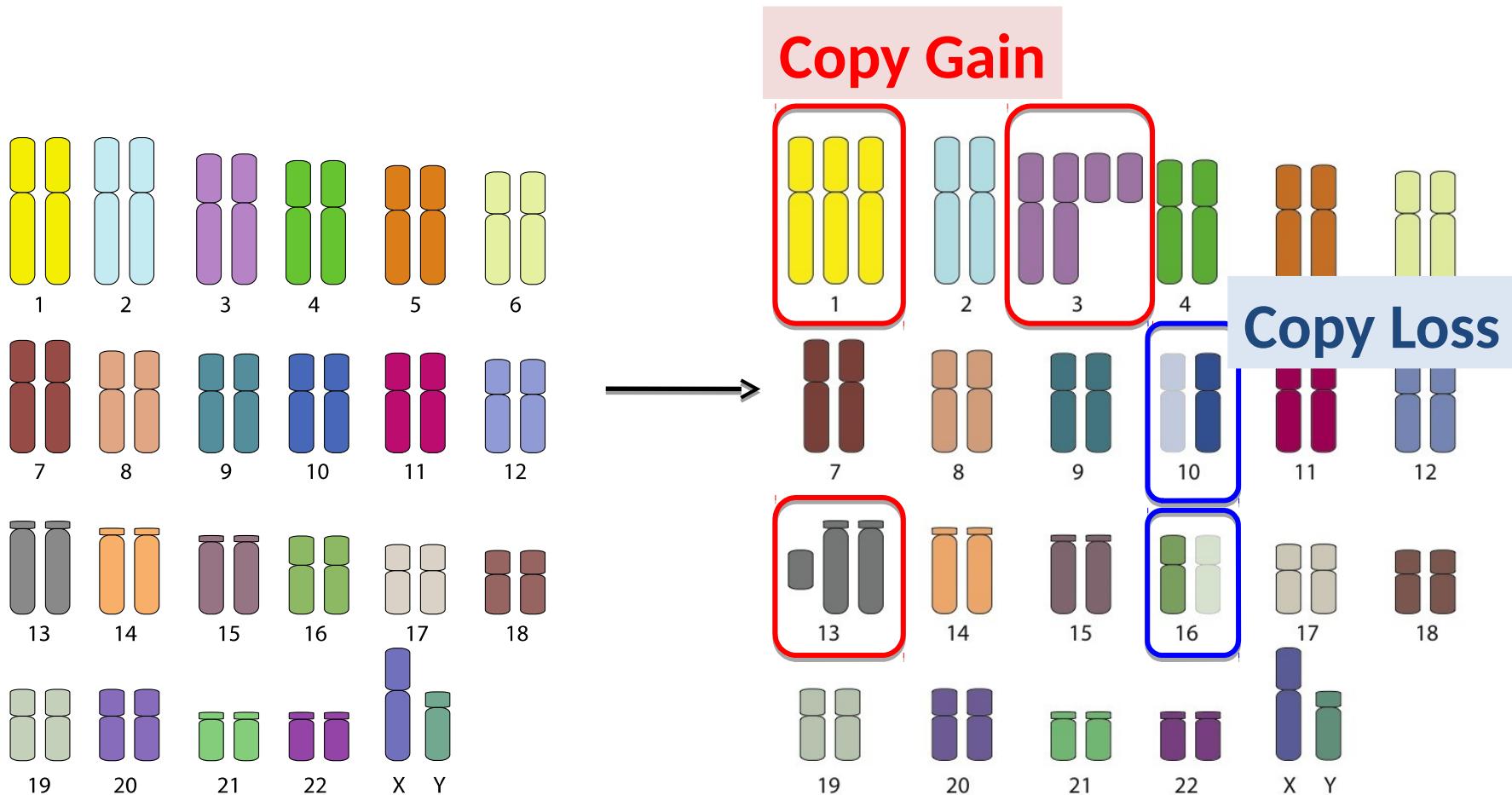
Patient 2 - SNVs



Patient 2 – SNVs

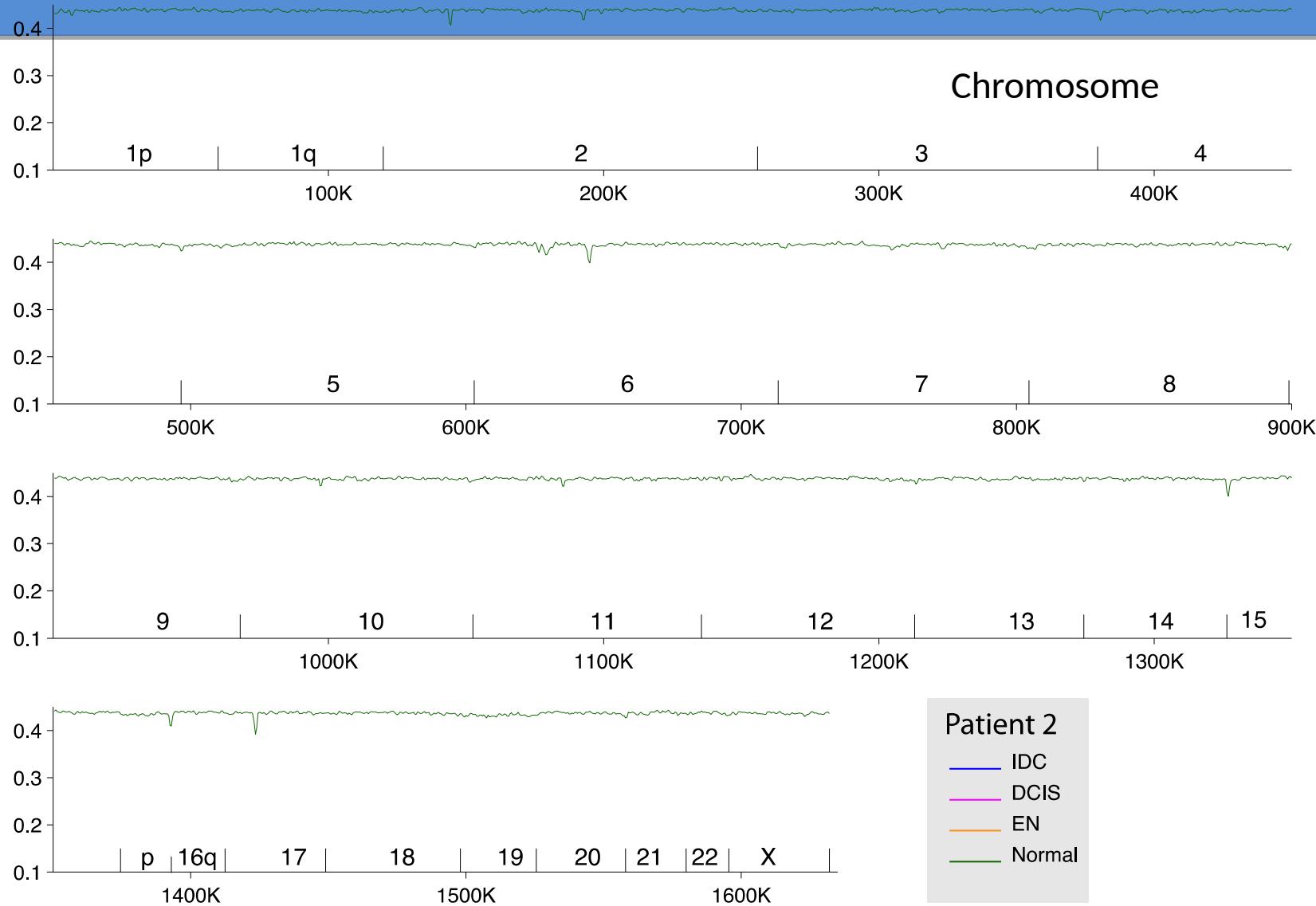


Aneuploidies



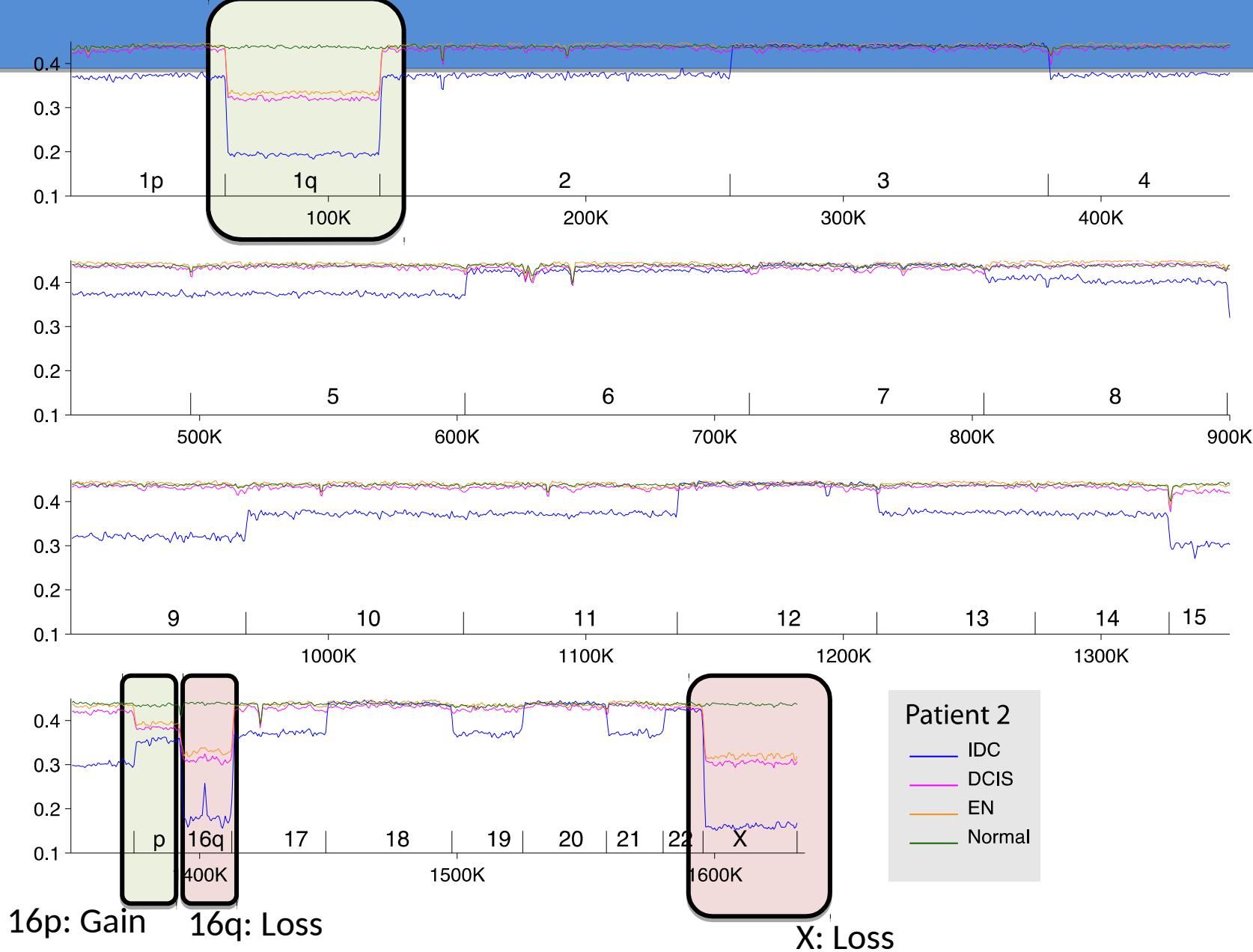
Patient 2 - Aneuploidies

Lesser allele fraction

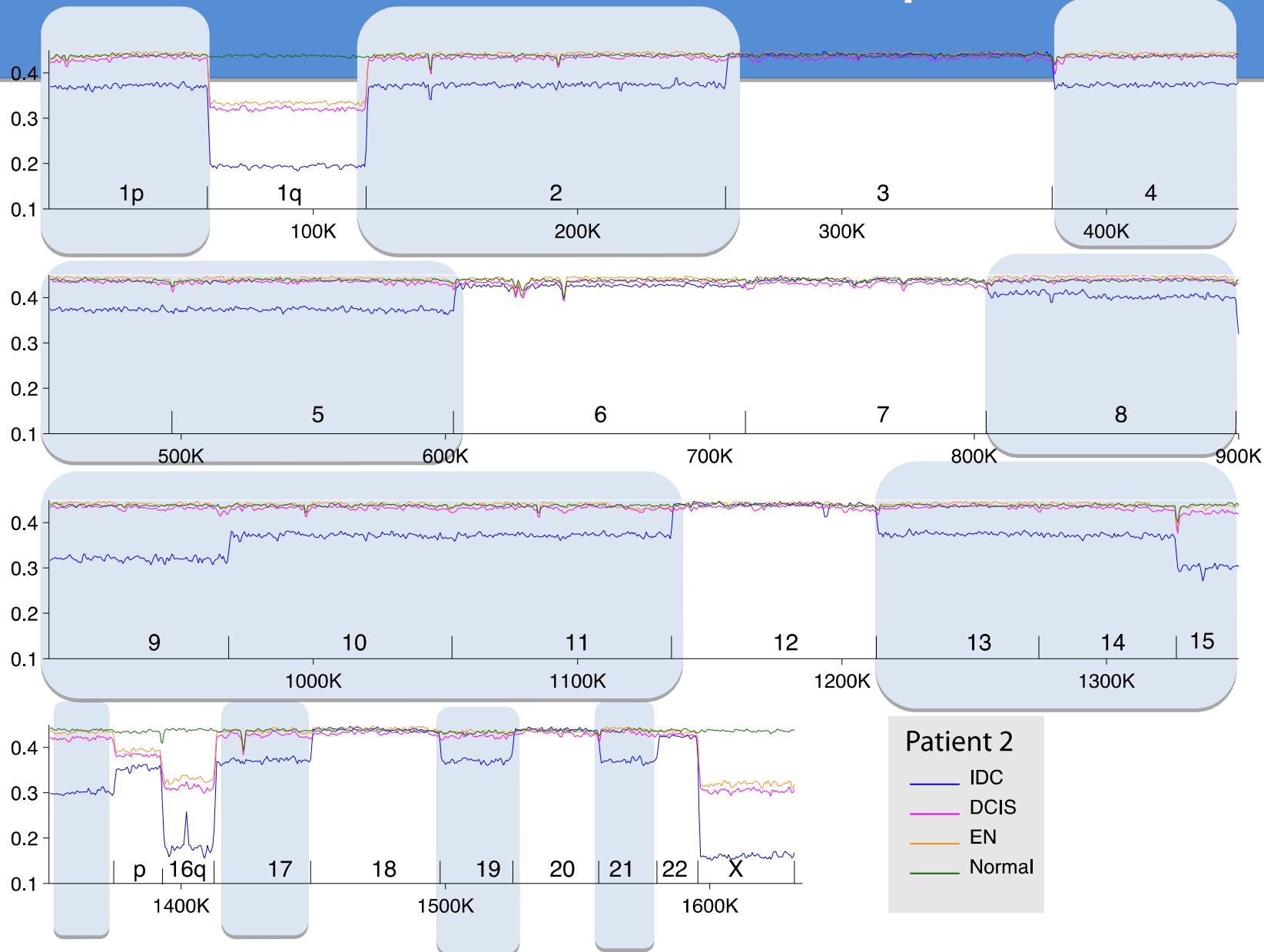


Plots are windows of 1000 SNPs overlapping by 500

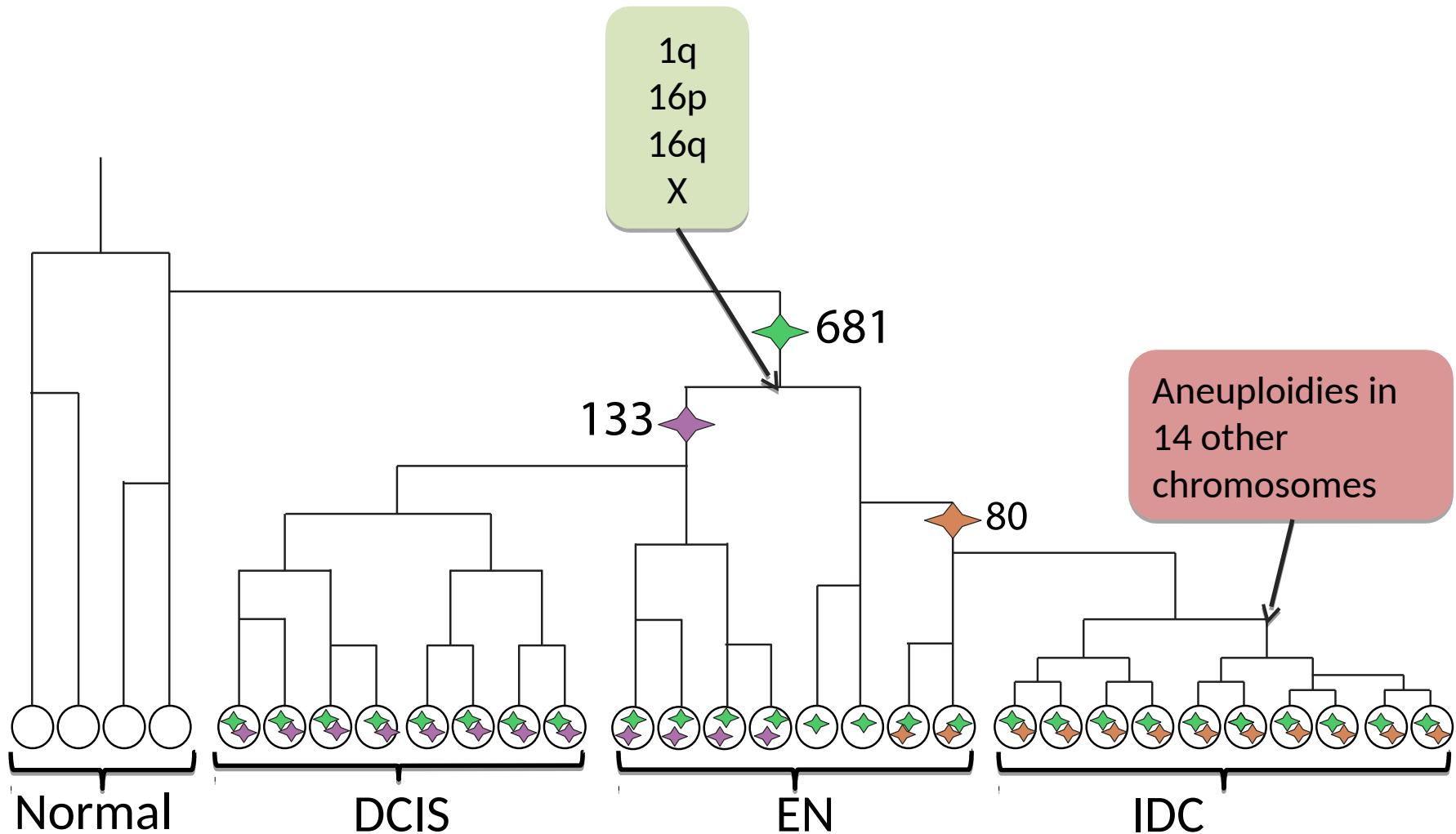
Patient 2 - Aneuploidies



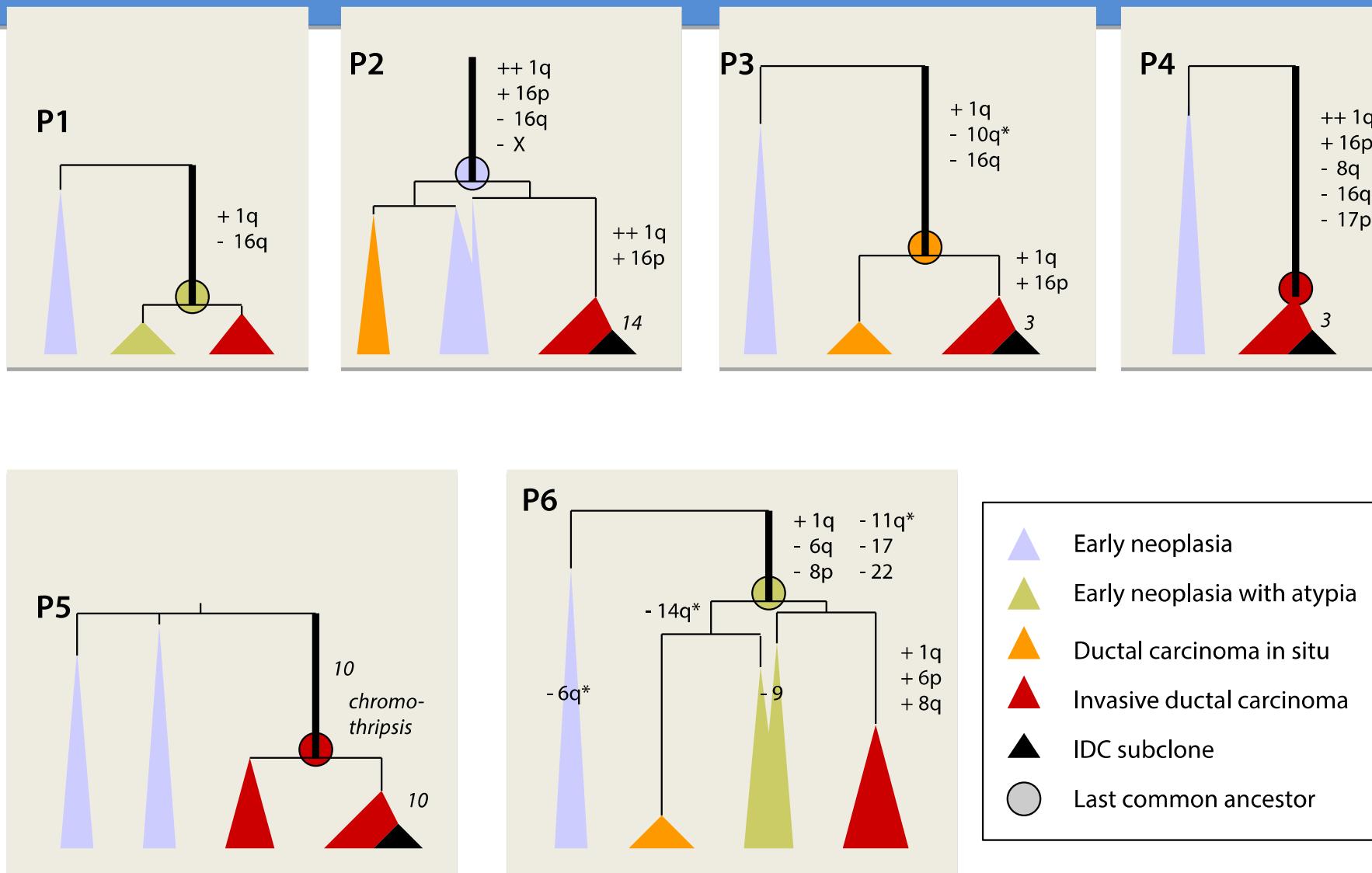
Patient 2 - Aneuploidies



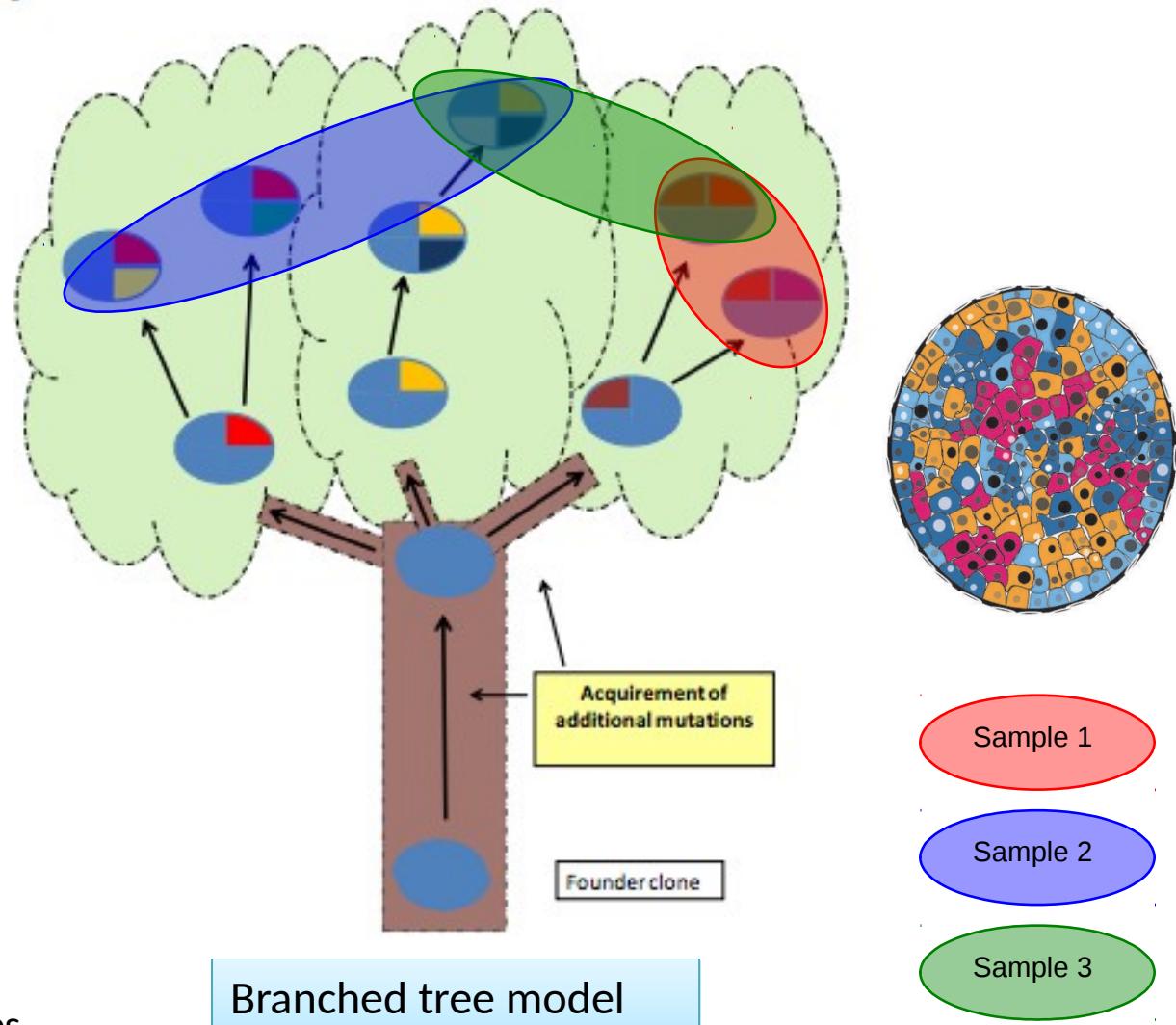
Patient 2 - Final Tree



Results



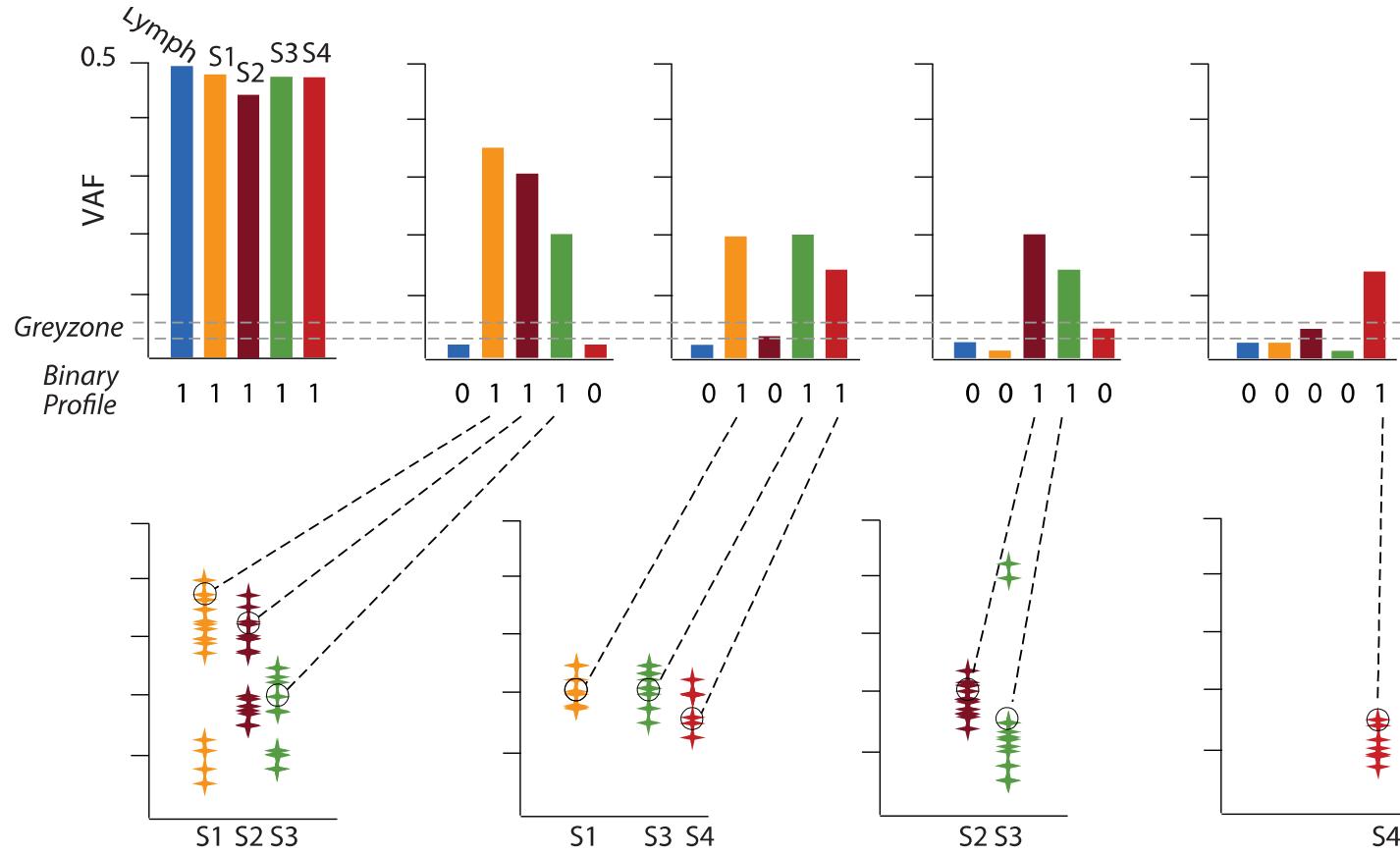
Automated Inference of Multi-Sample Cancer Phylogenies



Victoria Popic
Raheleh Salari

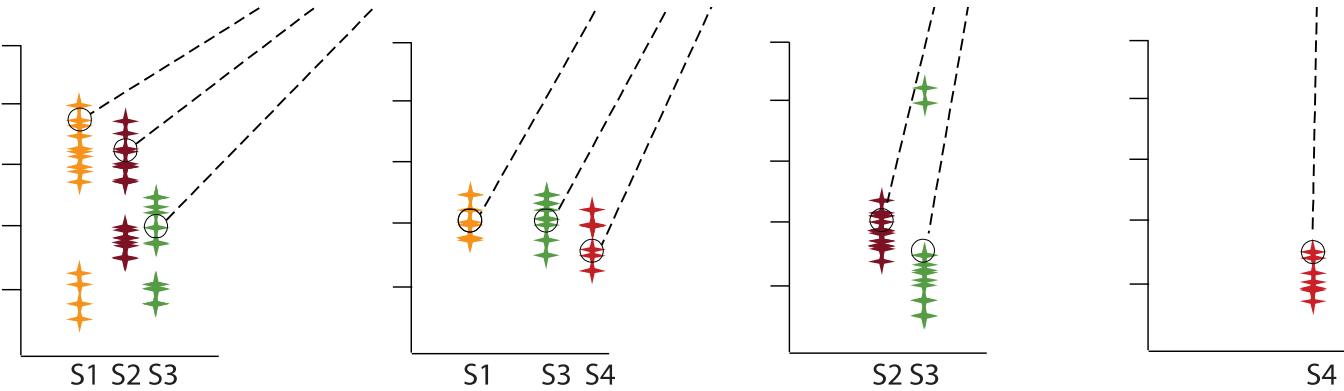
VAF profiles of SNVs across samples

SNV VAF
profiles
across
samples
S0 ... S4

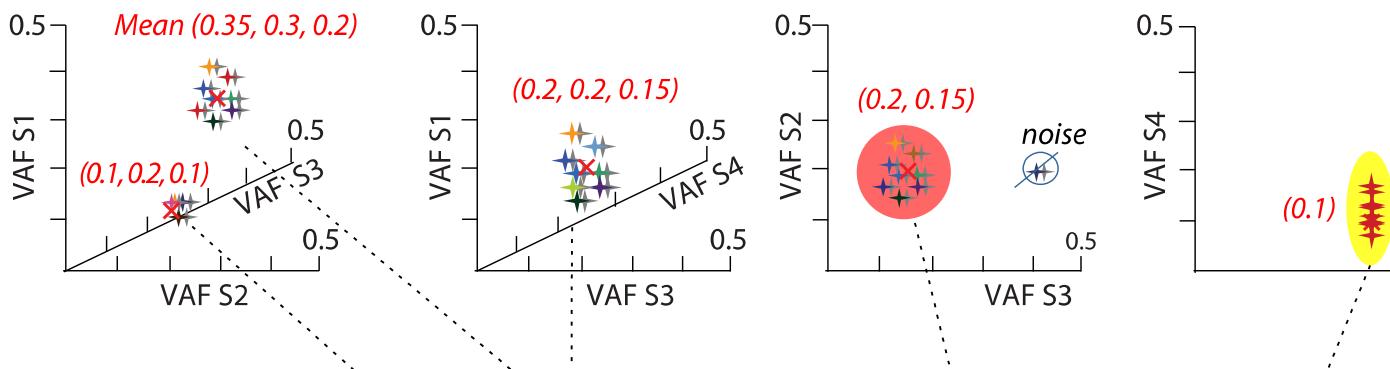


VAF profiles of SNVs – Clustering

Somatic
SNV
Groups



Subpopu-
lation
clusters

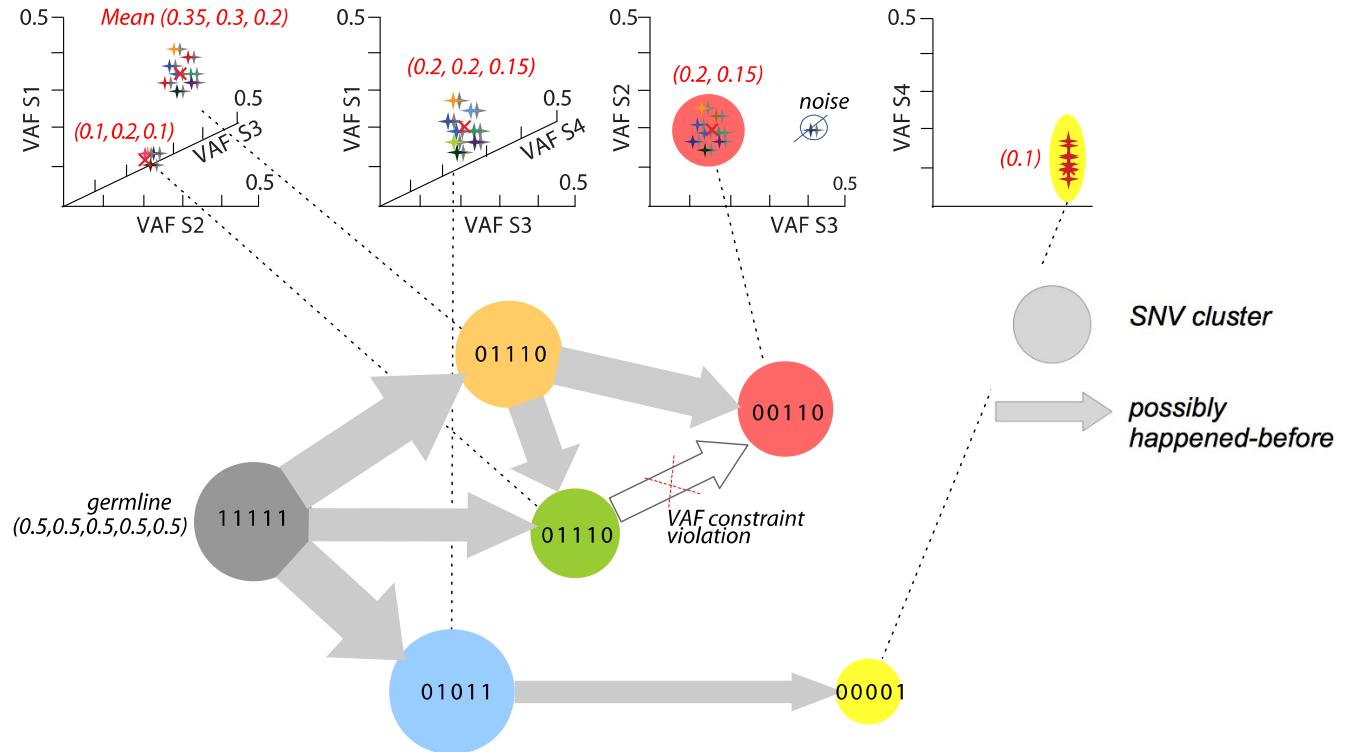


Cell-Lineage VAF Constraint

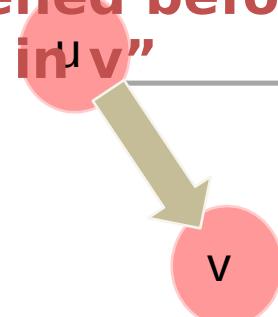
Subpopu-
lation
clusters



Evolutionary
constraint
network



"Possibly mutations in u happened before those in v"

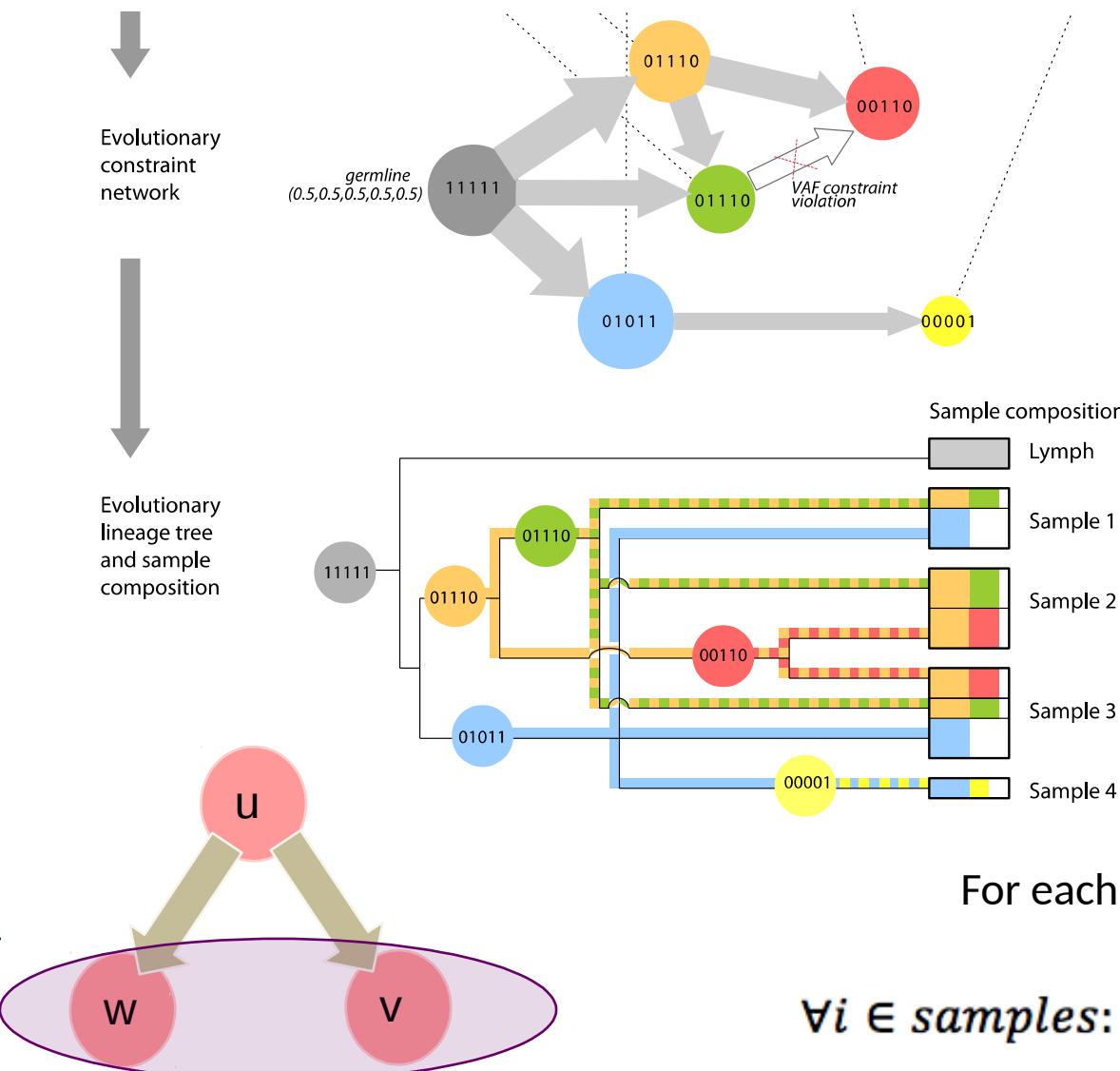


Edge $u \sqsubset v$: $\forall i \in samples:$

$$(1) u.\overline{VAF}[i] \geq v.\overline{VAF}[i] - \epsilon_{uv}$$

$$(2) v.\overline{VAF}[i] = 0 \text{ if } u.\overline{VAF}[i] = 0$$

Tree Construction



Find all spanning trees that satisfy VAF constraints

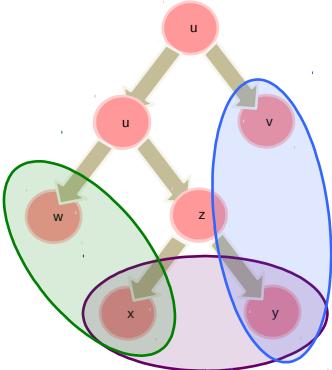
(extension of Gabow&Myers spanning tree search algorithm)

Rank trees according to their agreement with VAFs

For each node u and its children C :

$$\forall i \in samples: \sum_{v \in C} v.\overline{VAF}[i] \leq u.\overline{VAF}[i] + \epsilon$$

Simulation Results



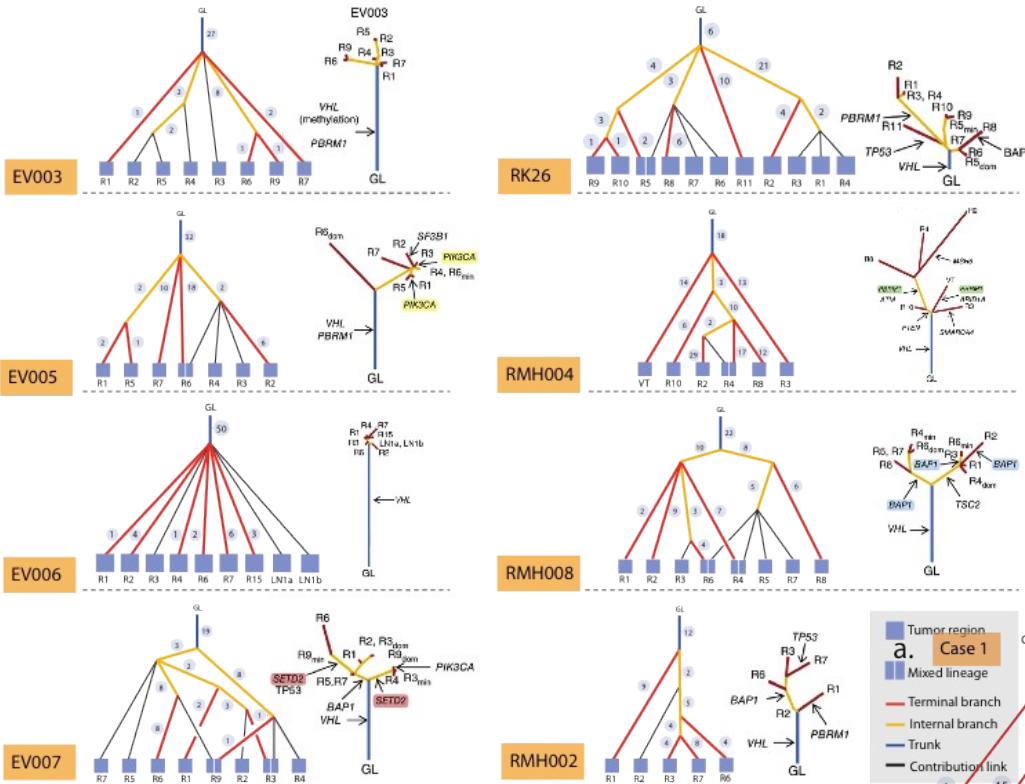
# Samples	VAF Stdev 0.03				VAF Stdev 0.1			
	Pred	Branch	Shared Edges	# Trees	Pred	Branch	Shared Edges	# Trees
5	0.95	0.91	0.81	85	0.95	0.85	0.80	43
6	0.91	0.93	0.82	77	0.91	0.91	0.79	43
7	0.89	0.91	0.83	67	0.88	0.89	0.83	38
8	0.89	0.96	0.84	64	0.93	0.95	0.87	43
9	0.88	0.95	0.81	60	0.87	0.97	0.82	34
10	0.89	0.94	0.82	50	0.87	0.95	0.85	24
11	0.89	0.94	0.85	46	0.86	0.97	0.77	26
12	0.86	0.97	0.81	39	0.85	0.97	0.80	20
13	0.91	0.97	0.85	50	0.91	0.87	0.84	26
14	0.88	0.94	0.83	42	0.89	0.90	0.89	15
15	0.88	0.95	0.85	38	0.82	0.91	0.79	16

Pred: pairs of nodes ordered correctly

Branch: pairs of nodes correctly assigned to separate branches

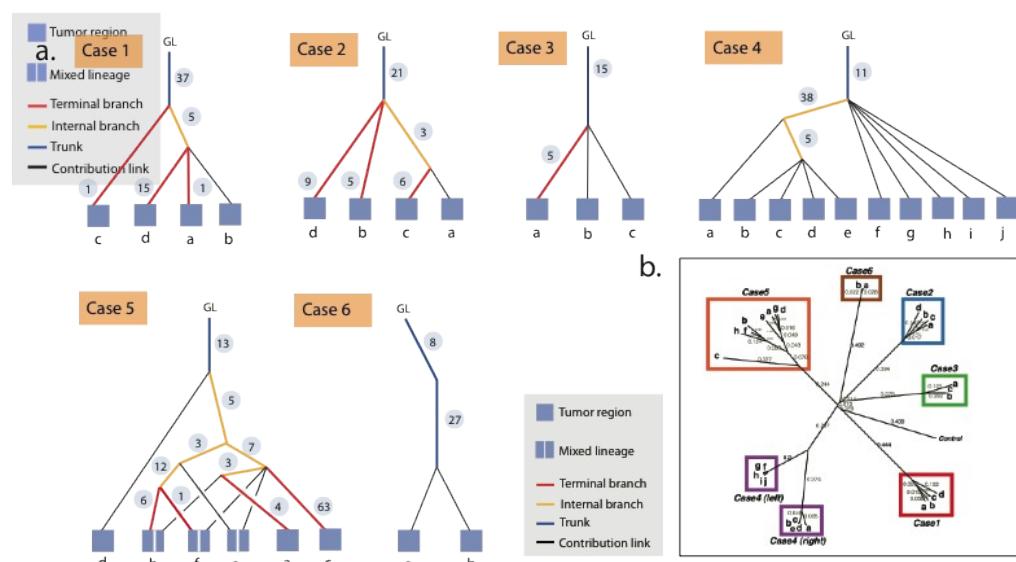
Shared edges: edges shared between true and reconstructed trees

Reconstruction of Lineage Trees in Recent Literature

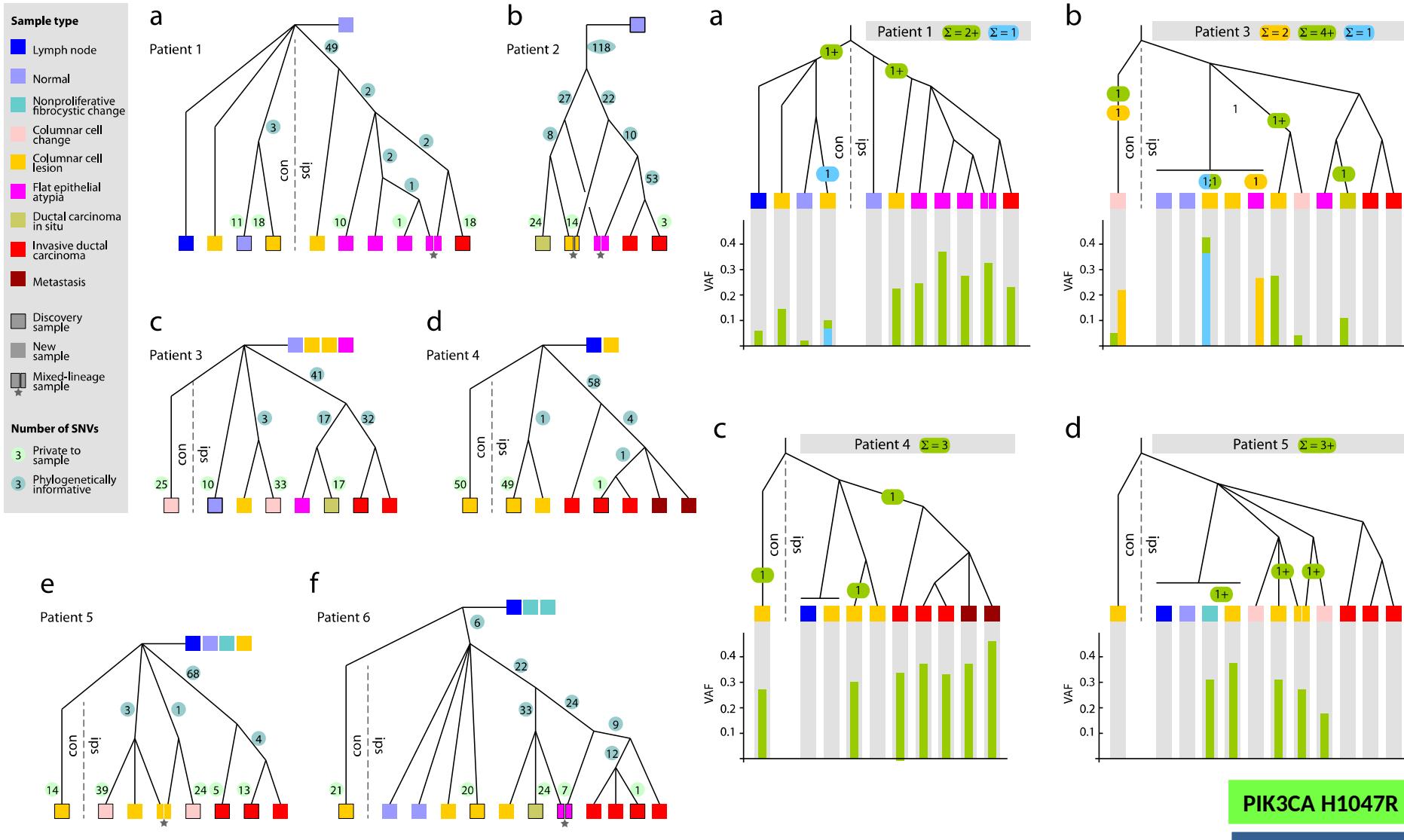


ccRCC Study of Renal Carcinoma by Gerlinger *et. al* (2014)

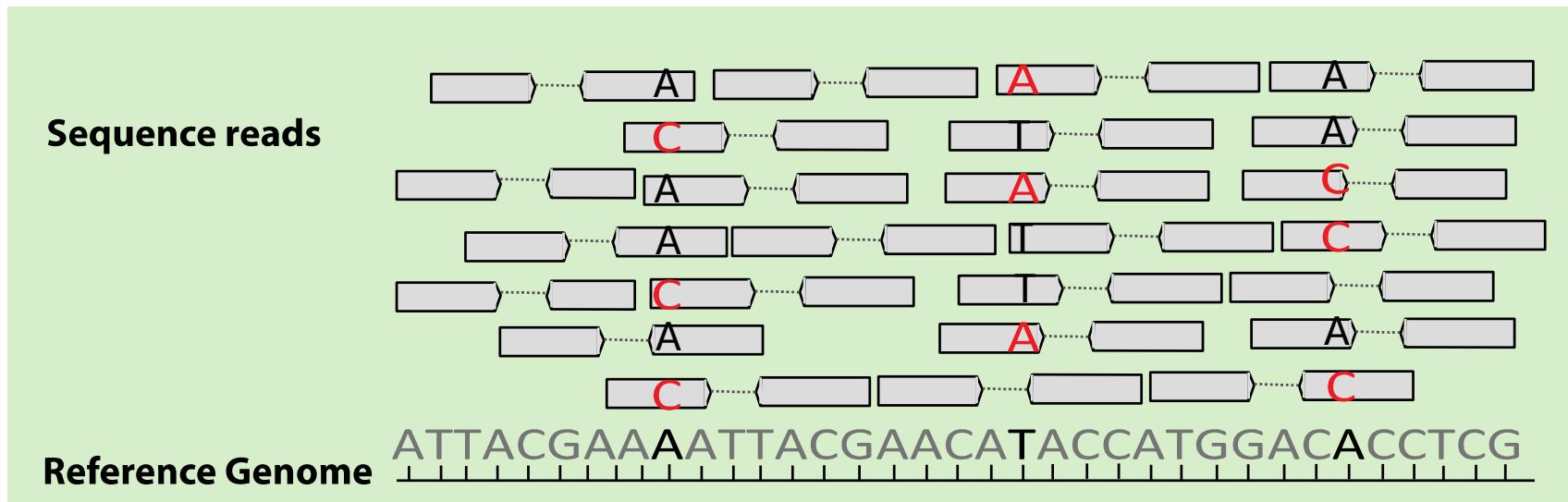
HGSC Study of Ovarian Cancer
Bashashati *et. al* (2013)



Expanded Breast Cancer Lineage Trees

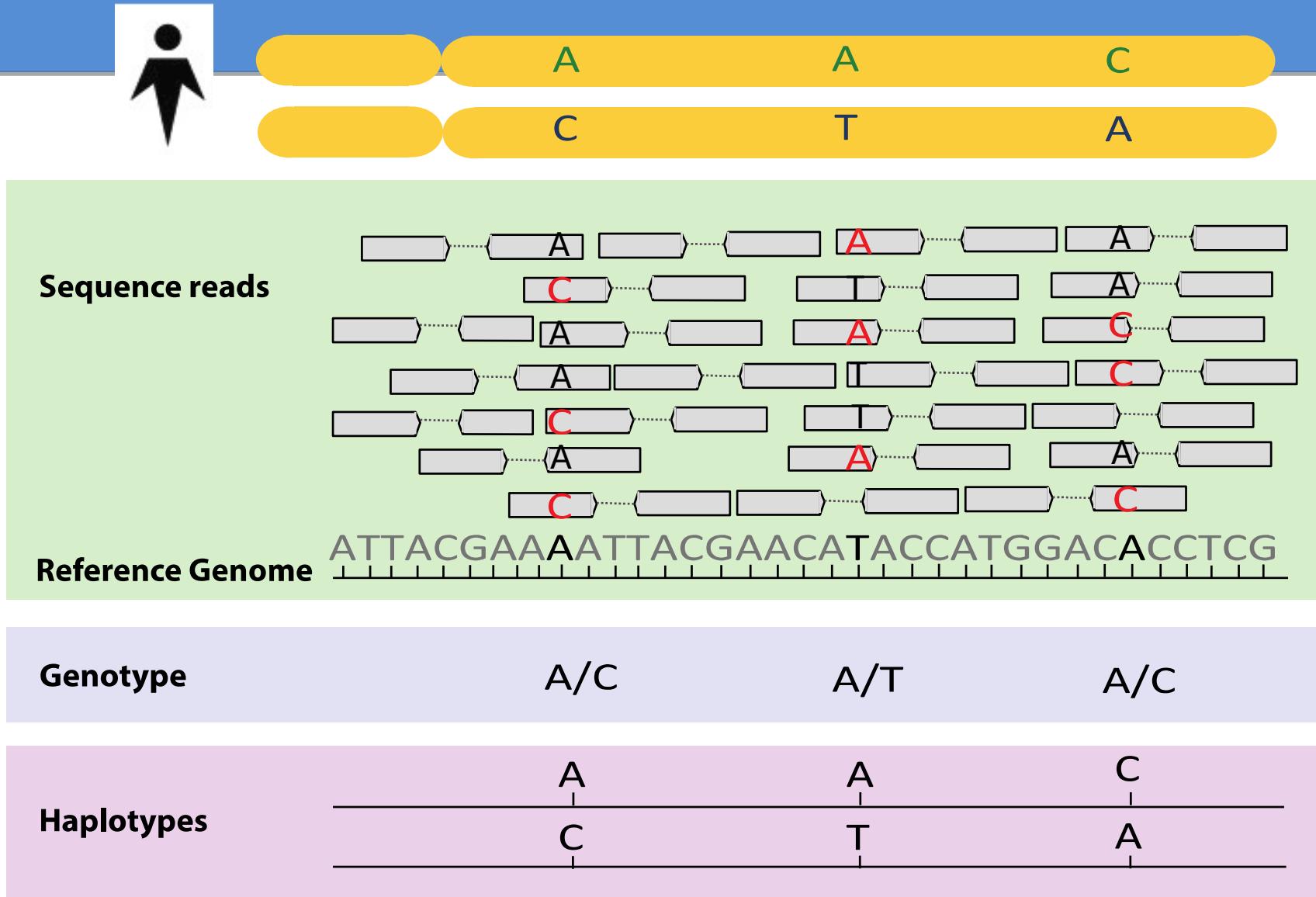


What is a haplotype?



Genotype	A/C	A/T	A/C				
A A C C T A	OR	A T C C A A	OR	A T A C A C	OR	A A A C T A	?
correct	wrong	wrong	wrong				

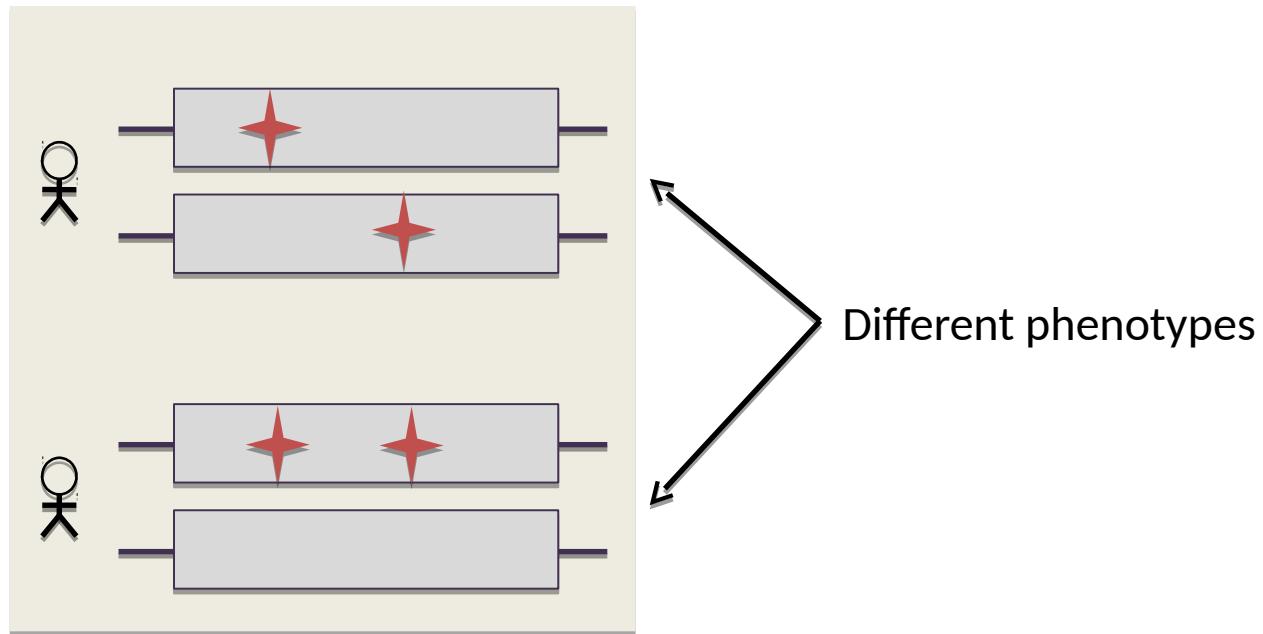
What is a haplotype?



Phasing is the process of recovering haplotypes from genotype data

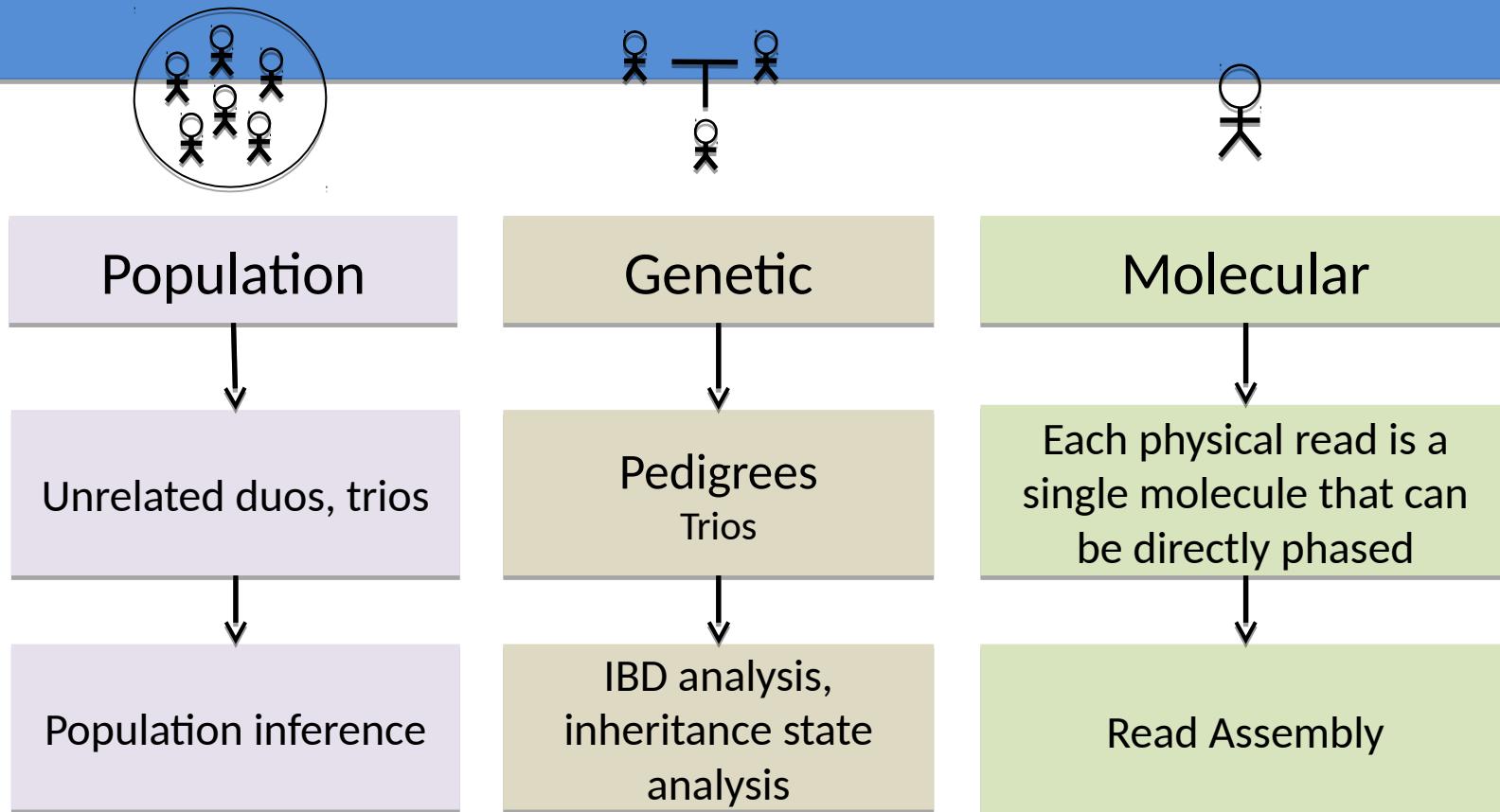
The importance of phase information

- Compound Heterozygosity



- Two-hit cancer model

Different Approaches for Phasing



	Population	Genetic	Molecular
Common variants	Yes	Yes	Yes
Private variants	No	Yes	Yes
Somatic variants	No	No	Yes