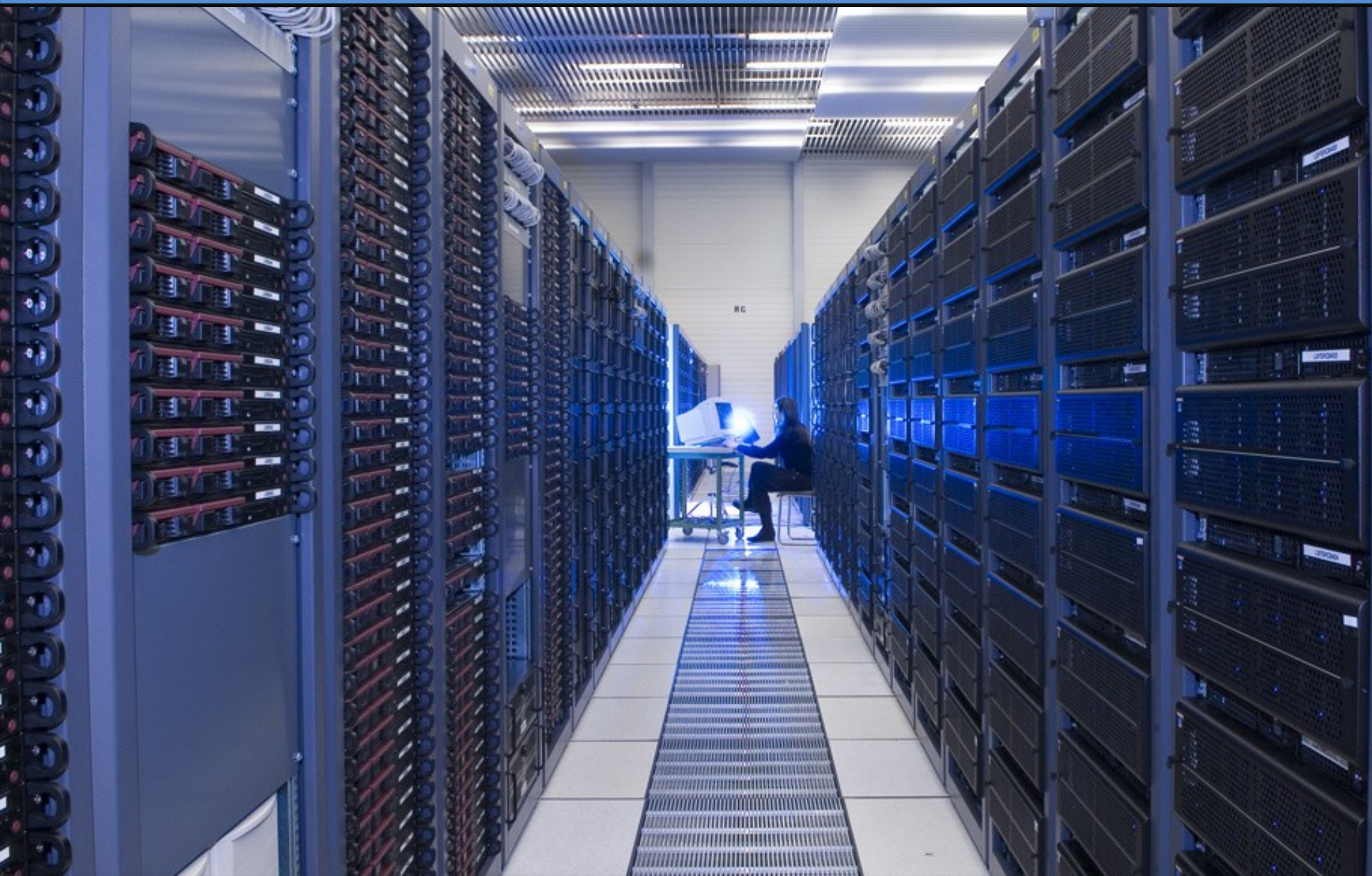
Cancer Sequencing



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

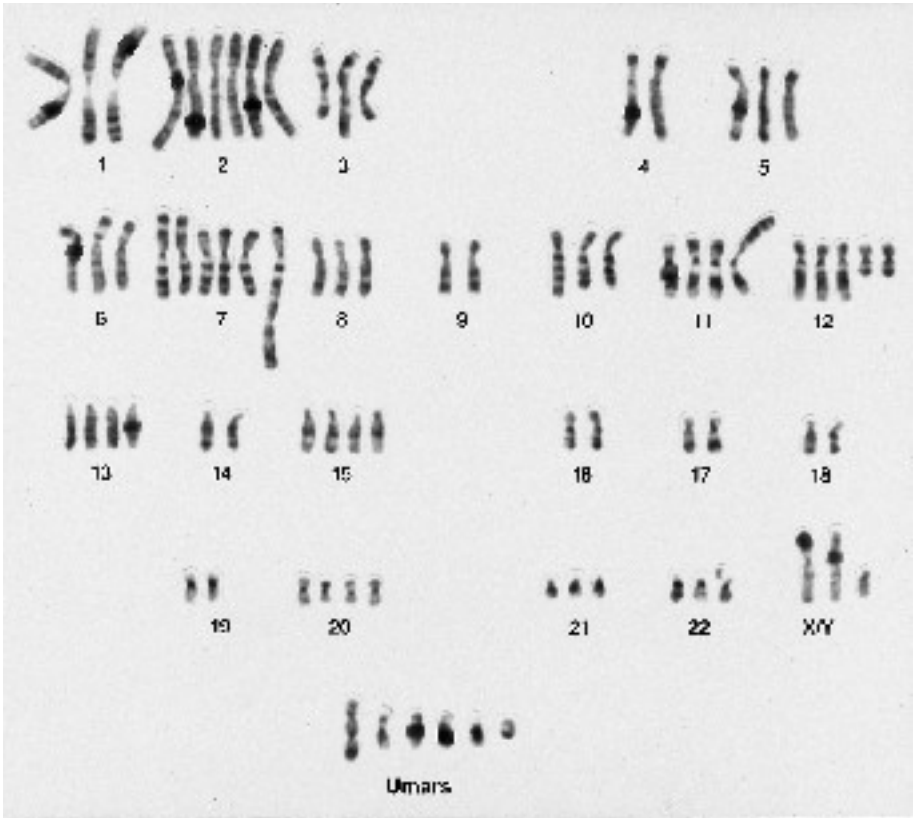
http://en.wikipedia.org/wiki/Cancer

http://faculty.ksu.edu.sa/tatiah/Pictures%20Library/normal%20male%20karyotyping.jpg

What is Cancer?

Definitions

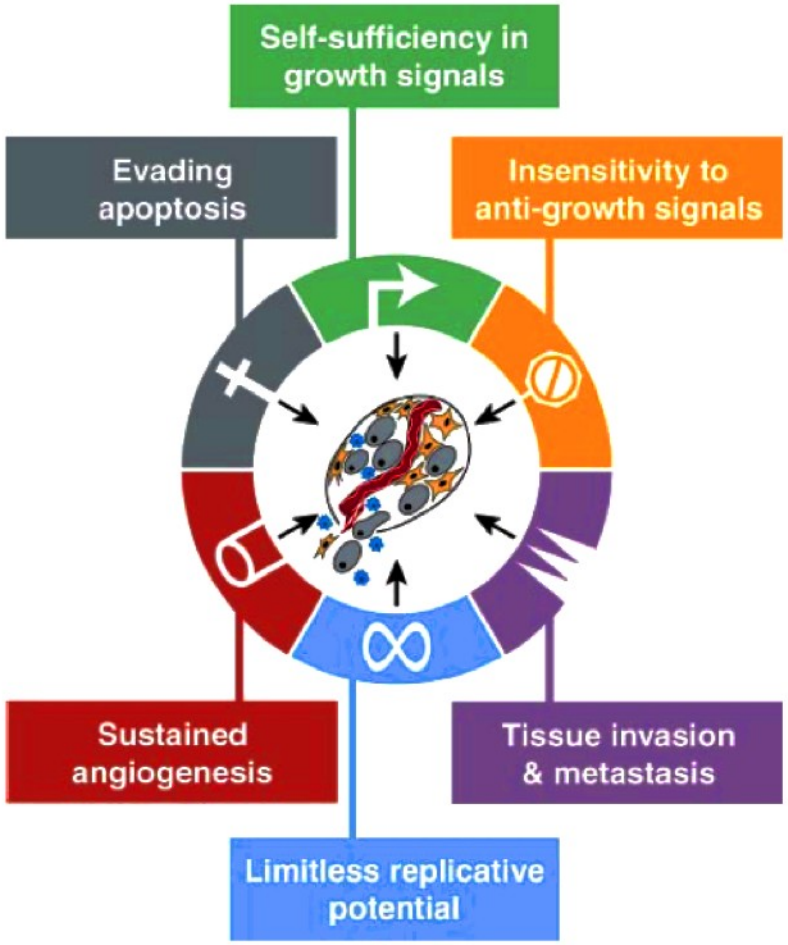
- A class of diseases characterized by malignant growth of a group of cells
  - Growth is uncontrolled
  - Invasive and Damaging
  - Often able to metastasize
- An instance of such a disease (a malignant tumor)
- A disease of the genome



http://en.wikipedia.org/wiki/Cancer

http://www.moffitt.org/CCJRoot/v2n5/artcl2img4.gif

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

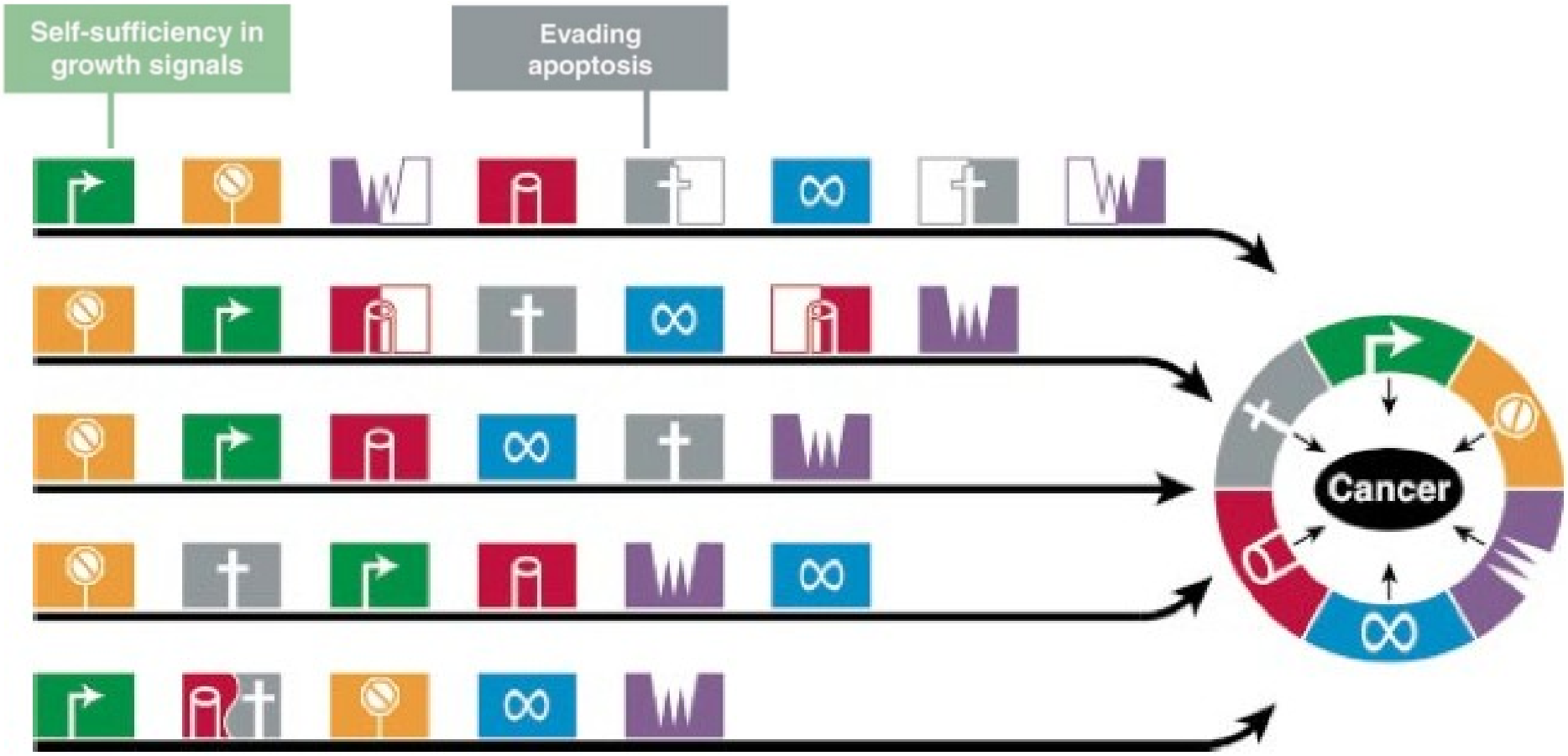
- Apoptosis (e.g. p53)
- Antigrowth signals (e.g. pRb)
- Cell Senescence

Acceleration of Cellular Evolution Via Genome Instability

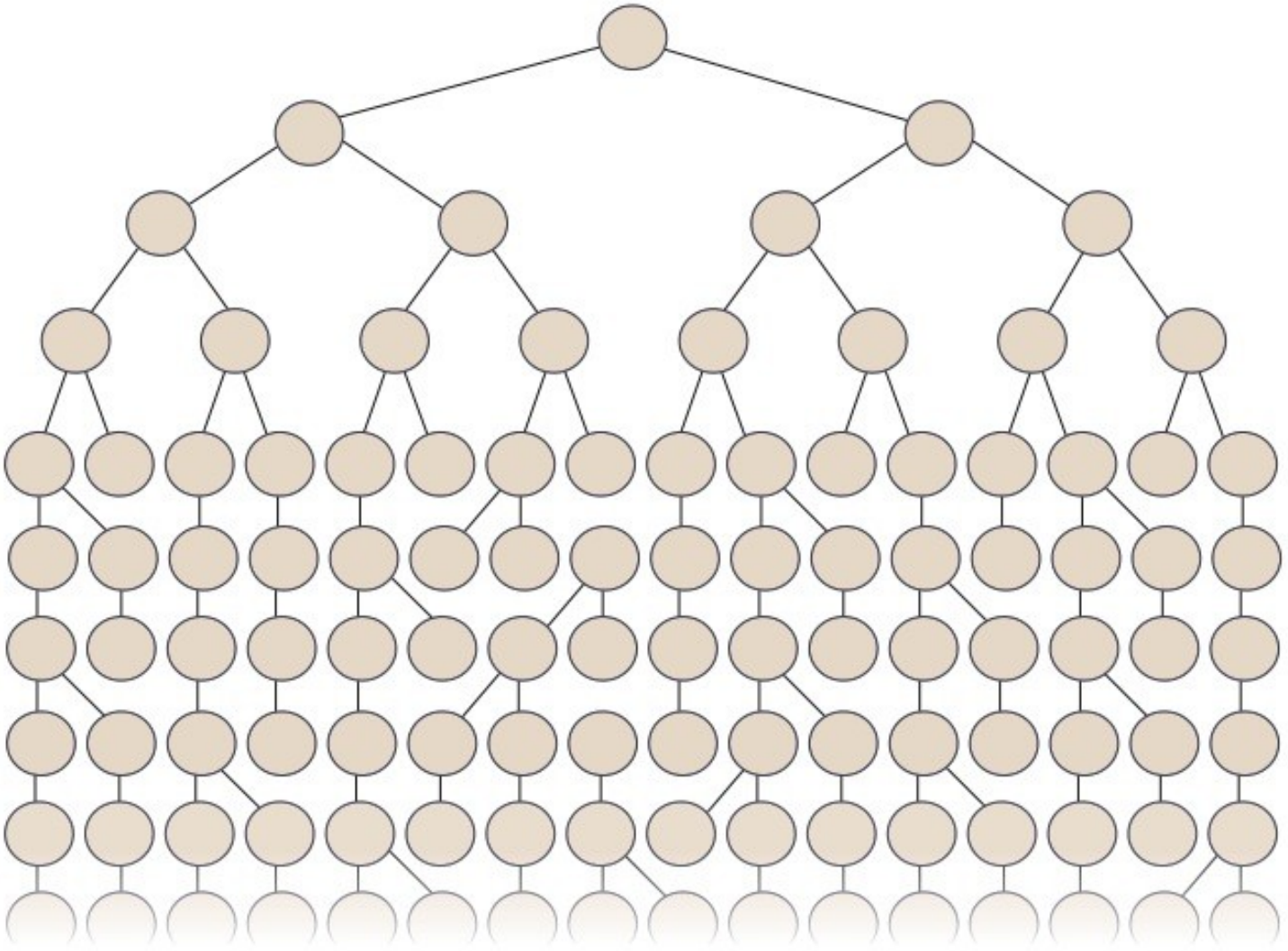
- DNA Repair
- DNA Polymerase

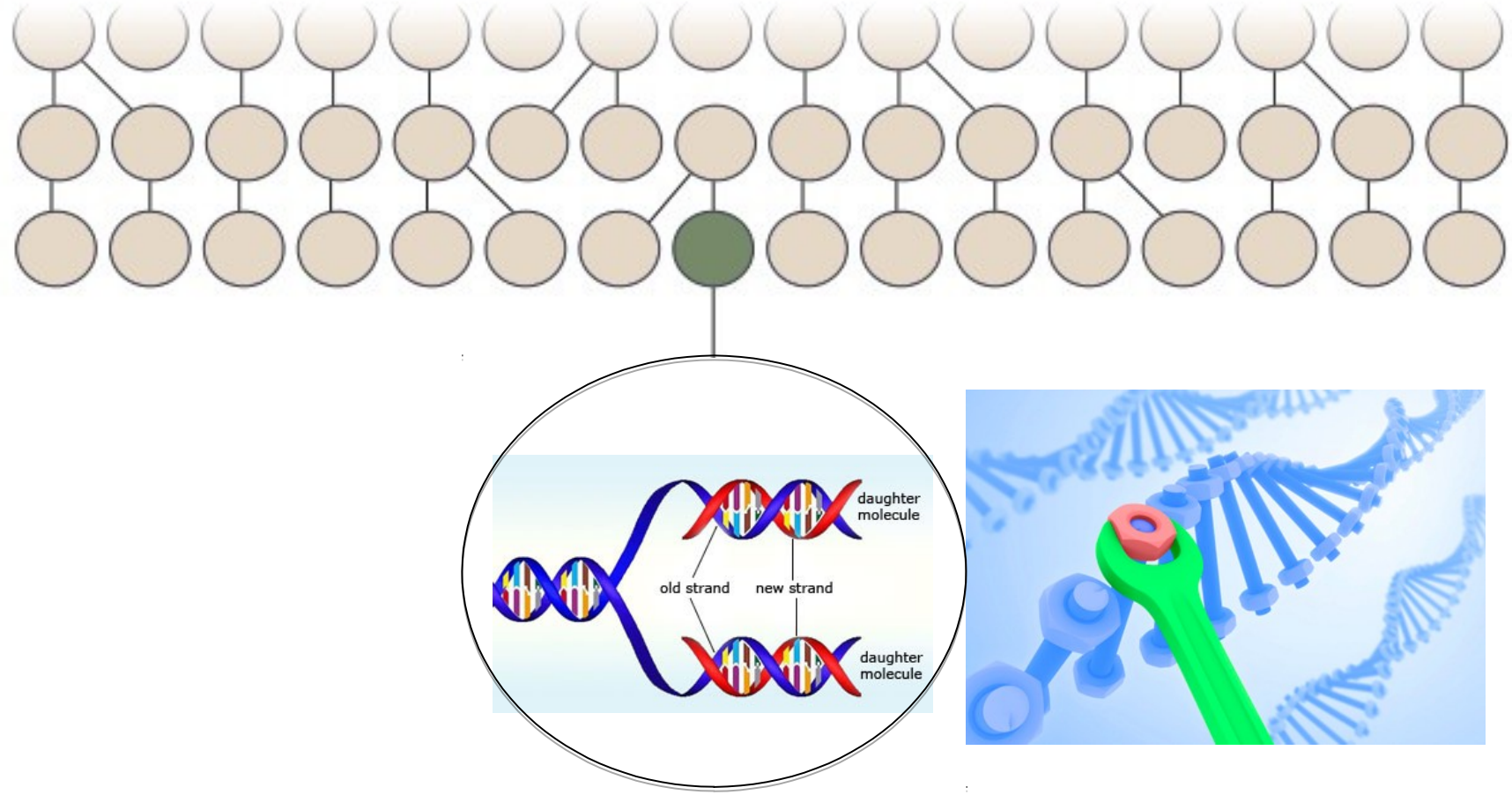
Hanahan and Weinberg. 2000. The hallmarks of cancer. Cell 100: 57-70.

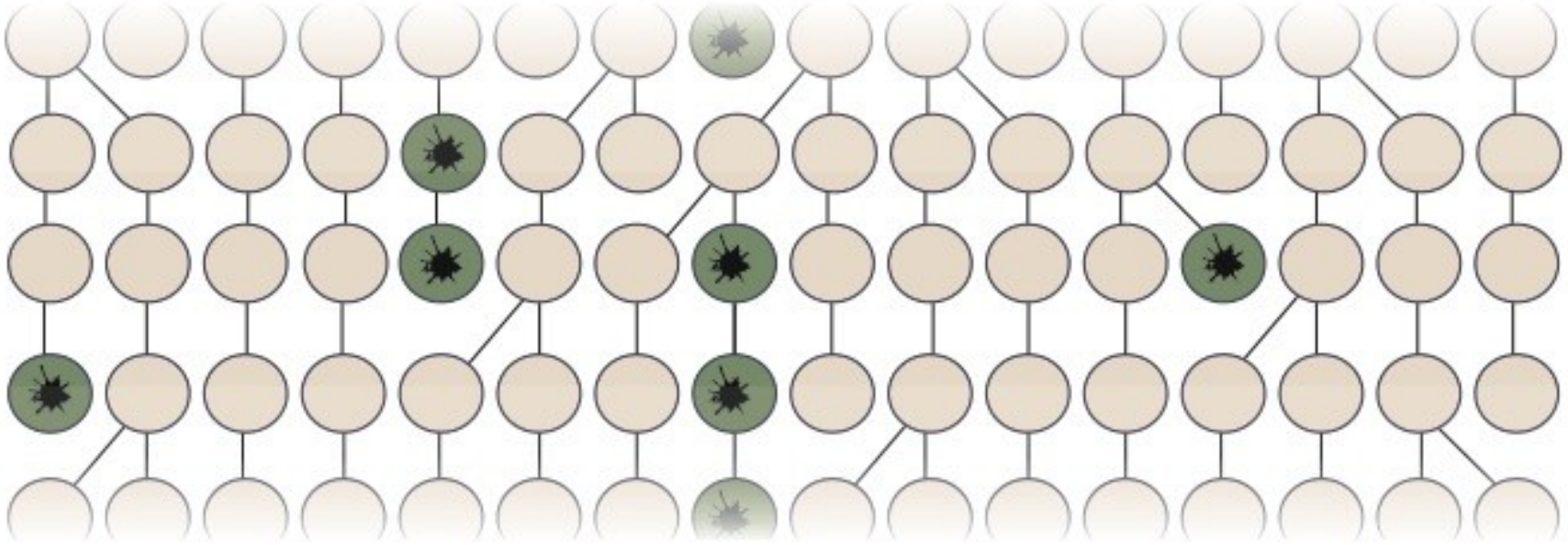
Many Paths Lead to Cancer Self-Sufficiency

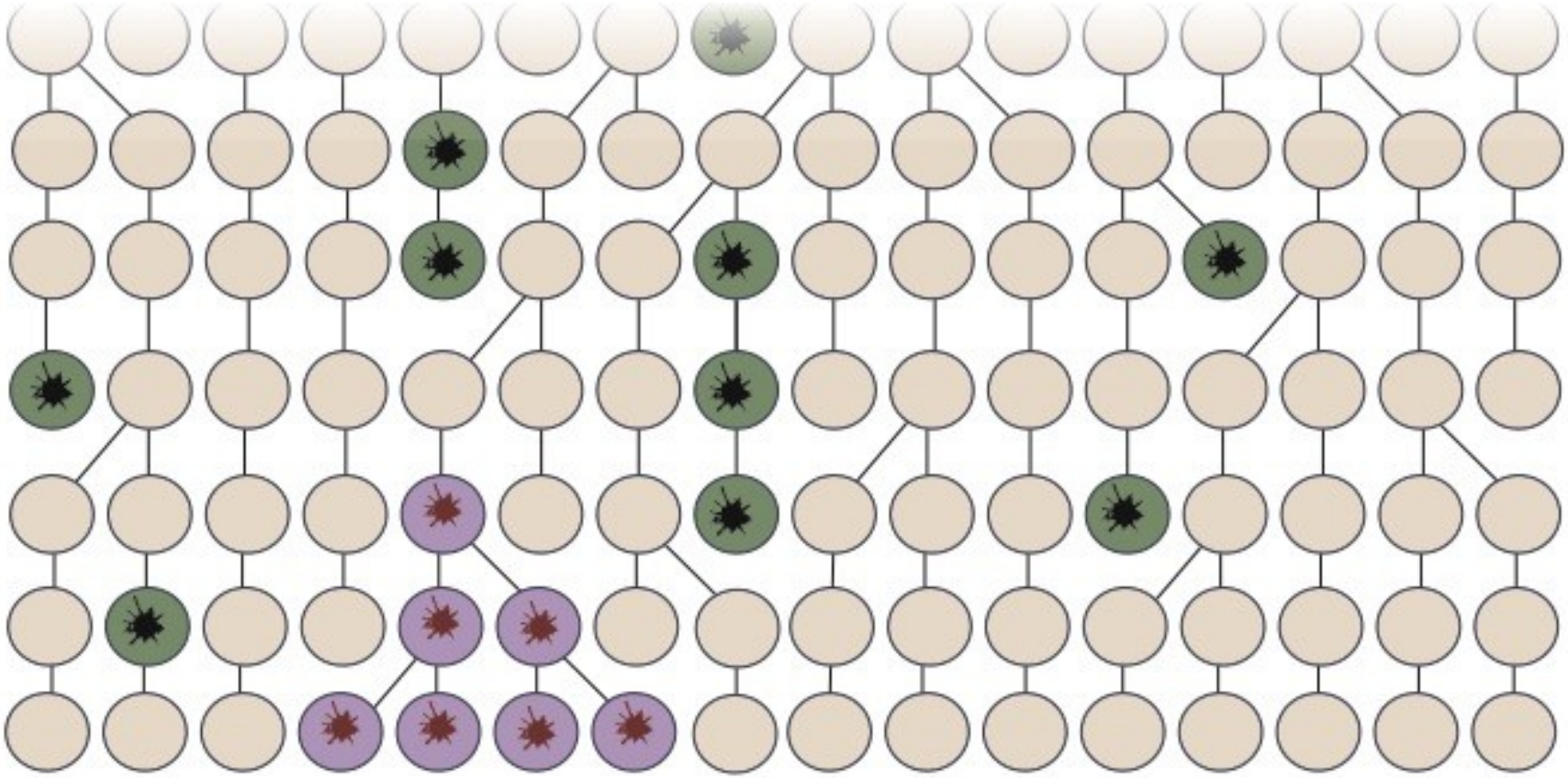


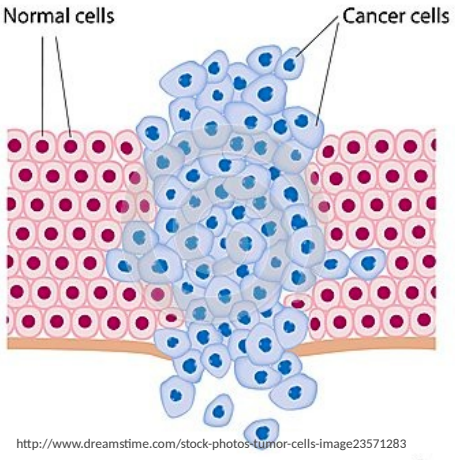
Hanahan, Douglas, and Ra Weinberg. 2000. The hallmarks of cancer. Cell 100: 57-70.



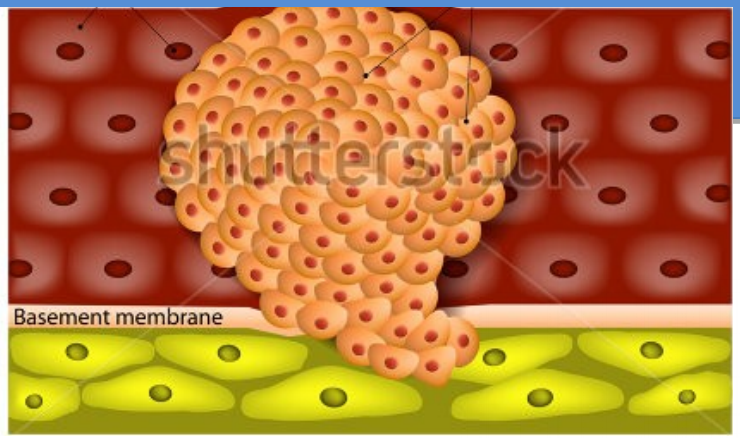

daughter molecule
old strand
new strand
daughter molecule

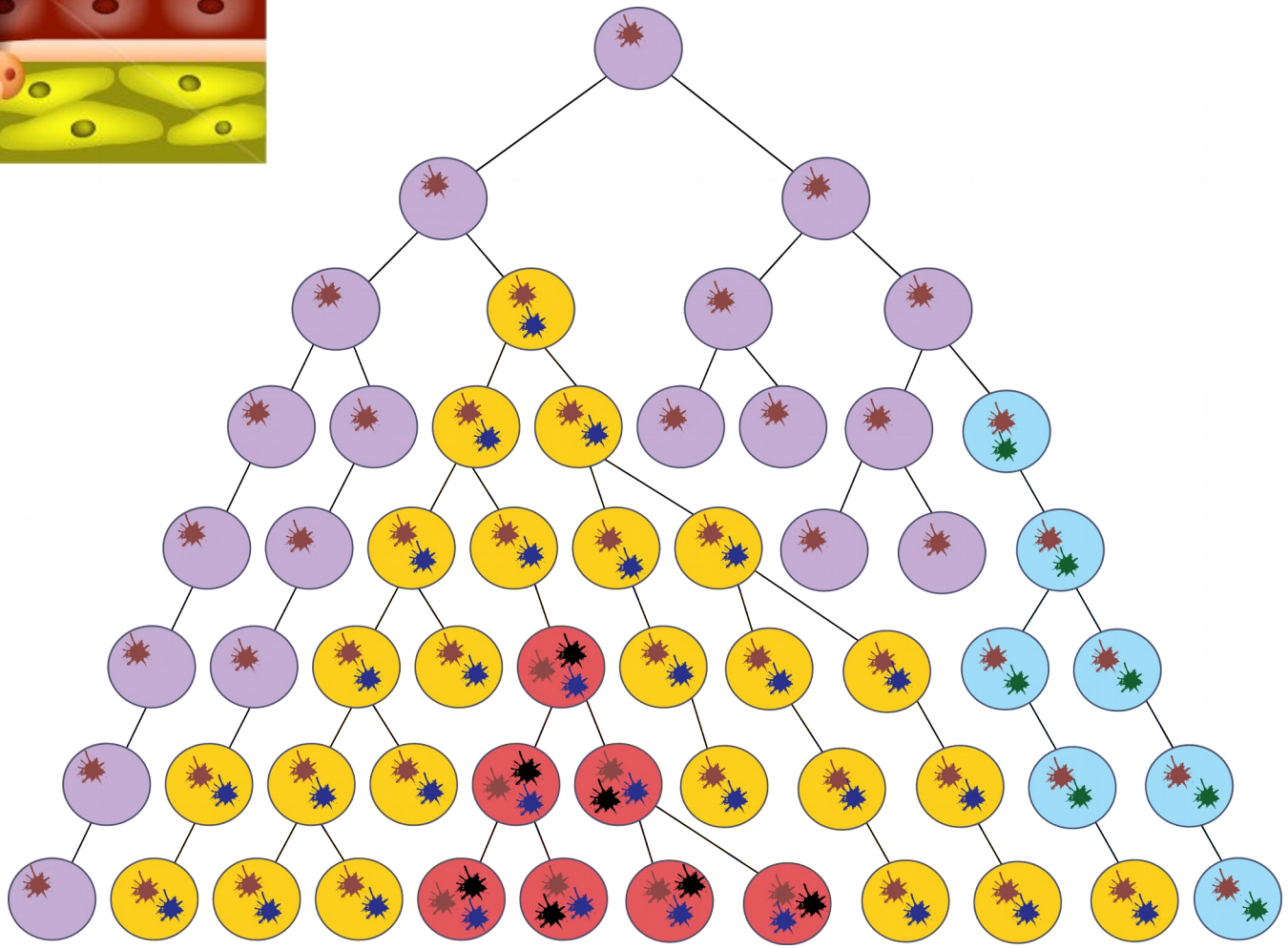


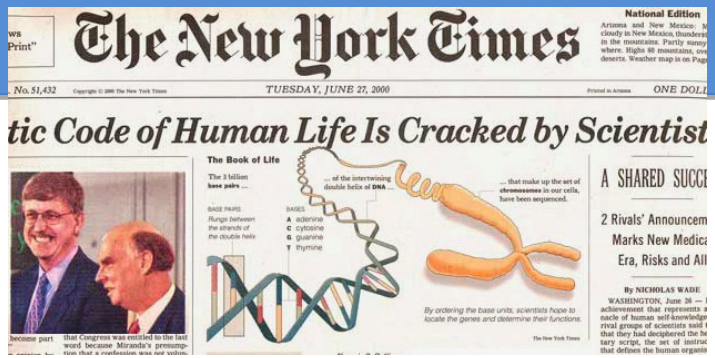


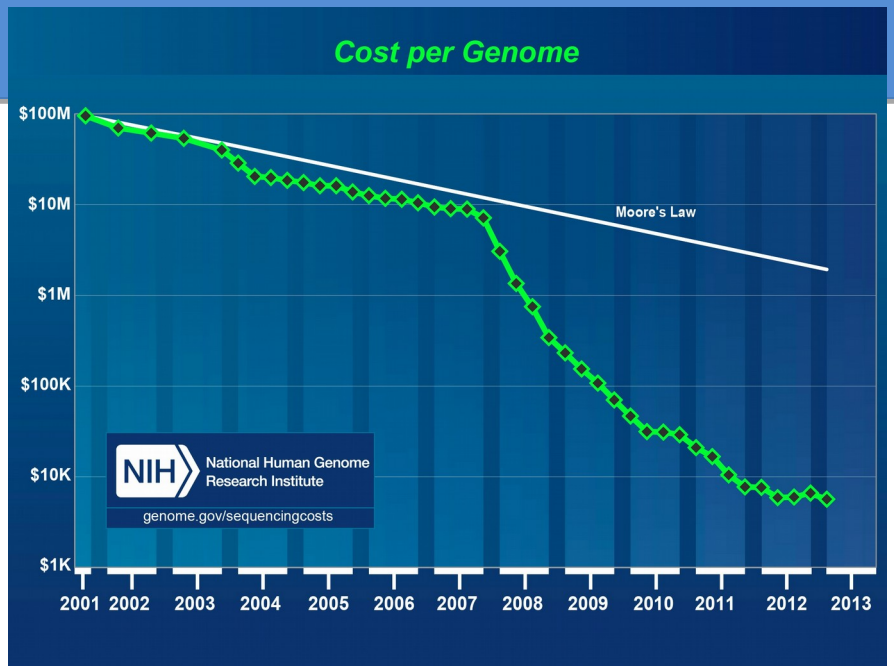


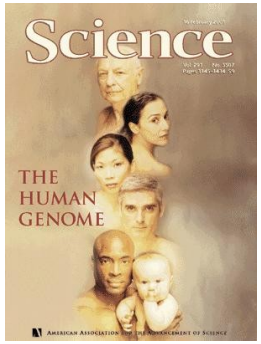
http://www.dreamstime.com/stock-photos-tumor-cells-image23571283


Basement membrane



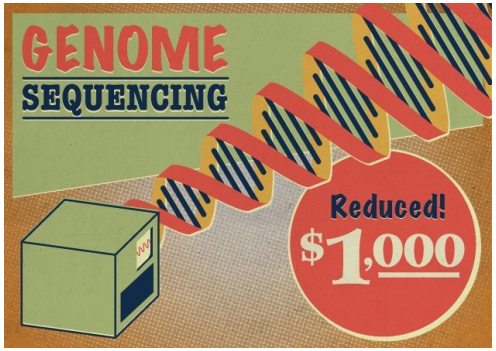






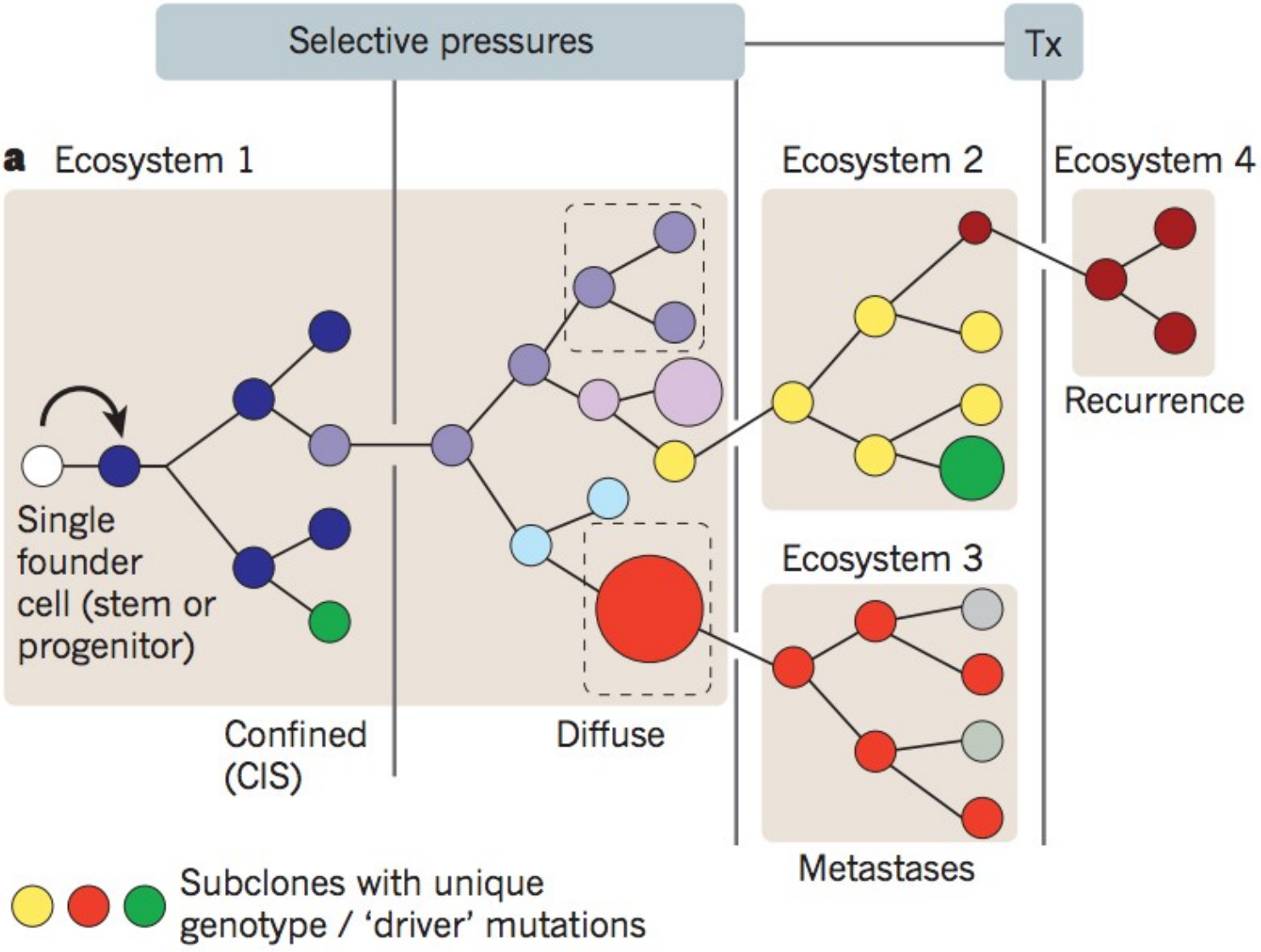
2000: initial draft

2003: Complete human reference genome ($3 billion)



2014

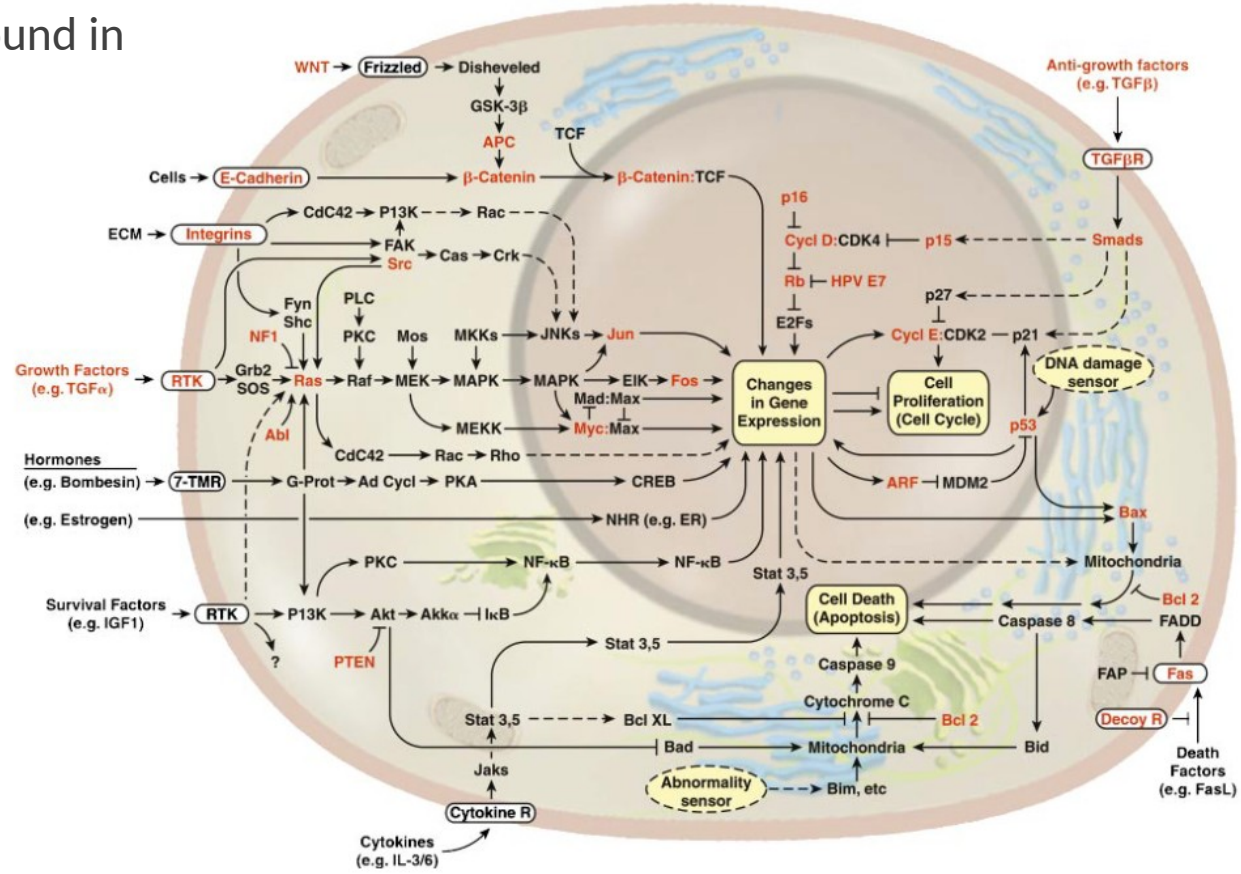
Cancer Heterogeneity



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

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/

Forbes et al. 2011. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Research 39: D945-D950

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/

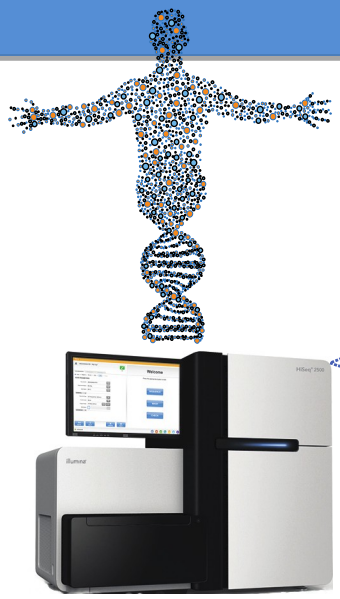
Why Sequence Cancer Genomes?

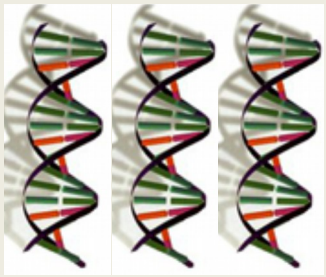
- Better understand cancer biology
  - Pathway information
  - Types of mutations found in different cancers
- Cancer Diagnosis
  - Genetic signatures of cancer types will inform diagnosis
  - Non-invasive means of detecting or confirming presence of cancer
- Improve cancer therapies
  - Targeted treatment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

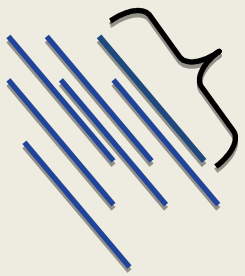
Genome Background





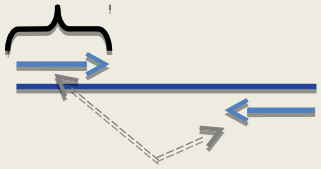
DNA is amplified

500 bp



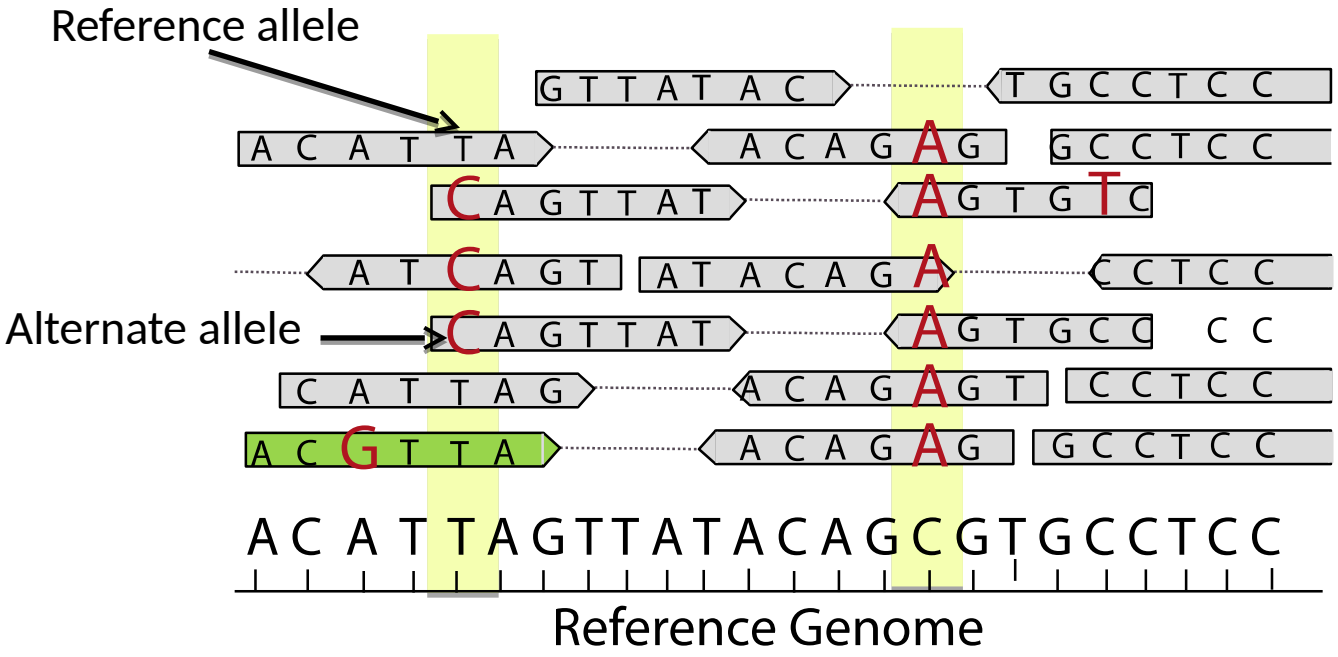
DNA is cut into small fragments

100 bp

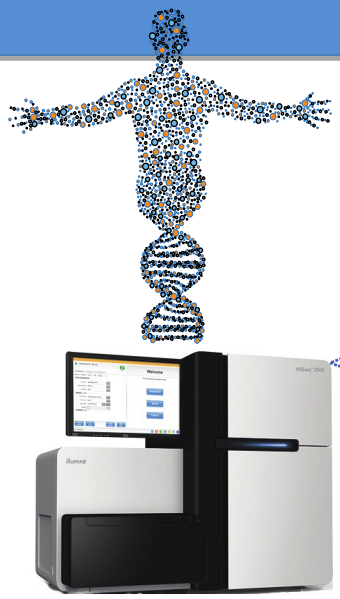


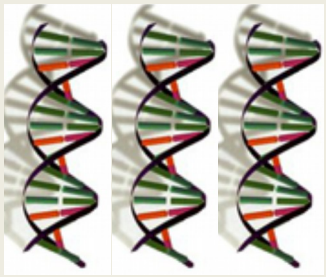
Sequenced reads

Two ends of the fragments are sequenced



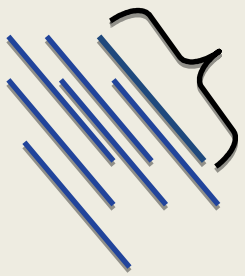
Genome Background





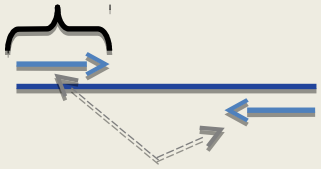
DNA is amplified

500 bp



DNA is cut into small fragments

100 bp

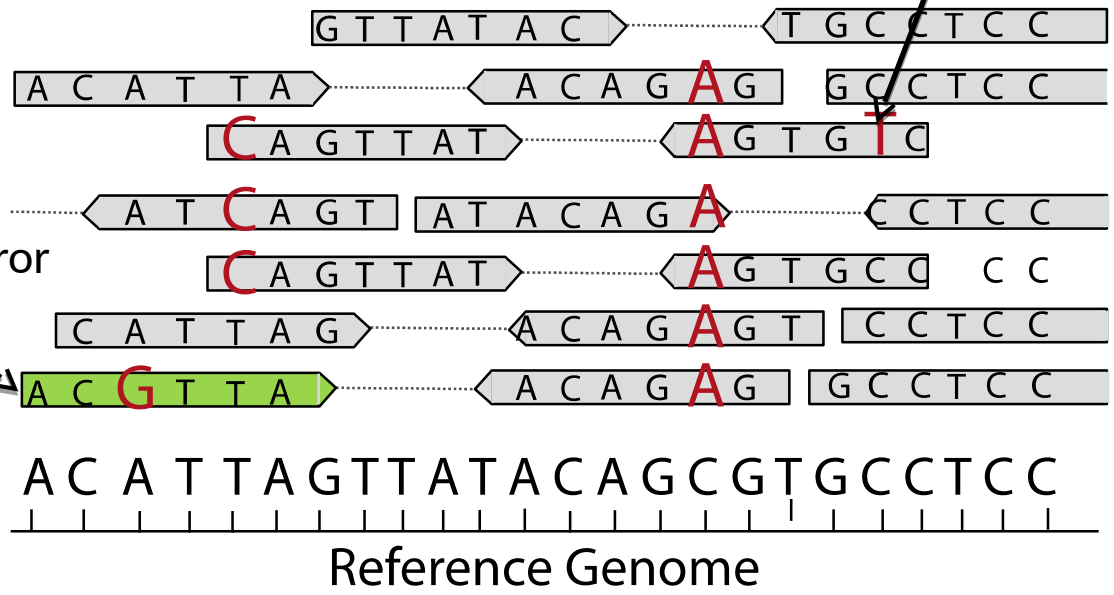


Sequenced reads

Two ends of the fragments are sequenced

Sequencing Error

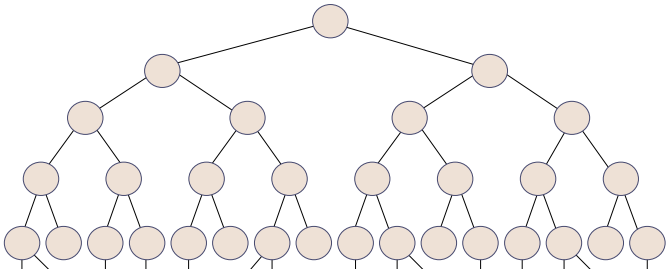
Alignment Error



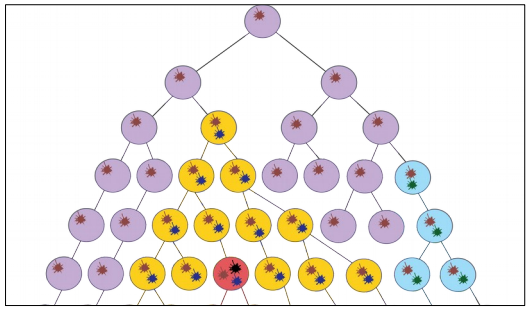
Genome Background

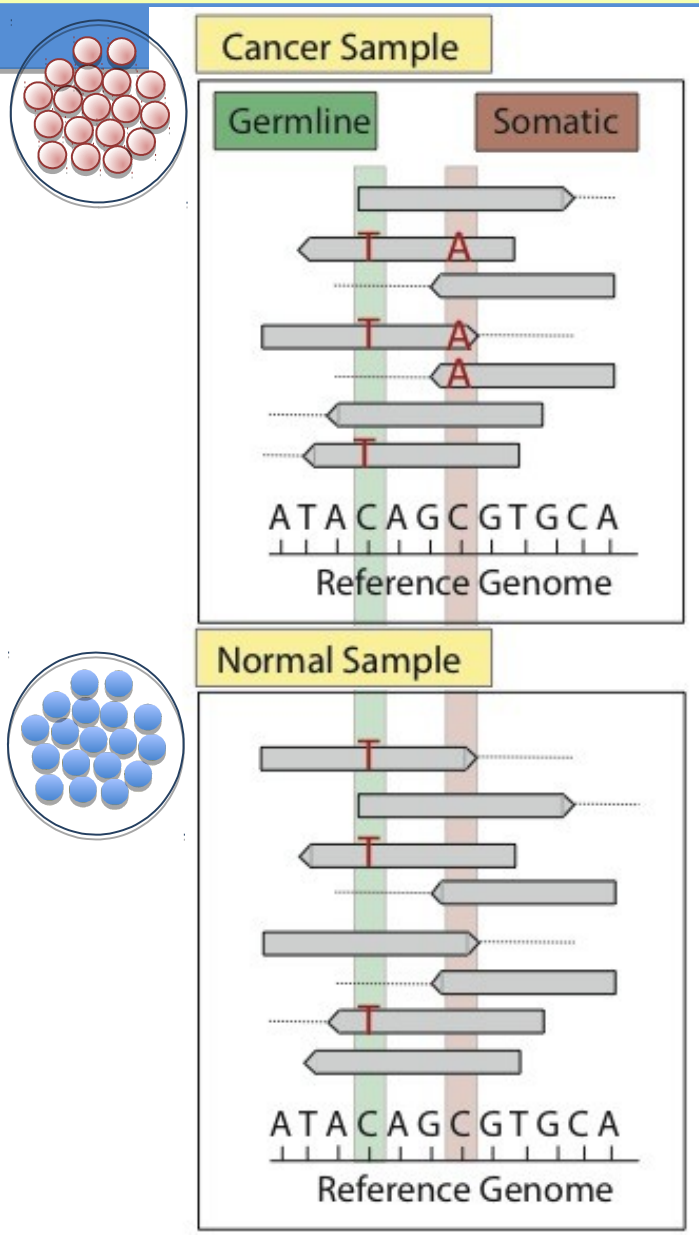
Types of SNVs in a cancer sample:

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

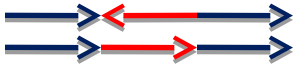
SNP

TGCTGAGA
TGCCGAGA

Novel Sequence

TGCTCGGAGA
TGC - - - GAGA

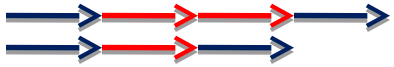
Inversion



Mobile Element or
Pseudogene Insertion

Translocation

Tandem Duplication



Microdeletion

TGC - - AGA
TGCCGAGA

Transposition

Large Deletion

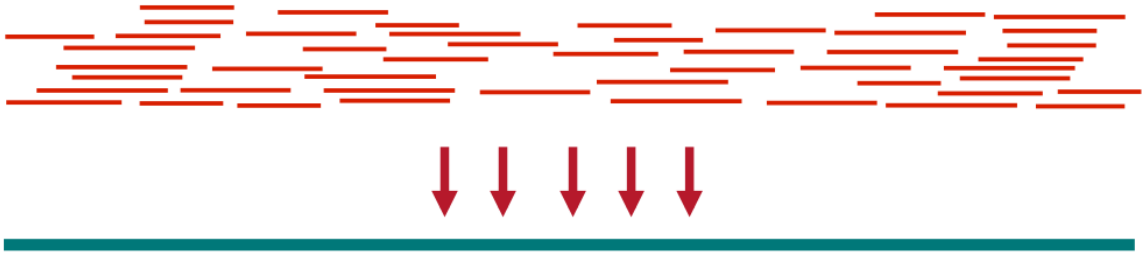
Novel Sequence
at Breakpoint

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

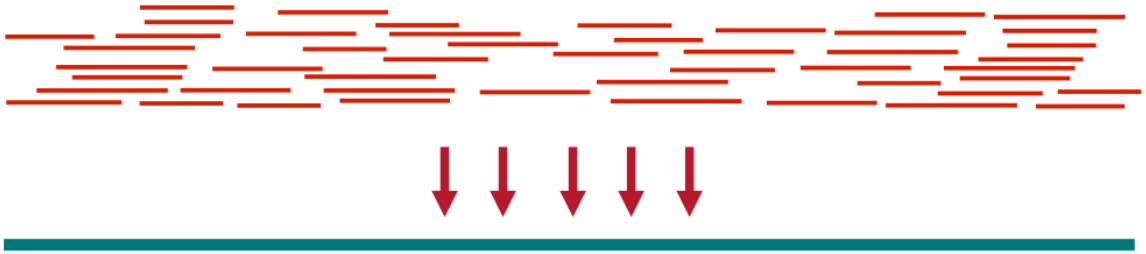


```
                    CATCGACCGAGCGCGATGCTAGCTAGGTGATCGT......
                 TGCCGCATCGACCGAGCGCGATGCTAGCTAGGTGATCGT...
              GCATGCCGCATCGACCGAGCGCGATGCTAGCTAGGTGATCGT
           GTGCATGCCGCATCGACCGAGCGCGATGCTAGCTAGGTGATC
......AGGTGCATGCCGCATCGATCGAGCGCGATGCTAGCTAGCTGATCGT......
```

- 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



```
                CATCGACCGAGCGCGATGCTAGCTAGGTGATCGT.....
             TGCCGCATCGACCGAGCGCGATGCTAGCTAGGTGATCGT...
          GCATGCCGCATCGACCGAGCGCGATGCTAGCTAGGTGATCGT
       GTGCATGCCGCATCGACCGAGCGCGATGCTAGCTAGGTGATC
.....AGGTGCATGCCGCATCGATCGAGCGCGATGCTAGCTAGCTGATCGT.....
```

Bayesian model

$$\Pr\{G|D\} = \frac{\overbrace{\Pr\{G\}}^{\text{Prior of the genotype}}\overbrace{\Pr\{D|G\}}^{\text{Likelihood of the genotype}}}{\Sigma_i \Pr\{G_i\}\Pr\{D|G_i\}}, \text{ [Bayes' rule]}$$

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

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

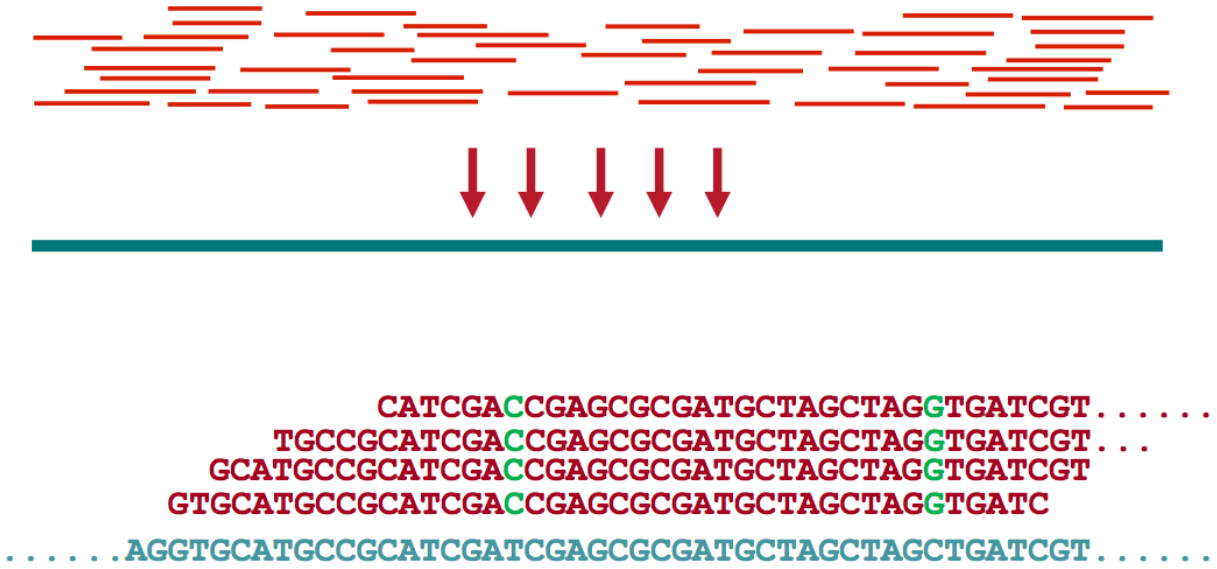
$$\Pr\{D_j|H\} = \Pr\{D_j|b\}, \text{ [single base pileup]}$$

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

http://www.broadinstitute.org/gatk//events/2038/GATKwh0-BP-5-Variant_calling.p

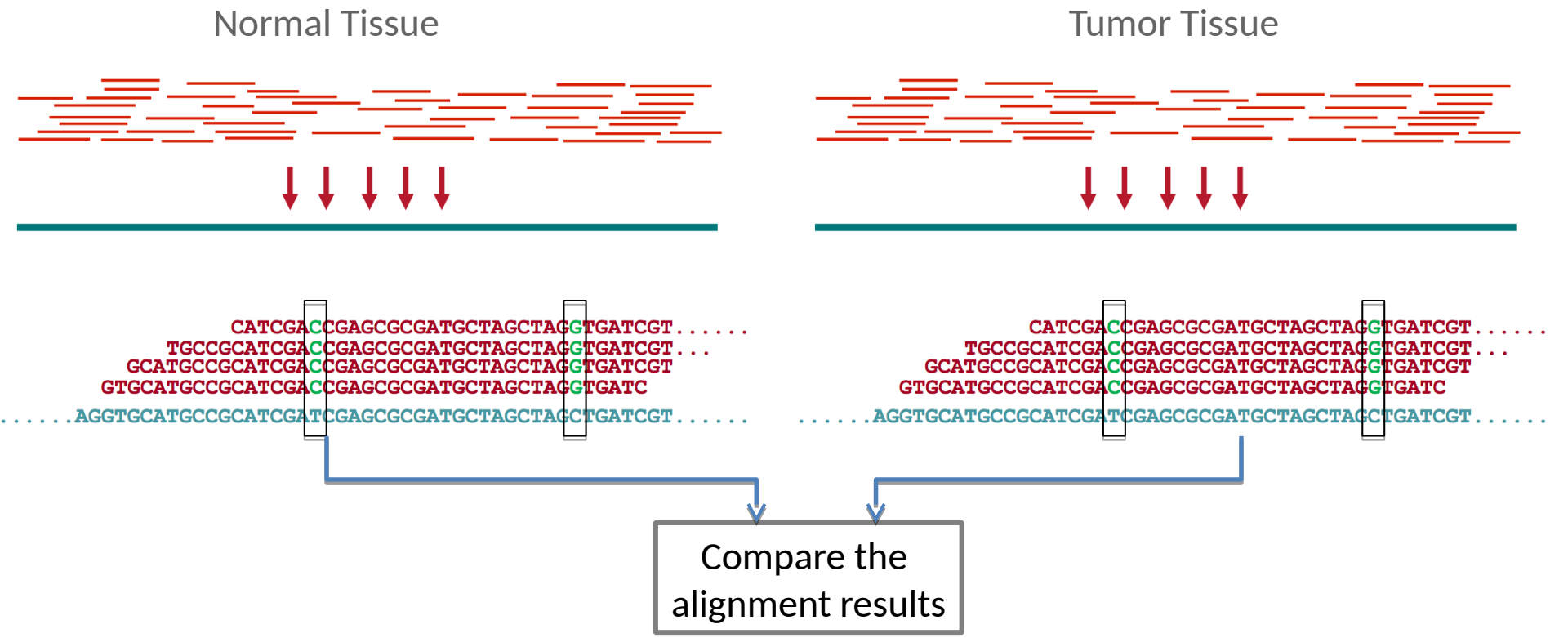
SNV Calling

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



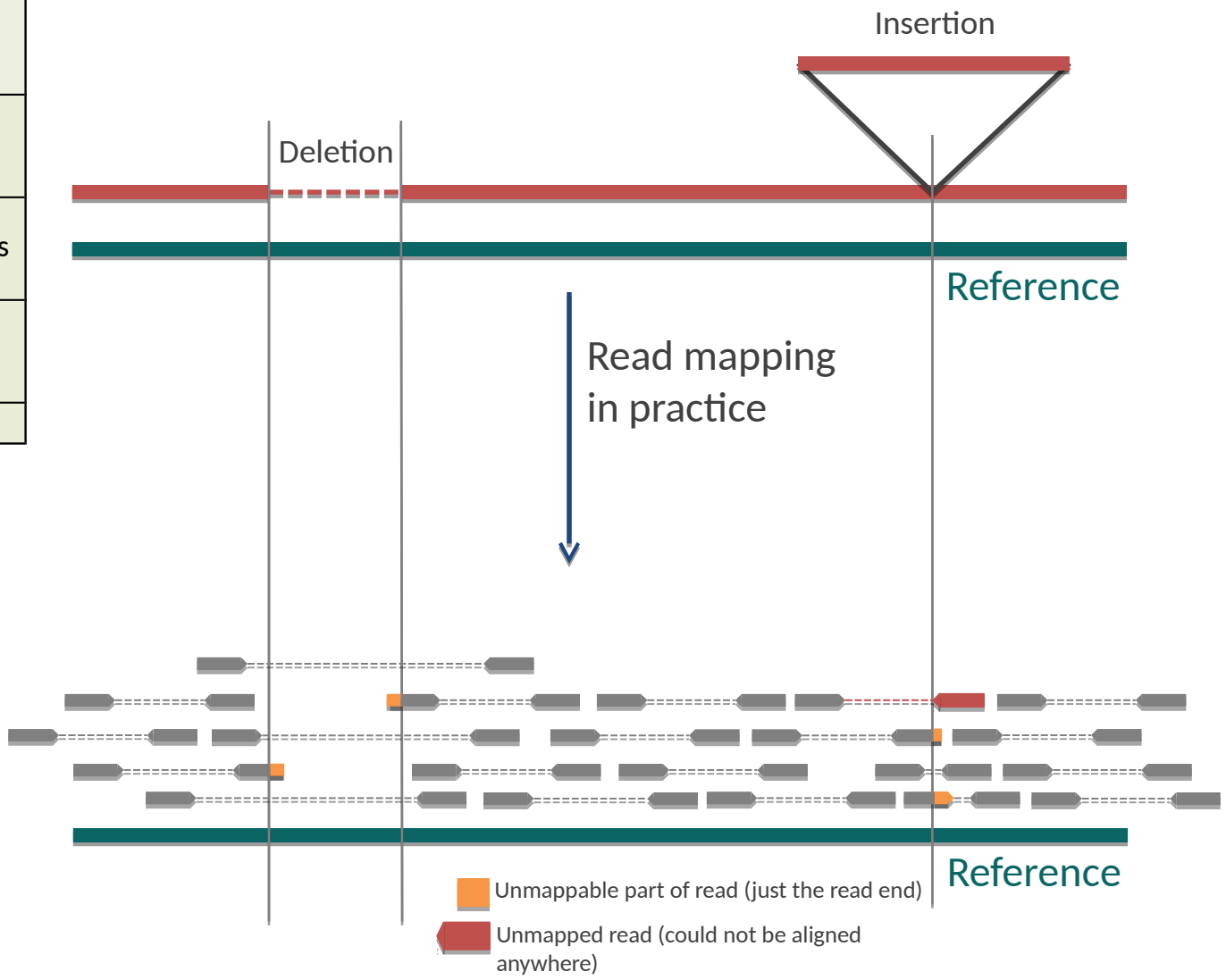
- Most naïve: use a standard SNV caller on both datasets. If there is a mutation found in the tumor sample but not the normal, it is somatic!

Short Indel Calling

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

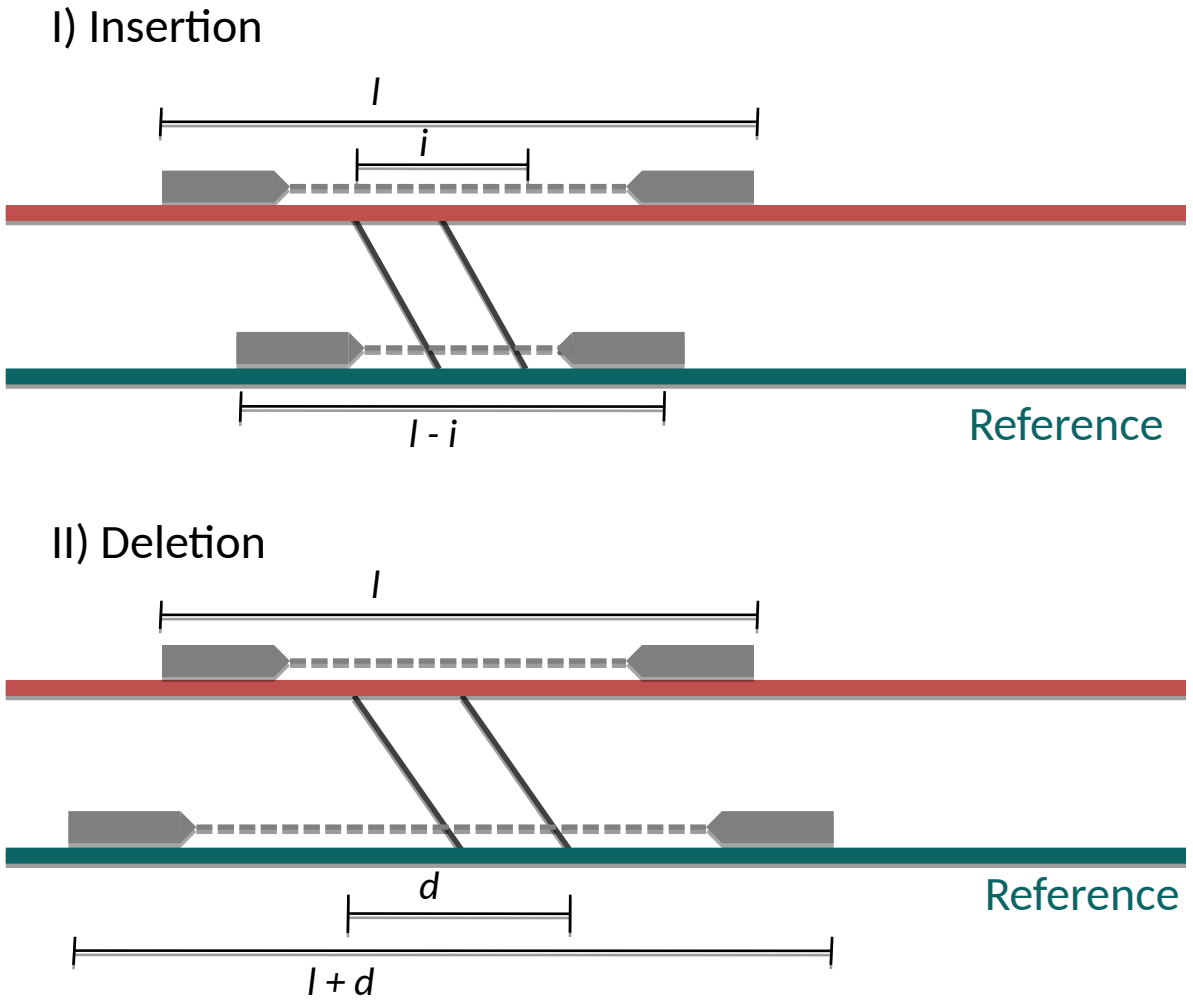
Short Indel Calling

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



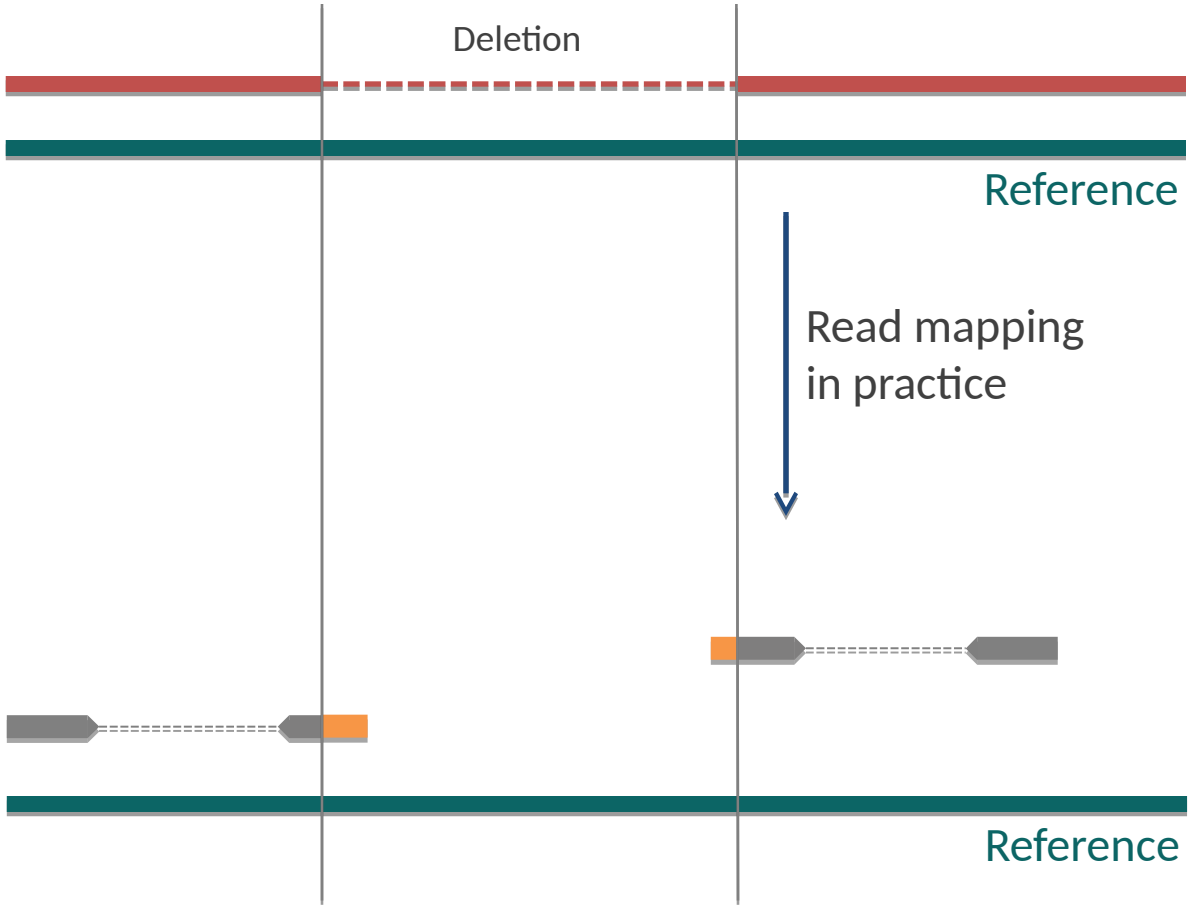
Short Indel Calling – Discordant Reads Pairs

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



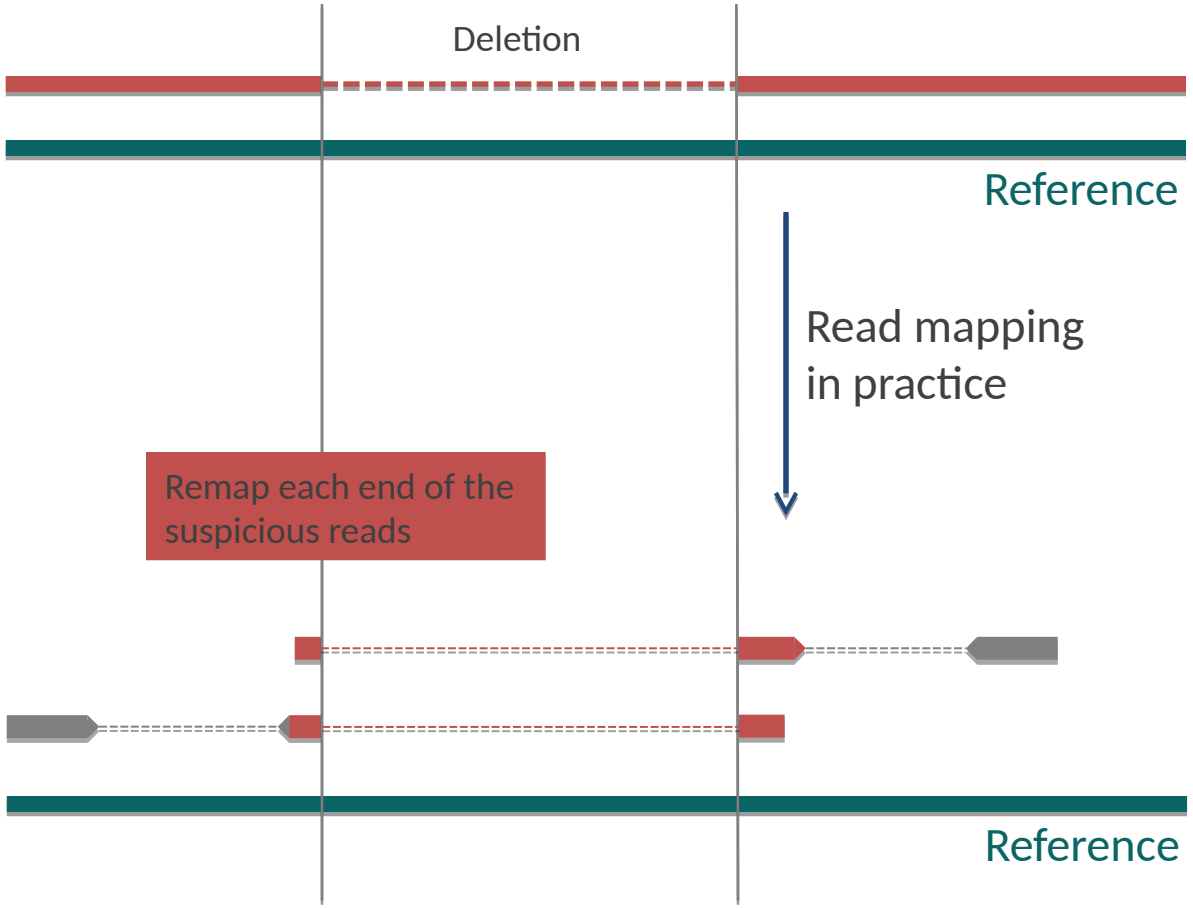
Short Indel Calling – Split Read Mapping

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



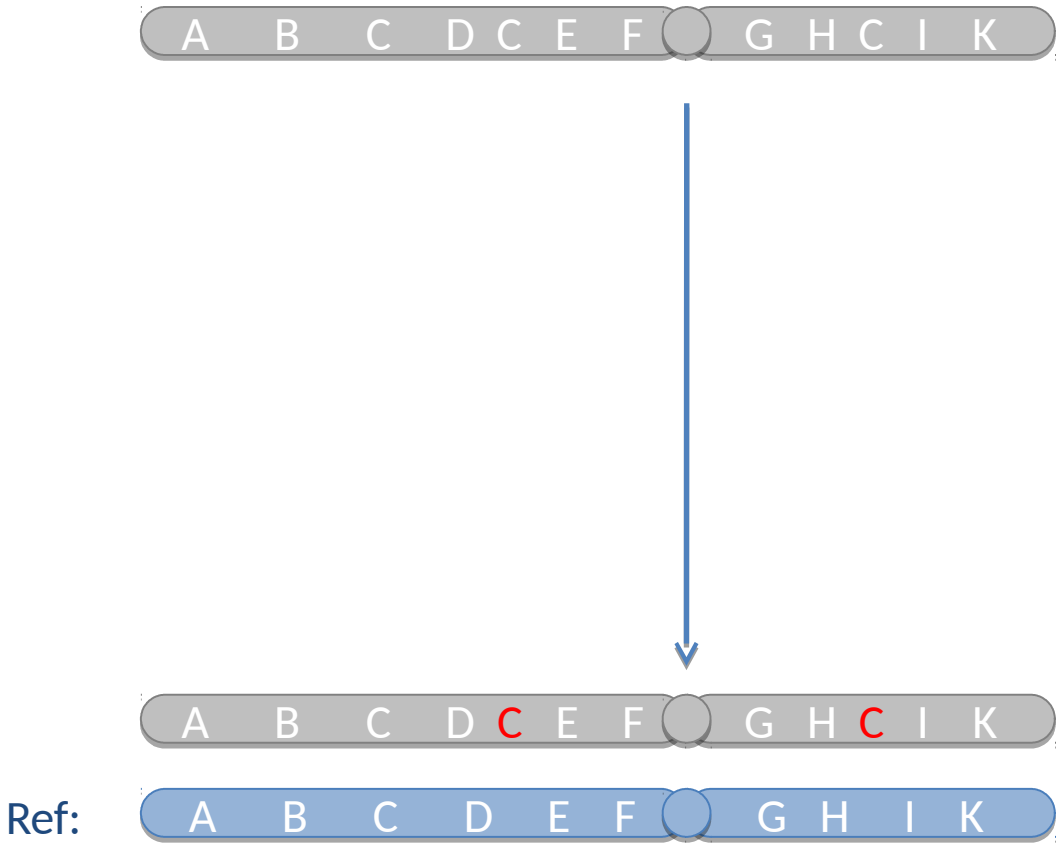
Short Indel Calling – Split Read Mapping

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



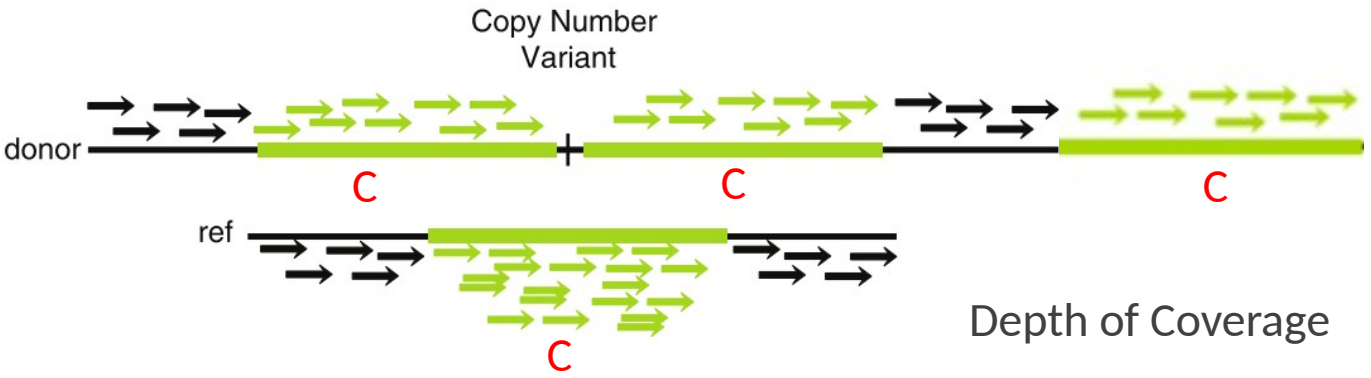
Copy Number Variants

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

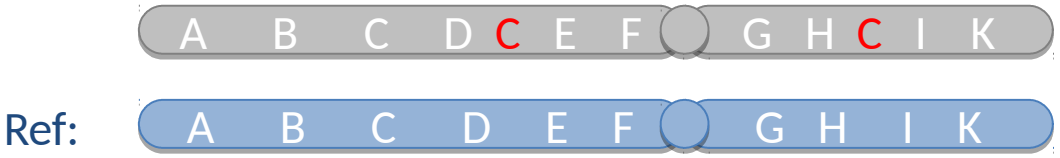


Copy Number Variants

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

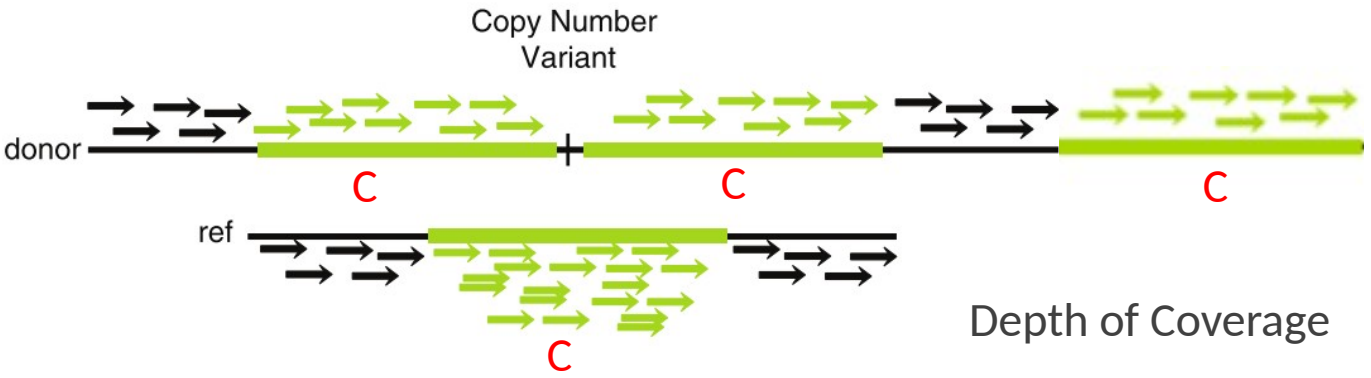


Modified from Dalca and Brudno. 2010. Genome variation discovery with high-throughput sequencing data. Briefings in bioinformatics 11, no. 1: 3-14

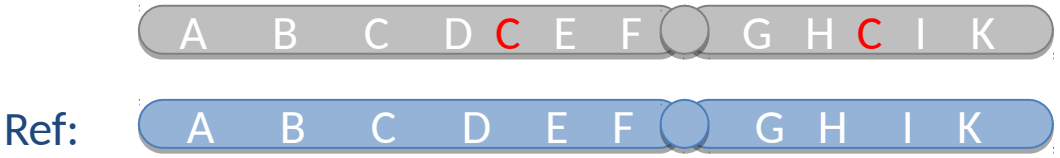


Copy Number Variants

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



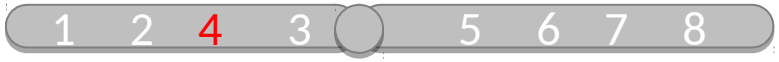
- Problems with DOC
  - Very sensitive to stochastic variance in coverage
  - Sensitive to bias coverage (e.g. GC content).
  - Impossible to determine non-reference locations of CNVs
- Graph methods using paired-end reads help overcome some of these problems



Variant Types

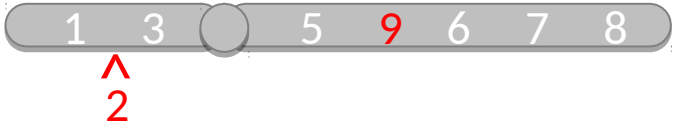
Variant Types
Single Nucleotide Variants(SNVs)
Short Insertion / Deletion (indels)
Copy Number Variants (CNVs)
Structural Variants (SVs)
Novel Sequence

Structural Rearrangement

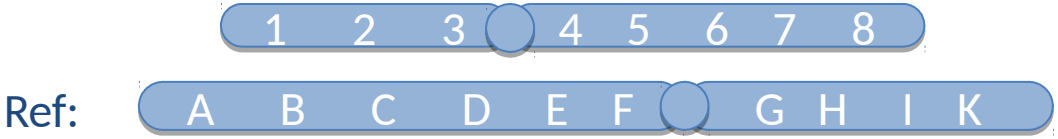


Translocation

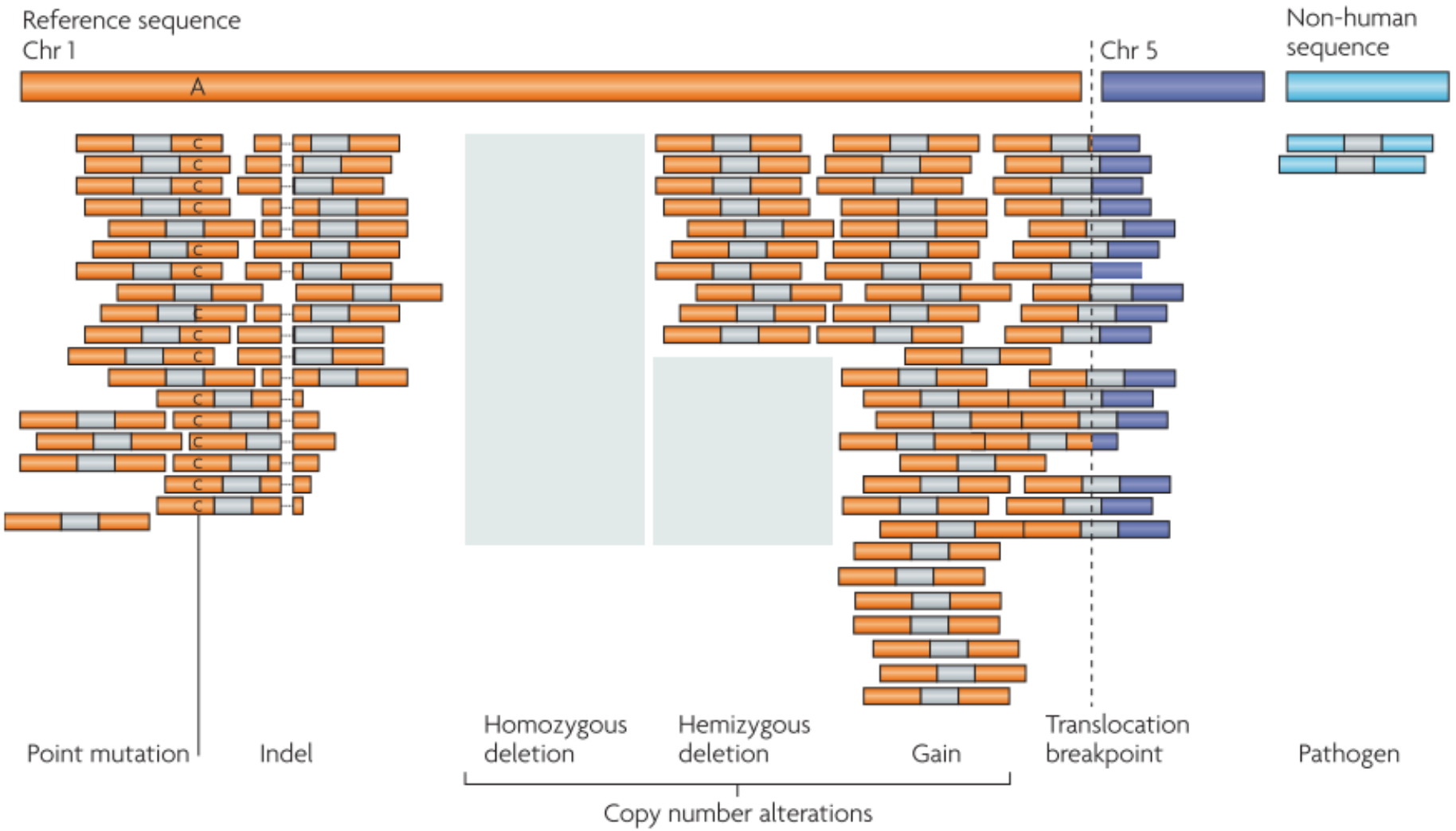
Inversion



Large Insertion / Deletion



Summary of Variant Types



Meyerson et al. . 2010. Advances in understanding cancer genomes through second-generation sequencing. Nature Reviews Genetics 11, no. 10 (October): 685-696

Tracking the Evolution of Cancer

Genome evolution during progression to breast cancer

Daniel E. Newburger, Dorna Kashef-Haghighi, 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

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] Manasa Ramakrishna,[1] Adam Shlien,[1] Susanna L. Cooke,[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] Elli Papaemmanuil,[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] Sancha Martin,[1] Goran Jönsson,[9] Odette Mariani,[7] Sandrine Boyault,[8] Penelope Miron,[9] Aquila Fatima,[9] 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] Sancha 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] Elli Papaemmanuil,[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,[6,12] Åke Borg,[13] Anne Vincent Salomon,[14] Gilles Thomas,[8] Anne-Lise Børresen-Dale,[9,15] Andrea L. Richardson,[7,16] 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

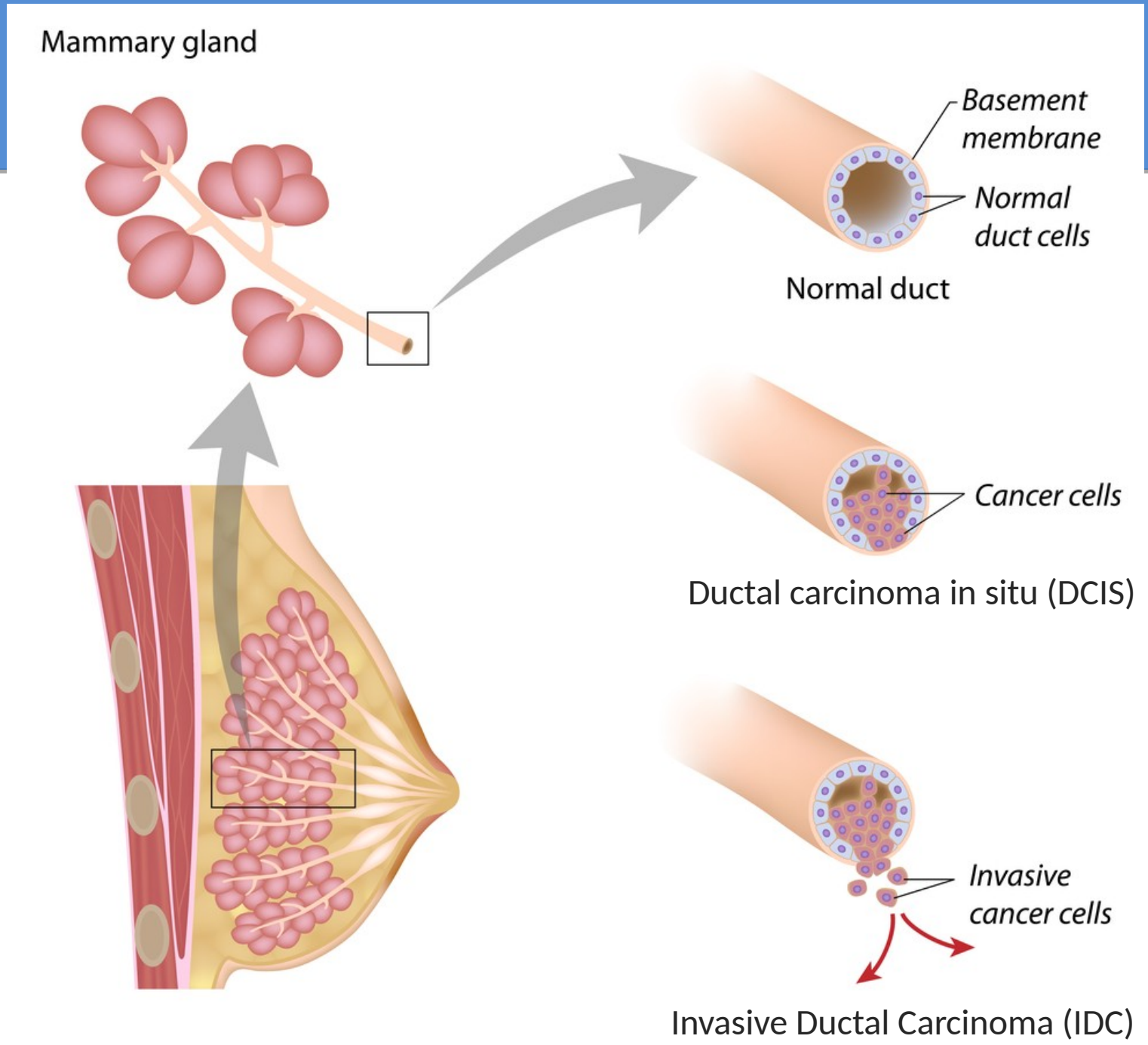
Lucy R. Yates[1] and Peter J. Campbell[1,2]

NATURE REVIEWS | GENETICS VOLUME 13 | NOVEMBER 2012

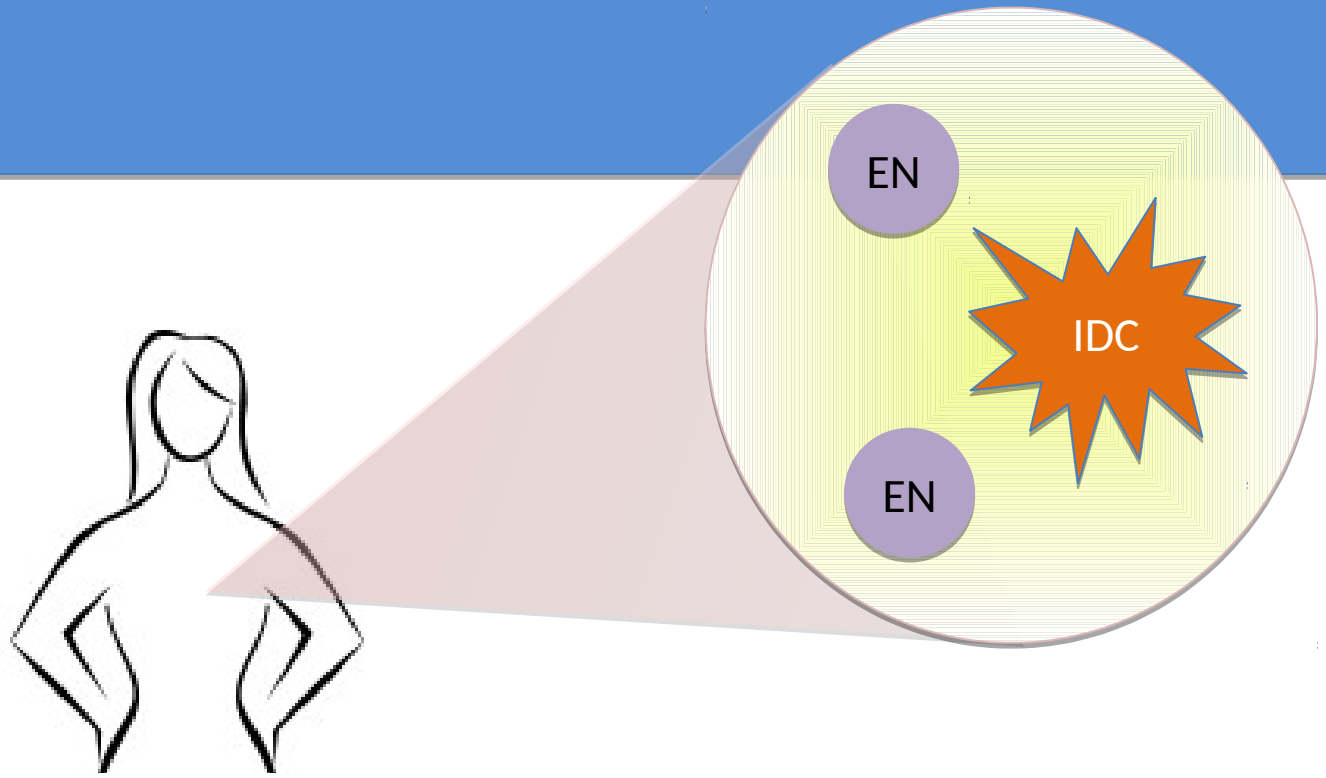
Clonal evolution in cancer

Mel Greaves[1] & Carlo C. Maley[2]

NATURE | VOL 481 | 19 JANUARY 2012



Source:http://marnieclark.com/category/types-of-breast-cancer/dcis/



Normal

Early Neoplasia

Ductal Carcinoma In situ

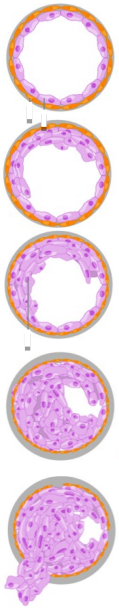
Invasive Ductal Carcinoma

Progression

Lesions

Patients

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



Types of Variants

- Somatic SNVs
- Aneuploidies

Questions

>50x whole genome coverage per sample

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

Variant calling process

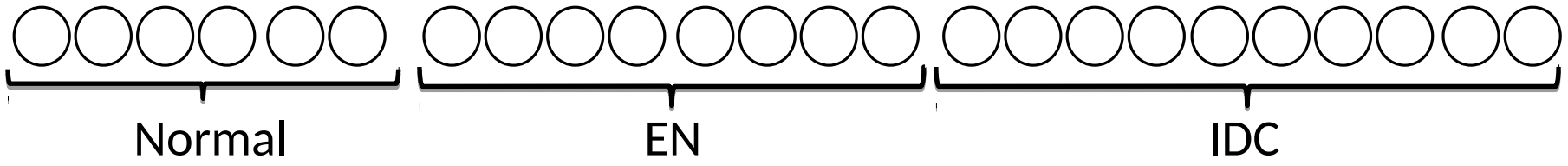
Challenges

- Sequencing Error
- Mapping Error
- Sequencing Coverage
- Normal Contamination

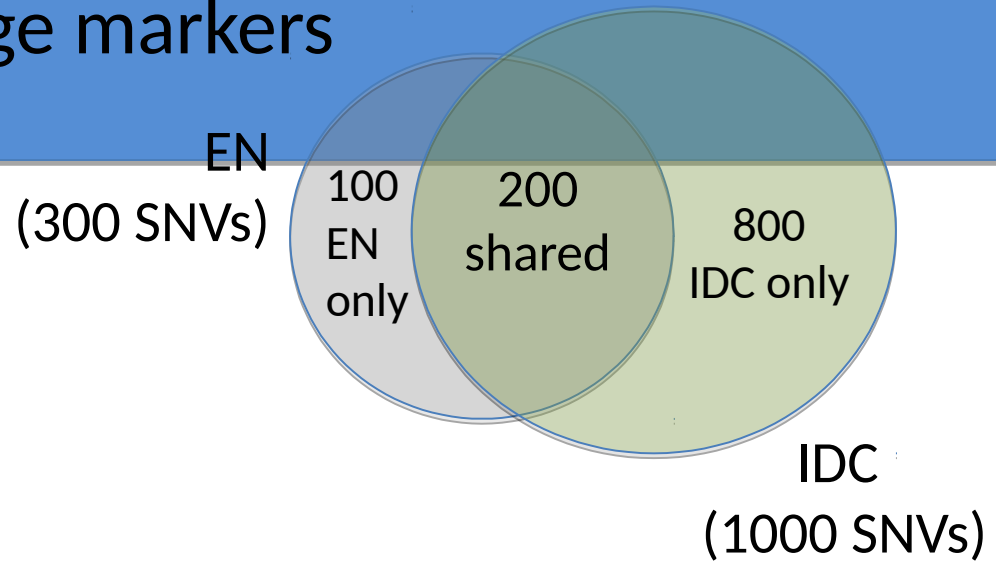
Pipeline

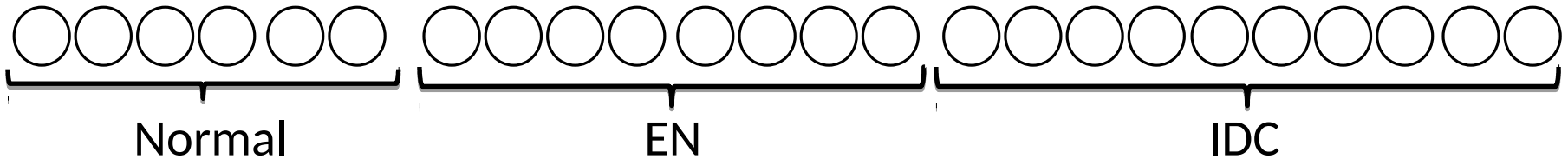
Alignment → PCR duplication removal → Base quality Recalibration → Local Realignment → GATK Unified-Genotyper → Filters → Rescue → Final Variants Set → Germline / Somatic

Somatic SNVs as lineage markers

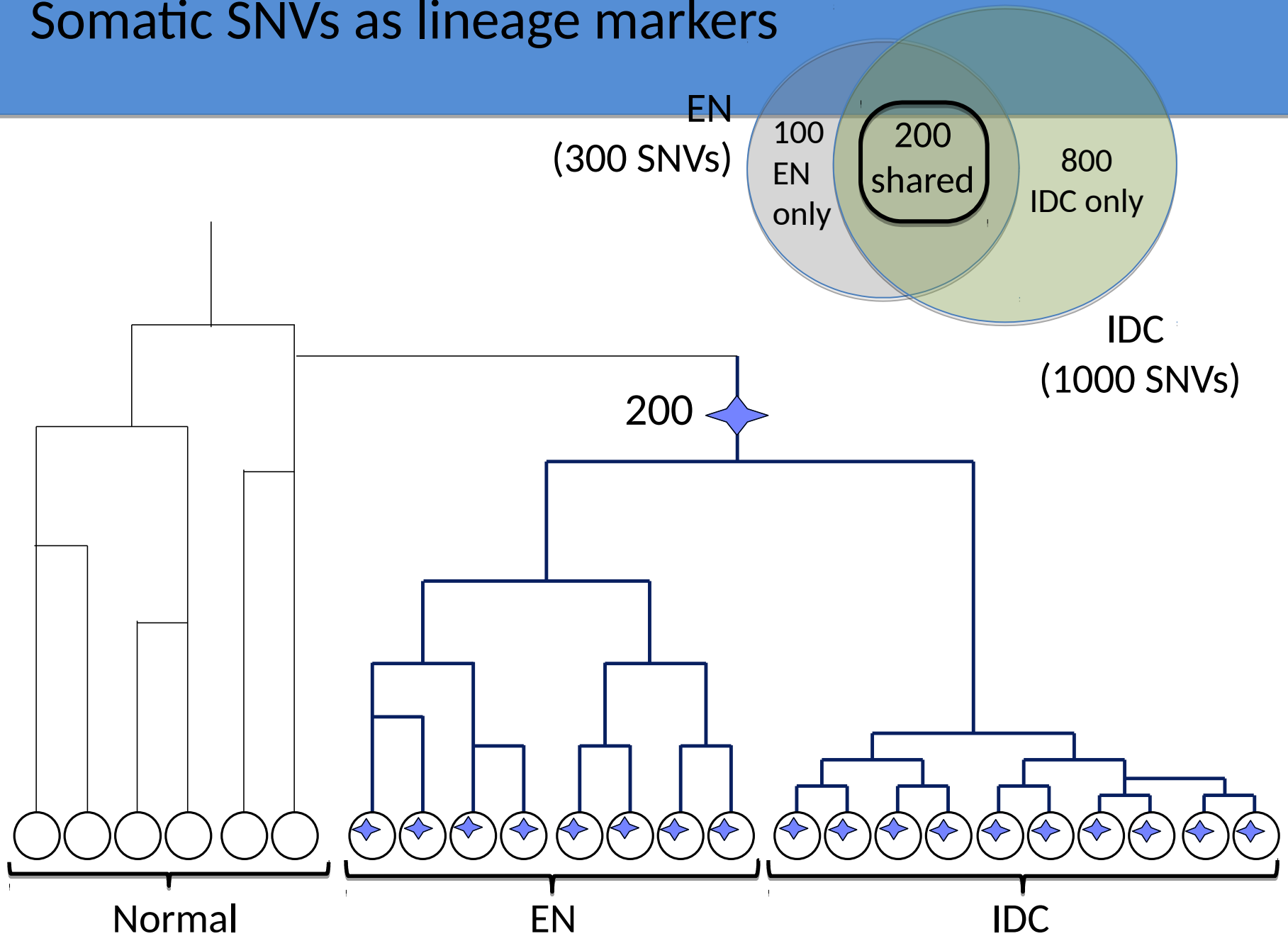


Somatic SNVs as lineage markers

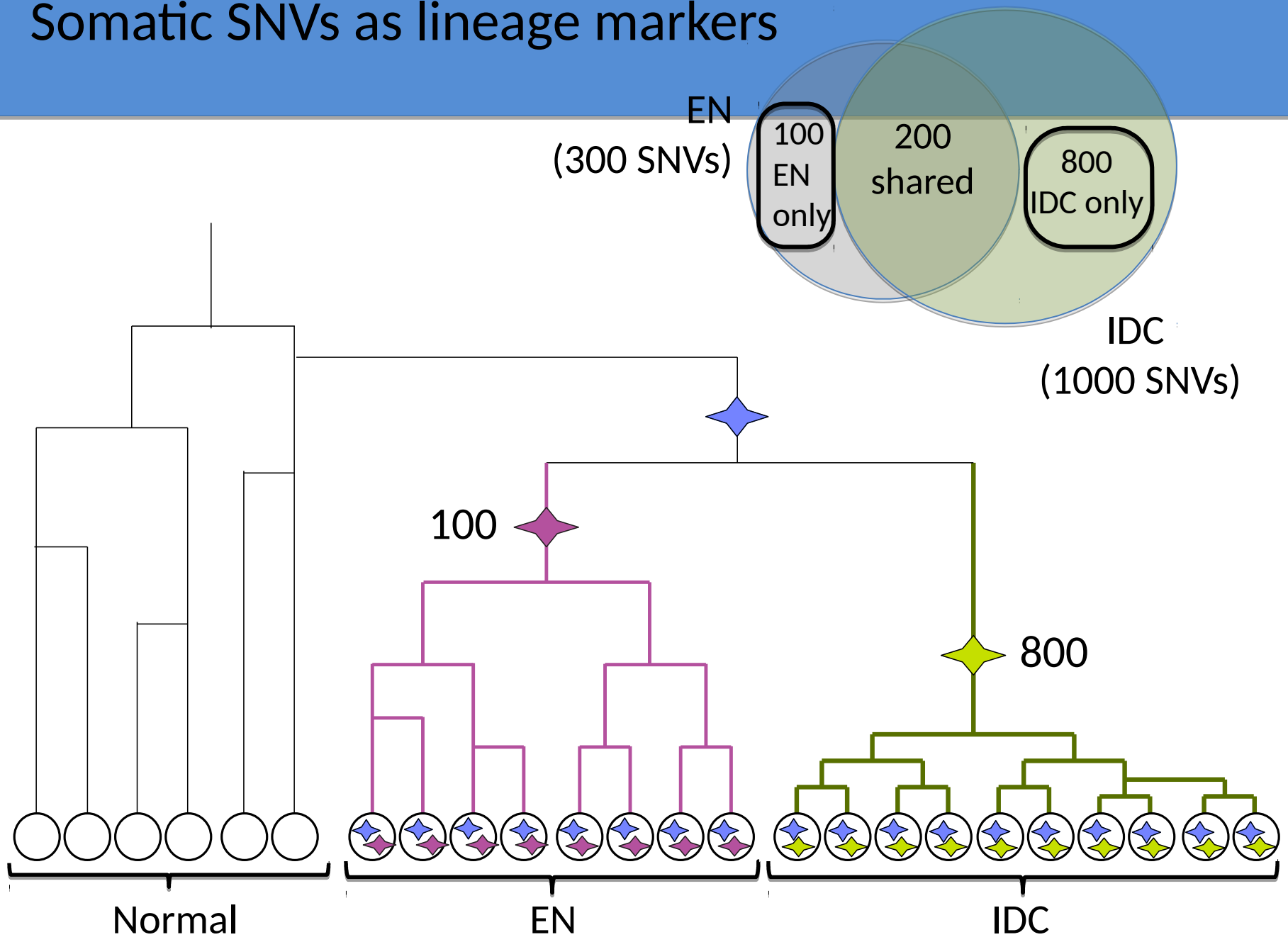




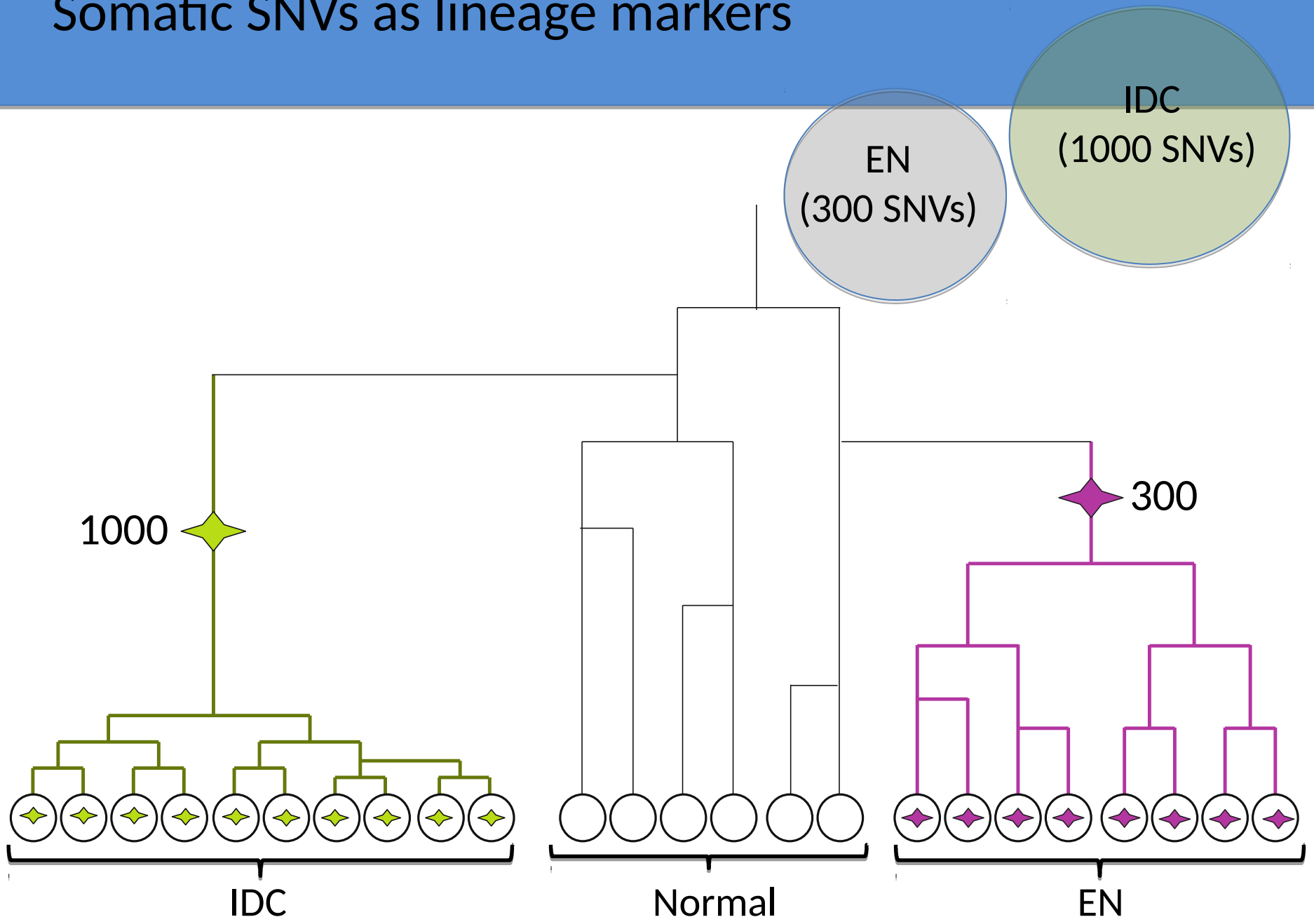
Somatic SNVs as lineage markers



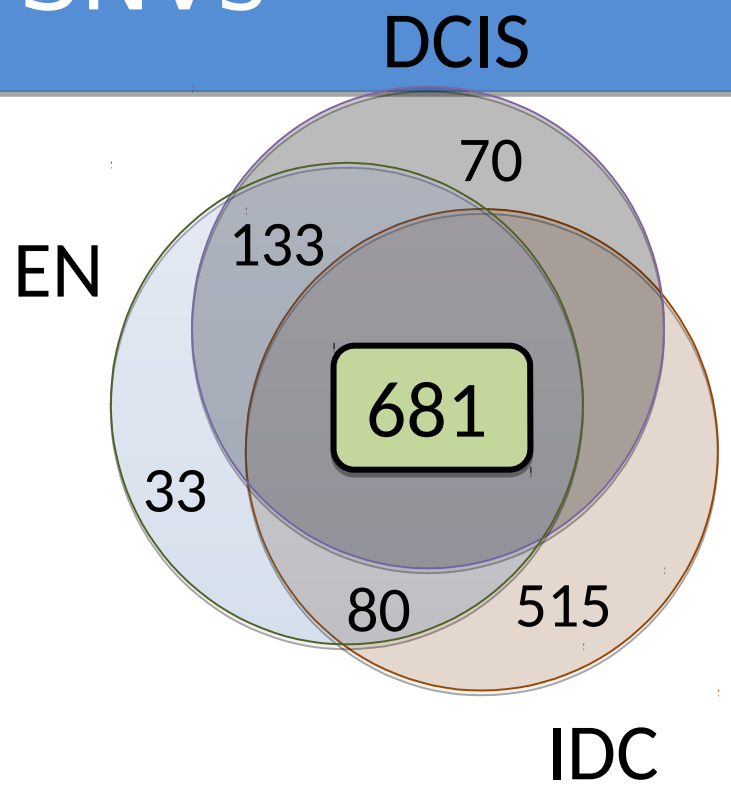
Somatic SNVs as lineage markers



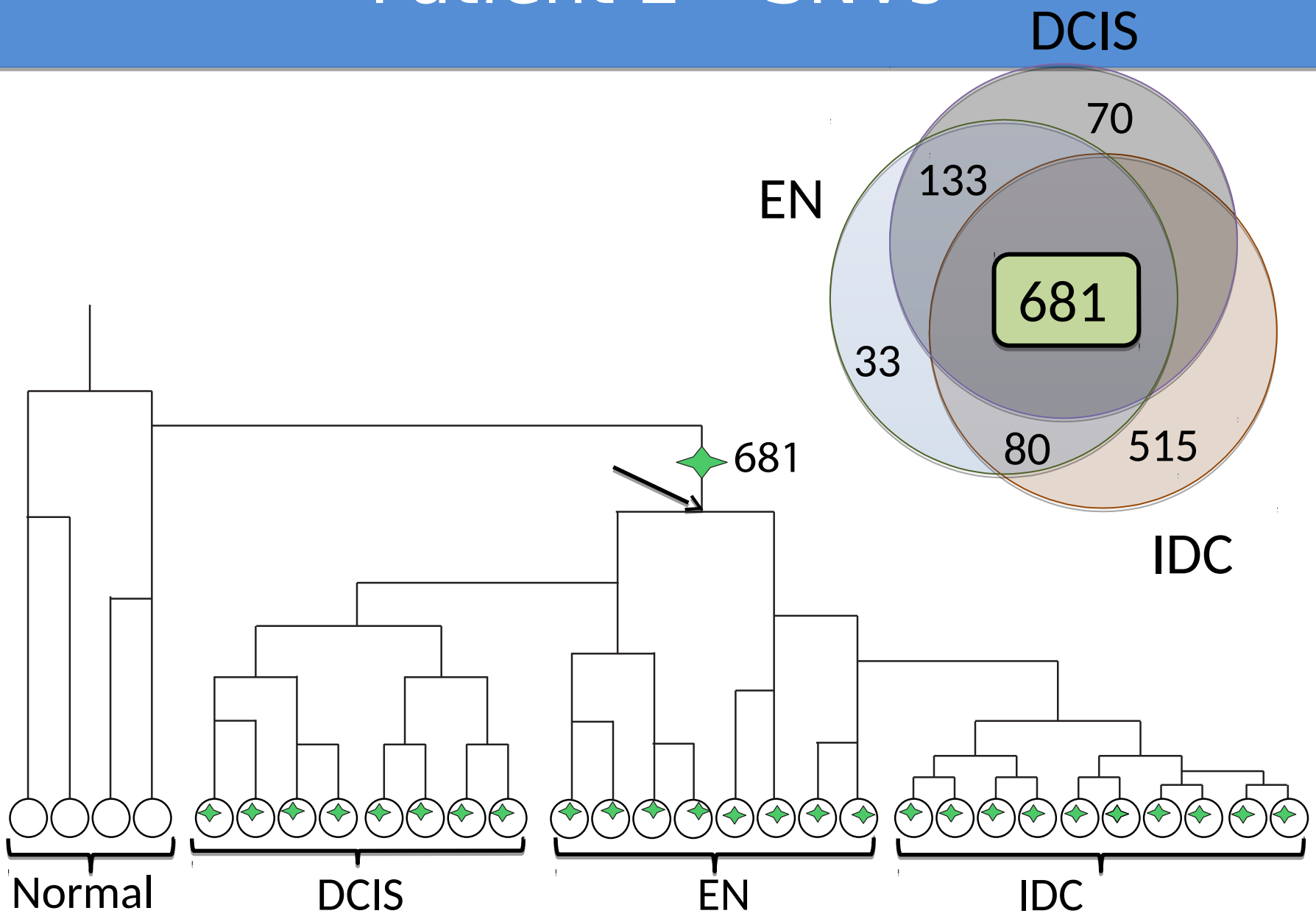
Somatic SNVs as lineage markers



Patient 2 – SNVs

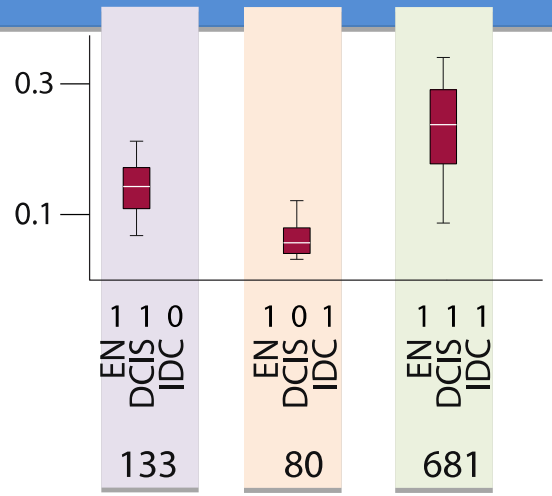


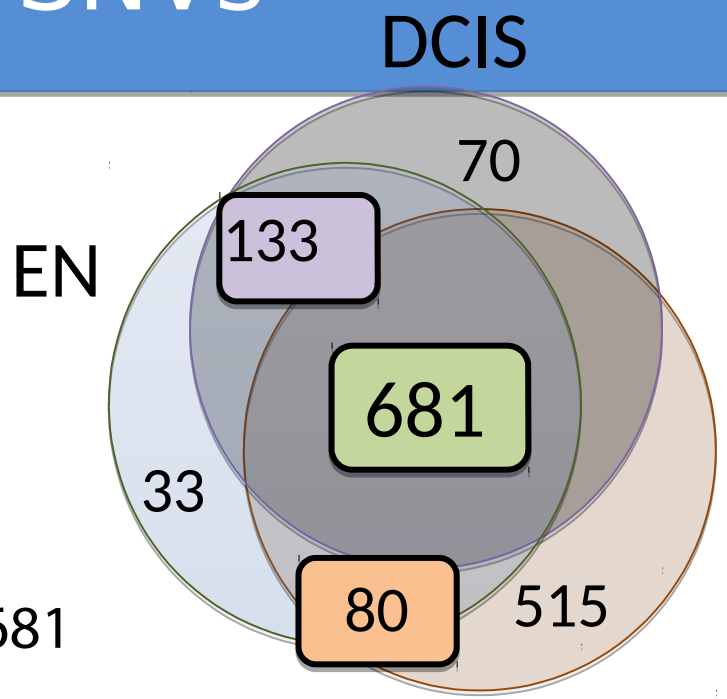
Patient 2 – SNVs

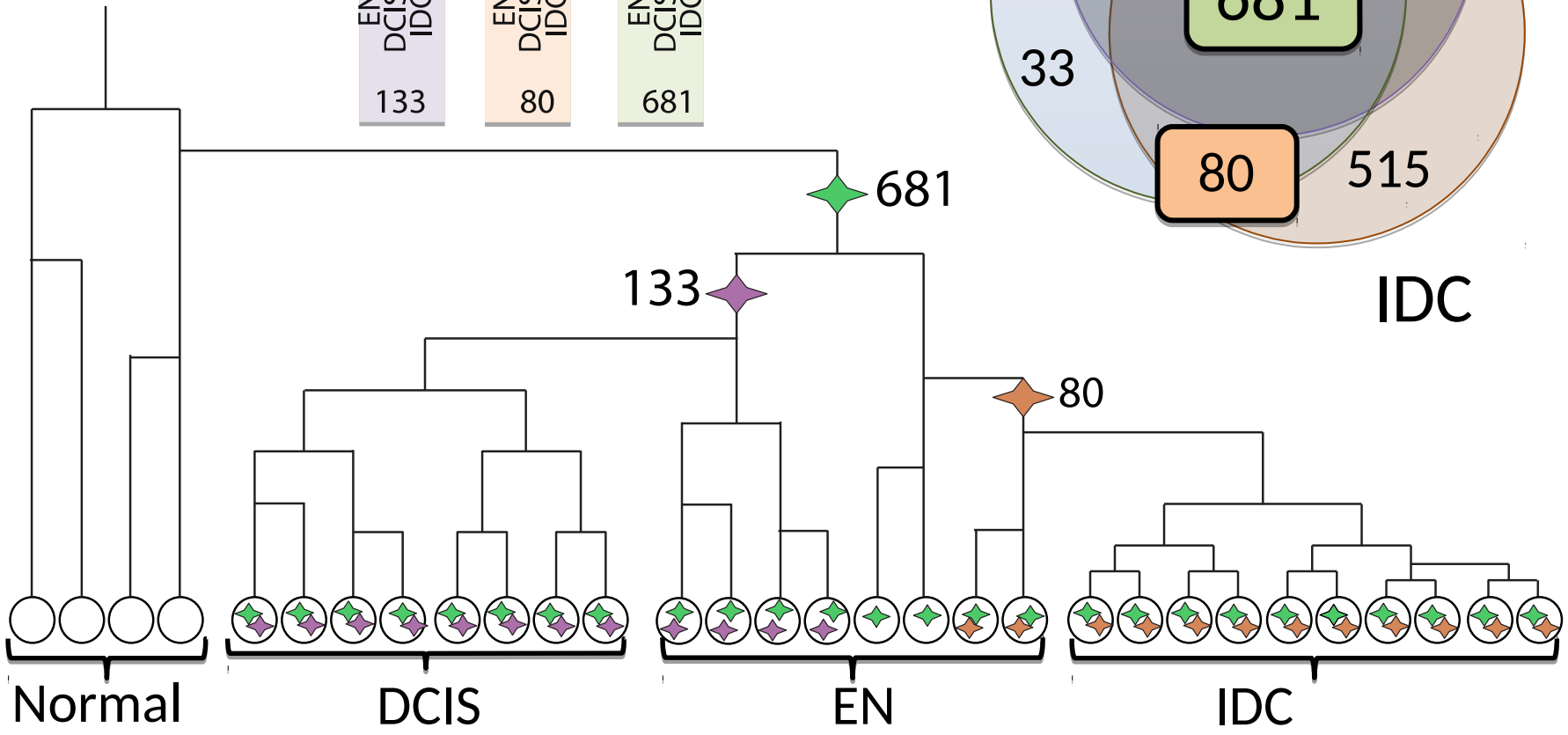


Patient 2 – SNVs

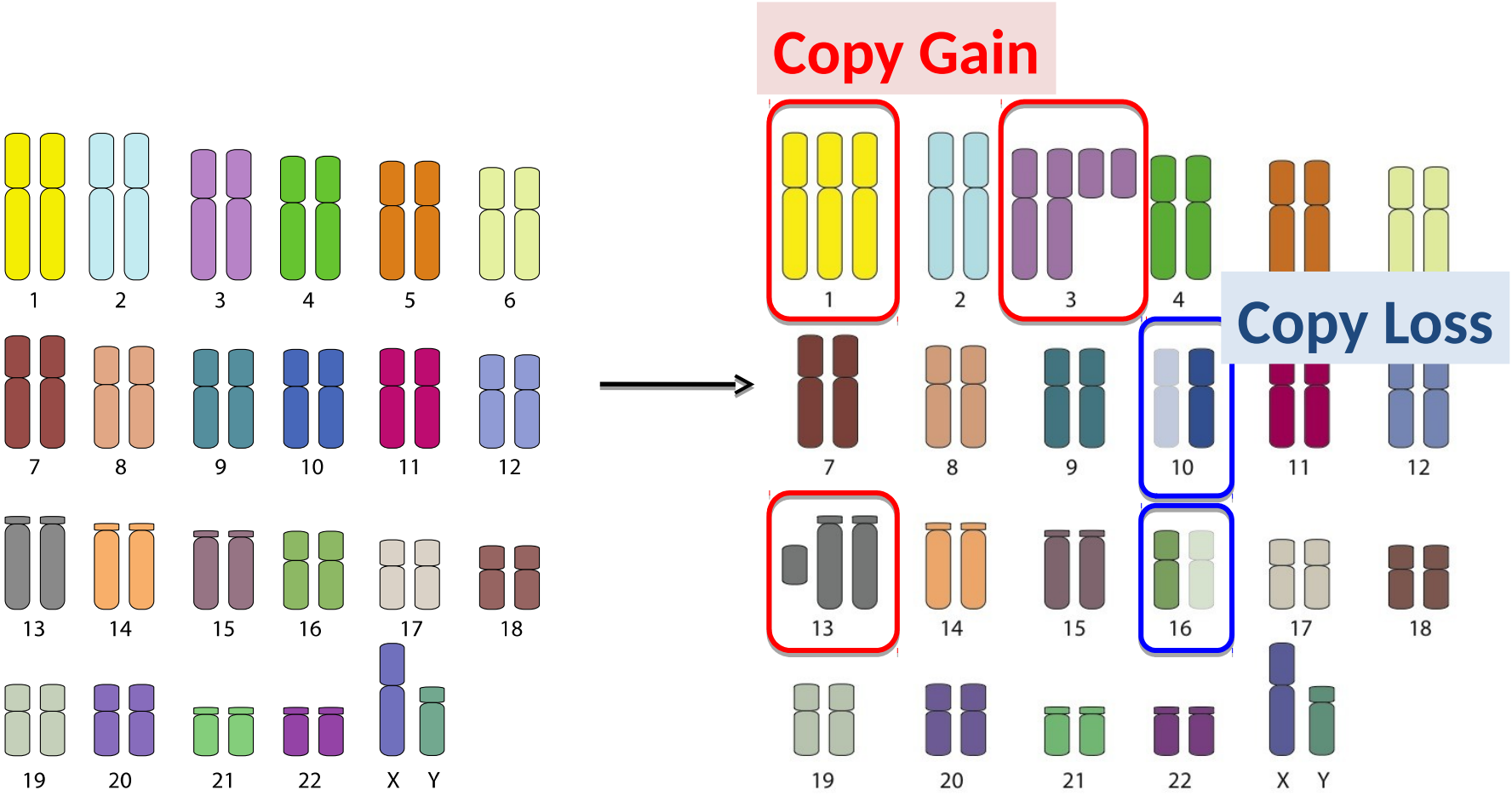
Alternate Allele Fraction in EN



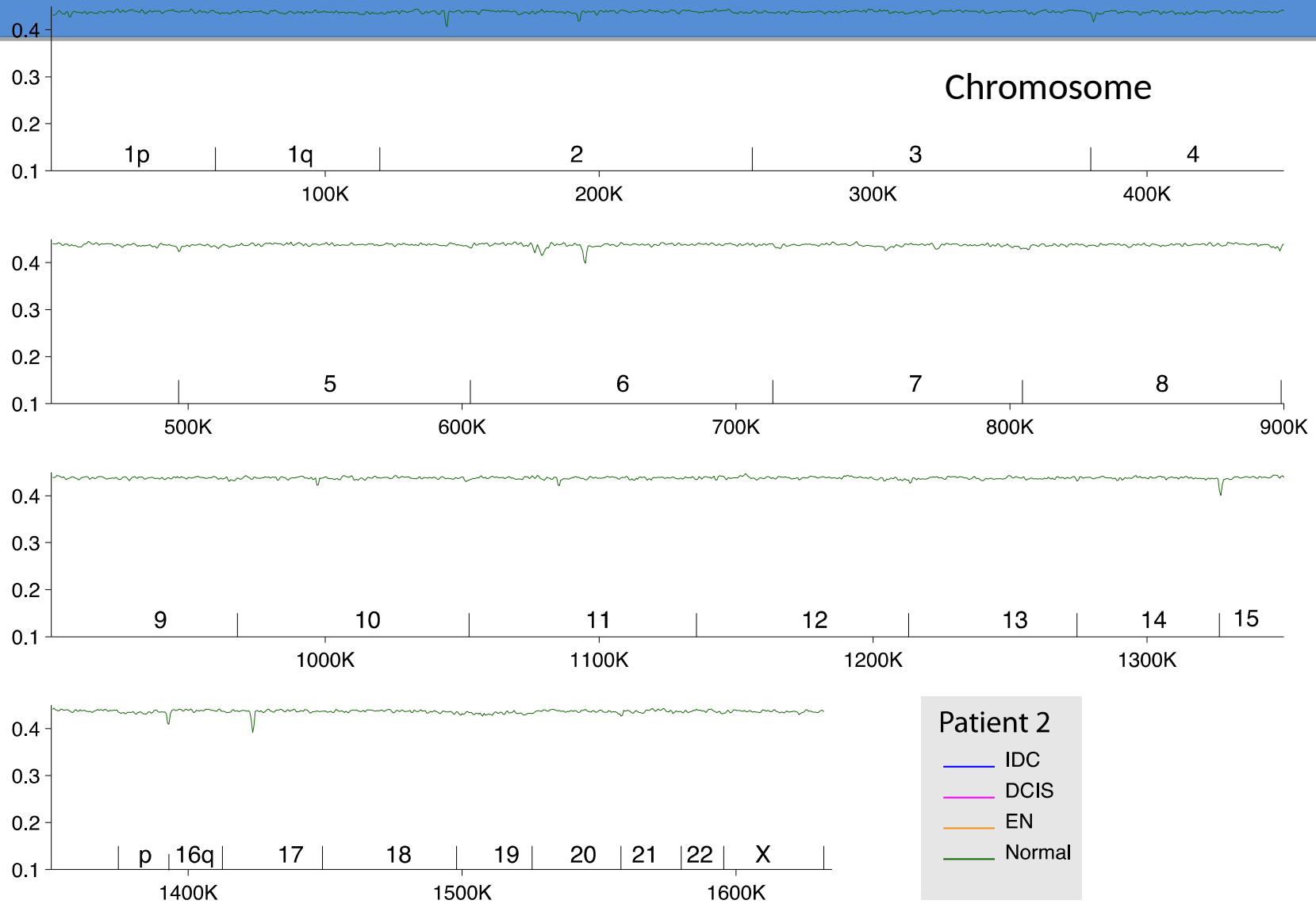




Aneuploidies

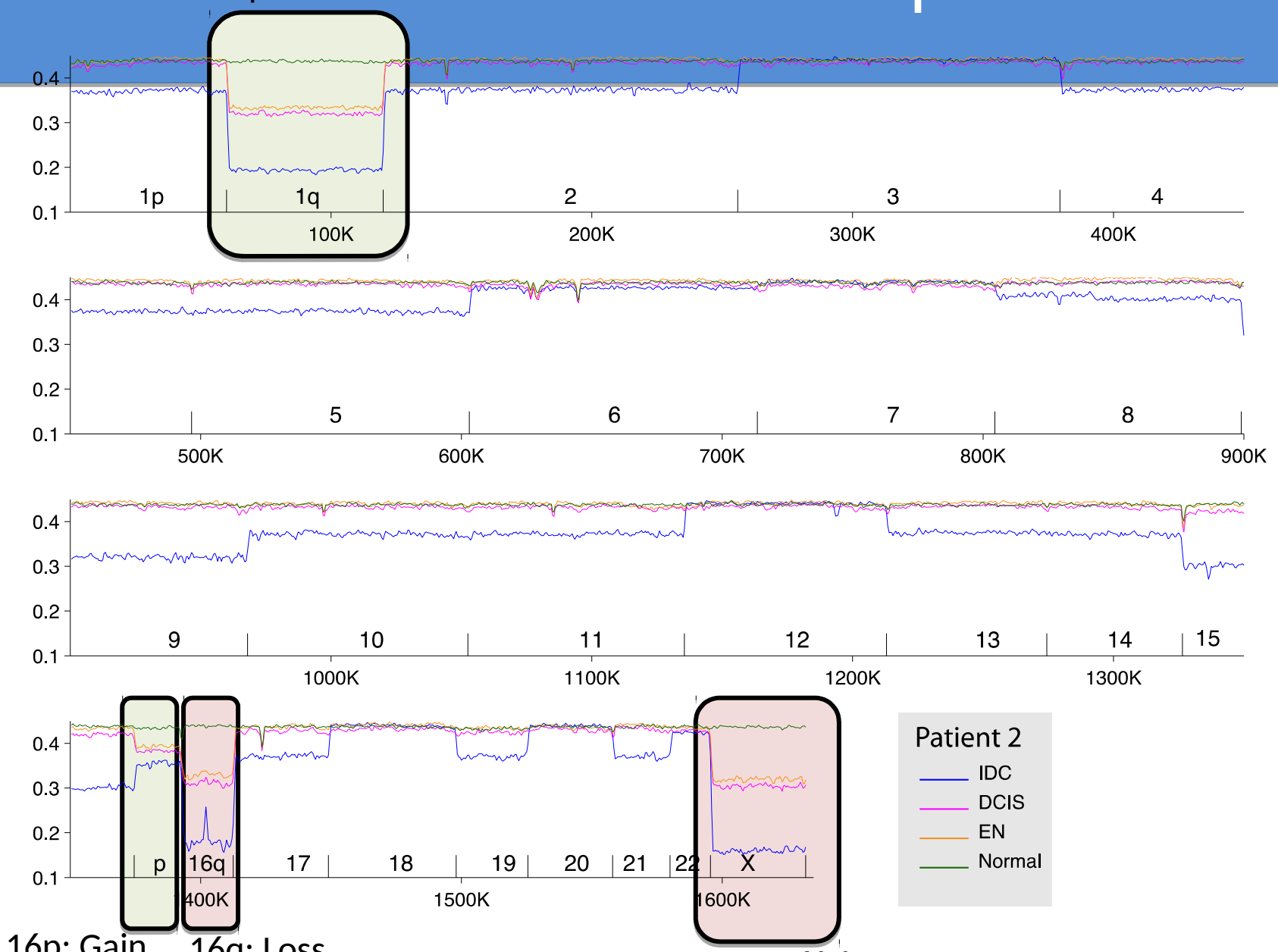


Patient 2 - Aneuploidies

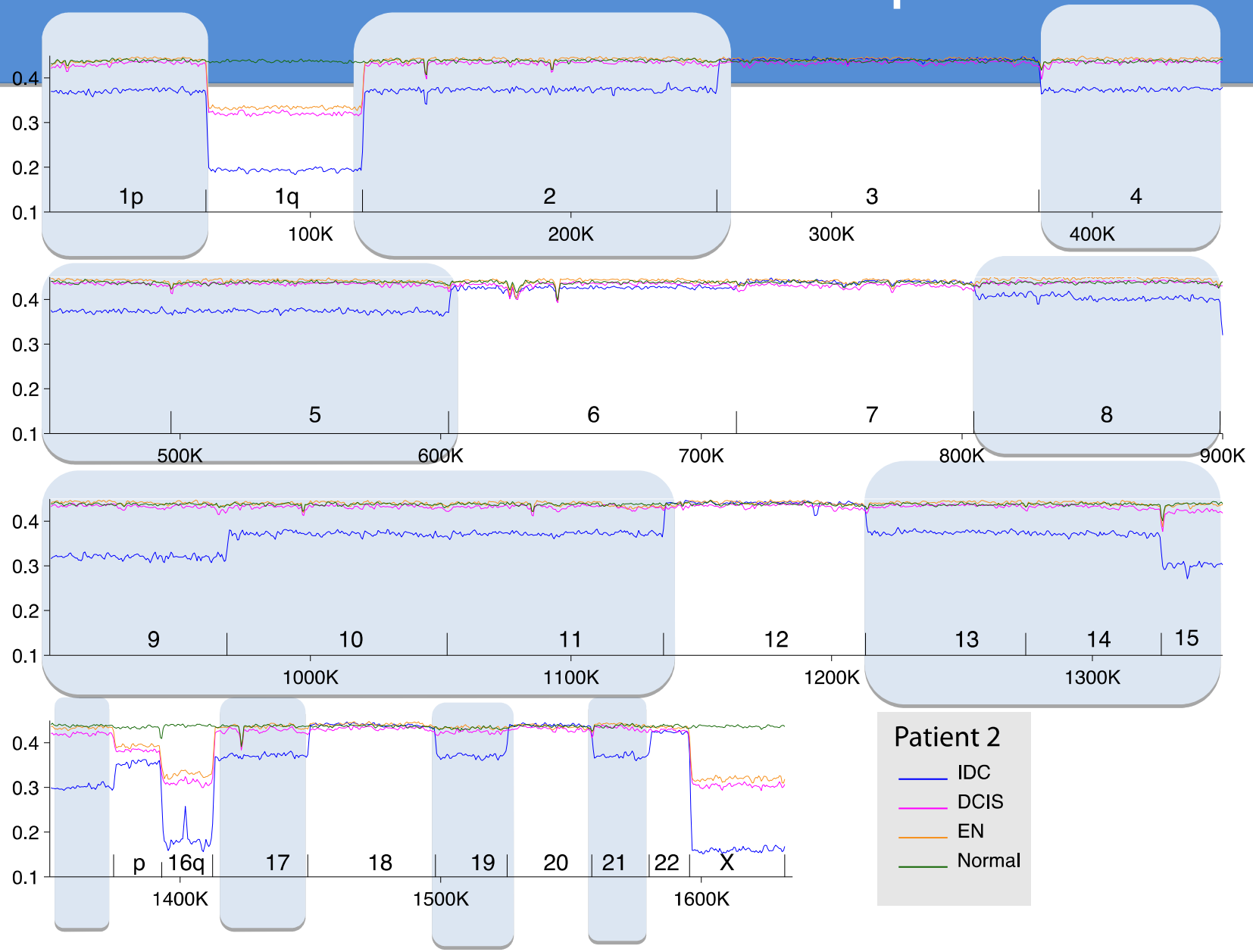


Plots are windows of 1000 SNPs overlapping by 500

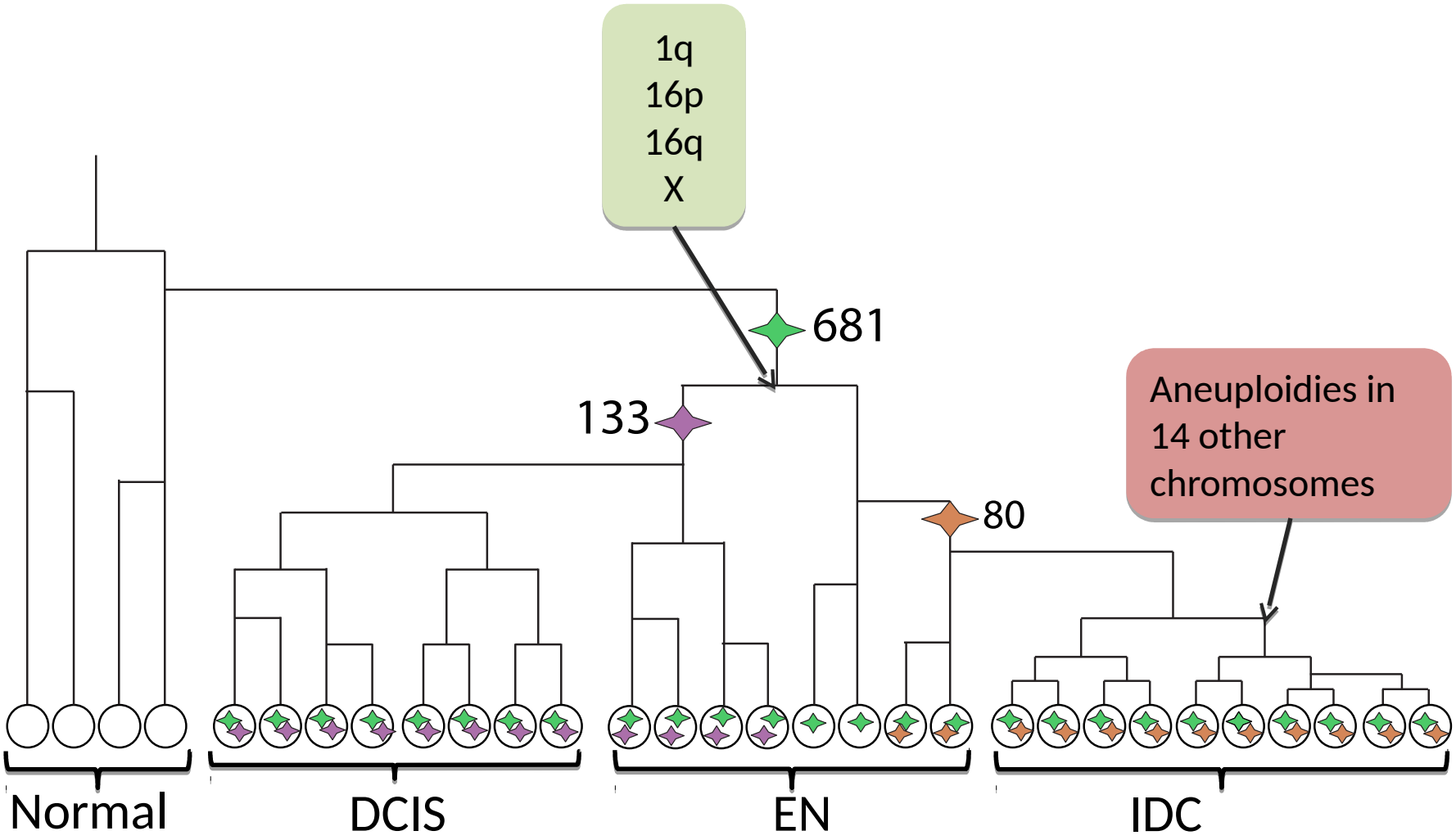
Patient 2 - Aneuploidies



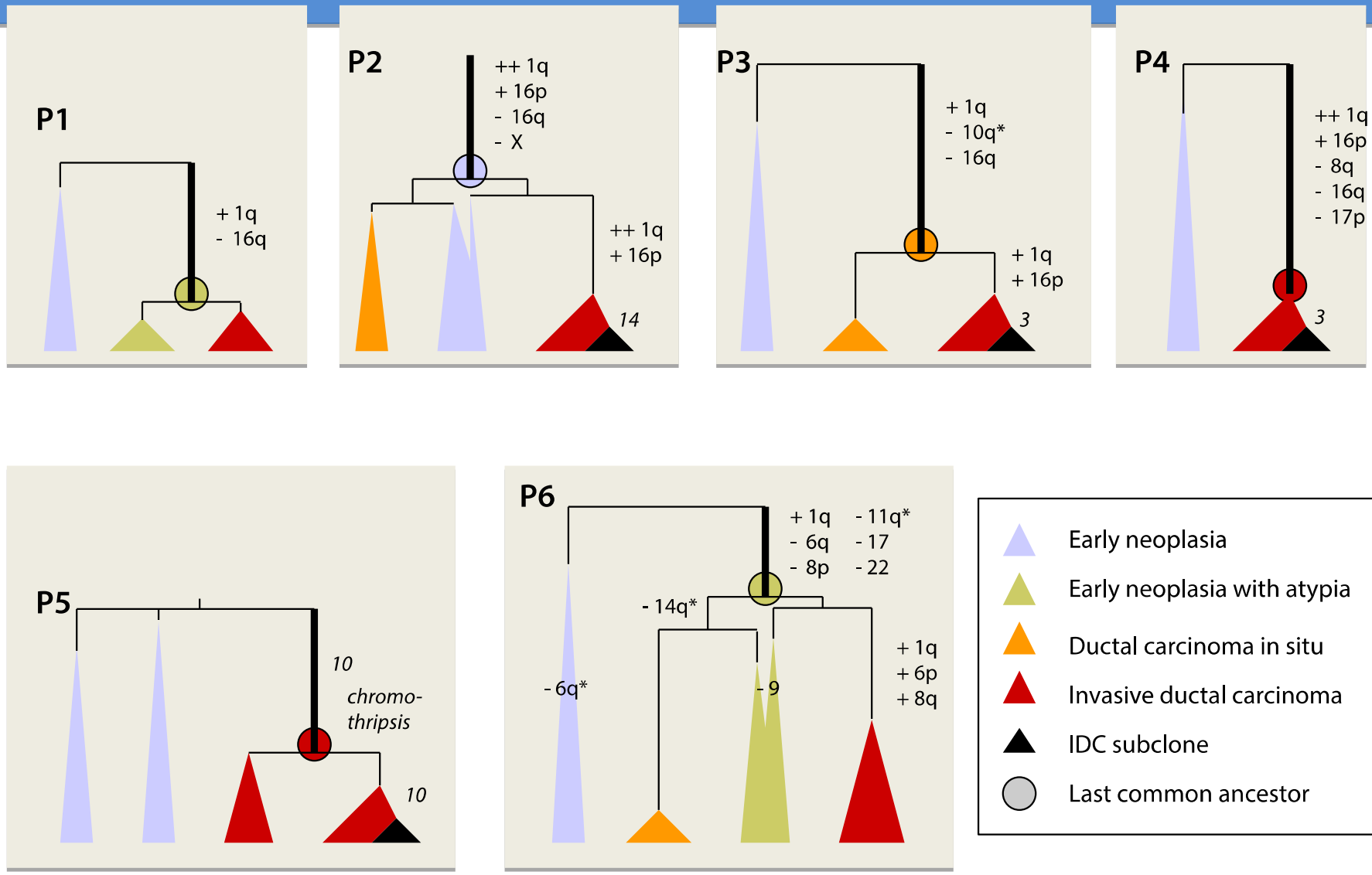
Patient 2 - Aneuploidies



Patient 2 – Final Tree

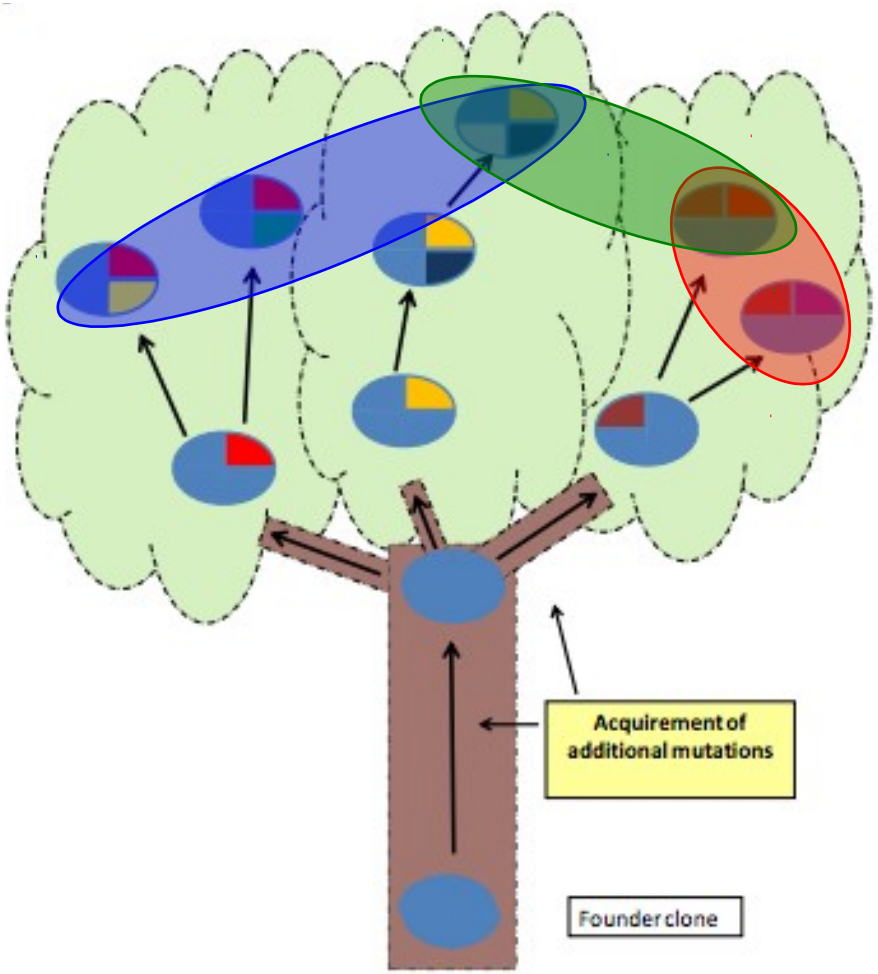


Results

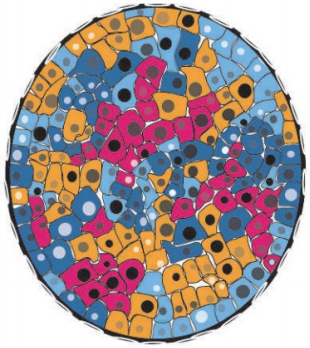

P1
+ 1q
- 16q
P2
++ 1q
+ 16p
- 16q
- X
++ 1q
+ 16p
14
P3
+ 1q
- 10q*
- 16q
+ 1q
+ 16p
3
P4
++ 1q
+ 16p
- 8q
- 16q
- 17p
3
P5
10
chromo-
thripsis
10
P6
+ 1q
- 6q
- 8p
- 11q*
- 17
- 22
- 14q*
- 6q*
- 9
+ 1q
+ 6p
+ 8q
Early neoplasia
Early neoplasia with atypia
Ductal carcinoma in situ
Invasive ductal carcinoma
IDC subclone
Last common ancestor

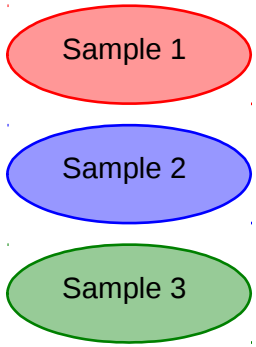
Automated Inference of Multi-Sample Cancer Phylogenies

Victoria Popic
Raheleh Salari



Branched tree model



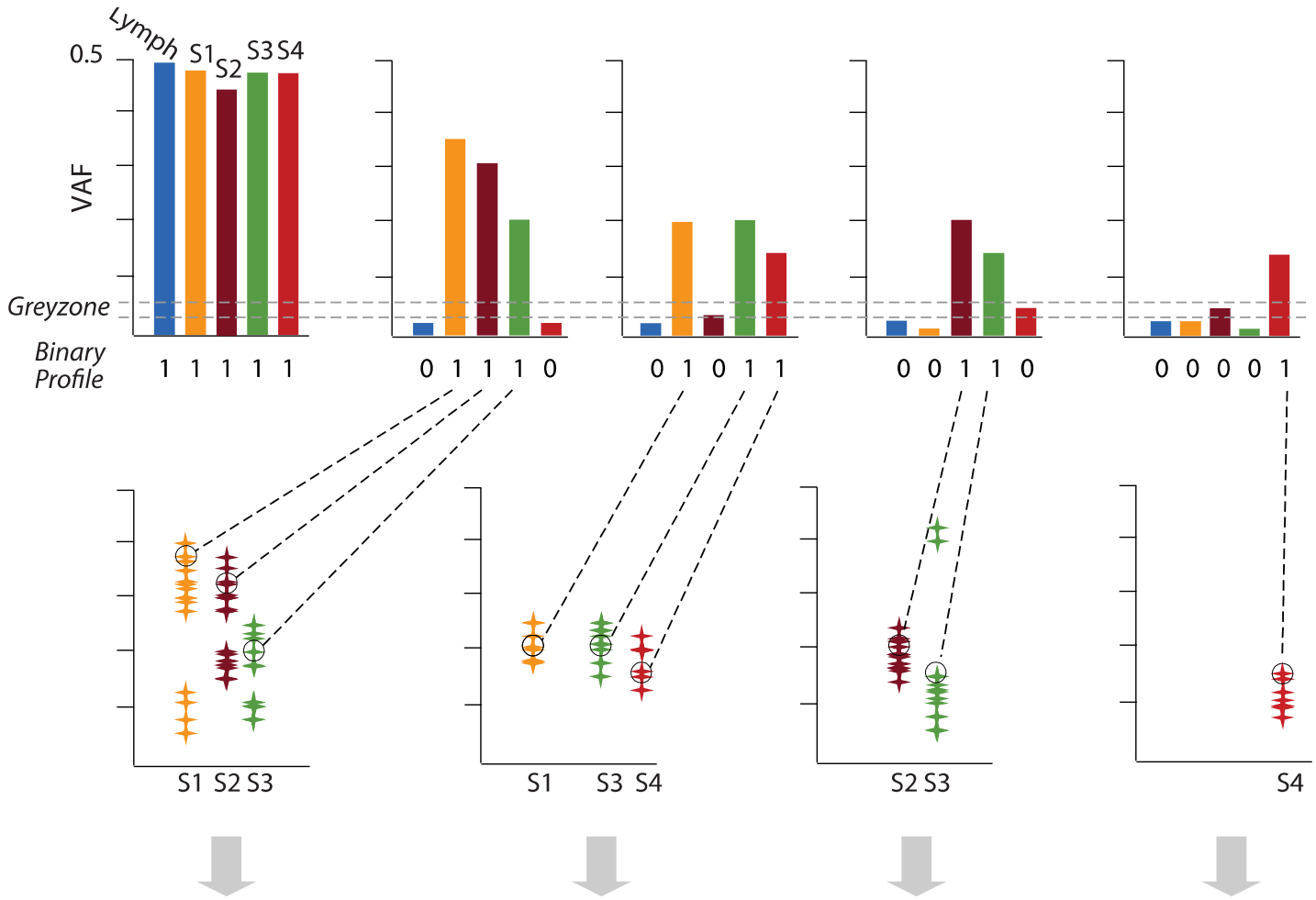


SMutH: Somatic Mutation Hierarchies

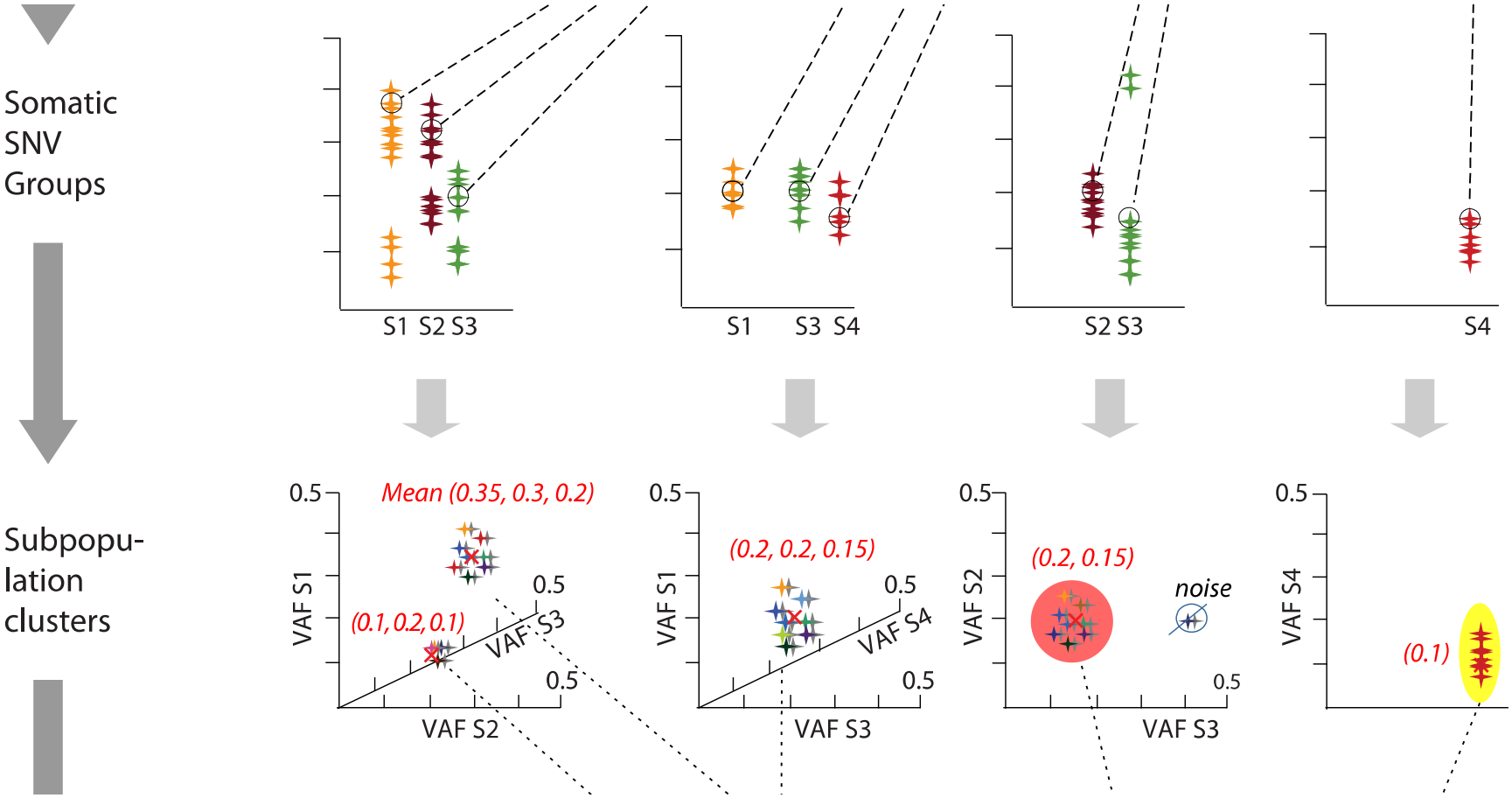
VAF profiles of SNVs across samples

SNV VAF profiles across samples S0 ... S4

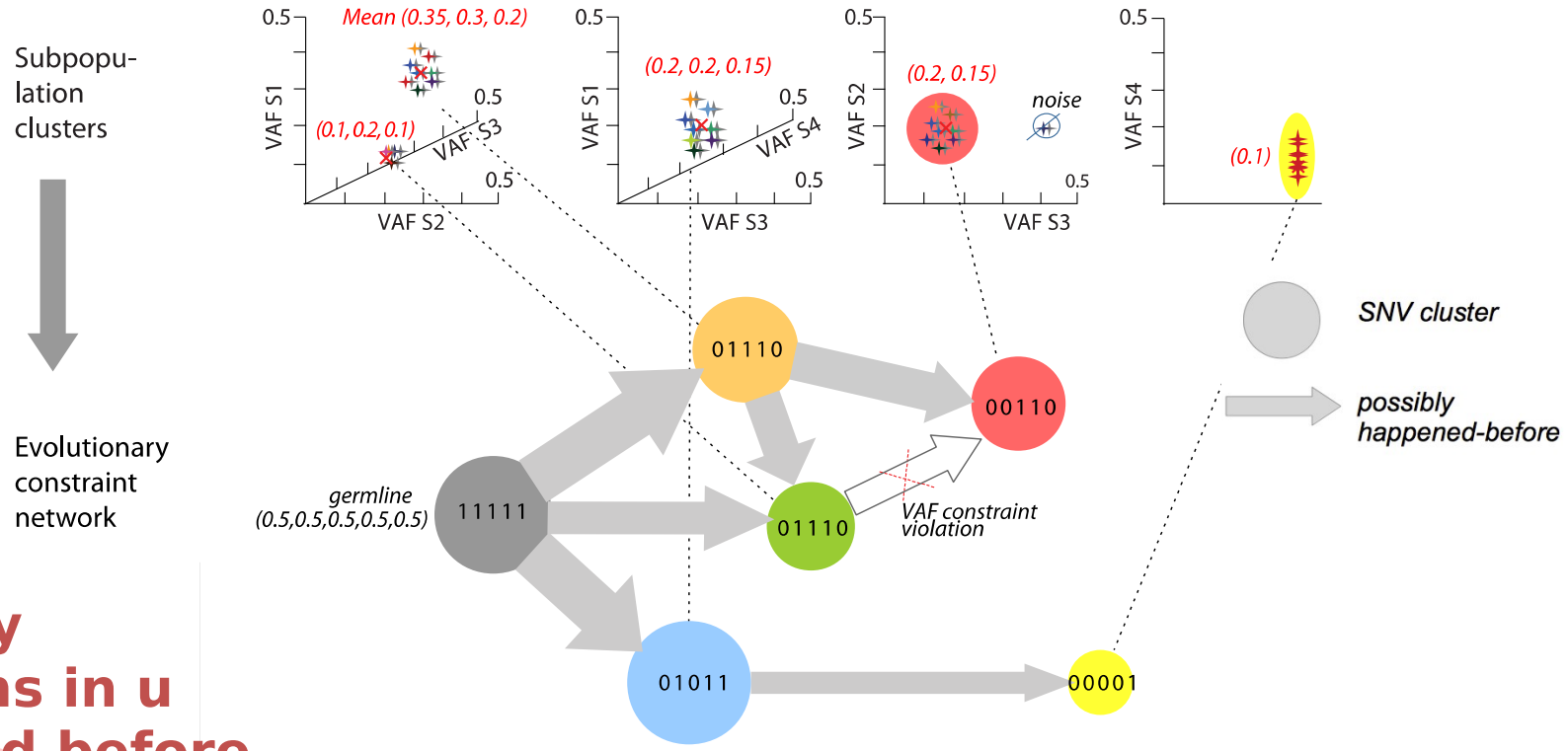
Somatic SNV Groups



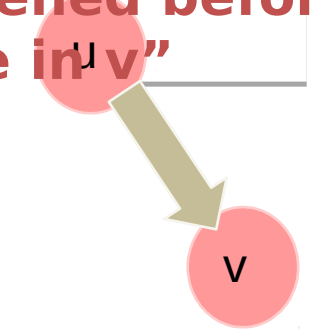
VAF profiles of SNVs – Clustering



Cell-Lineage VAF Constraint



"Possibly mutations in u happened before those in v"

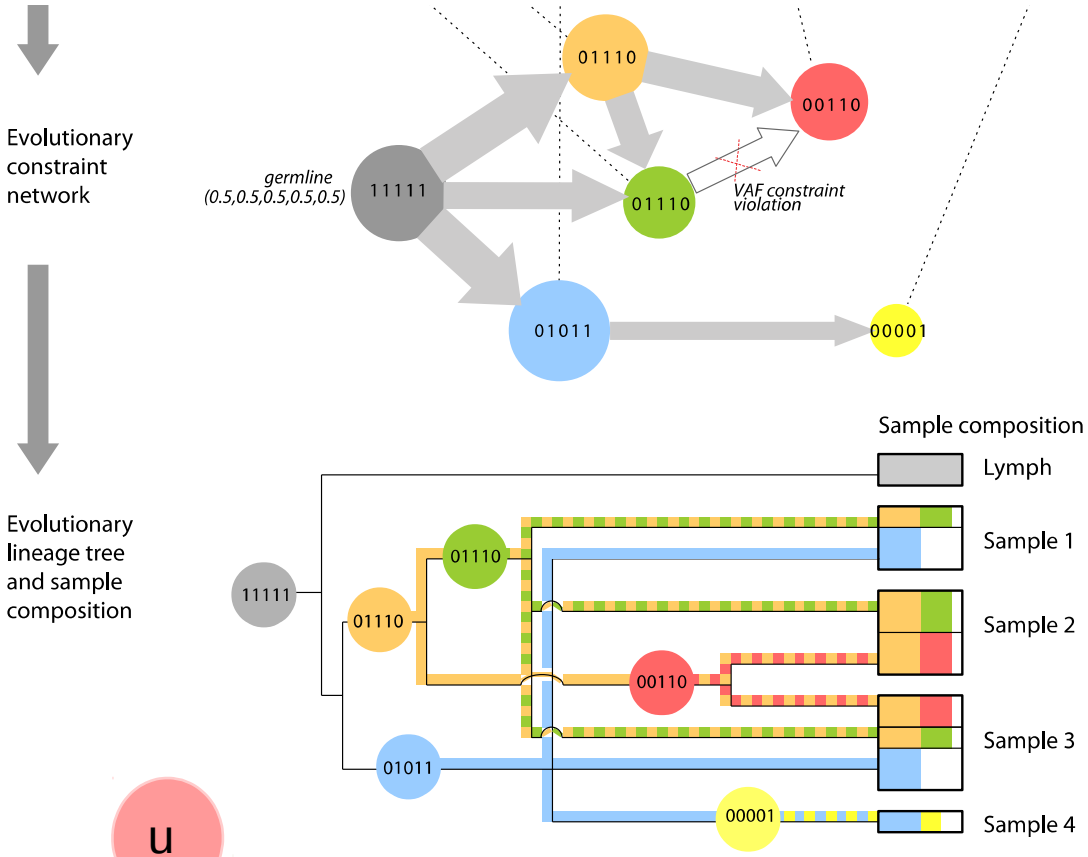


Edge u → v : $\forall i \in samples$:

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

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

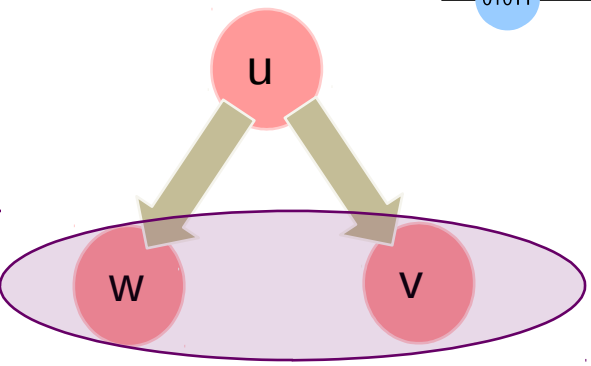
Tree Construction



Find all spanning trees that satisfy VAF constraints

(extension of Gabow&Myers spanning tree search algorithm)

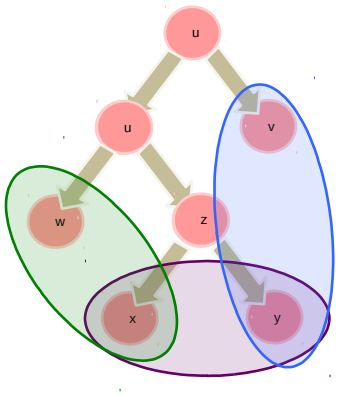
Rank trees according to their agreement with VAFs



For each node u and its children C :

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

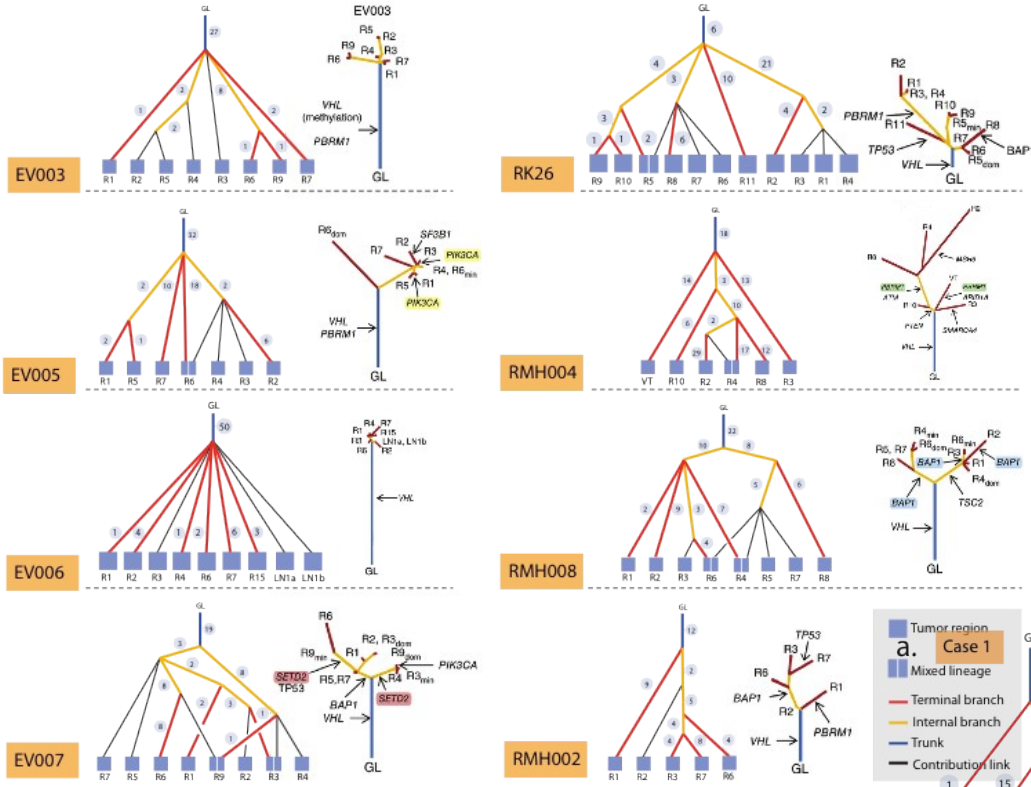
Simulation Results



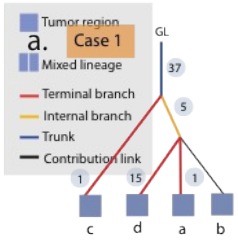
	VAF Stdev 0.03				VAF Stdev 0.1			
# Samples	Pred	Branch	Shared Edges	# Trees	Pred	Branch	Shared Edges	# Trees
5	0.95	0.91	0.81	85	0.95	0.85	0.80	43
6	0.91	0.93	0.82	77	0.91	0.91	0.79	43
7	0.89	0.91	0.83	67	0.88	0.89	0.83	38
8	0.89	0.96	0.84	64	0.93	0.95	0.87	43
9	0.88	0.95	0.81	60	0.87	0.97	0.82	34
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

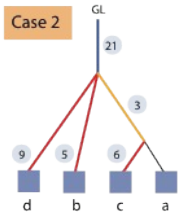
Pred: pairs of nodes ordered correctly
Branch: pairs of nodes correctly assigned to separate branches
Shared edges: edges shared between true and reconstructed trees

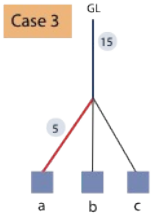
Reconstruction of Lineage Trees in Recent Literature

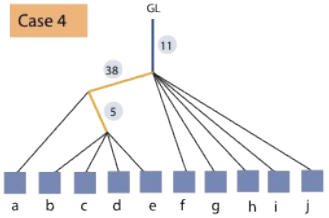


ccRCC Study of Renal Carcinoma by Gerlinger *et. al* (2014)

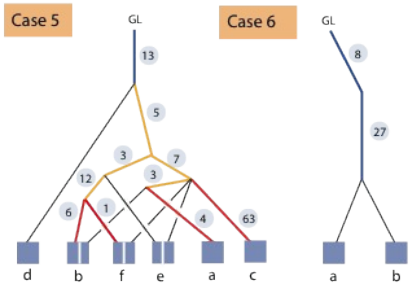


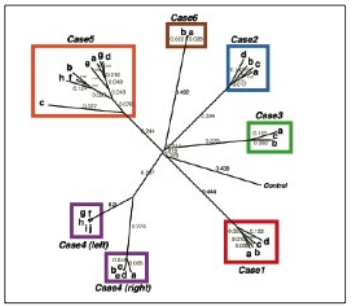




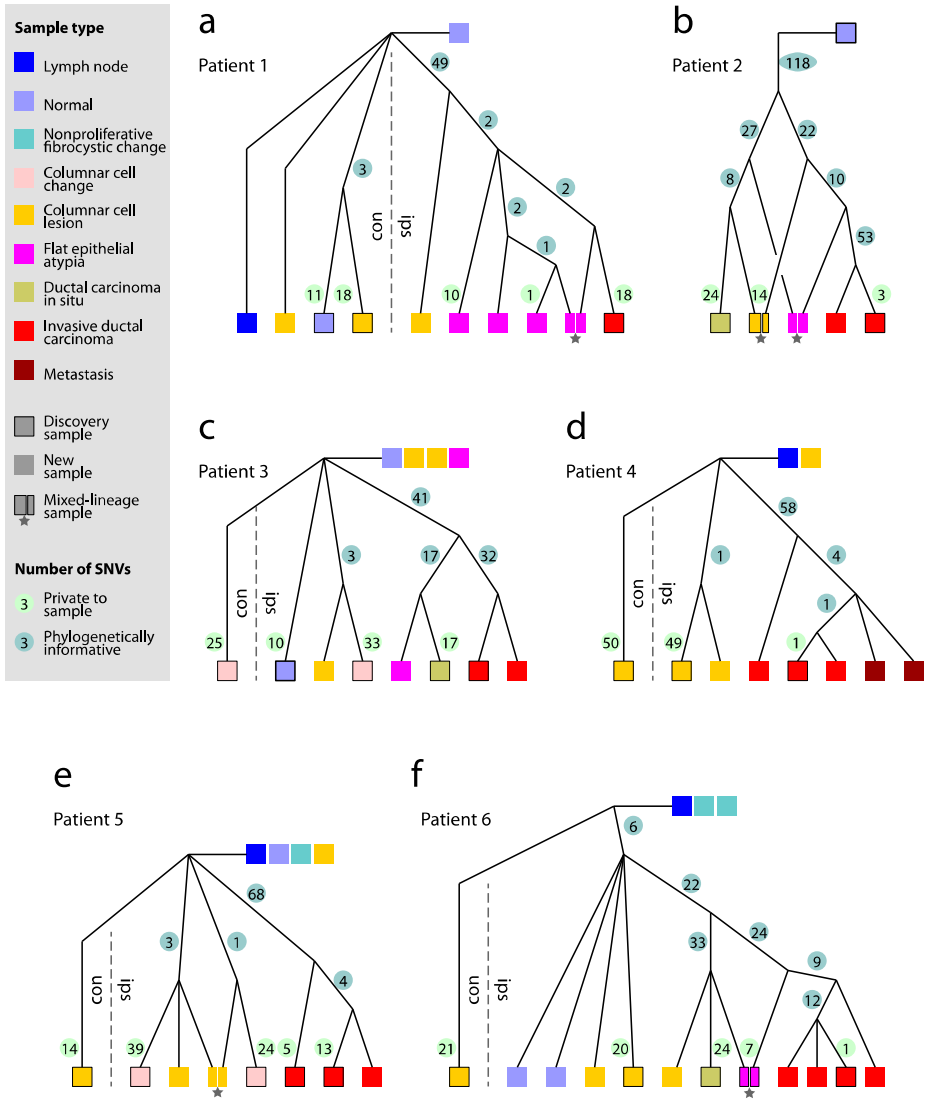


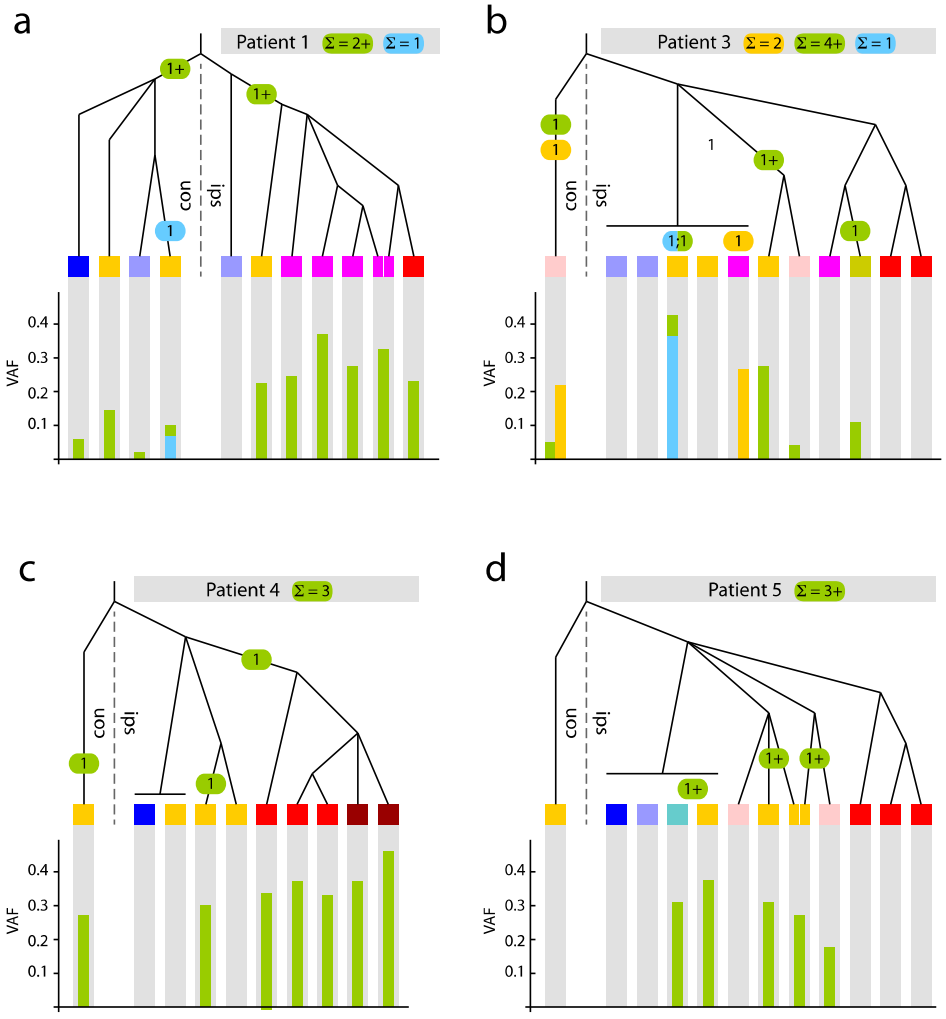
HGSC Study of Ovarian Cancer Bashashati *et. al* (2013)





Expanded Breast Cancer Lineage Trees

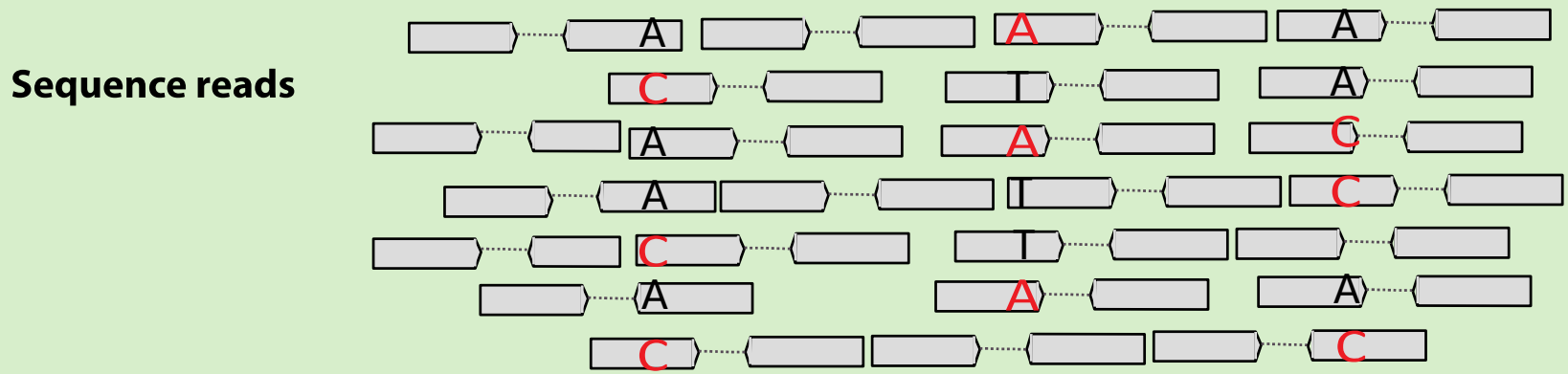




PIK3CA H1047R

PIK3CA H1047L

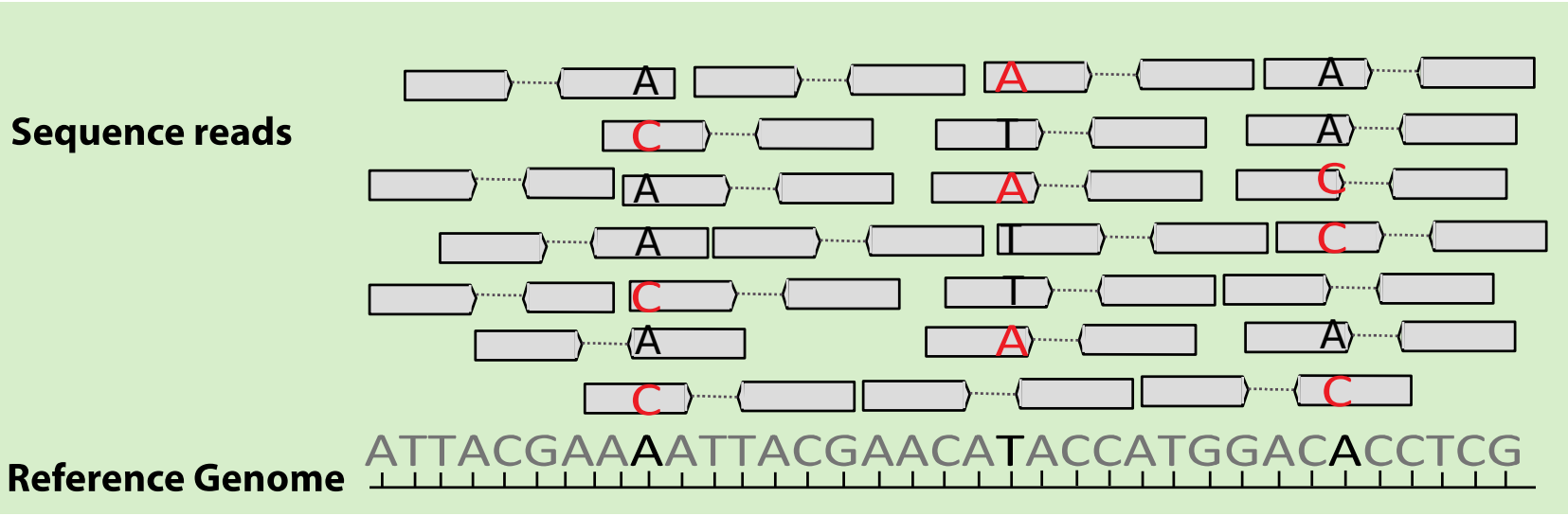
What is a haplotype?



Genotype A/C A/T A/C

A	A	C	OR	A	T	C	OR	A	T	A	OR	A	A	A	?
C	T	A		C	A	A		C	A	C		C	T	A	
correct				wrong				wrong				wrong			

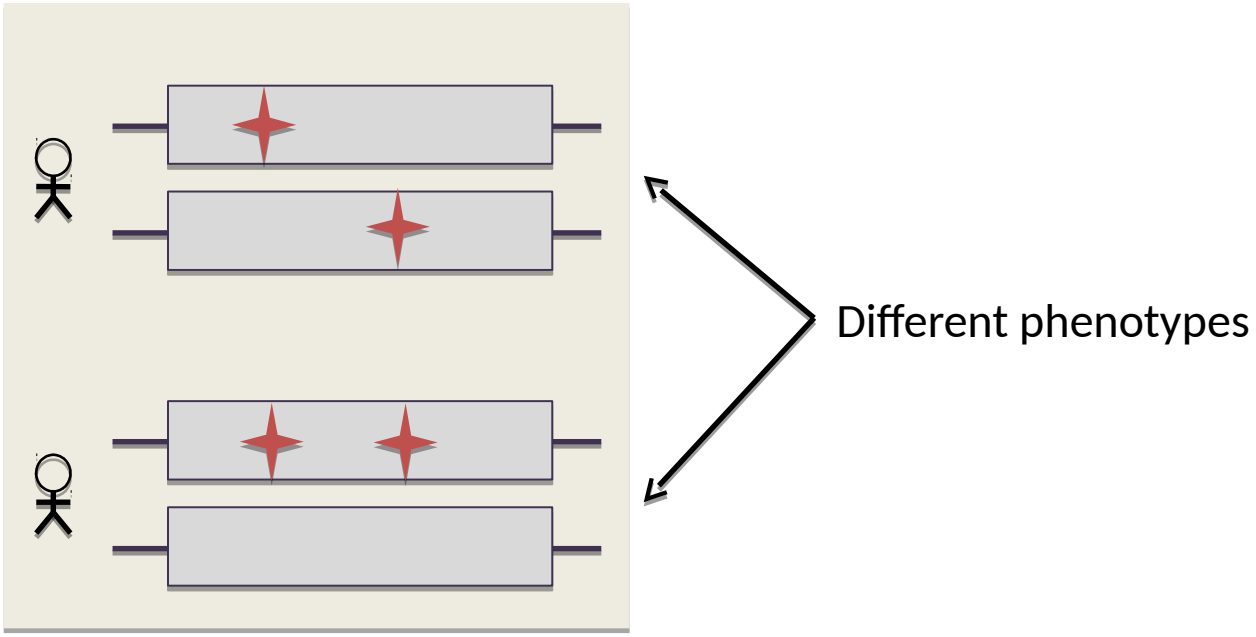
What is a haplotype?



Phasing is the process of recovering haplotypes from genotype data

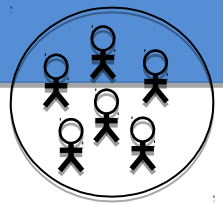
The importance of phase information

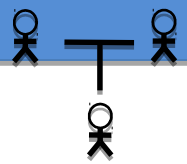
- Compound Heterozygosity



- Two-hit cancer model

Different Approaches for Phasing





Population	Genetic	Molecular
Unrelated duos, trios	Pedigrees Trios	Each physical read is a single molecule that can be directly phased
Population inference	IBD analysis, inheritance state analysis	Read Assembly

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