Session 7: Somatic Copy Number Alterations

Emerging Approaches For Tumor Analyses in Epidemiological Studies

March 08, 2023
9:30 AM- 12:00 PM

Session Overview

- Introduction to Somatic Copy Number Alterations (SCNAs)
- Methods for detecting SCNAs
- SCNA Validation & Visualization
- Challenges for SCNA detection
- Common SCNA analysis in cancer genomics
- Significance of SCNAs in cancer

Workshop website: https://nci-iteb.github.io/tumor_epidemiology_approaches/

About me (Azhar Khandekar)

- PhD student jointly supervised by **Ludmil Alexandrov (UCSD)** & **Tere Landi (DCEG)**

- Just arrived at the NCI in February

Research focus:

- Mutational Signatures
- Copy number and structural variants
- **Extrachromosomal DNA**

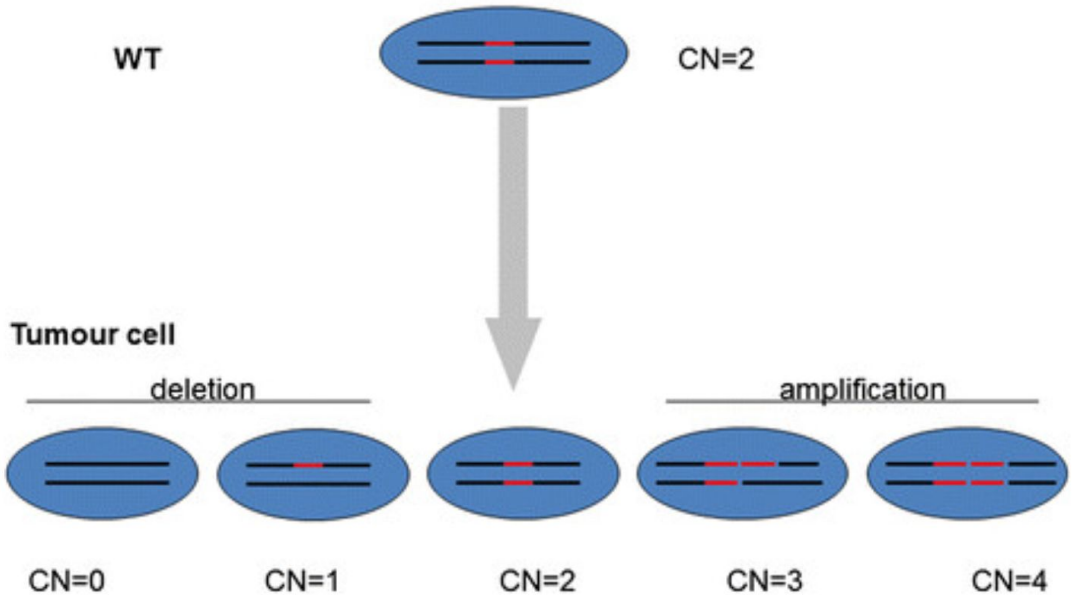
Contact:

akhandek@eng.ucsd.edu
khandekara2@nih.gov

Introduction to Somatic Copy Number Alterations (SCNAs)

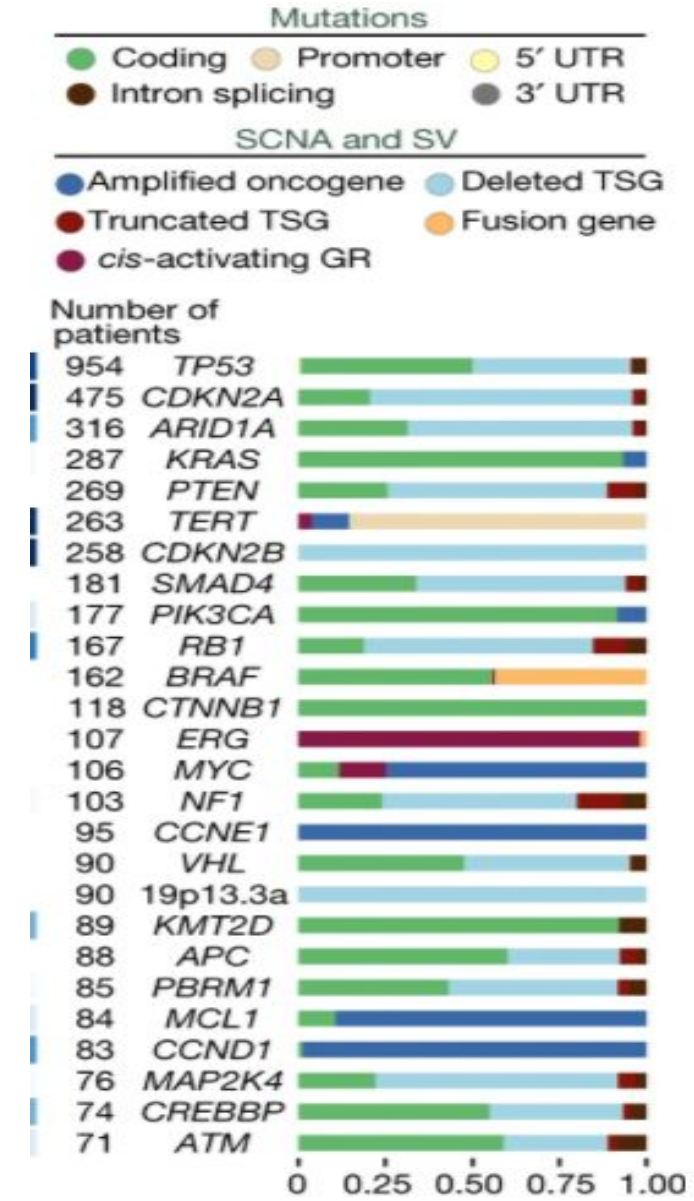
Somatic Copy Number Alteration (SCNA): Background and Motivation

- **Somatic Copy Number Alteration** (**SCNA**) – Operationally defined as **>50bp** alteration**,** post-zygotically in a somatic cell

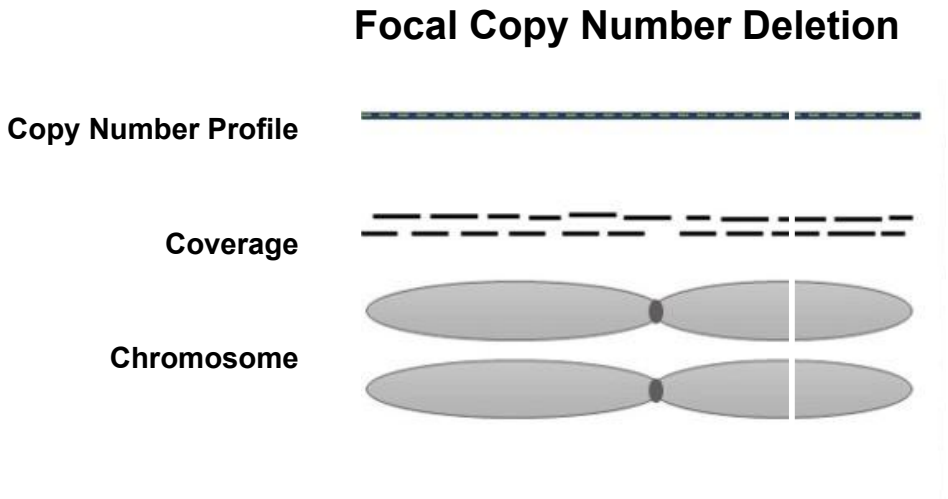


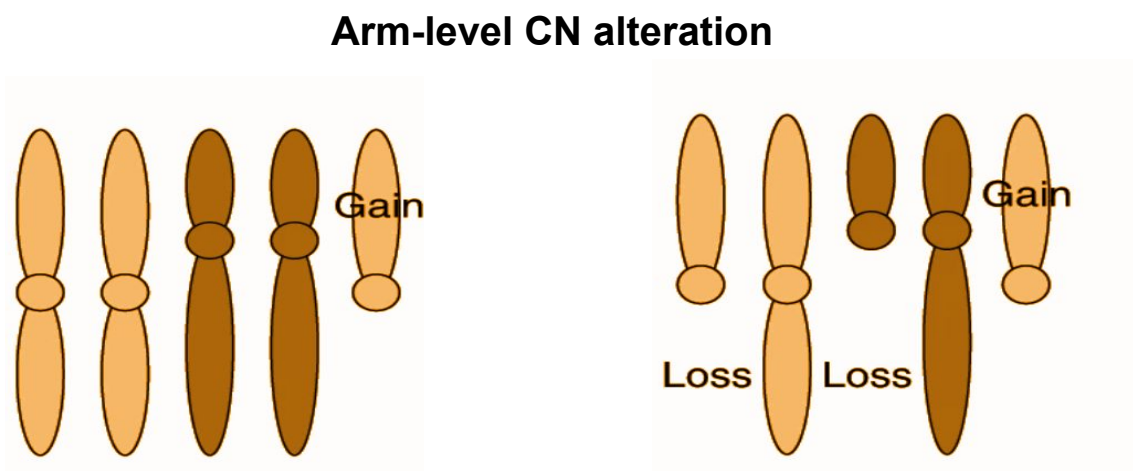
Cava et. al, BMC Systems Biology 2015

- More base pairs are altered as a result of SCNAs than as a result of point mutations in the majority of tumors
- **The field of SCNA detection is moving forward rapidly**
- Numerous previous studies have shown SCNAs drive tumor progression



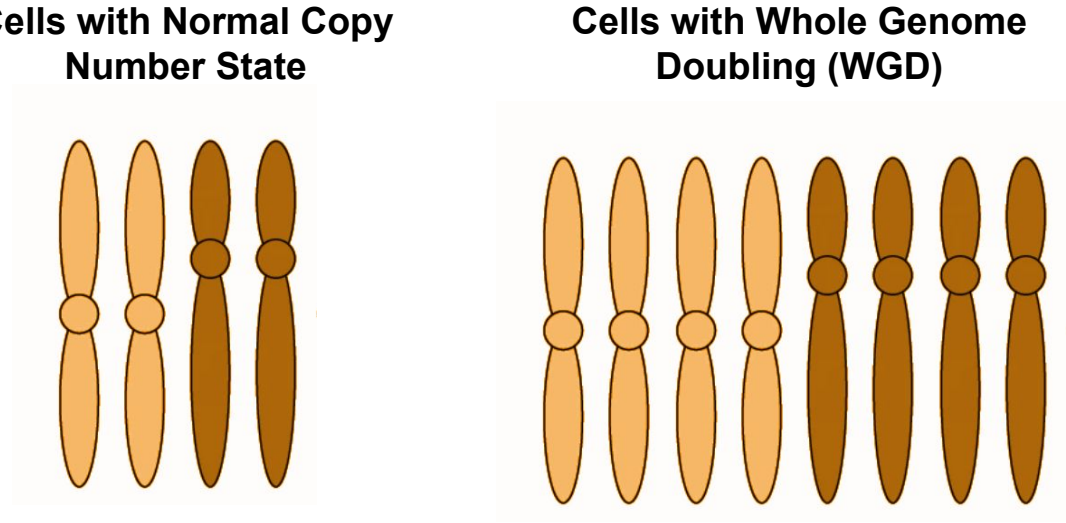
Focal vs arm-level SCNAs





- SCNAs are either **focal(smaller than the length of a chromosome arm)**, or almost exactly the length of a chromosome arm (**arm-level**).
- **Arm-level** gain or loss can include hundreds of genes, several of which are likely tumor suppressor genes (TSGs) or oncogenes.
- **Focal** SCNAs are often enriched at cancer driver genes.

Whole Genome Doubling (WGD)



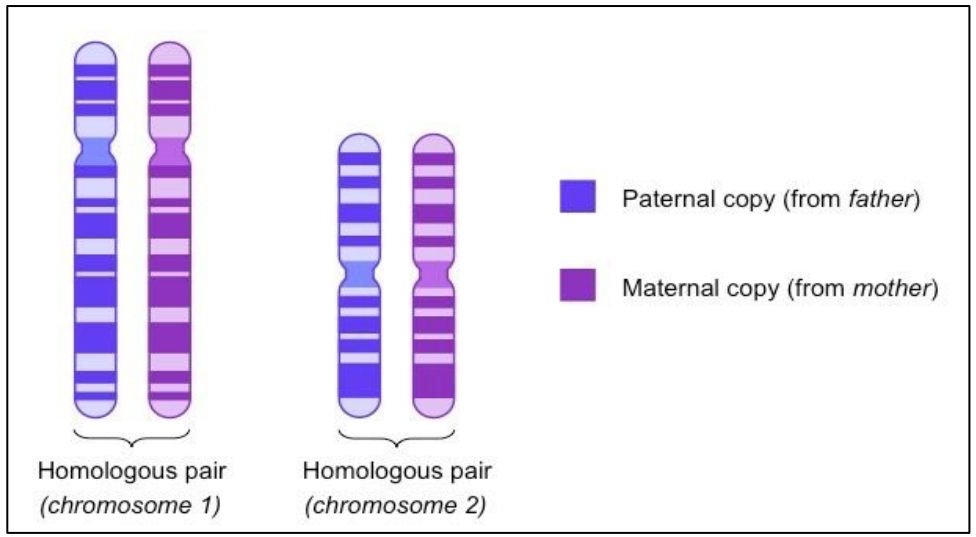
Whole-genome doubling (WGD), involving the duplication of a complete set of chromosomes, is a common feature of cancer genomes.

- WGD is the major source for changes in ploidy in tumor genomes.
- WGD is a prevalent event, found in nearly 30% of tumors through pan-cancer studies.
- WGD has been linked to increased tumor cell diversity, accelerated cancer genome evolution, and worse prognosis

Allele-Specific SCNA

- Total copy number (TCN)
  - The overall copy number (major + minor)
- Major copy number (MCN)
  - The larger copy number of the two alleles
- Minor copy number (LCN)
  - The smaller copy number of the two alleles

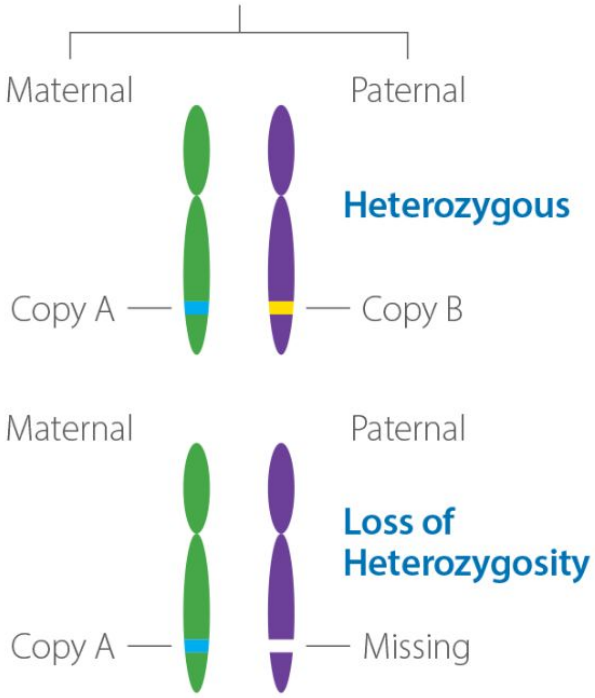
***if ABB, minor allele = A (1 copy) and major allele = B (2 copies)**



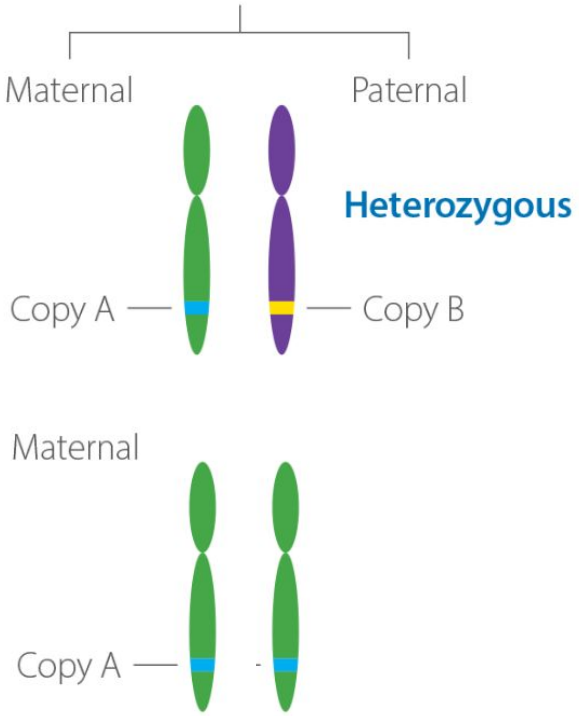
Total copy number	Minor copy number	Calling	Note
0	0	HOMD	homozygous deletion
1	0	DLOH	deletion LOH
2	0	NLOH	copy-neutral LOH
2	1	HET	heterozygous
2	2	WGD	whole genome duplications
3	0	ALOH	amplified LOH
3	1	ASCNA	allele-specific amplification
4	0	ALOH	amplified LOH
4	1	ASCNA	allele-specific amplification
4	2	BCNA	balanced-amplification

Somatic Loss of Heterozygosity (LOH)

- Somatic loss of heterozygosity (LOH) refers to the **loss of one of two alleles** in a tumor cell



(1) LOH with copy number loss (CNL-LOH) of a single allele (mom or dad's)



(2) Copy number neutral LOH (CNN-LOH)

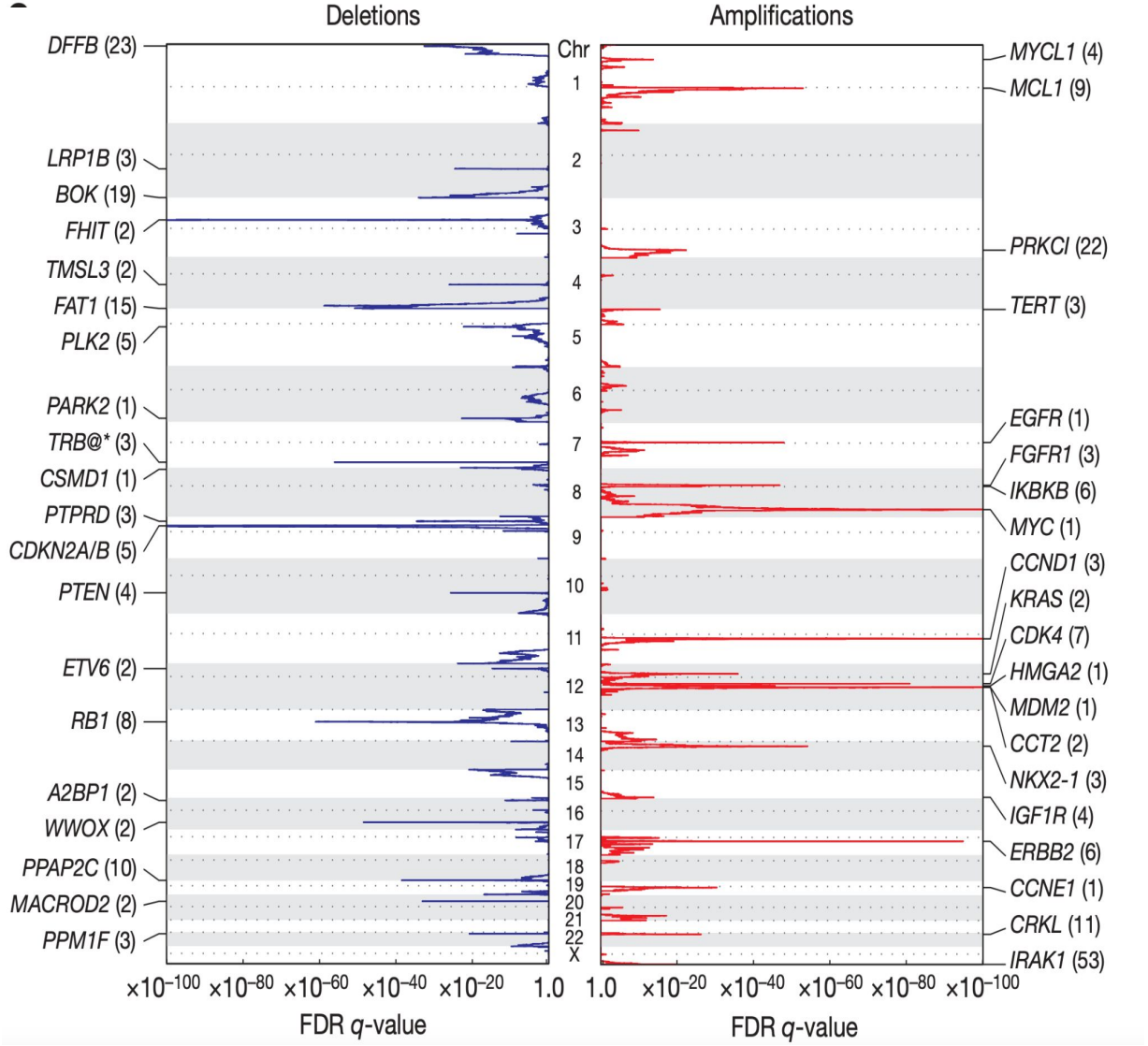
Overview of Common Focal SCNAs

Tumor suppressors are commonly deleted across samples:

(CDKN2A, PTEN, etc.)

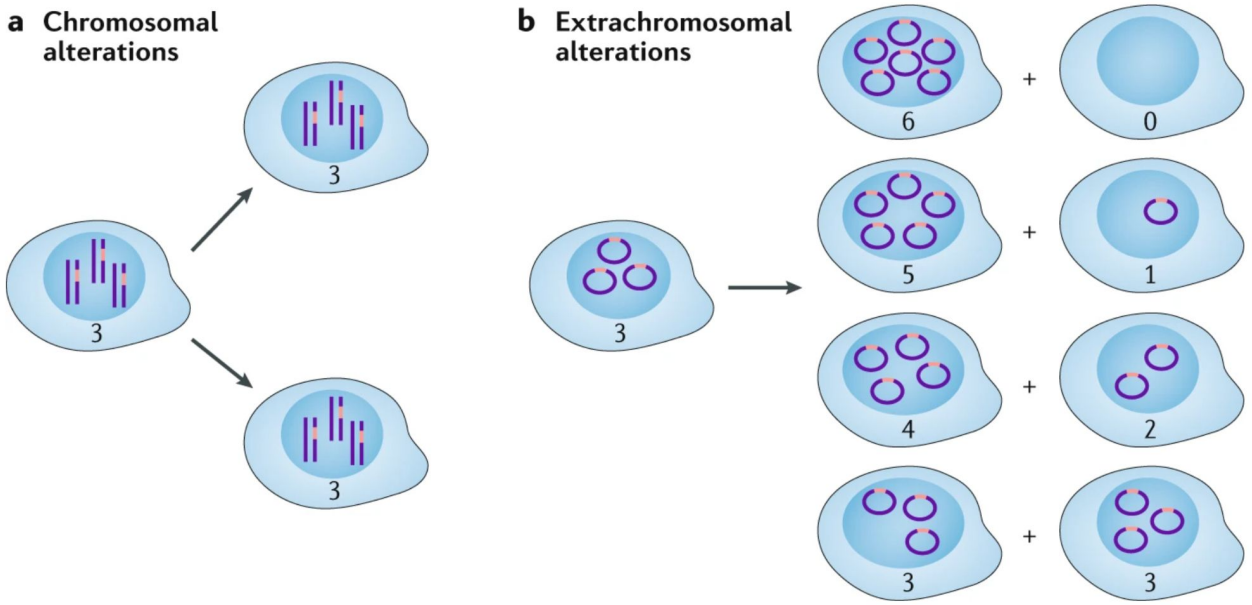
Oncogenes are commonly amplified across across samples:

(MYC, KRAS, CDK4, MDM2)

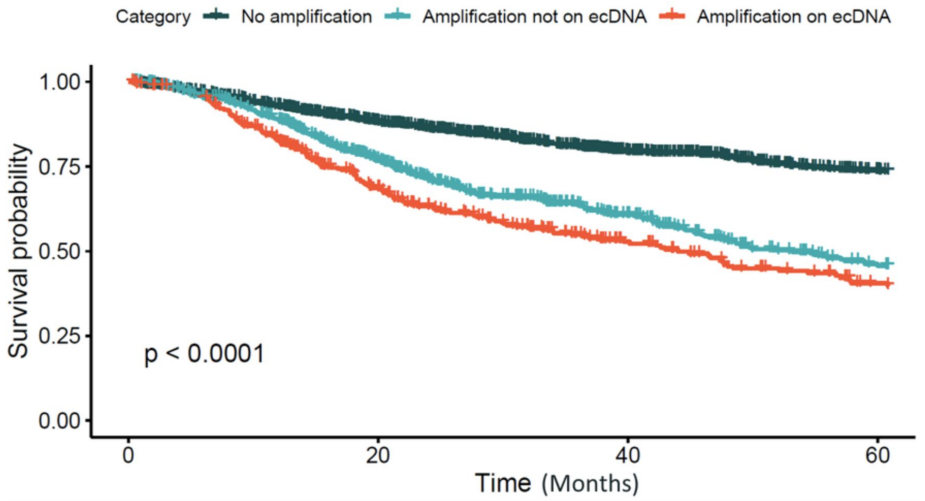


ecDNA: a potent form of focal amplification

- ***Extrachromosomal DNA*: Circular DNA** found outside of lin DNA
- Range between **0.2 and 5 Mb in length**
- ecDNA has a significant enrichment for oncogenes
- Lacks centromere ⟶ Asymmetric division
- Asymmetric division⟶ Intratumoral heterogeneity
- Intratumoral heterogeneity ⟶ therapeutic resistance



Mischel, Nature Reviews Cancer 2019



Methods for detecting SCNAs

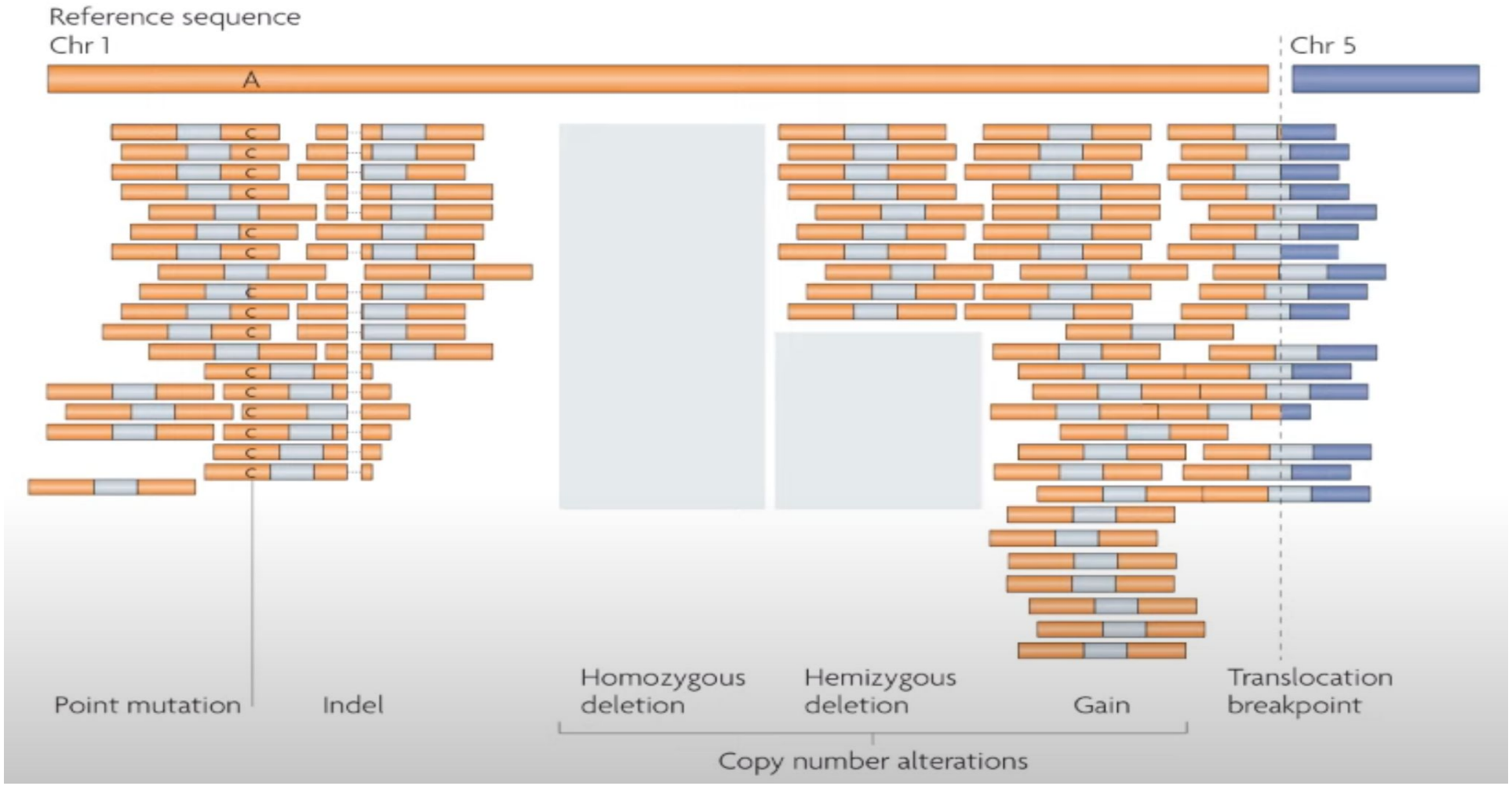
SCNA calling from WGS: Read Depth

Read Depth: Count of # of reads mapped to each genomic region

Hemizygous
Deletion: ½ the read depth

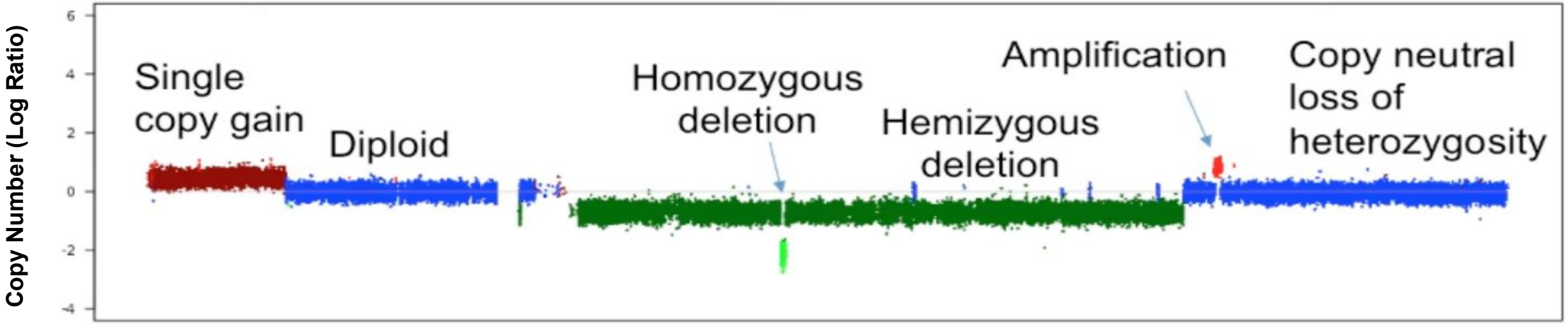
Homozygous
Deletion: 0 read depth

Gain: Increased read depth



Meyerson and Getz, 2010, Nat Rev Genet.

SCNA calling: Read Depth

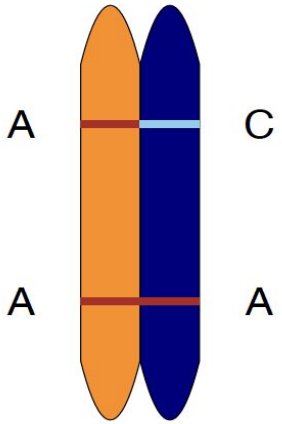


***y-axis: Log R – log ratio of tumor/normal read depth**

***Read Depth alone cannot distinguish between diploid and CN-LOH**

SCNA can be inferred from germline heterozygous SNPs

Germline



Allele frequencies

A	C
0.5	0.5
1	0

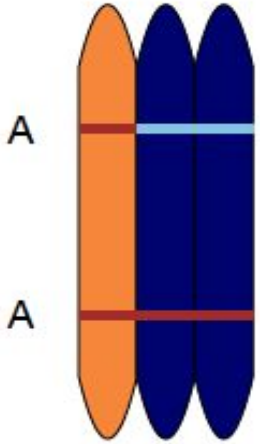
$$BAF = \frac{read\ depth_{allele\ A\ or\ B}}{total\ read\ depth_{allele\ A+B}}$$

Somatic loss

Allele frequencies

A	C
1	0
1	0

Somatic gain

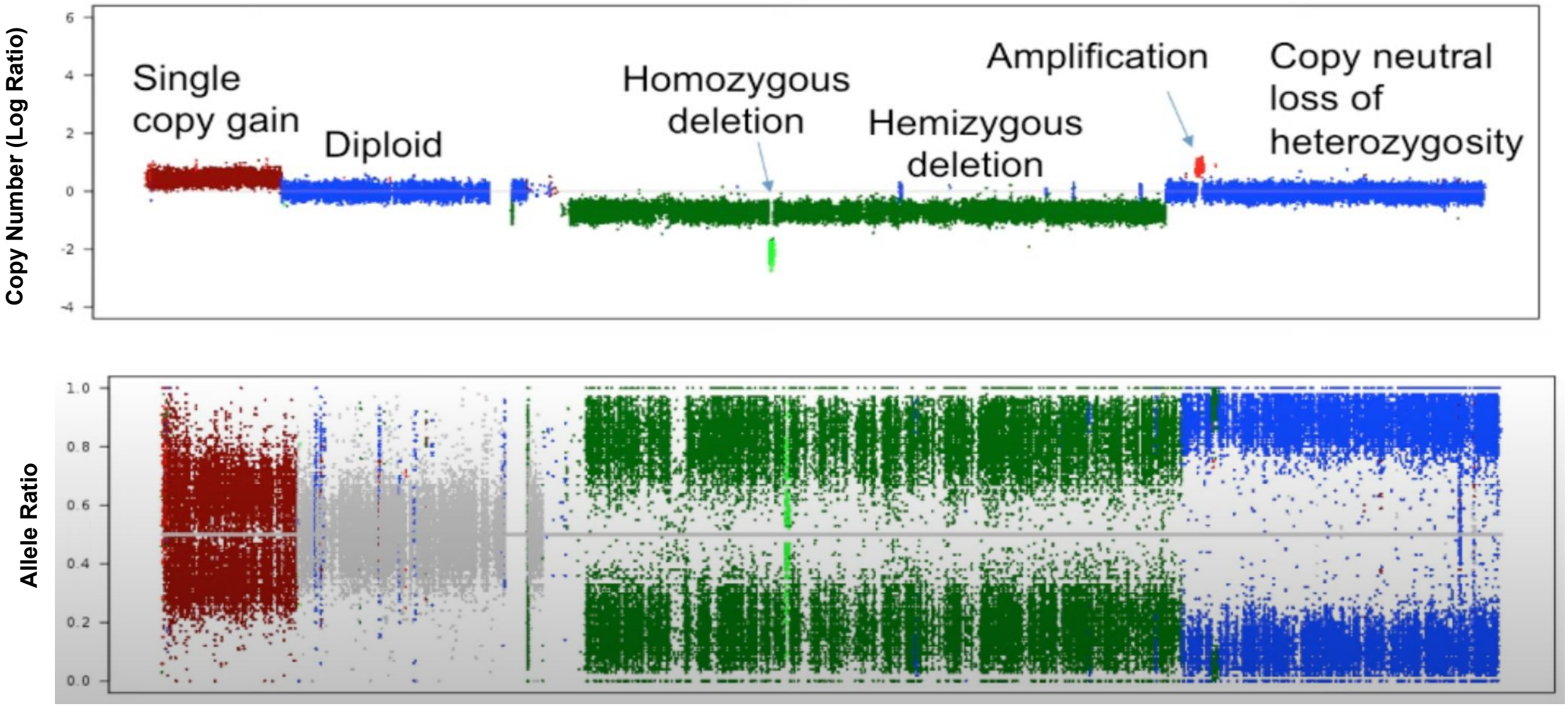


Allele frequencies

A	C
0.33	0.66
1	0

Chris Steele

Integrating BAF & Read Depth Signal

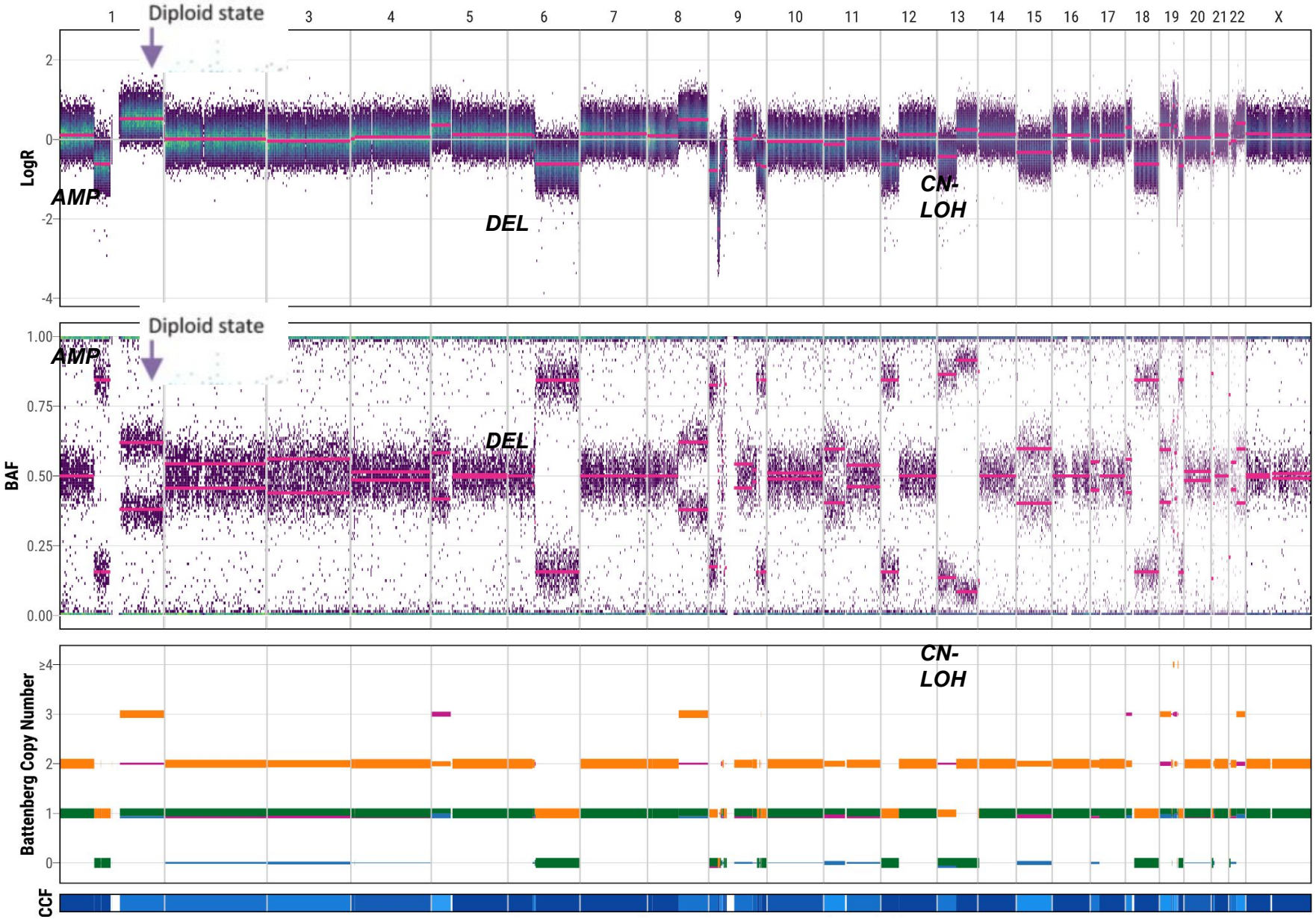


Joint segmentation using BAF and LogR

Log R – log ratio of tumor/normal read depth

$$BAF = \frac{read\ depth_{allele\ A\ or\ B}}{total\ read\ depth_{allele\ A+B}}$$

Total CN & Minor CN



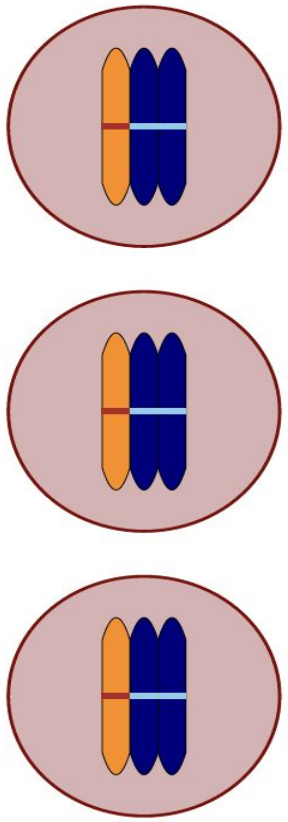
Log2 (Ratio) *v.s* B-allele frequency (BAF)

CNV detection approach	Advantages	Disadvantages
Read depth	1. Uses any locus with reasonable coverage, providing higher resolution. 2. Dependent on proper mapping, not on variant calls. 3. Relatively robust to small contamination events from another human sample.	1. Highly susceptible to technical bias both from batch to batch and from sample to sample due to variable sensitivity of the various regions to moderate or even slight changes in hybridization conditions and other factors (such as exome enrichment kit versions). When used by itself, this approach can yield a large number of false positives and negatives if not sufficiently validated. 2. Unable to identify complex copy number events such as CN-LOH or hyperdiploidy.
Variant allele frequency (VAF) **VAF = BAF**	1. Uses heterozygote positions to discern bands that indicate larger copy number changes. 2. More resistant to technical bias from batch to batch and from sample to sample.	1. Can confuse copy number events with contamination events. 2. Cannot discern concurrent gains in both alleles from normal state. 3. Misinterprets CN-LOH events as copy number loss events. 4. Requires that variant calling and heterozygous states be established. Resolution depends on the number and distribution of variants called.
Read depth combined with VAF	1. One method fills in the gaps in information from the other method. 2. Properly assesses CN-LOH events and concurrent gains in both alleles as well as copy number state when small levels of contamination occur.	1. Inherently different resolution levels complicate creation of individual segments from both sources of information.

VAF variant allele frequency; *CNV* copy number variation; *CN-LOH* copy-neutral loss of heterozygosity

Tumor purity and ploidy are essential to SCNA calling

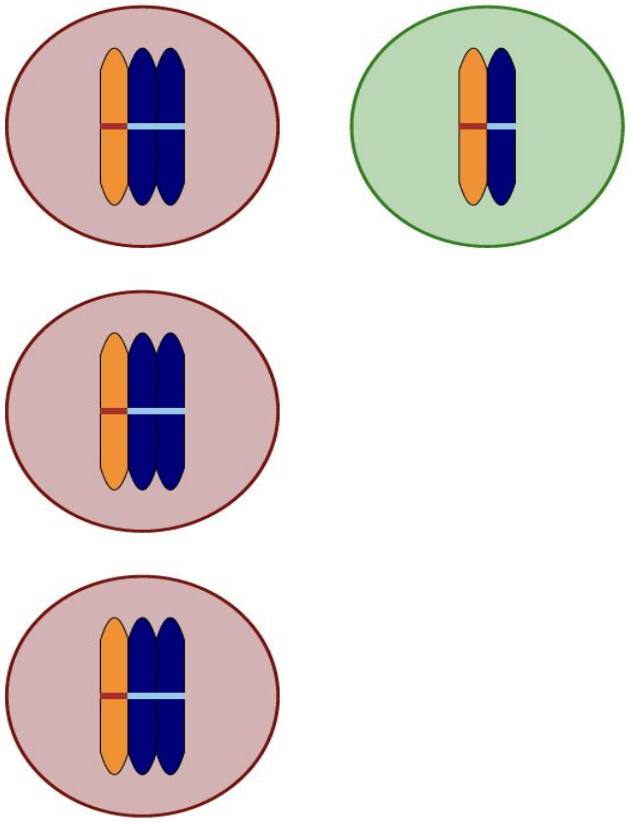
Purity: 100%



Allele frequencies

A: 0.33 C: 0.66

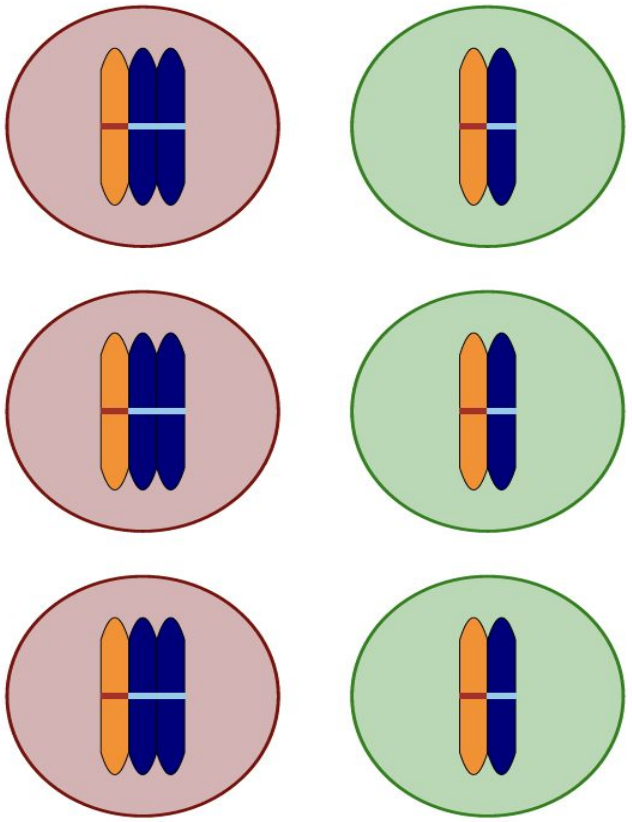
Tumor Purity: 75%



Allele frequencies

A: 0.363 C: 0.637

Tumor Purity: 50%



Allele frequencies

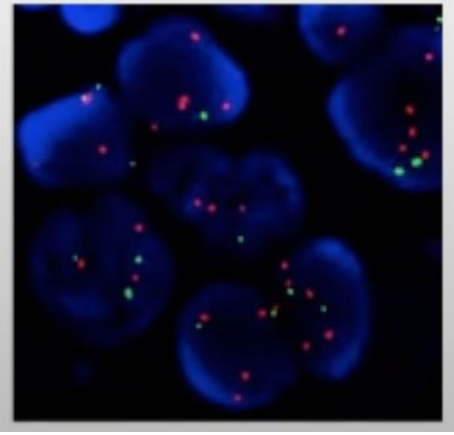
A: 0.4 C: 0.6

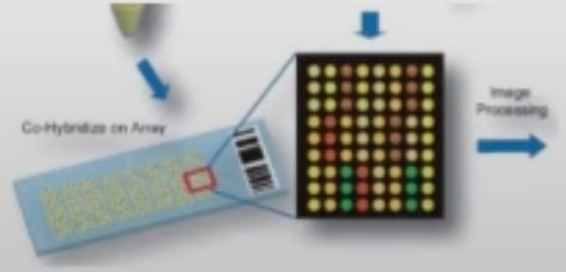
Chris Steele

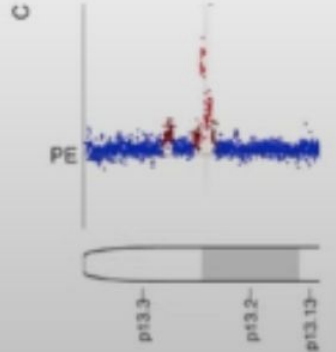
Measurement technologies for SCNA analysis

Tech:	FISH	Array CGH	Genotype arrays	WGSS
#:	<10	30-100K	100K-2M	3G!

Resolution

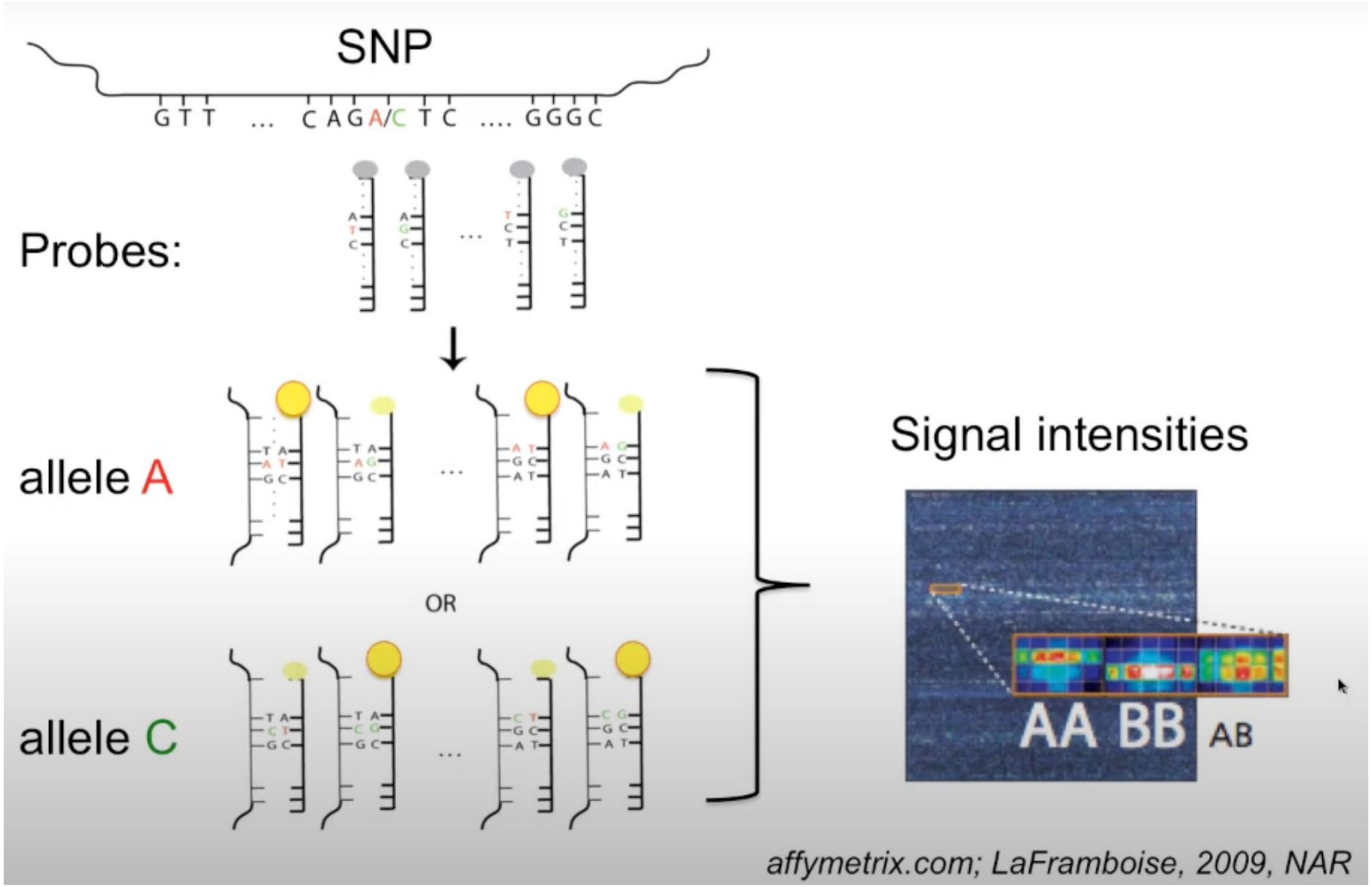






Note: CNV calls can also be derived from whole-exome sequencing data, not shown here.

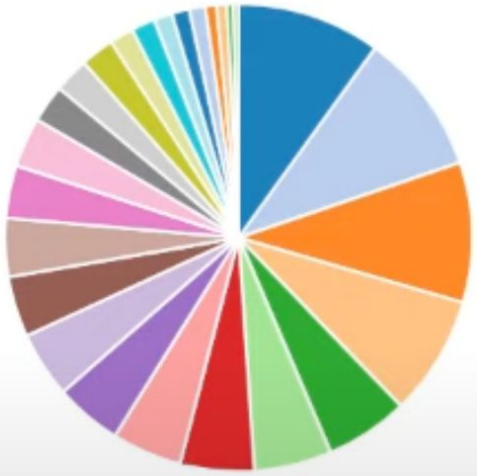
Affymetrix SNP 6.0



Sources of SCNA data

The Cancer Genome Atlas (TCGA): n=10,995 cases

NIH NATIONAL CANCER INSTITUTE
GDC Data Portal

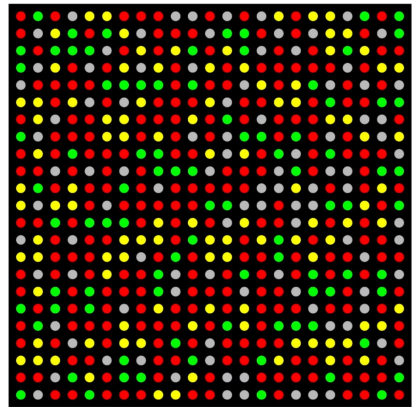


26 Primary Sites

Primary Site	Cases
Breast	1,096
Lung	1,022
Brain	1,107
Kidney	888
Colorectal	624
Uterus	604
Ovary	573
Head And Neck	521
Prostate	498
Thyroid	505
Skin	470
Stomach	443
Bladder	412
Liver	376
Cervix	302
Adrenal Gland	271
Soft Tissue	261
Esophagus	185
Pancreas	185
Bone Marrow	143
Testis	134
Thymus	124
Pleura	87
Eye	80
Lymph Nodes	48
Bile Duct	36

SCNA calling: SNP Array v.s Whole Genome Sequencing

Array based SCNA detection



Advantage:

- Reliable detection for SCNA with large size (>100 KB)
- Highly efficient with low computational burden, Low cost

Disadvantage:

- **Low Resolution, cannot detect short SCNA, limited by density of common SNPs**

WGS based SCNA detection

Advantage:

- Higher accuracy for breakpoint detection (base-pair resolution)
- High sensitivity for both large/small SCNA detection
- Additional information from read-depth
- FFPE (readily available in clinic) & single cell
- Disadvantage:
- **Higher cost.**

Popular SCNA callers

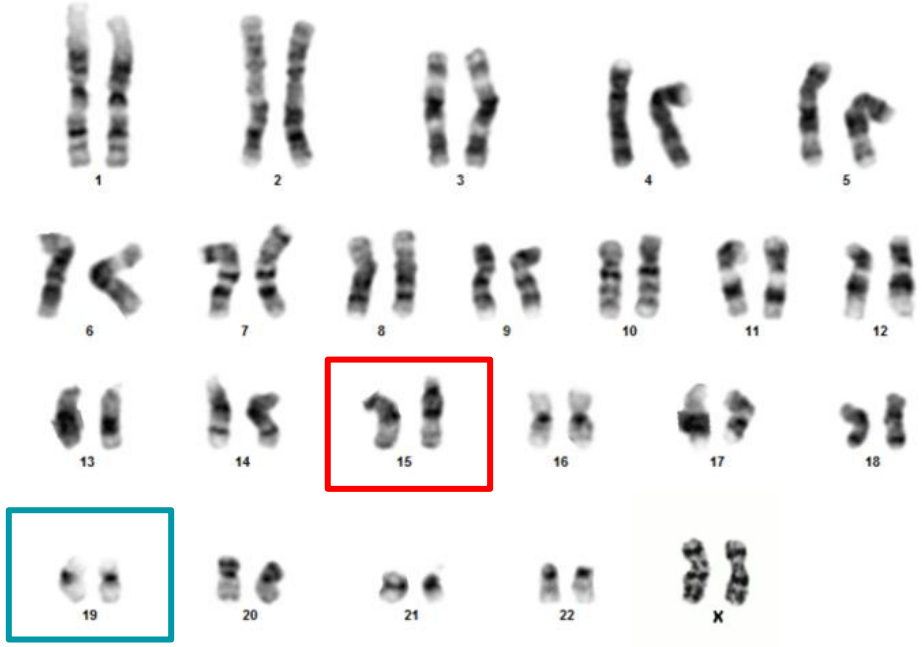
Copy number methods implementation choices and assumptions

Name	ABSOLUTE	Aceseq	Battenberg	cloneHD	Sclust
Event based	X	X	X		X
State based				X	
Allele counts for heterozygous SNPs	X	X	X	X	X
Binned read counts logR	X	X		X	X
logR from SNP positions			X		
Phasing of SNPs	X	X	X		
Replication timing correction of logR	X	X			
GC content correction of logR	X	X	X	X	
Assume GC artifact same in tumour and normal				X	
Assume raw data shape*				X	
Maximises genome with clonal copy number states	X	X	X	X	X
Purity/ploidy grid search fitting	X	X	X		
Hidden Markov model fitting				X	
Step-wise fitting					X
Estimates subclonal CNA	X	X	X	X	X
Fits subclonal CNA	X		X	X	X
Number of subclonal states allowed	3		2	many	2
Max. differences between subclonal states			1		1

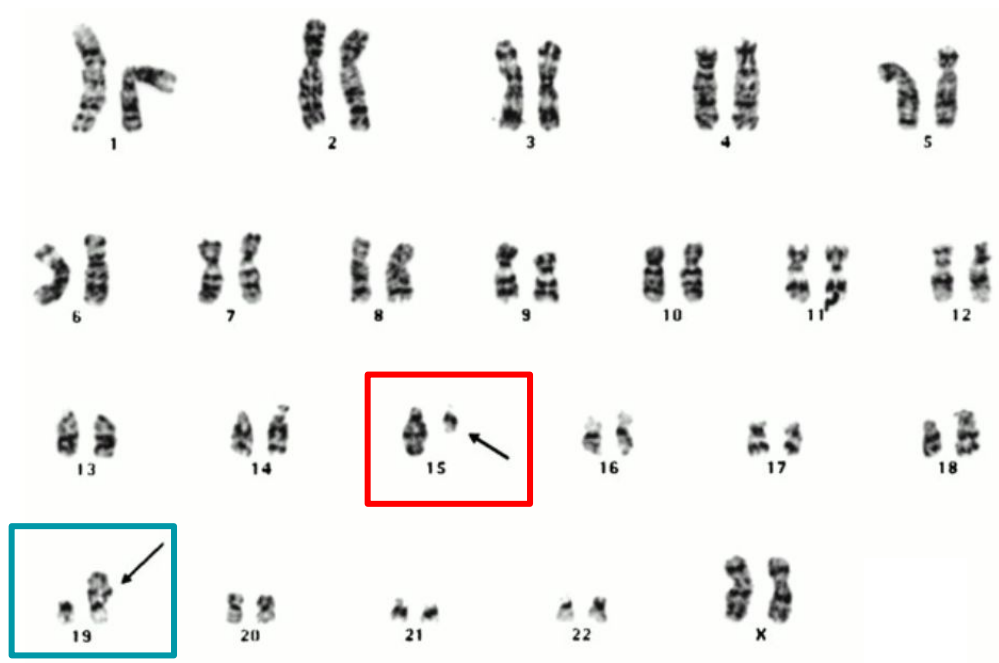
large-scale consortiums such as PCAWG will use an ensemble approach to SCNA calling

SCNA Validation

Validation of SCNAs by Karyotype analysis



Example of Normal Karyotype from control sample



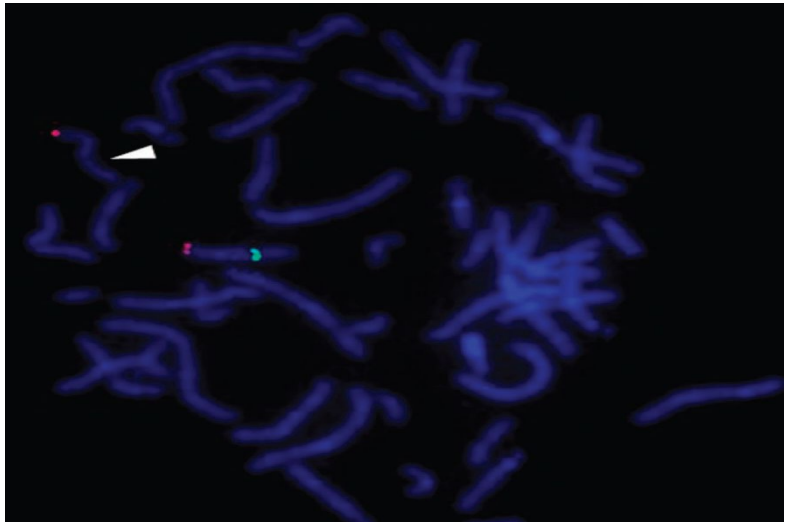
Abnormal Karyotype (19p and 15q loss) from tumor sample

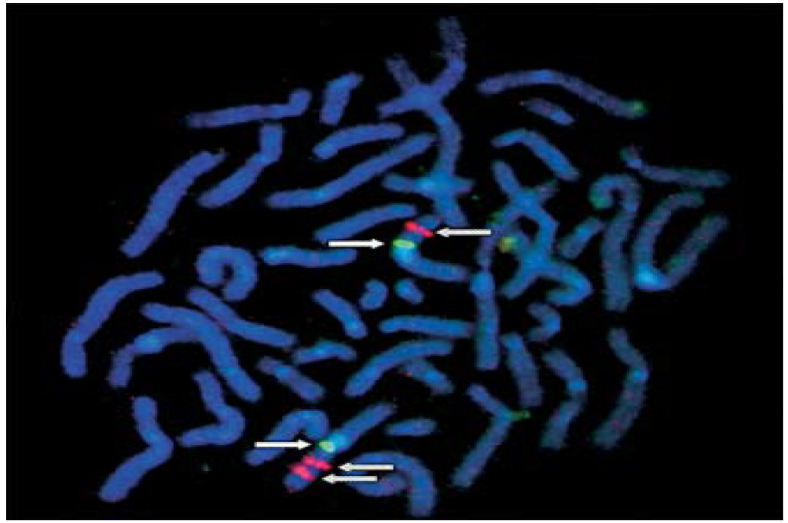
- Karyotyping is a technology **to examine chromosomes in targeted cells, mainly focused on their integrity, numbers and shapes**. This analysis can help to identify cytological characteristics of complex disease and cancers. With a resolution of 5 Mb, Karyotyping can detect chromosomal aneuploidies, unbalanced translocations, and large copy number alterations.

(PMID: **10944559)**

Fluorescence *in situ* Hybridization (FISH)

Red: Chr10 (control)
Green Signal: *PTEN*





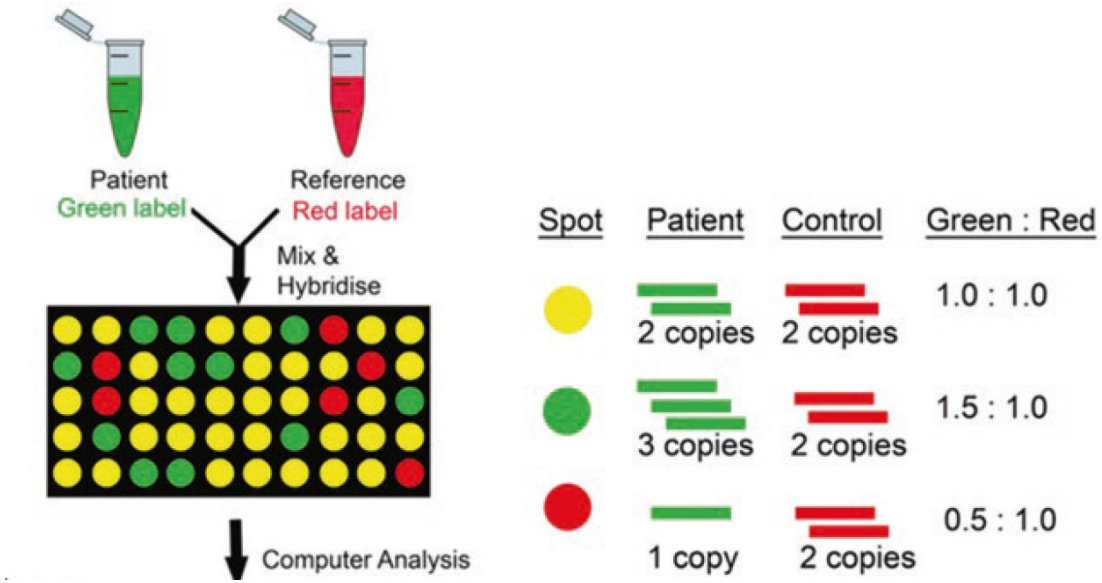
Green Signal: Chr9 (control)
Red: *CDKN2A*

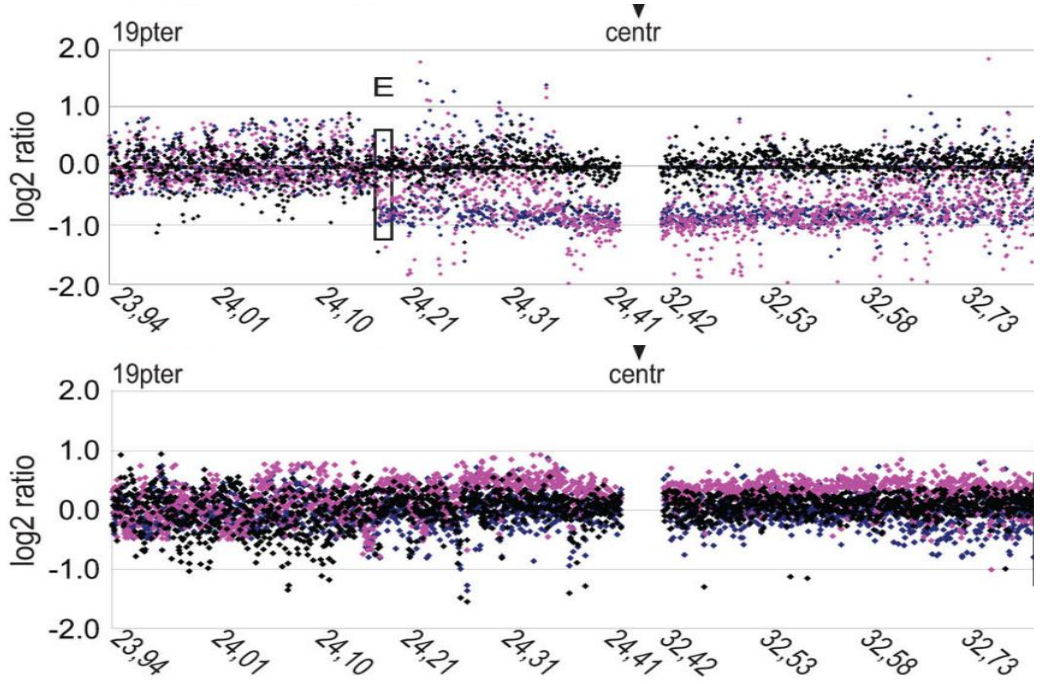
- FISH is **a cytological technique developed in the early 1980s**, with broad clinical applications in medical labs
  - FISH technology uses fluorescent DNA probes to target specific chromosome region, which can be detected using a specific fluorescent microscope.
  - Using carefully designed markers to hybridize a known CNV region (or gene), FISH can directly display copy number state information in both tumor and normal samples.

(PMID: 27579173, 18160777)

Array Comparative Genomic Hybridization (CGH)

- **array CGH** is the gold standard method for detecting both germline CNVs and SCNAs. It can provide detailed quantification of deletion or duplication.
- The intensity of the fluorescence color will be used to calculate logRatio and estimate copy number state of the targeted sample.





Case/Normal: 0.5

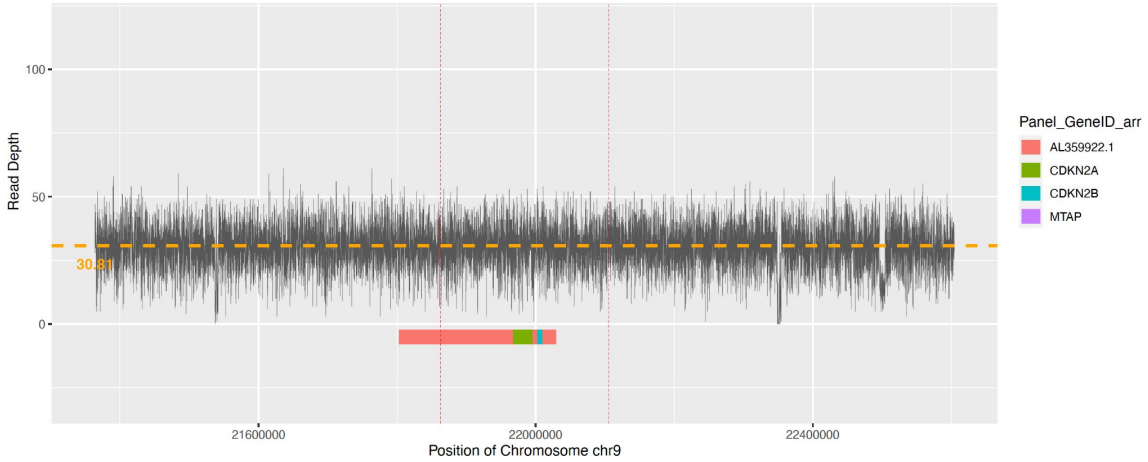
Case/Normal: 1

(chr 19 loss)

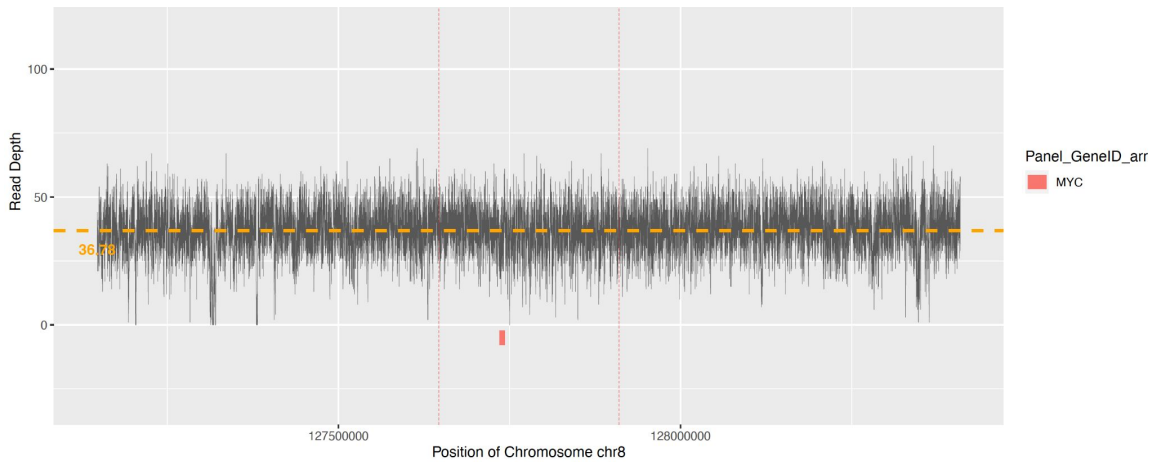
(PMID: 19119320)

Direct validation by Read-Depth - Typical Focal SCNAs

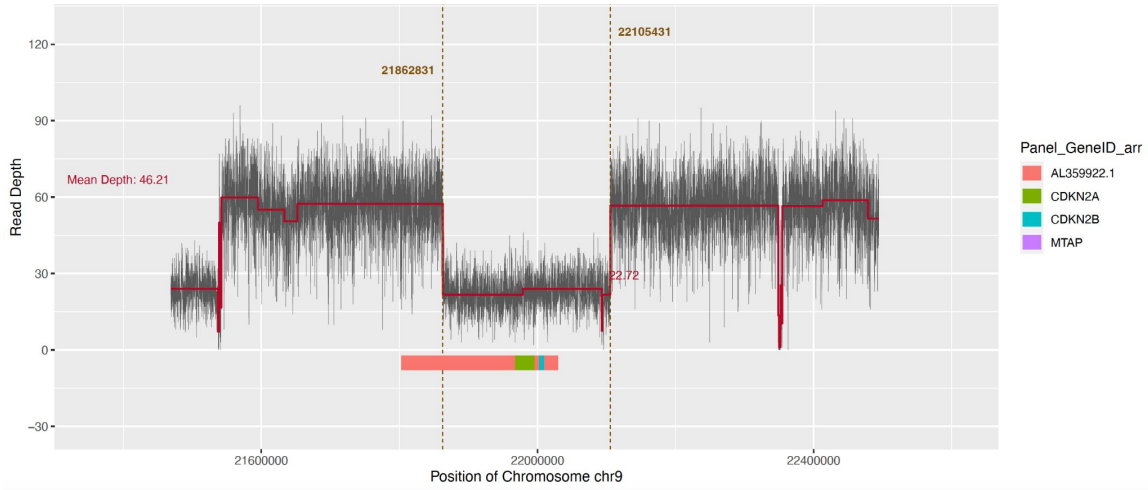
chr9:21862831-22105431 [9131_Normal]



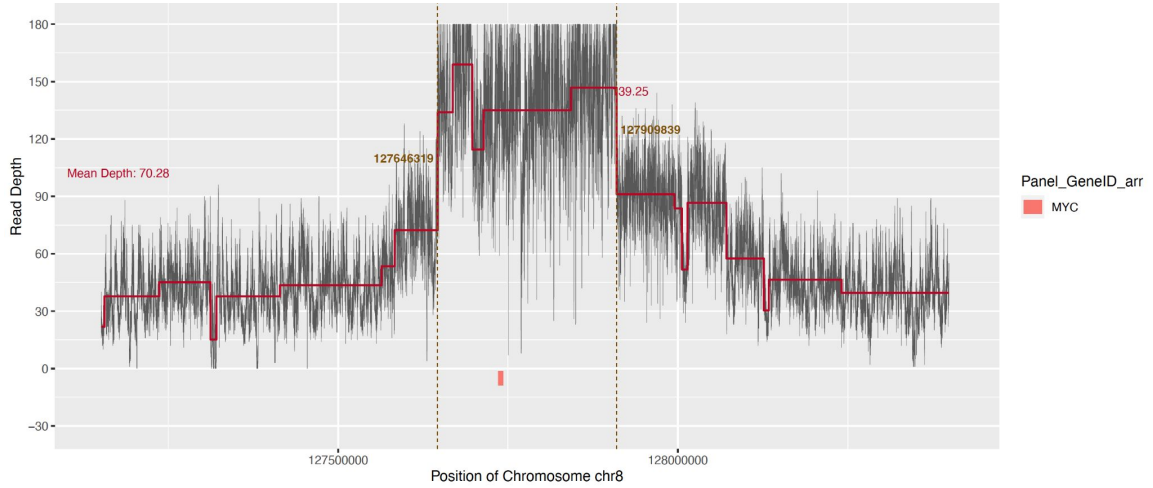
chr8:127646319-127909839 in [4630_Normal]



chr9:21862831-22105431 in [9131_Tumor]



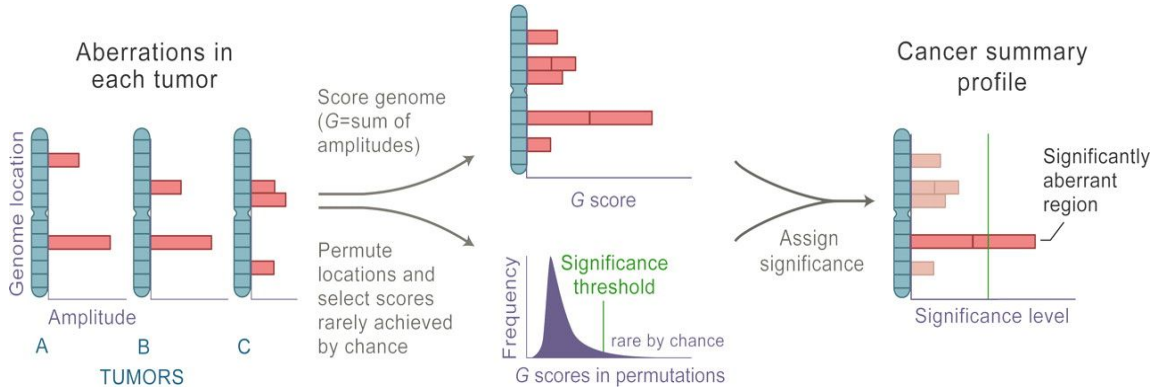
chr8:127646319-127909839 in [4630_Tumor]



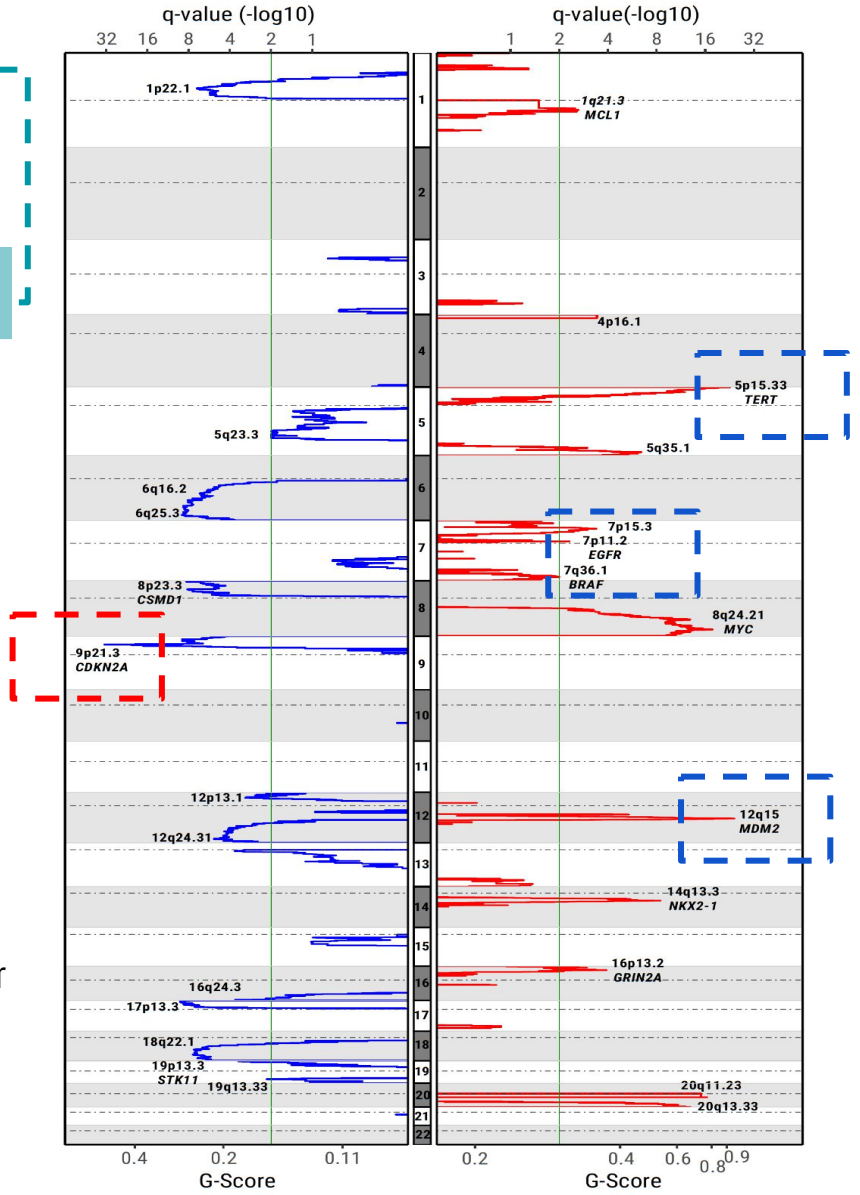
Common SCNA-related analyses in cancer genomic studies

Identification of significant SCNA regions in cancer genome

- Somatic copy number alterations are distributed across the whole genome, while only a few 'driver' alterations can promote cancer development.
- How to distinguish those driver SCNAs from random passenger SCNAs is a critical question.



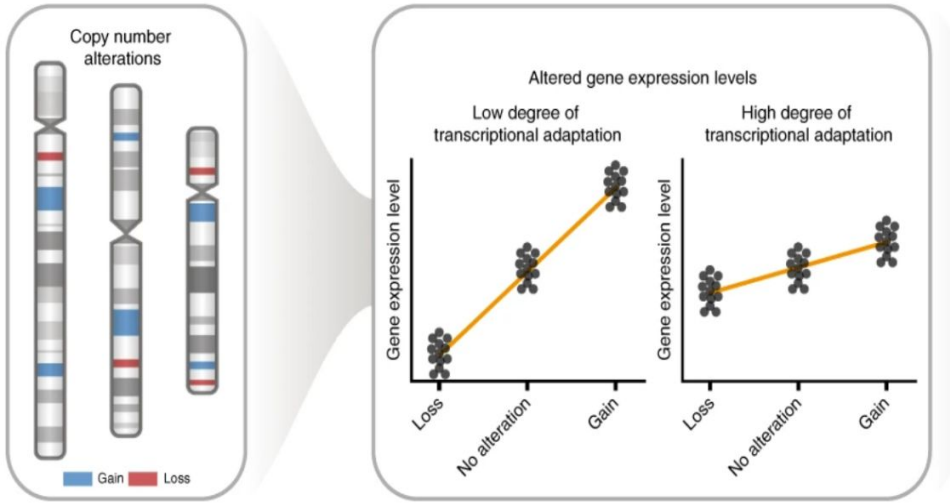
- GISTIC2 (**Genomic Identification of Significant Targets In Cancer**) is an algorithm for analyzing copy number alterations in cancer genomes.
- It identifies recurrent focal events that are likely to be drivers of tumorigenesis based on their frequency.



(PMID: 21527027, 34493867)

Impact of SCNAs on gene expression

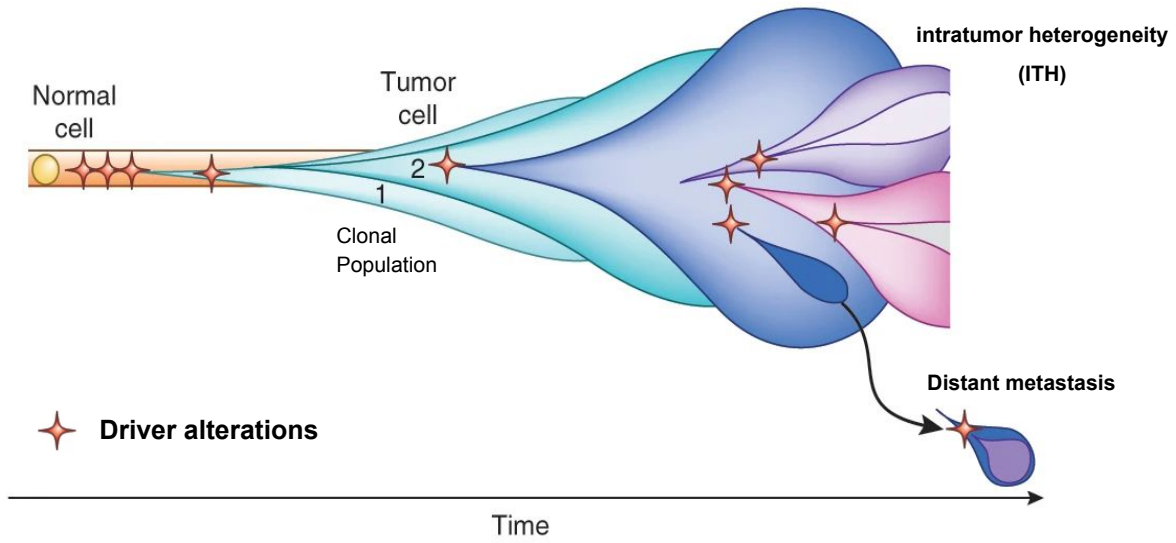
- Amplification of oncogenes can lead to increased gene expression and activate pathways that drive tumor growth and progression.
- Deletion of tumor suppressor genes can result in decreased gene expression and inactivate normal cellular processes that control cell growth and division.



Gene	SCNA Type	Cancer Types	Description	PMID
PIK3CA	Amplification	lung cancer	Amplification of *PIK3CA* was found to be associated with increased expression of this gene in lung cancer, which can develop potential therapeutic targets with PI3K pathway inhibitors	PMID: 17992665
NPM1	Amplification	liver cancer	Amplification of *NPM1* was found to be associated with increased expression of this gene in liver cancer, which can greatly promote cancer progression compared to patients without the amplification.	PMID: 28874807
CDKN2A	Deletion	melanoma	Deletion of tumor suppressor *CDKN2A* was found to decrease expression of this gene in melanoma, and is reported to be associated with metastasis.	PMID: 29990501

Tumor evolution and ITH analysis based on SCNA profile

(Cancer is an evolutionary process)



- Tumor evolution begins when a cell in normal tissue expands to form tumor tissue. During the process, clonal population diverge and form different subclones, resulting in intratumor heterogeneity (ITH)

- Exploring tumor evolutionary history can help us understand the processes that drive tumor development and design potential targets for therapeutic intervention.

- Researchers investigate ITH of SCNA, DNA methylation, and point mutations in lung cancer driver genes from 84 patients with LUAD.

- Average pairwise ITH index was developed to quantify ITH level. APITH indexes for both SCNAs and methylation changes show significant associations with poor prognosis

nature COMMUNICATIONS

ARTICLE

Check for updates

https://doi.org/10.1038/s41467-020-16295-5 OPEN

Genetic and epigenetic intratumor heterogeneity impacts prognosis of lung adenocarcinoma

Xing Hua[1,9], Wei Zhao[2,9], Angela C. Pesatori[3,4], Dario Consonni[4], Neil E. Caporaso[5], Tongwu Zhang[2], Bin Zhu[1], Mingyi Wang[6], Kristine Jones[6], Belynda Hicks[6], Lei Song[6], Joshua Sampson[1], David C. Wedge[7,8], Jianxin Shi[1] & Maria Teresa Landi[2]✉

Speaker: **David Wedge, Ph.D., University of Manchester**

Date: **April 20th, 2023**

Title: **Tumour evolution in diverse human populations**

Host: Dr. Maria Teresa Landi, Senior Investigator, ITEB and Senior Advisor for Genomic Epidemiology, TDRP, DCEG

The Wedge group have pioneered the development of computational methods to study heterogeneity in primary and metastatic cancers. Recently, the focus of the lab has shifted towards understudied populations, including the genomics of breast, ovarian, oesophageal and prostate cancers in ethnically diverse populations globally and the genomics of lung cancer in non-smokers.

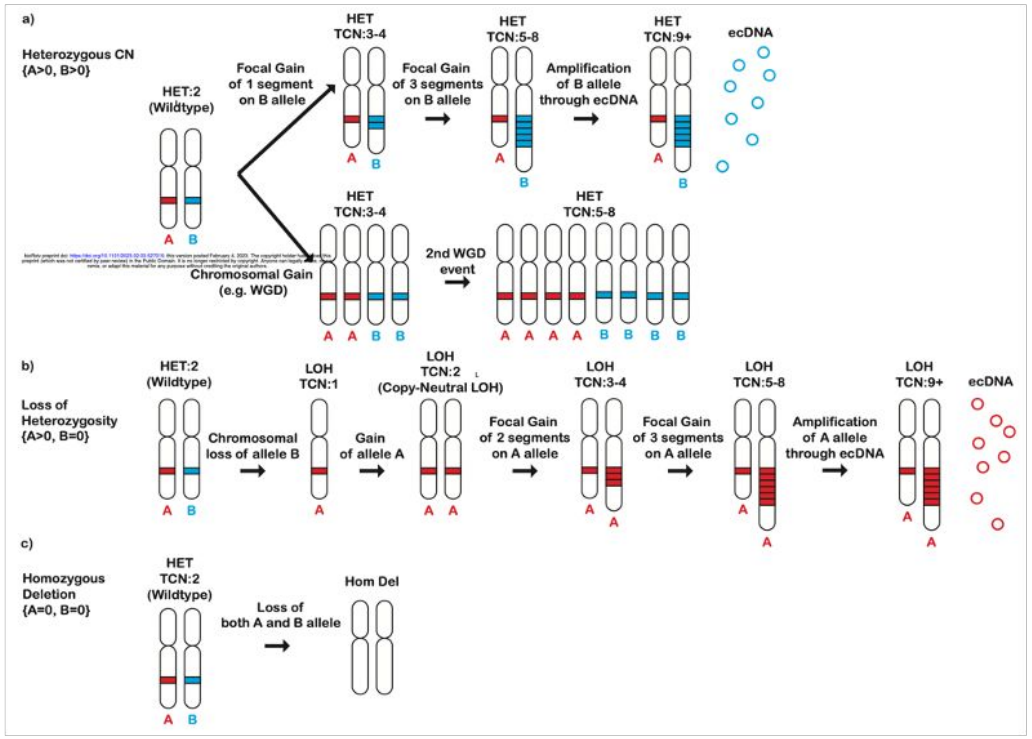
To learn more about tumor evolution, please visit Dr. Wedge's lab (https://wedge-group.netlify.app/)

Exploration of somatic copy number alterations signatures

Visualizing and exploring patterns of large mutational events with SigProfilerMatrixGenerator

Azhar Khandekar, Raviteja Vangara, Mark Barnes, Marcos Díaz-Gay, Ammal Abbasi, Erik N. Bergstrom, Christopher D. Steele, Nischalan Pillay, Ludmil B. Alexandrov

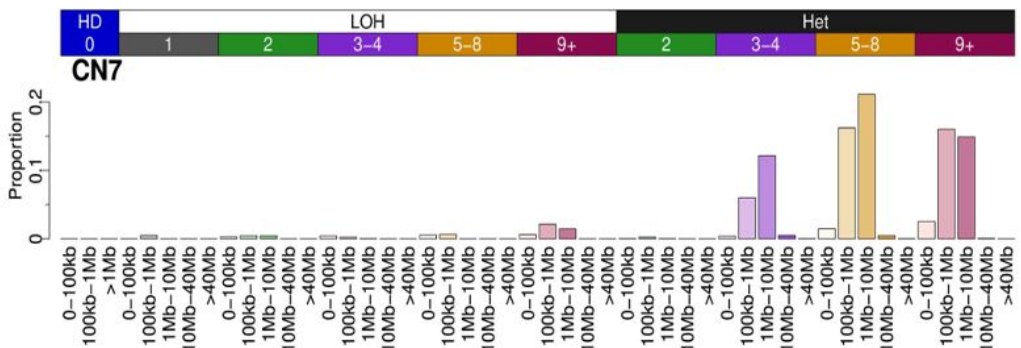
doi: https://doi.org/10.1101/2023.02.03.527015



COPY NUMBER MATRIX GENERATION

In order to generate a copy number matrix, provide the an absolute path to a multi-sample segmentation file obtained from one of the following copy number calling tools (if you have individual sample files, please combine them into one file with the first column corresponding to the sample name):

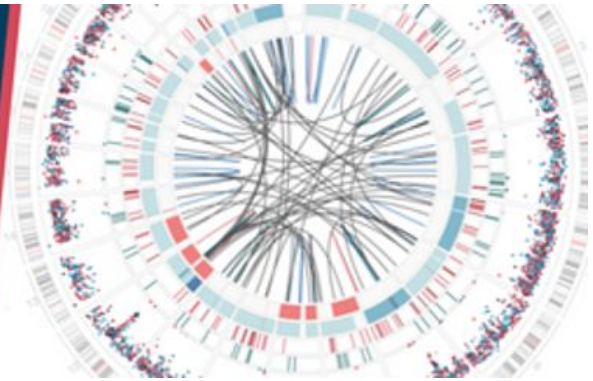
1. ASCAT
2. ASCAT_NGS
3. SEQUENZA
4. ABSOLUTE
5. BATTENBERG
6. FACETS
7. PURPLE

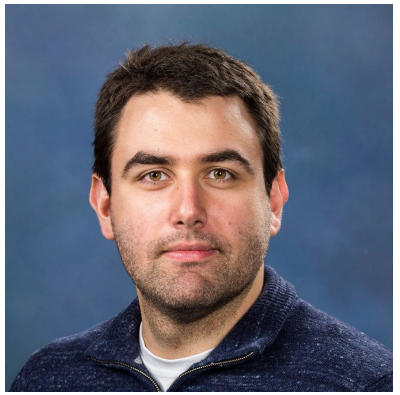


EMERGING APPROACHES FOR TUMOR ANALYSES IN EPIDEMIOLOGICAL STUDIES

A DCEG Annual Course

2022-23





Speaker: **Ludmil Alexandrov, M.Phil., Ph.D.**

Date: **January 19th, 2023**

Title: **Anthology of unusual patterns of somatic mutations in cancer genomes**

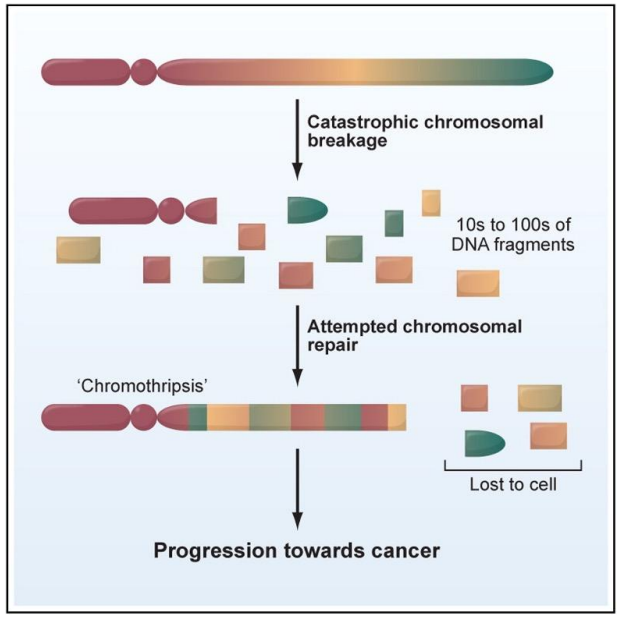
Host: Dr. Maria Teresa Landi, Senior Investigator, ITEB and Senior Advisor for Genomic Epidemiology, TDRP, DCEG

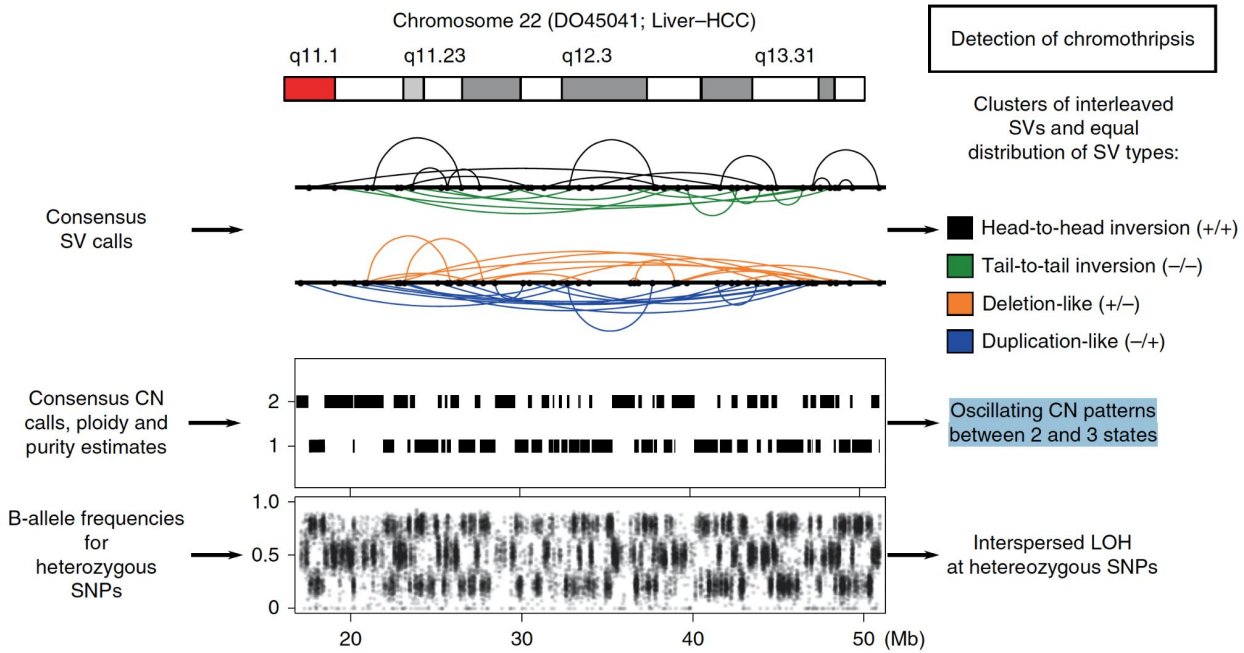
Dr. Alexandrov received his Ph.D. in 2014 from the University of Cambridge researching mutational processes and signatures in human cancers at the Wellcome Sanger Institute. Dr. Alexandrov then went on to research as an Oppenheimer Fellow at the Los Alamos National Laboratory from 2014 to 2017 before becoming an Assistant Professor of Bioengineering and of Cellular and Molecular Medicine at UCSD in 2018.

To learn more about SCNA signatures, please visit Dr. Alexandrov's lab (https://alexandrov.cloud.ucsd.edu/)

Characterization of chromothripsis based on SCNA analysis

- Chromothripsis is a phenomenon in which a complete chromosome is fragmented into hundreds of small pieces which are then repaired incorrectly by DNA repair system.
- Unlike the traditional view of tumorigenesis as a gradual process of accumulating mutations, chromothripsis provides a mechanism for rapid generation of many rearrangements through only a few cell divisions.





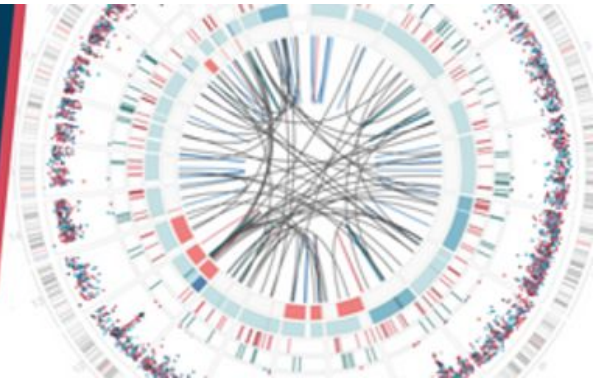
A hallmark of chromothripsis is multiple oscillations between two or three copy-number states at whole genome level.

(**PMID: 32025003**)

EMERGING APPROACHES FOR TUMOR ANALYSES IN EPIDEMIOLOGICAL STUDIES

A DCEG Annual Course

2022-23



Speaker: **Peter Park, Ph.D.**

Date: **February 16th, 2023**

Title: **Structural alterations in cancer genomes**

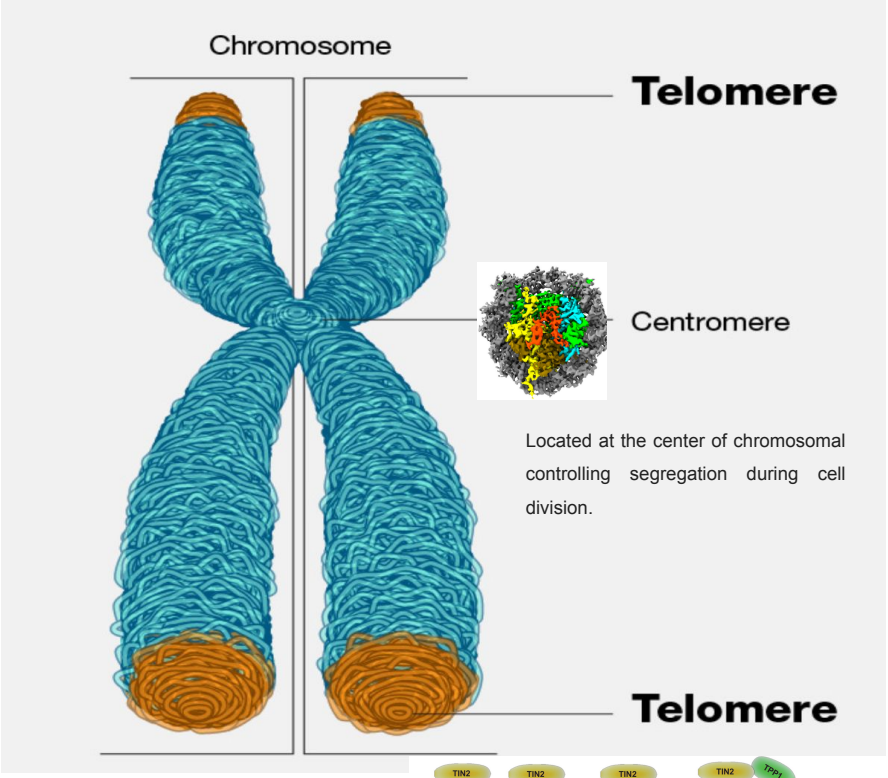
Host: Dr. Maria Teresa Landi, Senior Investigator, ITEB and Senior Advisor for Genomic Epidemiology, TDRP, DCEG

Dr. Park is a Professor of Biomedical Informatics at Harvard Medical School and the director of its PhD program in Bioinformatics and Integrative Genomics. His research group specializes in computational and statistical analysis of large-scale DNA sequencing data to understand genetic and epigenetic mechanisms related to disease processes.

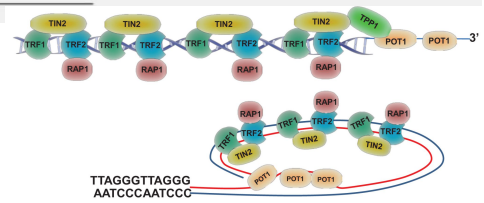
To learn more about chromothripsis, please visit Dr. Park's lab (https://compbio.hms.harvard.edu/)

Limitations of SCNA analysis

SCNAs at centromere and telomere regions are difficult to detect



Located at the center of chromosomal controlling segregation during cell division.



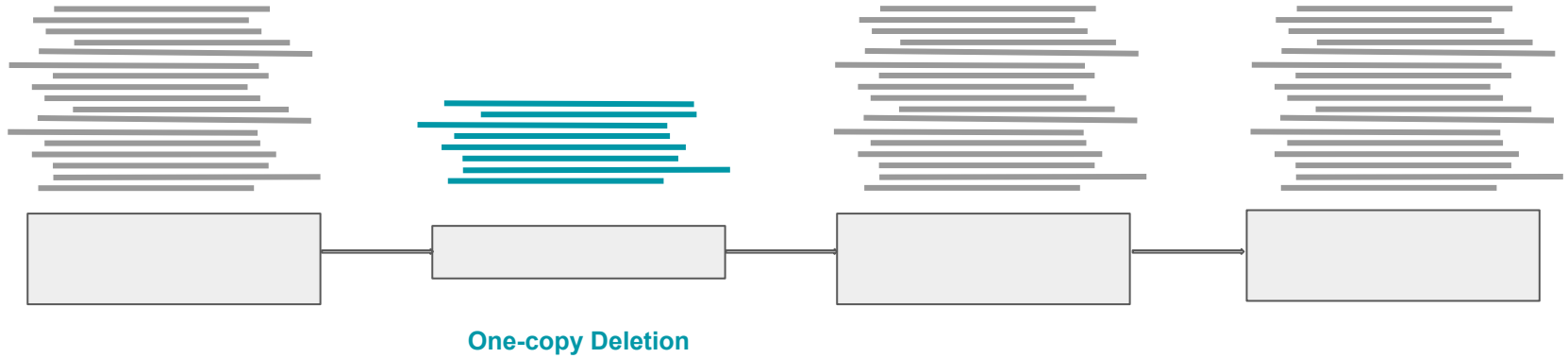
Located at the ends of chromosomal protecting chromosomes from becoming frayed or tangled.

- Telomeres and centromeres have complex structures composed of specific DNA motifs and proteins, making it very difficult to achieve sufficient reads for estimating SCNA changes.
- Telomeres and centromeres consist of high-level repeated sequences which can cause strong bias during NGS library prep and sequencing.
- Centromeres are large regions (hundreds of kilobases), which are difficult to directly sequence with short-read sequencing.

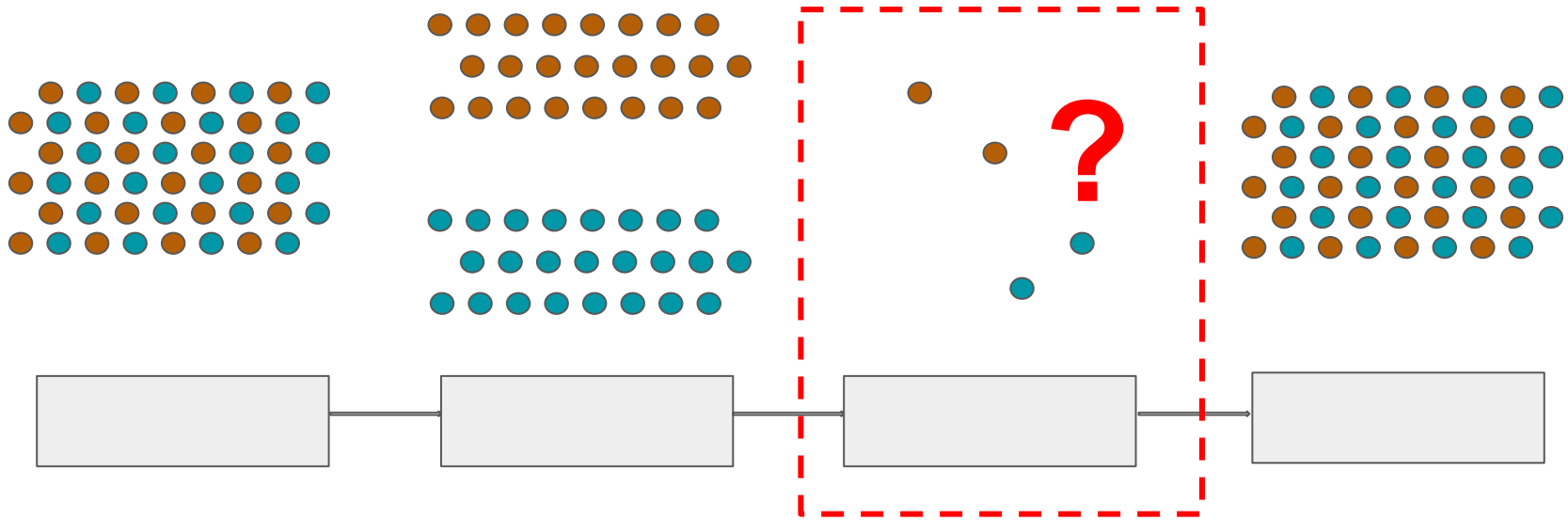
(**PMID: 24683489**)

Resolution of SCNA detection is limited by SNP density

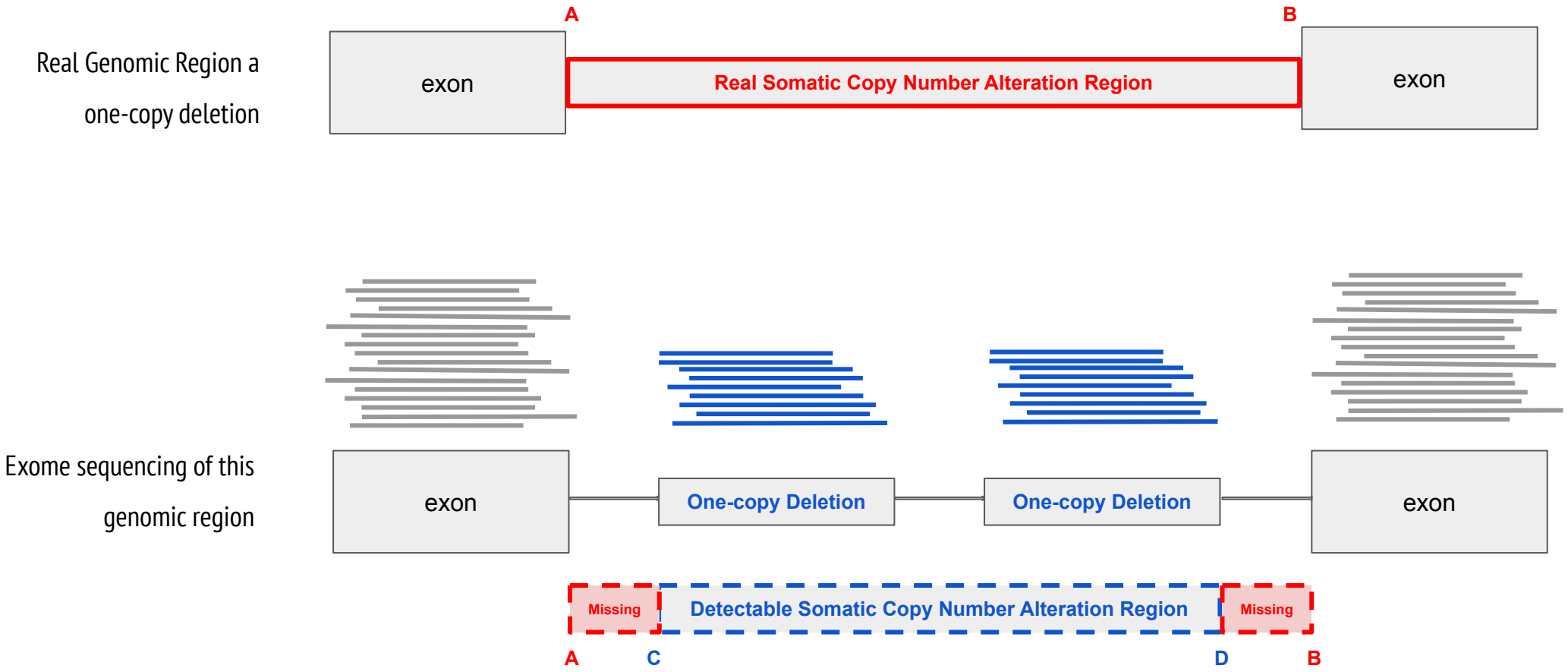
Real Genomic Region with a one-copy deletion



SNPs density is inconsistently distributed at human genome

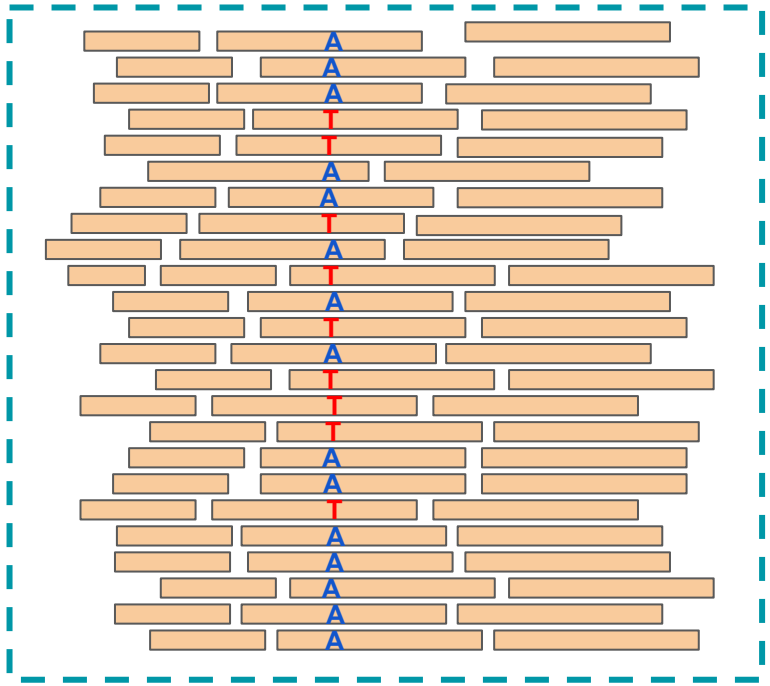


The SCNA breakpoints are outside of the exome region



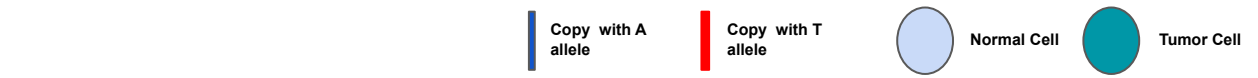
As the real breakpoints (A and B) fall outside reads covered regions, the detected breakpoint is C and D

Multiple fit with different purity/ploidy combinations

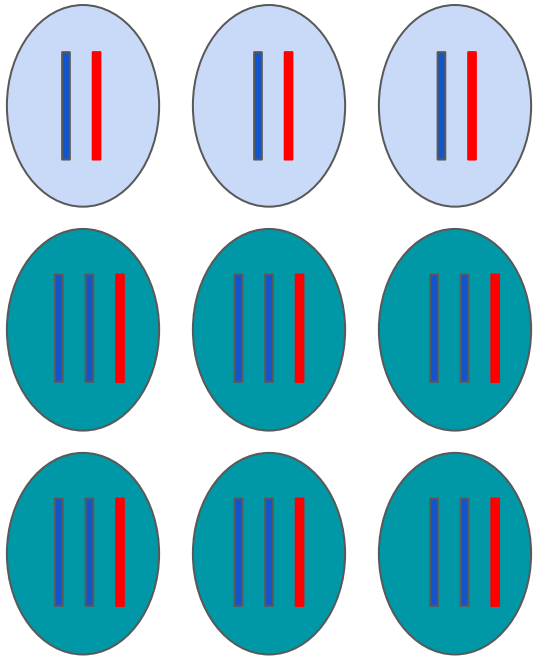


heterozygous mutation A->T (9 / 24)

Expected frequency: (12 / 24)



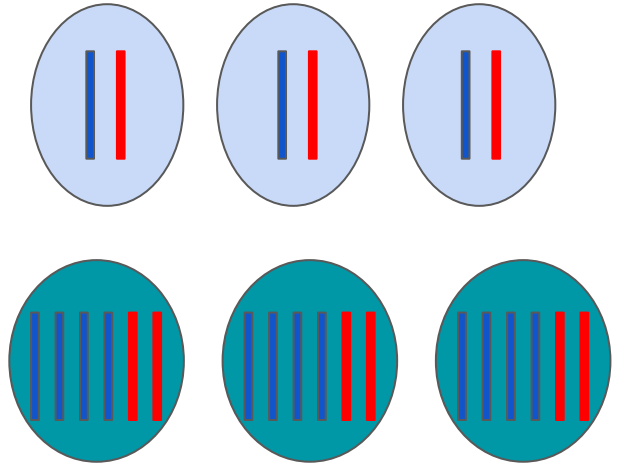
Purity: 67% & Ploidy: 3 (triploid)



A->T mutation (9 / 24)

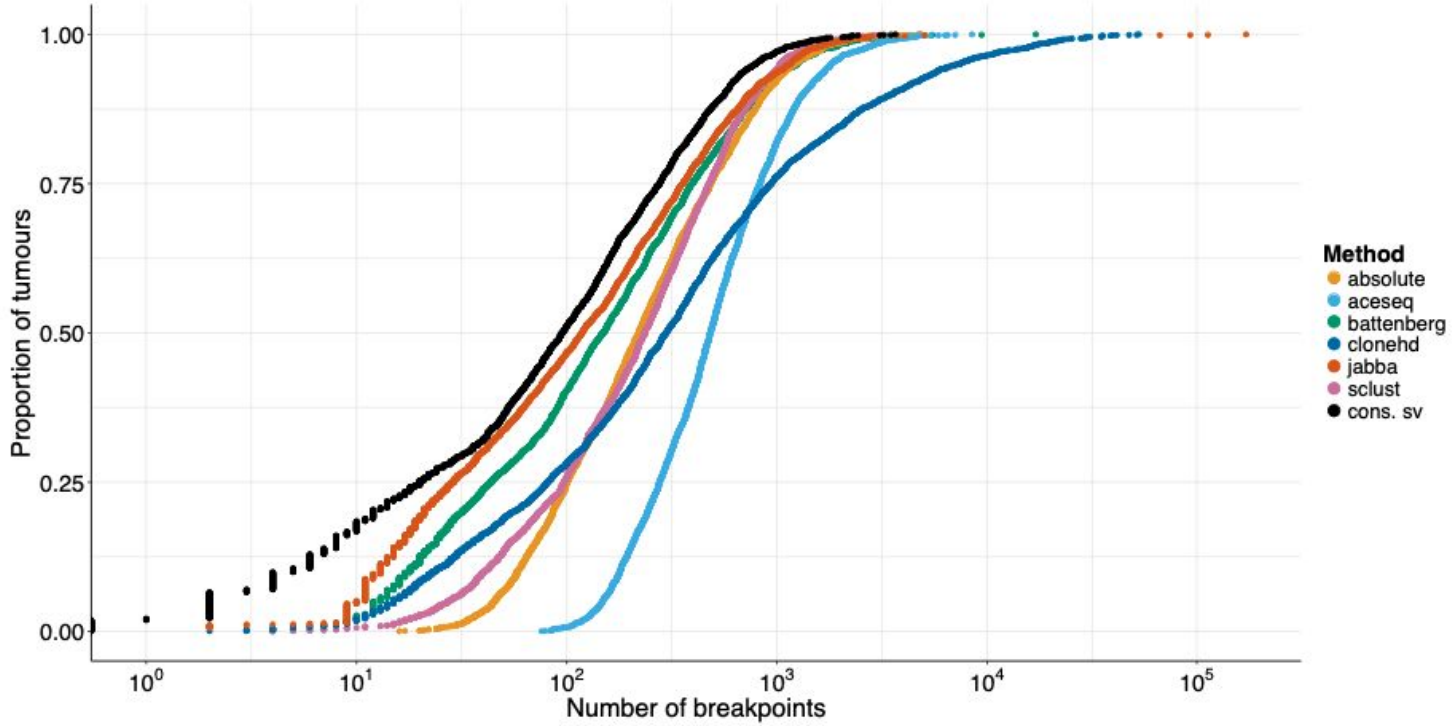
OR

Purity: 50% & Ploidy: 6 (hexaploid)



A->T mutation (9 / 24)

Inconsistent SCNAs calling results from different algorithms

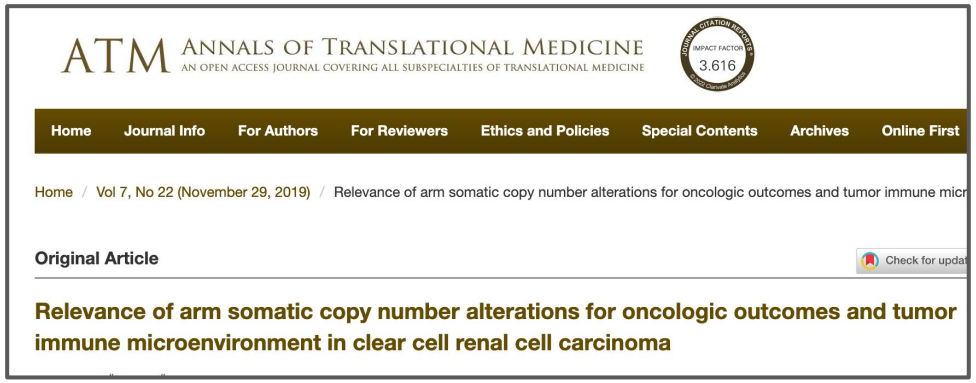


- Integrates results from multiple algorithms to increase the robustness of SCNA detection and account for the variability inherent in different algorithms.
- Cross-validation to compare different algorithms and determine which perform best for a particular dataset

Significance of SCNAs in clinical practice

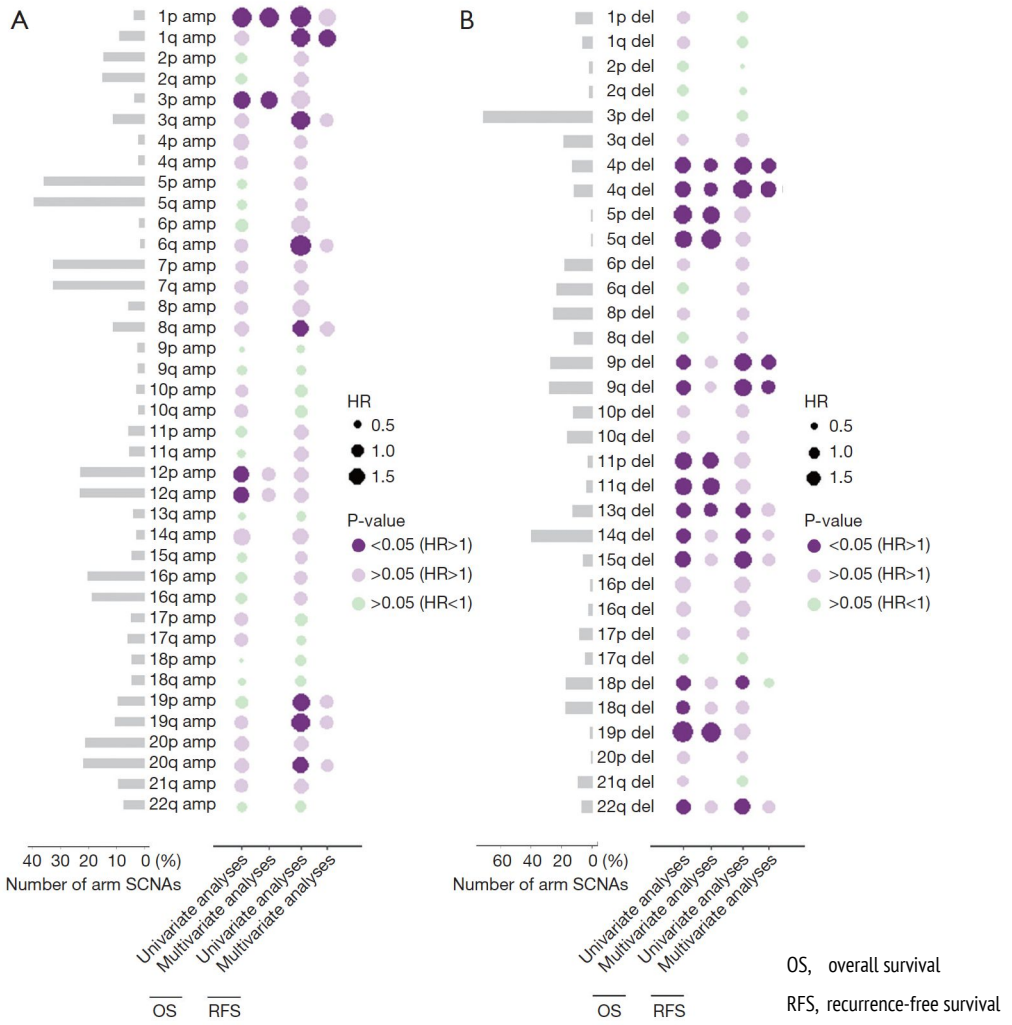
Clinical impacts of arm-level SCNAs in cancer

- Arm-Level SCNAs are prognostic markers associated with poorer prognosis and increased risk of recurrence in several cancer types, including breast cancer and ovarian cancer



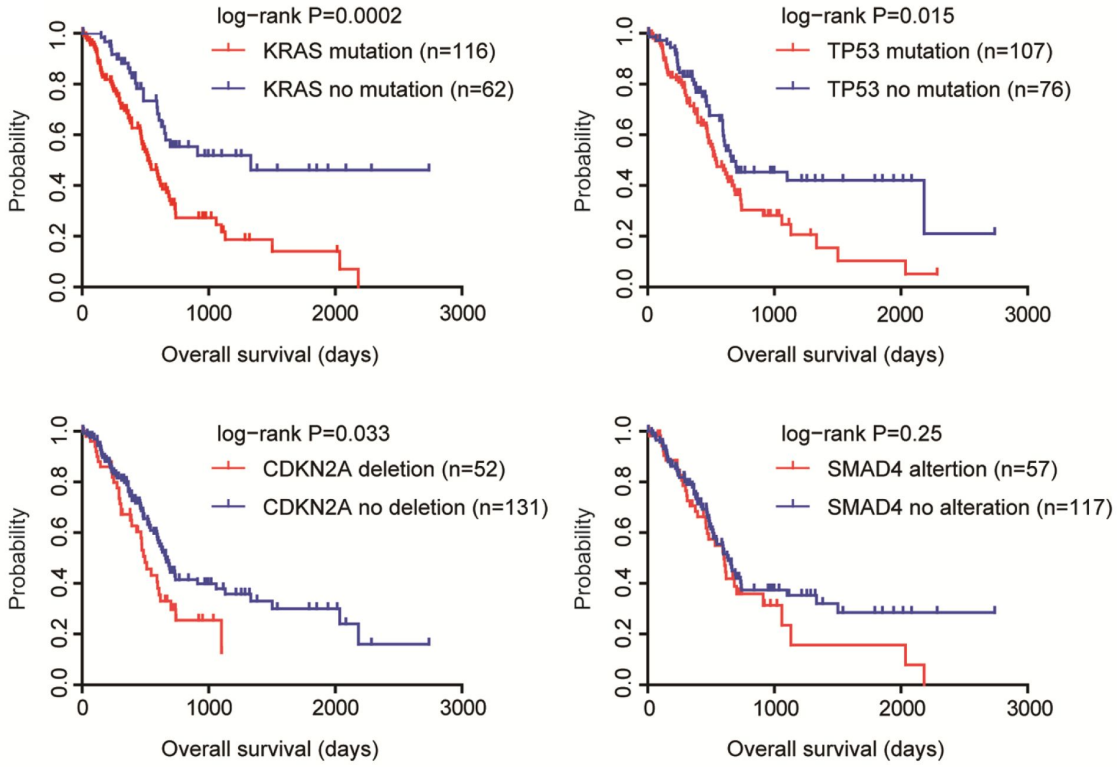
Several arm-level amplification and deletion were identified to be associated with poor survival in clear cell renal cell carcinoma, including amplification of 1p, 1q, 3p and 3q; and deletion of 4p, 4q, 9p, 9q, 13q, and 14q

(PMID: 31930047)



Clinical impacts of focal SCNAs in cancer

- In a pancreatic cancer genomic study, *CNDKA2A* deletion as well as *KRAS*, *TP53* and *SMAD4* mutation or alteration were identified to be associated with poor survival.



(PMID: 31930047)

- Through a large-scale genomic study in acute myeloid leukemia (AML), a group of four focal CNAs on chr11 and 21 were identified to be a prognostic marker with poor survival.
- The result was validated using the TCGA-LAML datasets

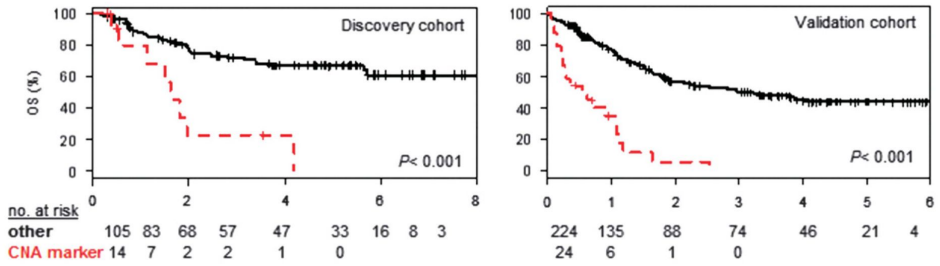


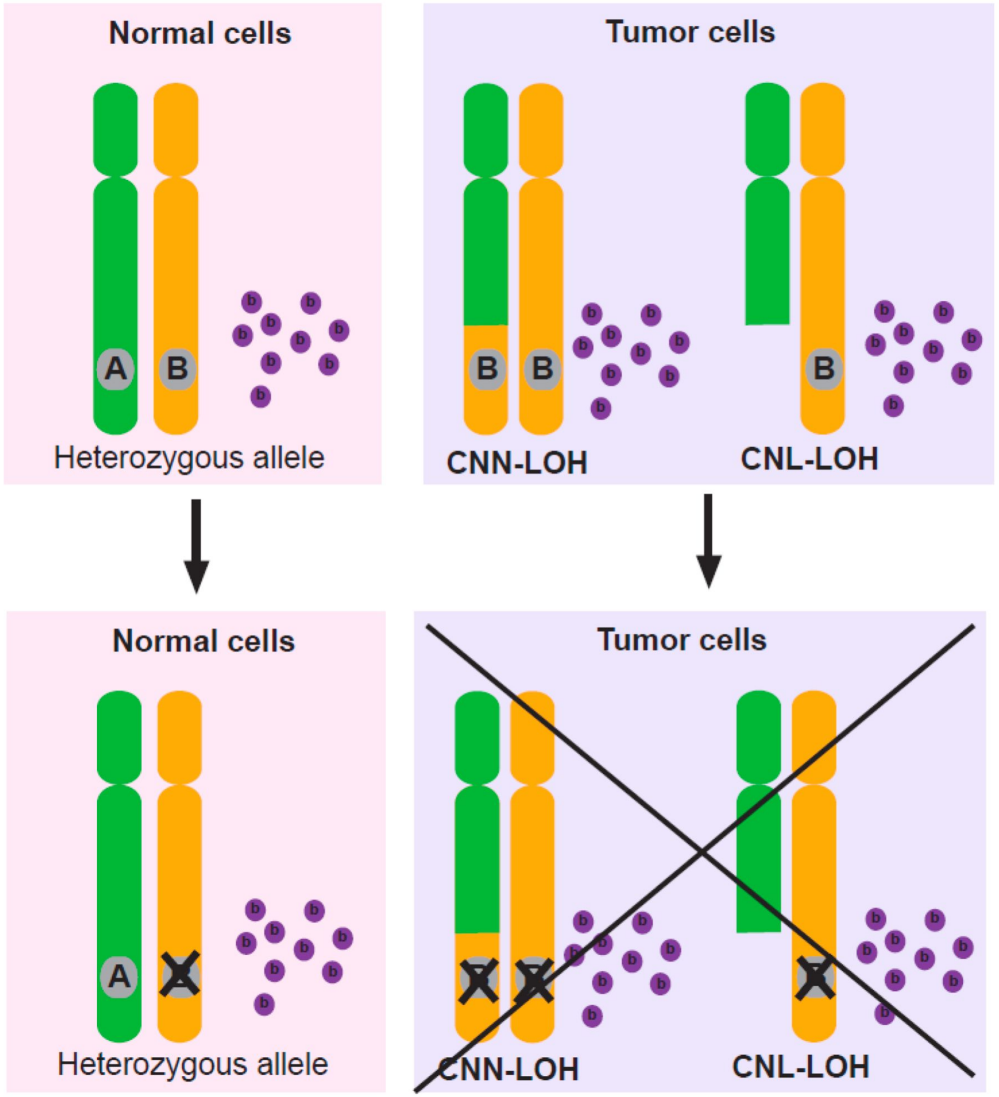
Table 1. Genes located in CNA marker on chromosomes 11 and 21

CNA marker	*Size (kb)*	*GS*	*Chr.*	*Start*	*End*	*Gene name*
1. Chr21:39,837,306-39,953,387 Cytoband: 21q22.2	116	ERG	21	39739182	40033704	*v-ets Erythroblastosis virus E26 oncogene homolog*
2. Chr21:43,388,231-44,744,122 Cytoband: 21q22.3	1356	ZNF295	21	43406939	43430496	*Zinc finger and BTB domain containing 21*
		ZNF295-AS1	21	43442112	43445060	*ZNF295 antisense RNA 1*
		UMODL1	21	43483067	43563105	*Uromodulin-like 1*
		C21orf128	21	43522243	43528644	*Chromosome 21 open reading frame 128*
3. Chr11:3,641,337-3,767,456 Cytoband: 11p15.4	126	TRPC2	11	3647689	3658789	*Transient receptor potential cation channel, subfamily C*
		ART5	11	3659735	3663546	*ADP-ribosyltransferase 5*
		ART1	11	3666360	3685646	*ADP-ribosyltransferase 1*
		CHRNA10	11	3686816	3692614	*Cholinergic receptor, nicotinic, alpha 10 (neuronal)*
		NUP98	11	3696239	3819022	*Nucleoporin 98 kDa*
4. Chr11:55,362,386-55,384,209 Cytoband: 11q11	22	OR4C11	11	55370829	55371874	*Olfactory receptor, family 4, subfamily C, member 11*

Abbreviations: Chr, chromosome; CNA, copy-number alteration; GS, gene symbol.

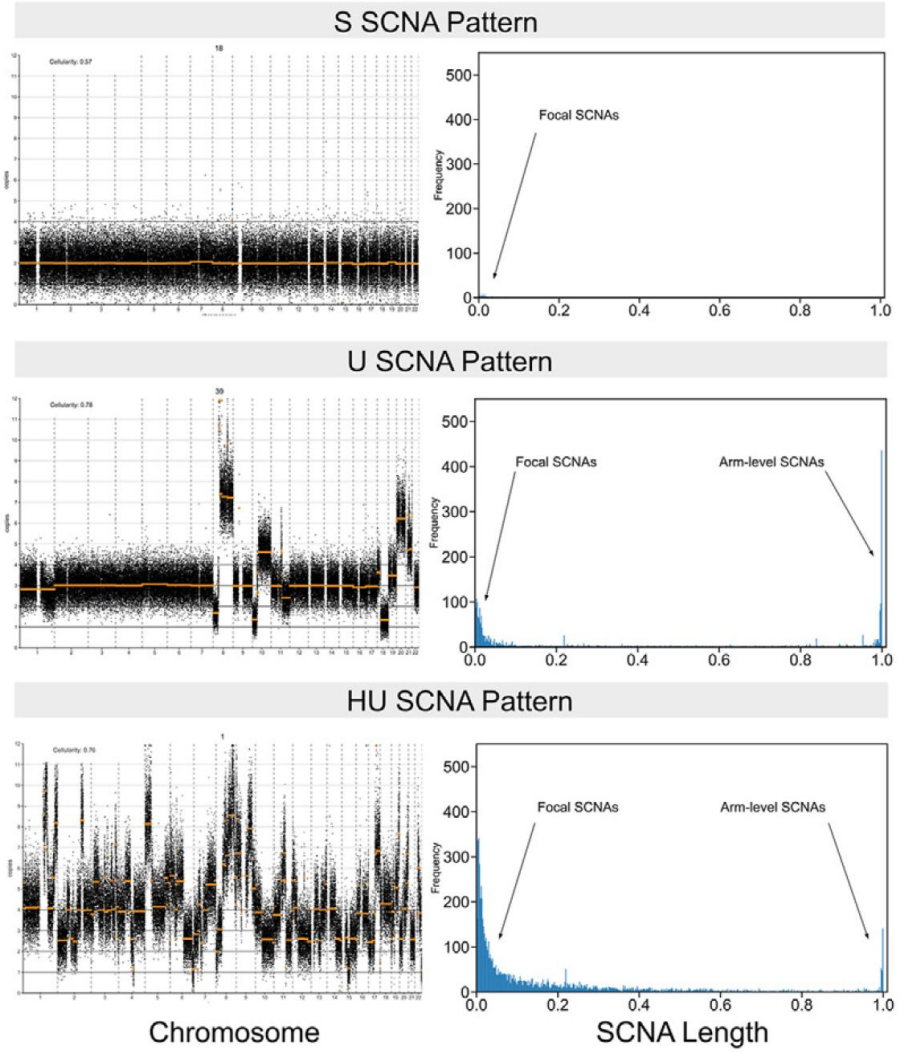
(PMID: 27686867)

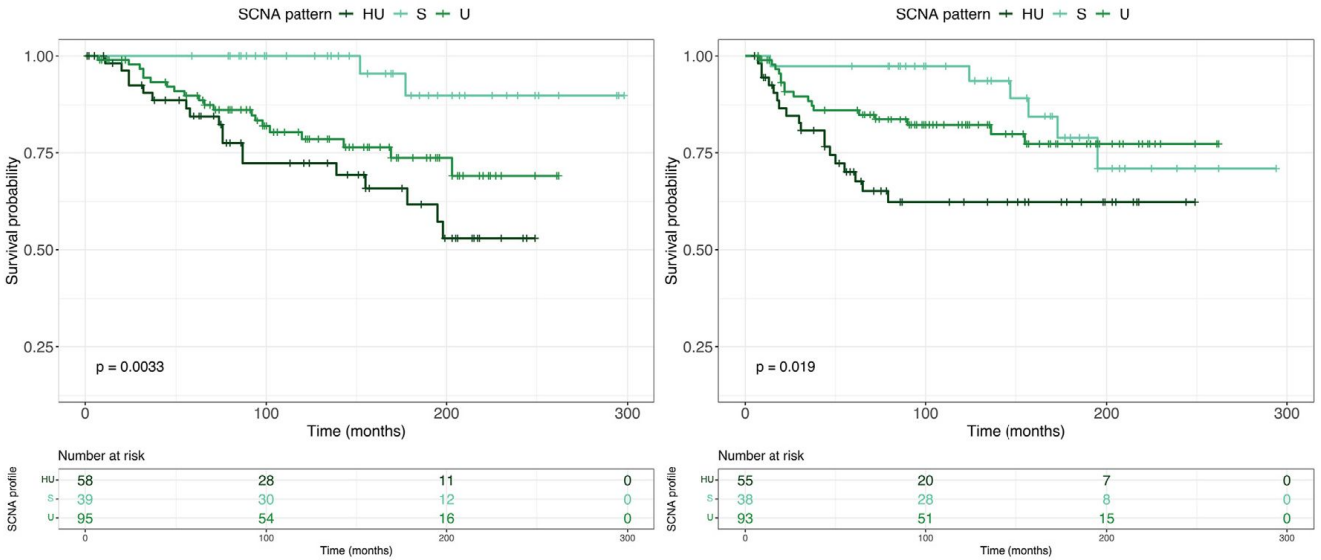
SCNAs provide potential novel therapeutic targets for cancer treatment



(PMID: 33450833)

SCNA profile is an effective marker for tumor molecular classification





- Researchers performed identification of SCNAs in 225 ovarian cancer patients, and found three patterns based on SCNA burden
- The patterns were named *S*, *U*, and *HU*, and correspond to low, median, and high SCNA burdens, respectively.
- Increasing SCNA burden results in lower survival odds.

(PMID: 35714451)

THANKS FOR YOUR ATTENTION!

Questions?

Next: Practical session 7 (10:45am)

- Identification of somatic copy number alterations
- Identification of significant SCNA regions
- Perform somatic copy number signature analysis