

CANCER GENOMICS

Lecture 4:

Additional Topics

GENOME 541
Spring 2020



Gavin Ha, Ph.D.
Public Health Sciences Division
Human Biology Division



@GavinHa



gha@fredhutch.org



<https://github.com/GavinHaLab>

GavinHaLab.org

Outline

1. Additional Copy Number Analysis Features

- Allelic copy number analysis

2. Estimating tumor heterogeneity

- Modeling tumor-normal admixture
- Modeling tumor clonality and heterogeneity

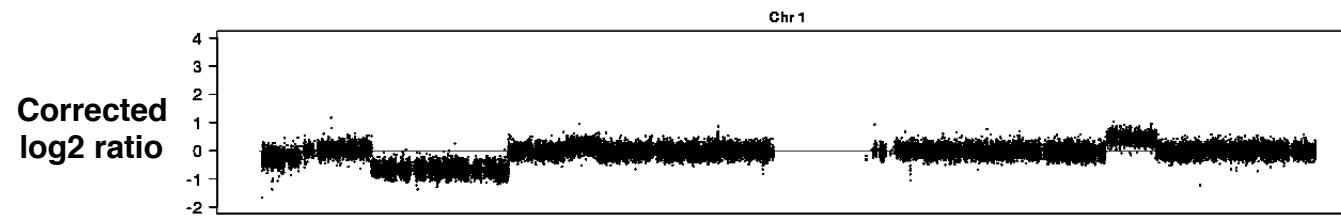
3. Assessing Statistical Power for Variant Discovery

- Power calculation
- Calibrating sequencing depth for variant discovery

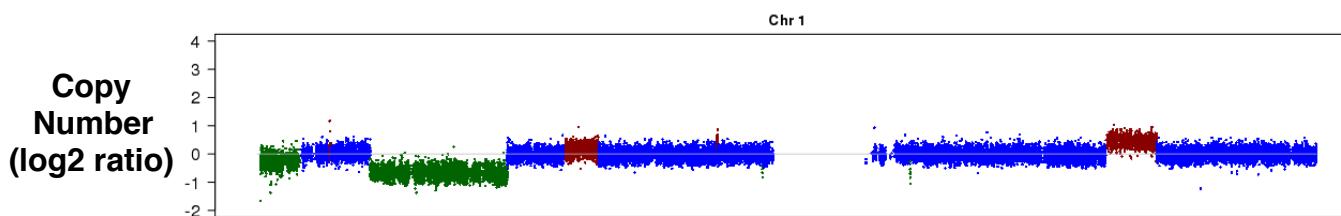
4. Structural Rearrangement Analysis in Cancer Genomes

- Structural variant types predicted from sequencing analysis
- Complex genomic structural rearrangement patterns

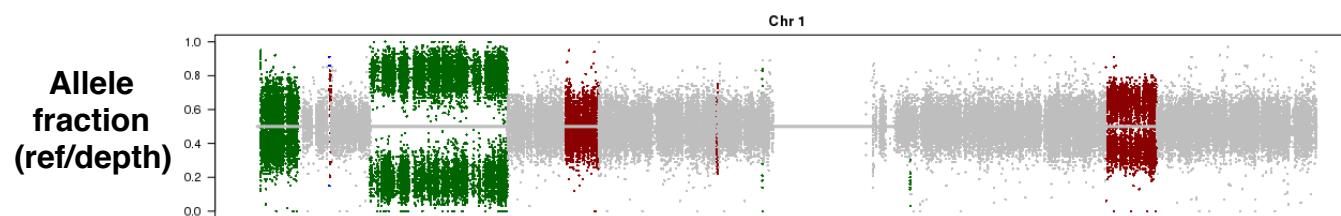
Allele-based Copy Number Analysis



Data normalization

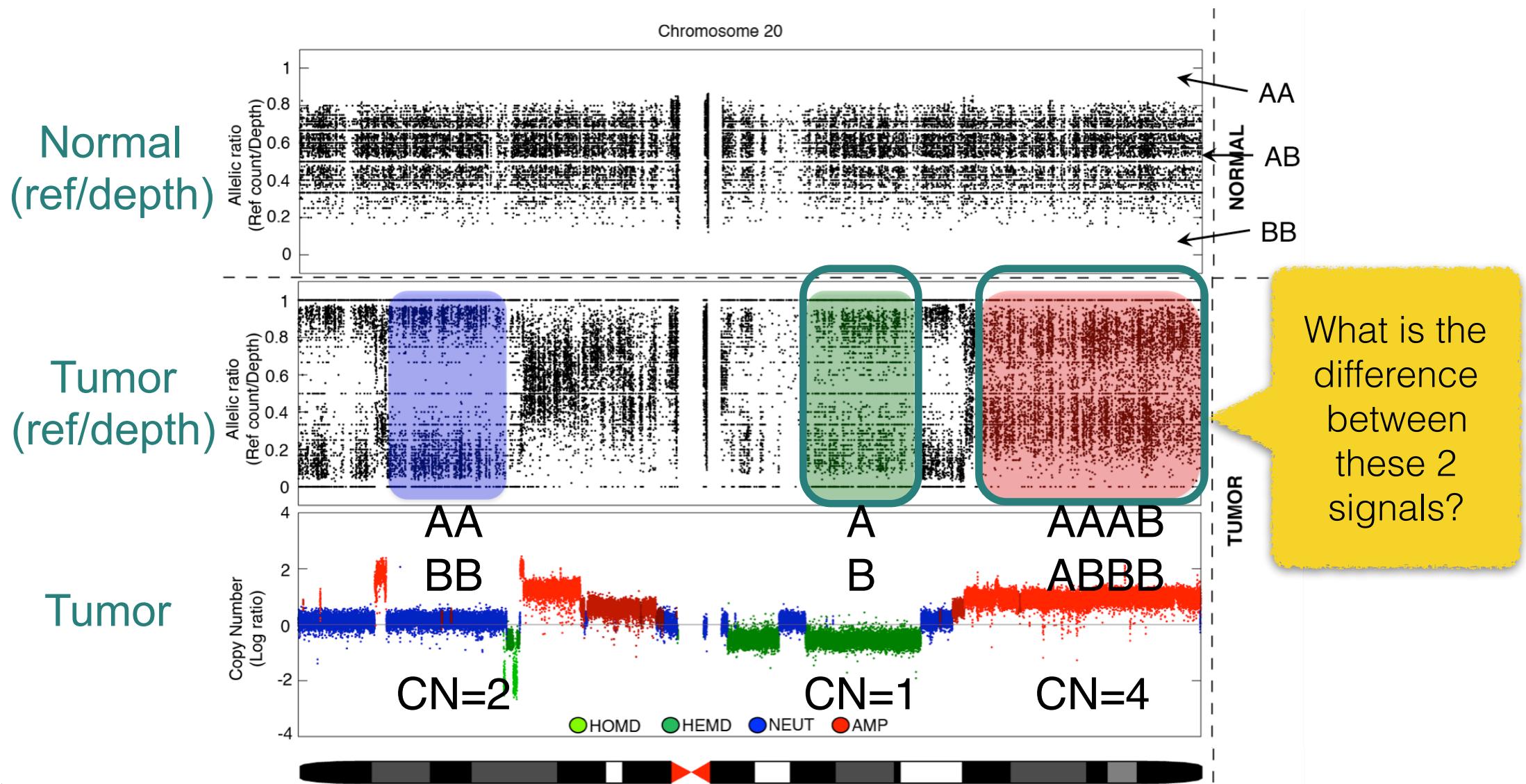


Total Copy Number Only

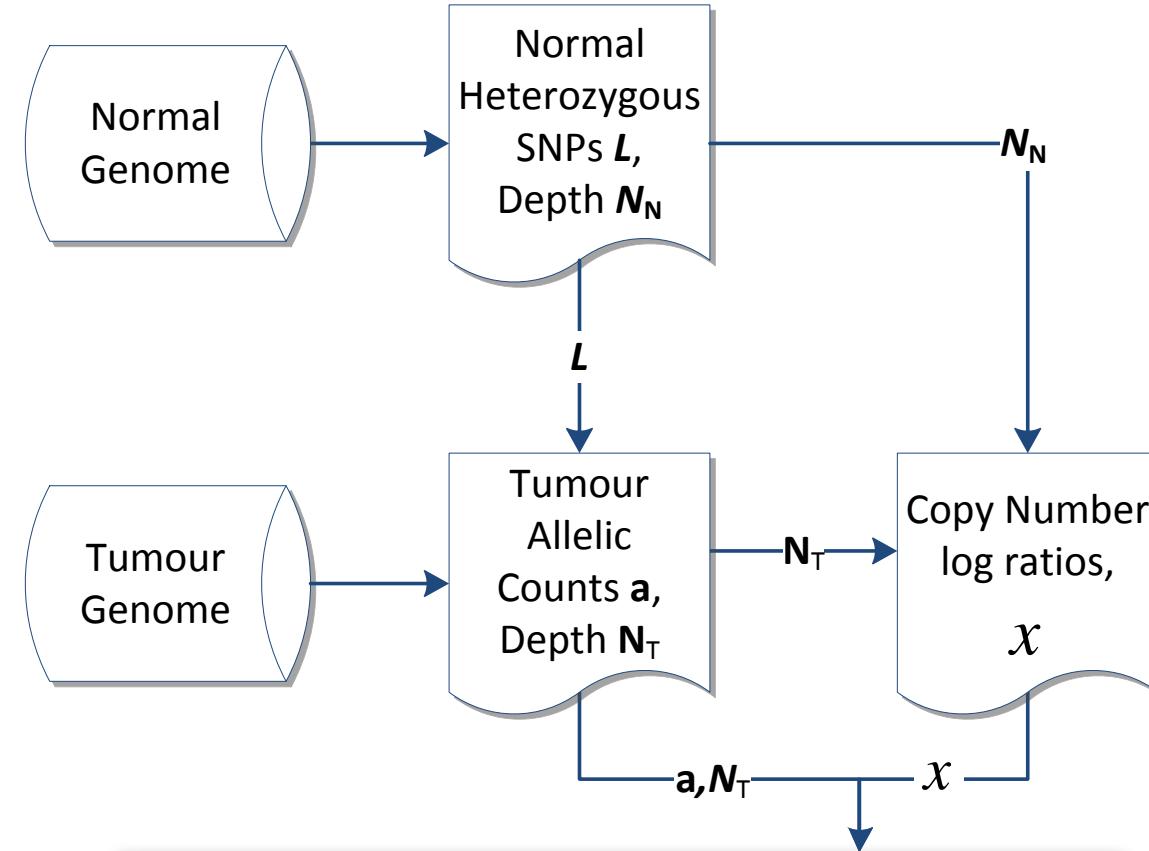


Allelic Copy Number

Copy Number Analysis: Allelic Features

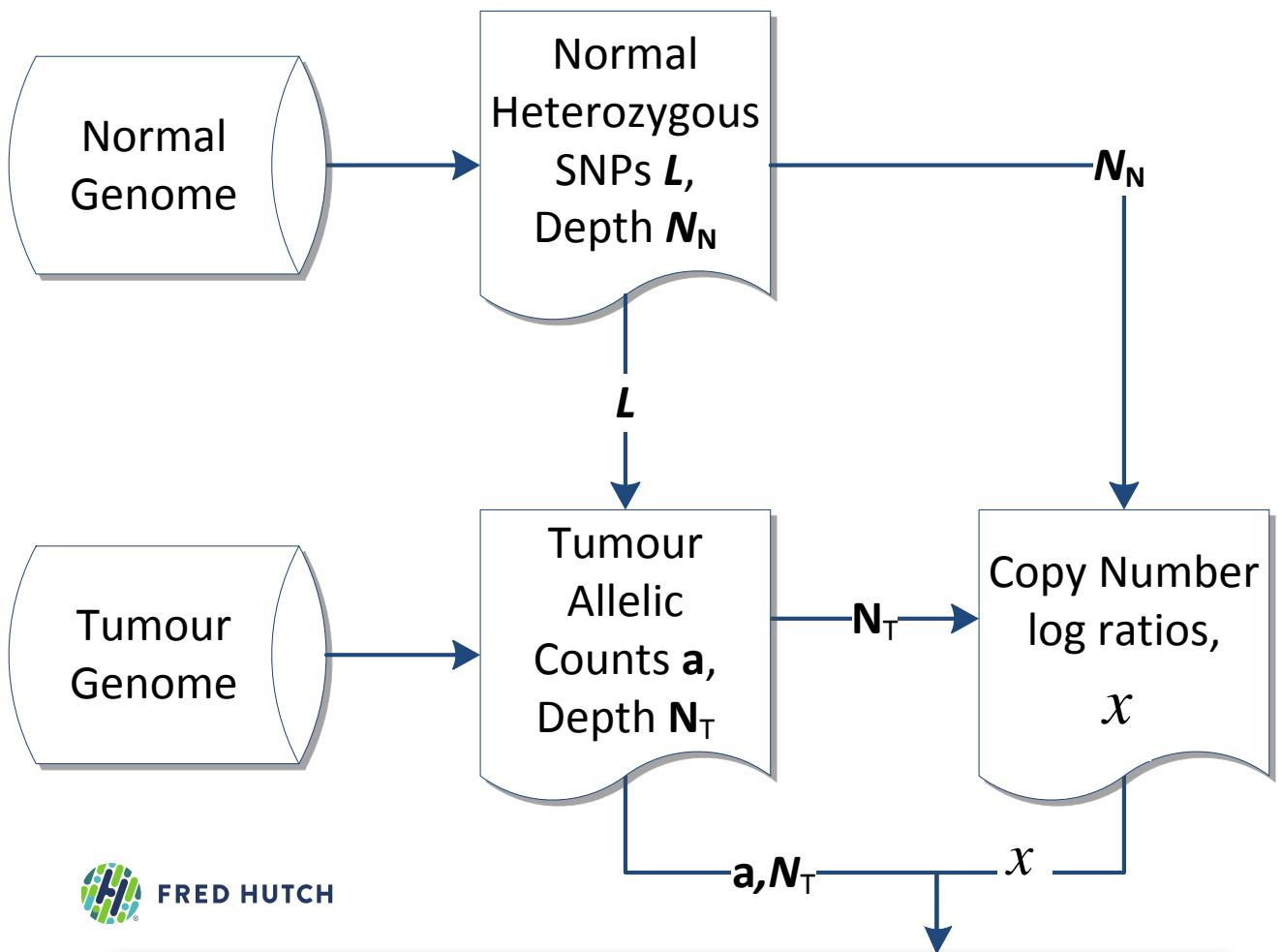


Cancer Genome Copy Number Analysis Workflow



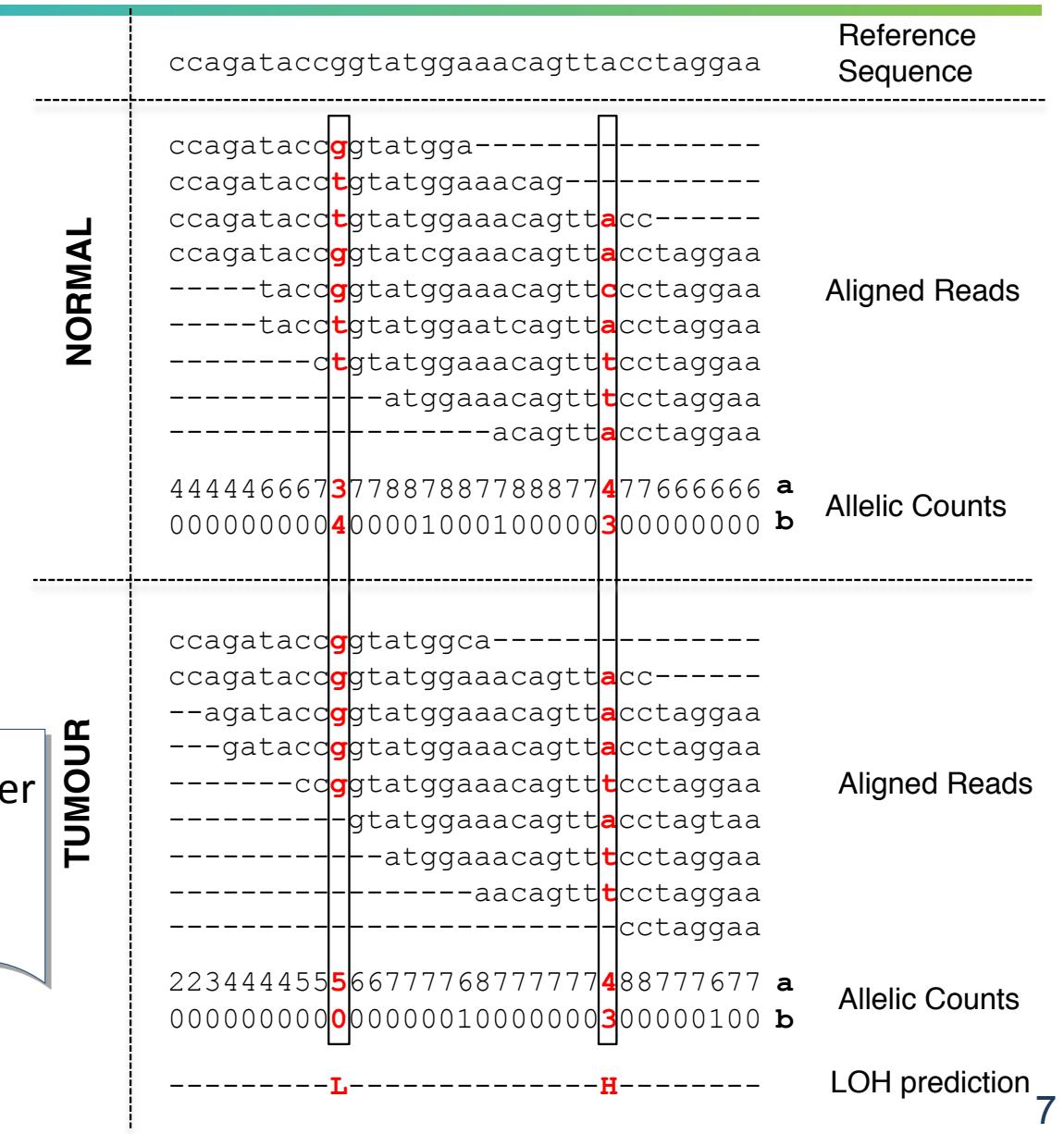
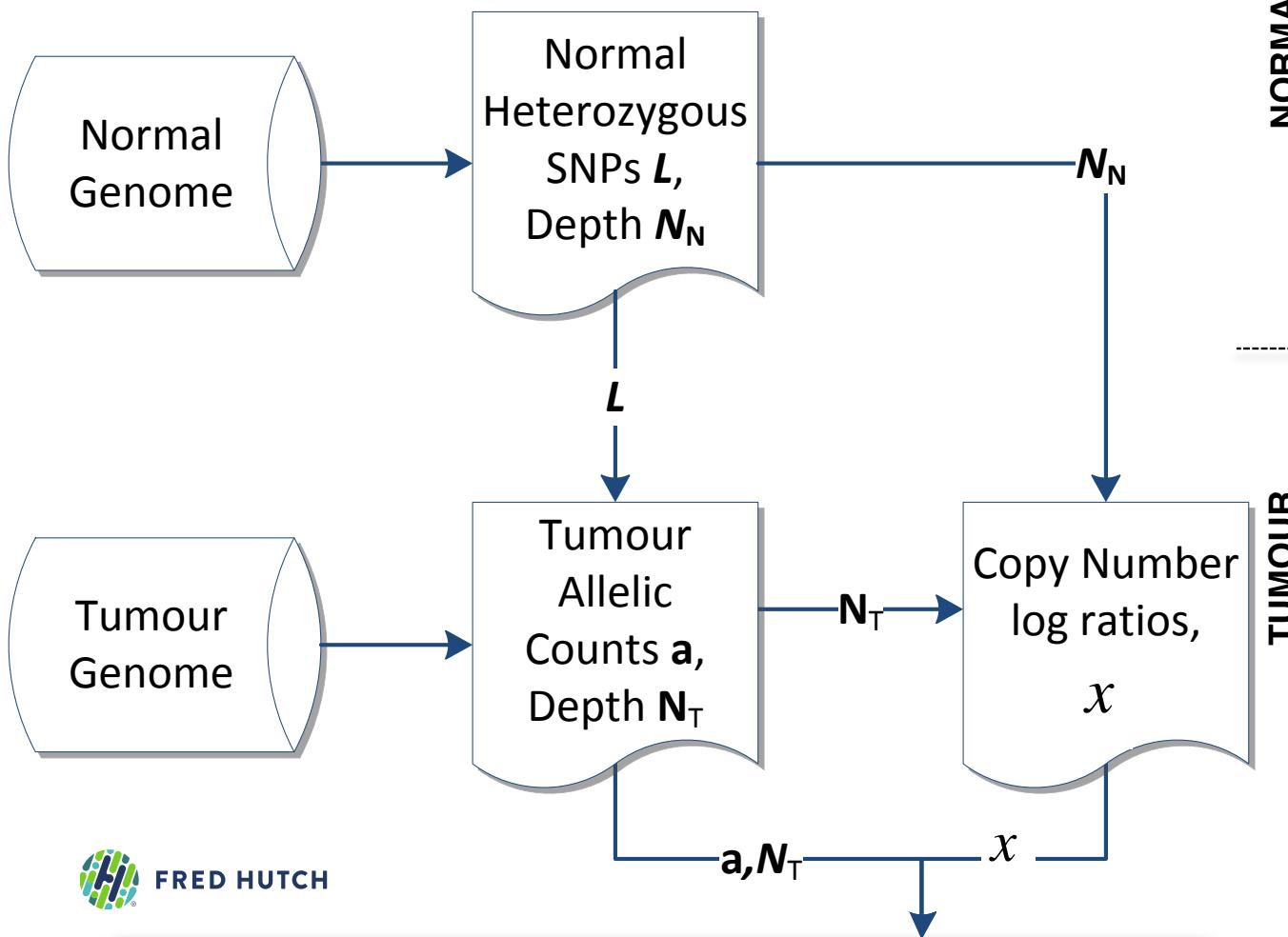
Copy Number Segmentation &
Prediction

Copy Number Analysis Workflow: Allele Features

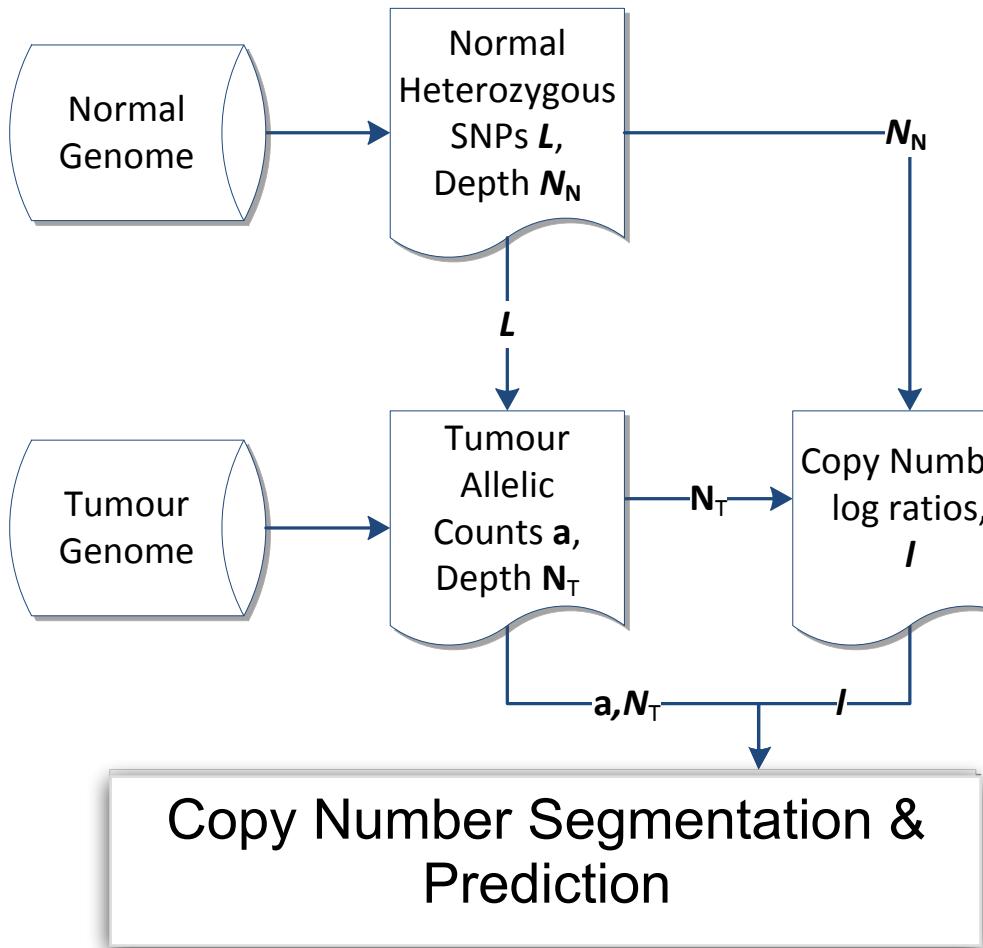


1. Correct GC/mappability biases for tumor read depth
2. Identify germline heterozygous SNPs from normal
3. Extract read counts at SNPs from tumor
4. Perform segmentation and copy number prediction

Copy Number Analysis Workflow: Allele Features



Copy Number Analysis Workflow: Segmentation



1. Correct GC/mappability biases for tumor read depth
2. Identify germline heterozygous SNPs from normal
3. Extract read counts at SNPs from tumor
4. **Perform segmentation and copy number prediction**

Probabilistic Model for Allelic Copy Number Analysis

Input Data: T different genomic loci

- log ratio data $x_{1:T}$
- reference counts $a_{1:T}$ and read depth $N_{1:T}$ for SNP data

Latent State Model: copy number states

There are 8 possible joint copy number state and allele genotype states.

Transition Model

The transition model is similar to before for matrix $A \in \mathbb{R}^{K \times K}$

Emission Model: likelihood for log ratio and allele data

The **emission model** is a mixture of bi-variate Gaussians for the joint likelihood

$$p(x_i, a_i | Z_i = k, N_i, \mu^c, \sigma^2, \mu^a) = \mathcal{N}(x_i | \mu_k, \sigma_k^2) \times \text{Bin}(a_i | \mu_k^a)$$

Prior Model

$$p(\boldsymbol{\pi} | \boldsymbol{\delta}^\pi) = \text{Dirichlet}(\boldsymbol{\pi} | \boldsymbol{\delta}^\pi)$$

$$p(\mu_k^c | m_k, s_k) = \mathcal{N}(\mu_k^c | m_k, s_k)$$

$$p(\sigma_k^2 | \alpha_k, \beta_k) = \text{InvGamma}(\sigma_k^2 | \alpha_k^c, \beta_k^c)$$

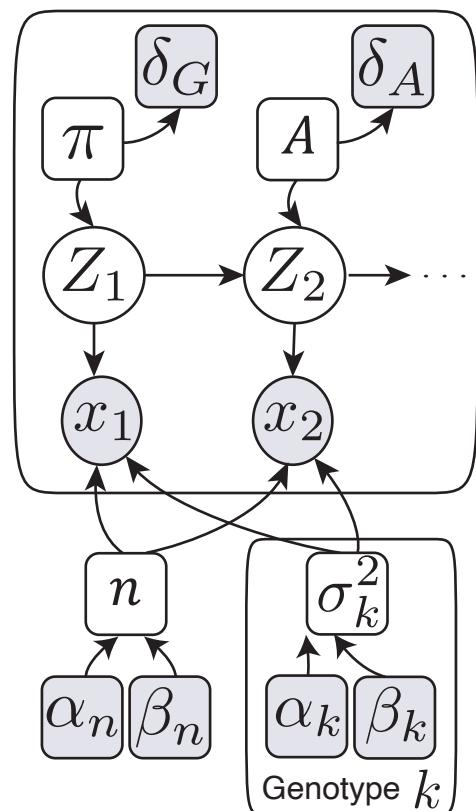
$$p(\mu_k^a | \alpha_k, \beta_k) = \text{Beta}(\mu_k^a | \alpha_k^a, \beta_k^a)$$

$$p(A_{k,1:K} | \boldsymbol{\delta}^A) = \text{Dirichlet}(A_{k,1:K} | \boldsymbol{\delta}_k^A)$$

K	Genotype	CN
1	A/B	1
2	AA/BB	2
3	AB	2
4	AAA/BBB	3
5	AAB/ABB	3
6	AAAA/BBB	4
7	AAAB/ABBB	4
8	AA/BB	4

2. Estimating tumor heterogeneity

- Estimating tumor heterogeneity from copy number analysis
- References:

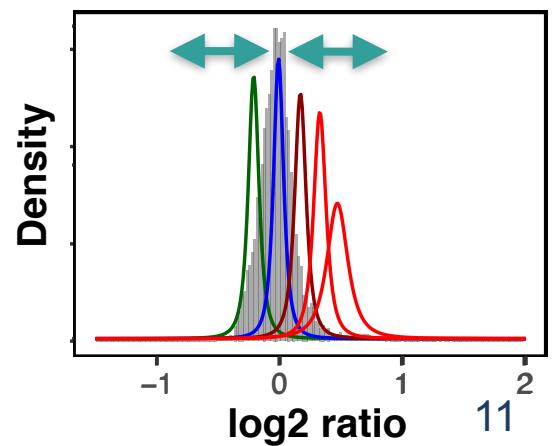
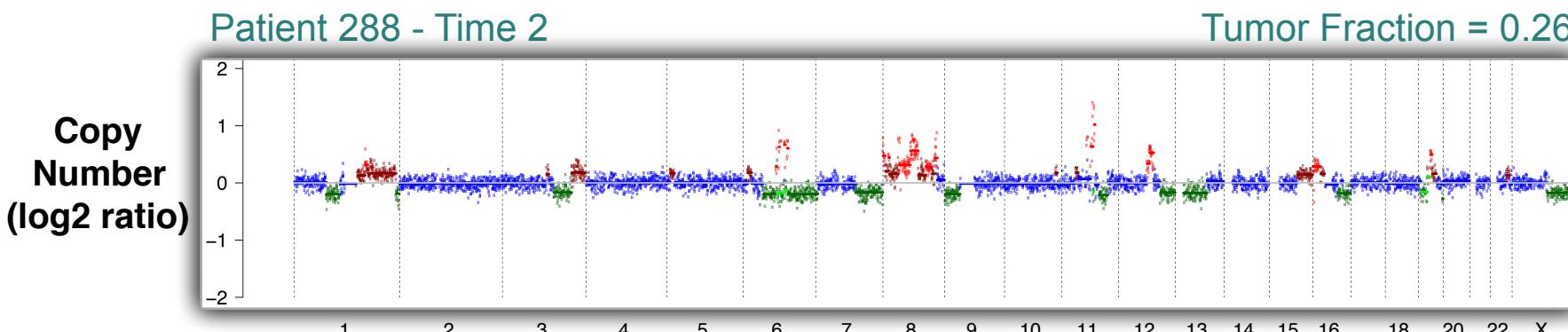
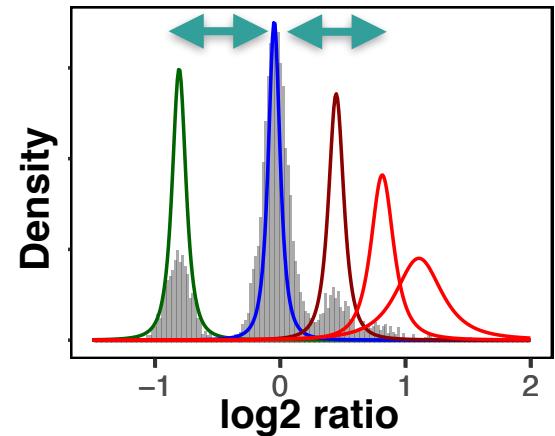
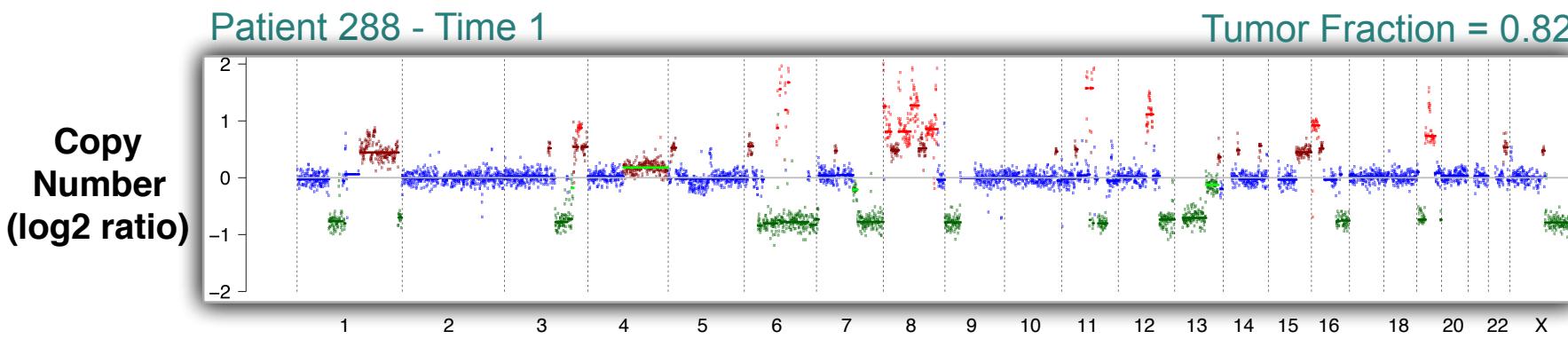


- **ichorCNA** - Adalsteinsson*, Ha* Freeman* et al. *Nature Communications* **8**:1324 (2017).
- **HMMcopy** - Ha et al. *Genome Research* **22**:1995-2007 (2012).
- **TitanCNA** - Ha et al. TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequencing data. *Genome Research* **24**:1881-1893 (2014).
- Murphy, K. (2012). Machine Learning: A Probabilistic Perspective. MIT Press. ISBN: 9780262018029
- Bishop, C. M. (2006). Pattern Recognition and Machine Learning (Information Science and Statistics). Springer. ISBN: 0387310738

Modeling tumor-normal admixture

Why estimate the model parameters $\mu = \{\mu_0, \dots, \mu_5\}$ and $\sigma^2 = \{\sigma_0^2, \dots, \sigma_5^2\}$?

- Data variability due to sequencing depth (technical) and *tumor heterogeneity* (biological)



Modeling tumor-normal admixture

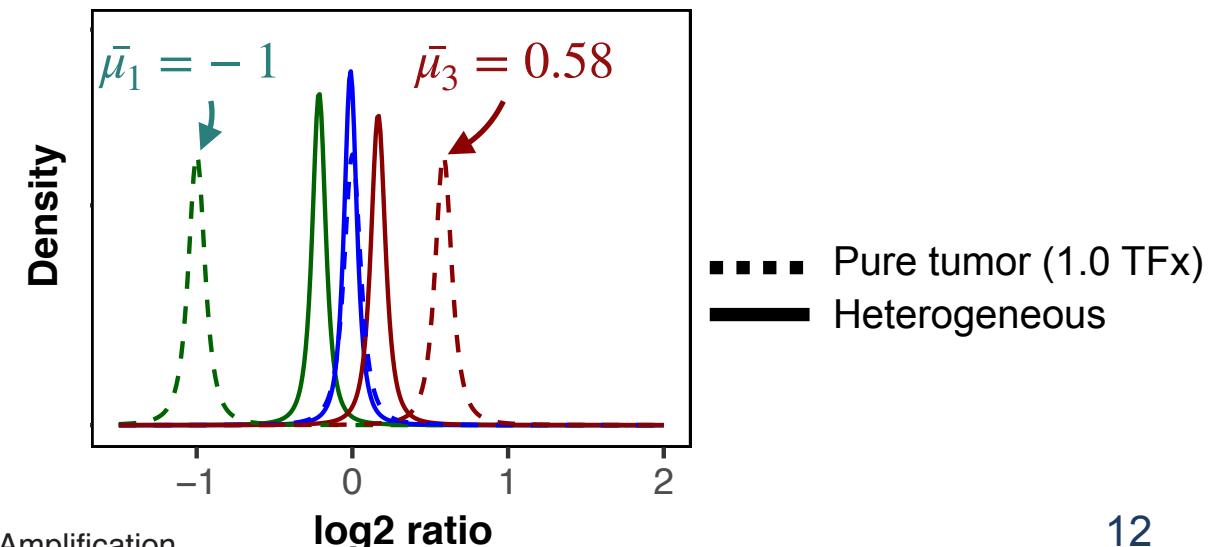
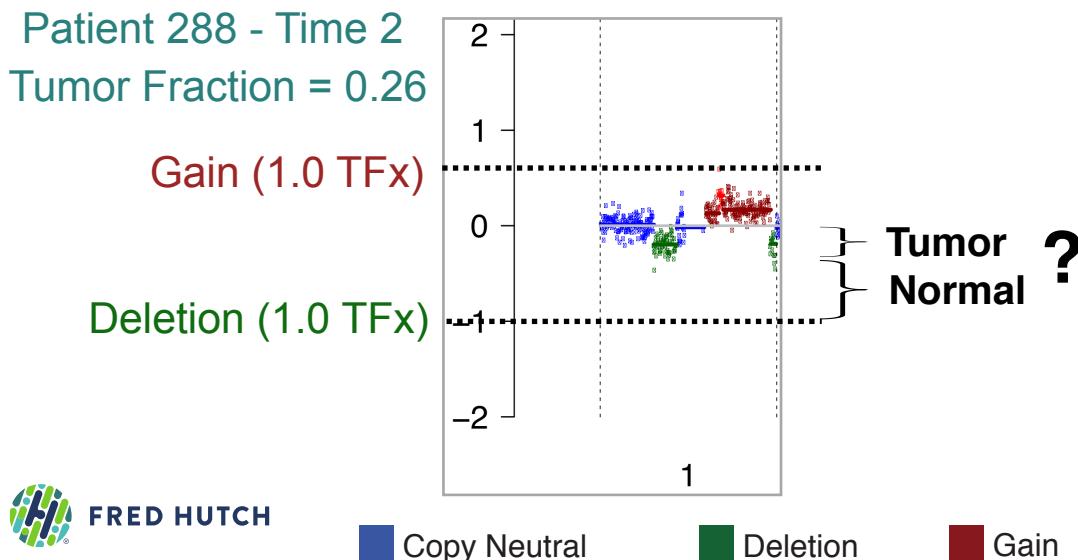
The mean (μ) of the copy number state mixture components can inform the tumor fraction.

- Recall: the log ratio input data is computed as

$$x_t = \log_2 \left(\frac{\hat{r}_t^{\text{Tumor}}}{\hat{r}_t^{\text{Normal}}} \right)$$

- For number $c_k \in \{1, 2, 3, 4, 5\}$, a pure tumor with 1.0 tumor fraction copy will have log ratios $\bar{\mu}_{1:K}$

$$\bar{\mu}_{1:K} = \log \left(\frac{c_{1:K}}{2} \right) = \left\{ \log_2 \left(\frac{1}{2} \right), \log_2 \left(\frac{2}{2} \right), \log_2 \left(\frac{3}{2} \right), \log_2 \left(\frac{4}{2} \right), \log_2 \left(\frac{5}{2} \right) \right\}$$



Modeling tumor fraction as a parameter

- A tumor biopsy contains both tumor and normal cells

$$\text{tumor signal} \approx [(1 - n) \times \text{tumor CN}] + [n \times \text{normal CN}]$$

- n is the fraction of non-cancer cells
- $(1 - n)$ is the fraction of cancer cells
- Typically $\text{normal CN} = 2$

- Then, the expected log ratio can be written as

$$\bar{\mu}_k = \log_2 \left(\frac{c_k}{2} \right)$$

Pure tumor

$$\mu_k = \log_2 \left(\frac{2n + (1 - n)c_k}{2} \right)$$

Tumor-normal admixture
(Heterogeneous)

where $c_k \in \{1, 2, 3, 4, 5\}$ is the tumor copy number for state k

- Let's use some examples of *deletions* (CN=1) from the Slide 11:

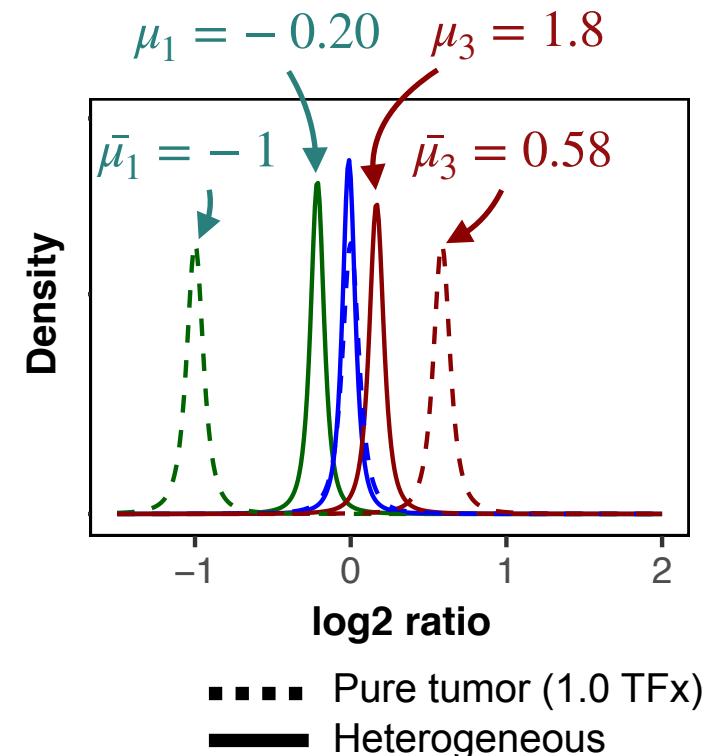
$$\bar{\mu}_1 = \log_2 \left(\frac{2(0) + (1 - 0)(1)}{2} \right) = -1$$

Pure tumor
($n = 0$)

$$\log_2 \left(\frac{2(0.74) + (1 - 0.74)(1)}{2} \right) = -0.20$$

Tumor-normal admixture
($n = 0.74$)

- Note that this formulation does not account for genome doubling in the tumor which would involve a tumor ploidy parameter ϕ and denominator of the ratio would be $2n + (1 - n)\phi$ instead of just 2



Modeling tumor fraction as a parameter

- The expected log ratio for copy number state k is

$$\mu_k = \log_2 \left(\frac{2n + (1 - n)c_k}{2} \right), \text{ where } c_k \in \{1, 2, 3, 4, 5\}$$

- Recall the likelihood model:

$$p(x_i | Z_i = k, \mu, \sigma^2) = \mathcal{N}(x_i | \mu_k, \sigma_k^2)$$

- Since μ_k is now a function of n , we no longer need to estimate μ_k .
- However, the non-cancer proportion n is what we want to estimate to obtain the tumor fraction $(1 - n)$.

$$p(\mu_k | m_k, s_k) = \mathcal{N}(\mu_k | m_k, s_k)$$

$$p(n | \alpha_n, \beta_n) = \text{Beta}(n | \alpha_n, \beta_n) \quad \text{Prior for } n$$

**Log Posterior
(with n terms)**

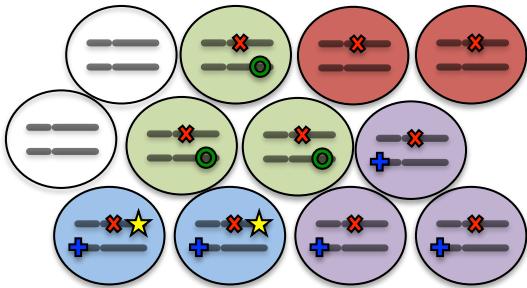
- Take the derivative wrt to n
- Equate to 0
- Find the roots to estimate n

$$\log \mathbb{P}(n) = \sum_{t=1}^T \sum_{k=1}^K \gamma(Z_t = k) \log \mathcal{N}(x_t | \mu_k, \sigma_k^2) + \sum_{k=1}^K \log \text{Beta}(\mu_k | \alpha_n, \beta_n)$$
$$\frac{\partial(\log \mathbb{P}(n))}{\partial \mu} \times \frac{\partial \mu}{\partial n} = \frac{\partial(\log \mathbb{P}(n))}{\partial n} = 0, \text{ then find } n$$

Since the Beta distribution is not conjugate with the Gaussian, we can use numerical optimization to find \hat{n} that maximizes the log Posterior

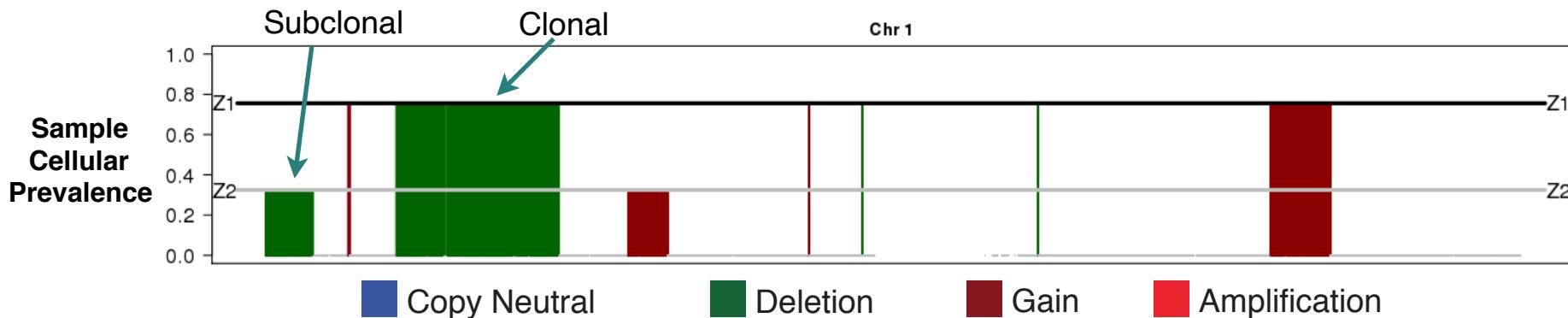
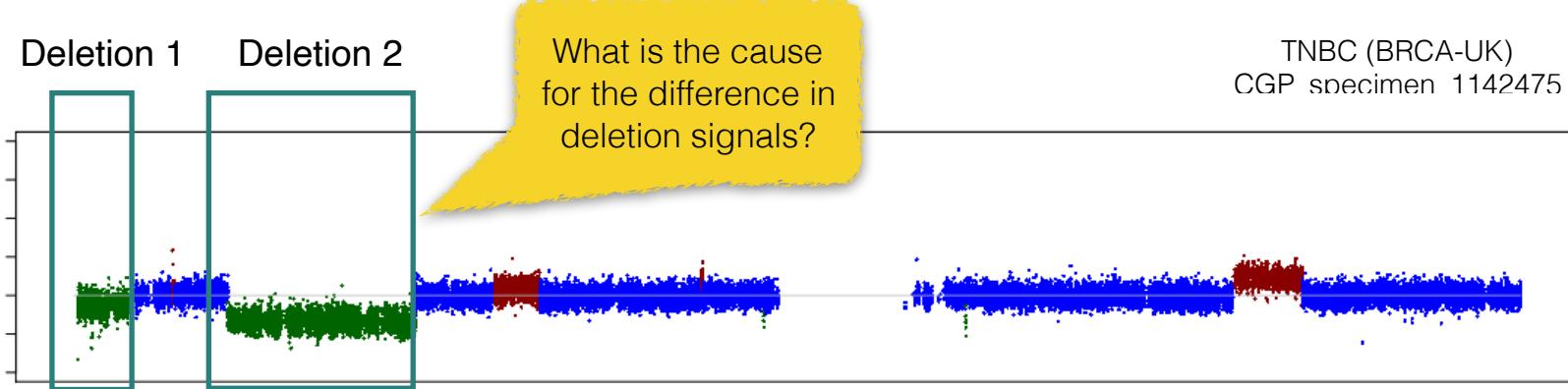
Copy Number Analysis: Clonality

Heterogeneous tumour biopsy



Cellular prevalence

✖ 100%	Clonal
✚ 50%	
● 30%	Subclonal
★ 20%	



- Subclonal CNA events have weaker signals compared to clonal CNAs because of contribution from *non-tumor cells* with normal copy number signals

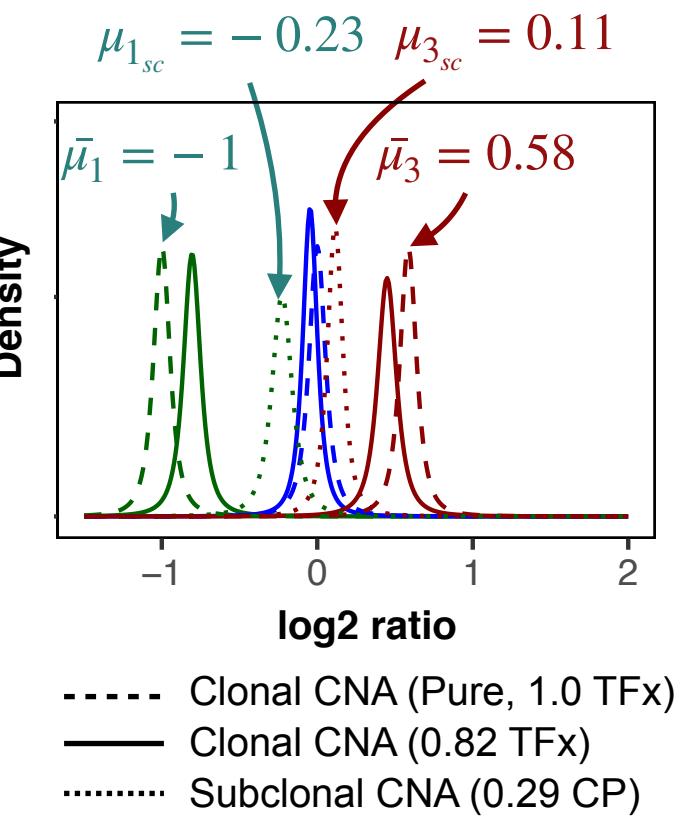
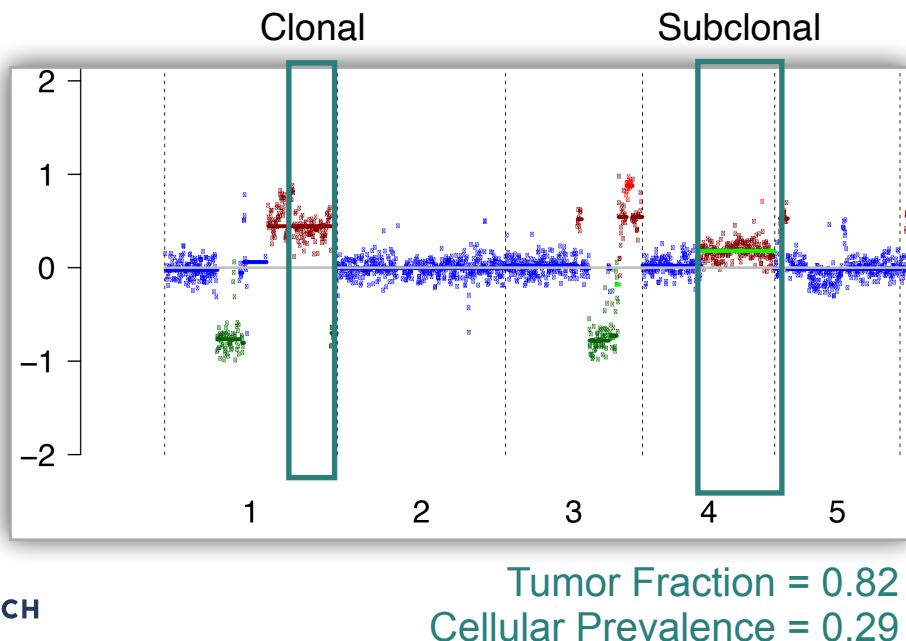
ichorCNA: Modeling subclonal copy number

- Add two additional states for subclonal deletion and subclonal gain, $K_{sc} = \{1, 3\}$, $K = \{0, 1, 2, 3, 4, 5, K_{sc}\}$
- The expected log ratio for subclonal copy number state $k_{sc} \in \{1, 3\}$ is

$$\mu_{k_{sc}} = \log_2 \left(\frac{2n + (1-n)s + (1-n)(1-s)c_{k_{sc}}}{2} \right)$$

Normal Tumor w/o event Tumor w/ event

- s is the fraction of **cancer cells without** CNA event
- $(1 - s)$ is the fraction of **cancer cells with** CNA event (aka tumor cellular prevalence)



3. Assessing Statistical Power for Variant Discovery

- Power calculation
- Calibrating sequencing depth for variant discovery
- References:
 - Cibulskis et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology* **31**:213-19 (2013)
 - Adalsteinsson et al. *Nature Communications* **8**:1324 (2017). DOI: 10.1038/s41467-017-00965-y

Sensitivity of Mutation Calling is Subject to Heterogeneity

- Tumor biopsy samples may exhibit intra-tumor heterogeneity
 - The tumor fraction (aka tumor content) influences our ability to detect an SNV at a specific locus
- Here are some statistical considerations:
 - What is our power (sensitivity) to detect an SNV given the read depth?
 - What read depth is required to detect an SNV at a specific power?
 - If we do not detect a mutation, is it because (1) there is no mutation? Or (2) we do not have sufficient power to make a confident call?
- Answering these questions with theoretical power calculations can help to calibrate the required sequencing depth and the expectation to detect mutations.

Power Calculation for Mutation Detection

- Let μ be the expected probability of observing a variant read at a locus
- Tumor fraction α , copy number c , and multiplicity M

$$\mu = \frac{\alpha M}{ac + 2(1 - \alpha)}$$

average tumor copies average normal copies

“average # of chromosomes with the variant tumor cells in the sample”

“average # of chromosomes from all cells in sample”

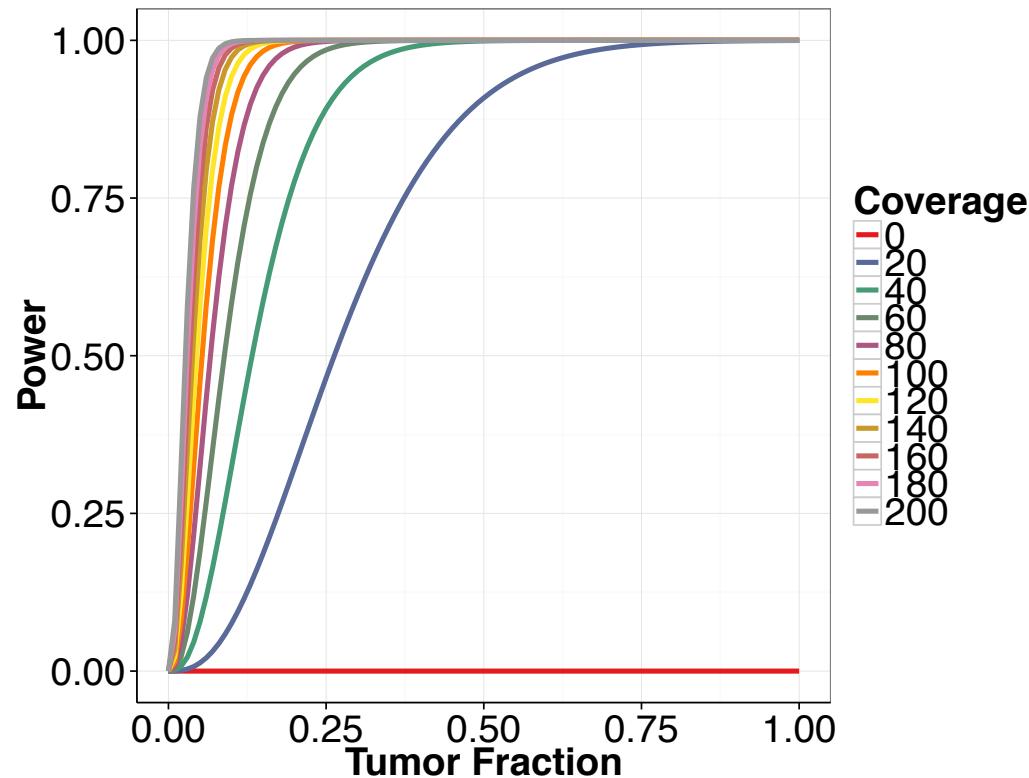
- $\mu = \frac{\alpha}{2}$ for tumor copy number $c = 2$ and multiplicity $M = 1$ (for heterozygous SNV, e.g. AB)
- The power to detect ≥ 3 variant reads at locus i with N_i total read depth is estimated using a binomial exact test

$$p(X \geq 3) = \sum_{k=3}^N \text{Bin}(k | N, \mu)$$

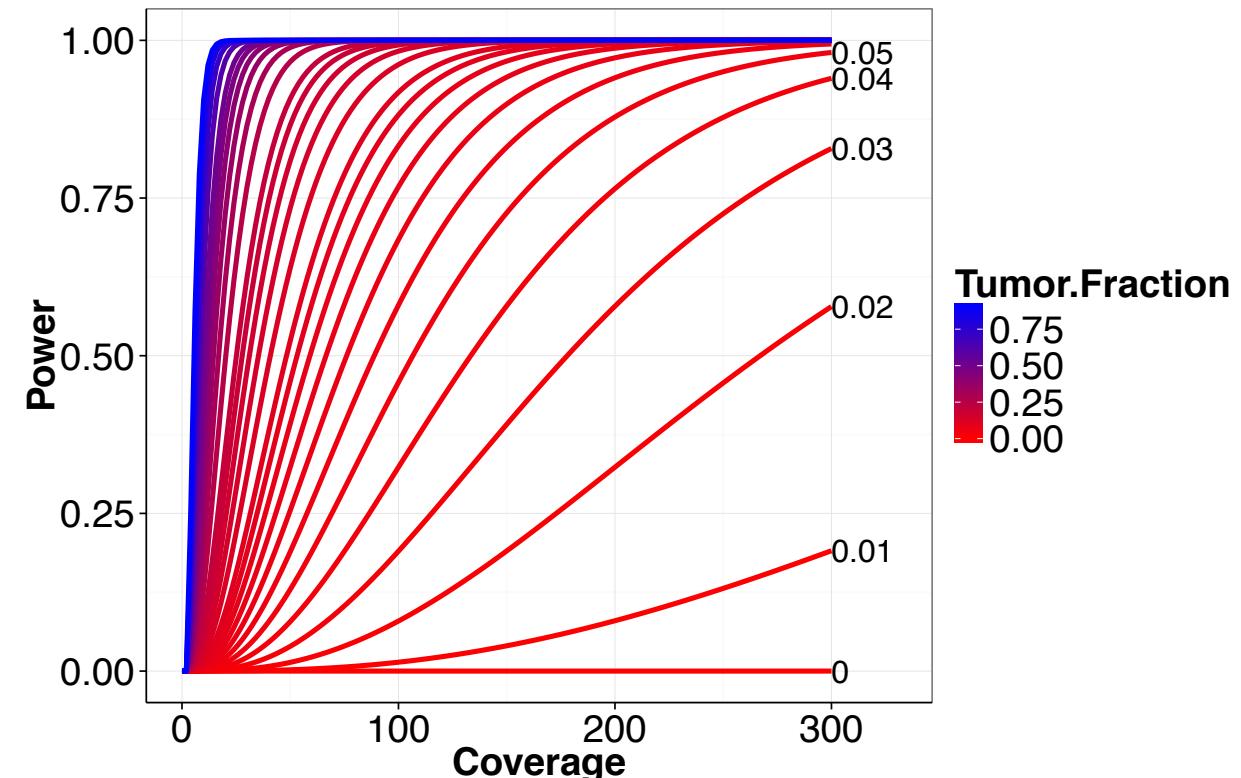
$$p(X \geq 3) = 1 - [\text{Bin}(0 | N, \mu) + \text{Bin}(1 | N, \mu) + \text{Bin}(2 | N, \mu)]$$

Power Calculation for Mutation Detection

What is our power (sensitivity) to detect an SNV at a specific tumor fraction?

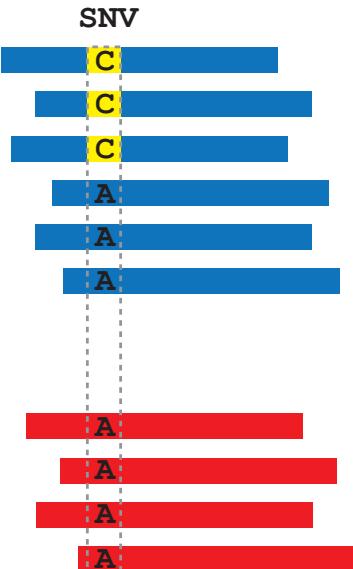


What read depth is required to detect an SNV at a specific power?

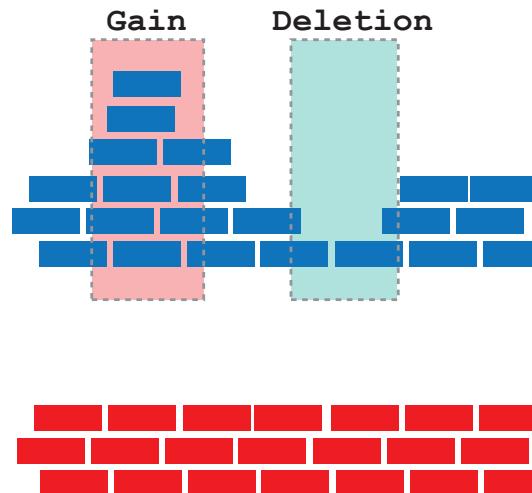


4. Structural Rearrangement Analysis in Cancer Genomes

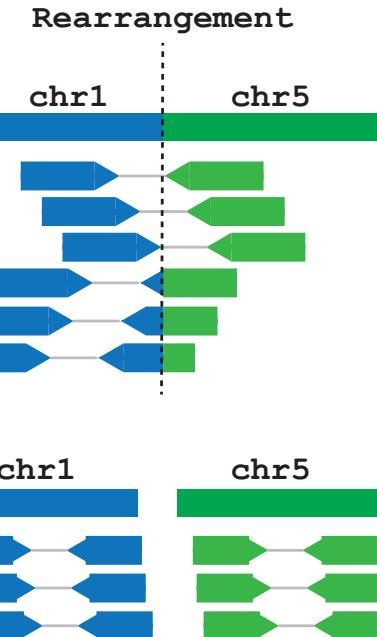
Mutations (SNV, INDEL)



Copy Number Alterations



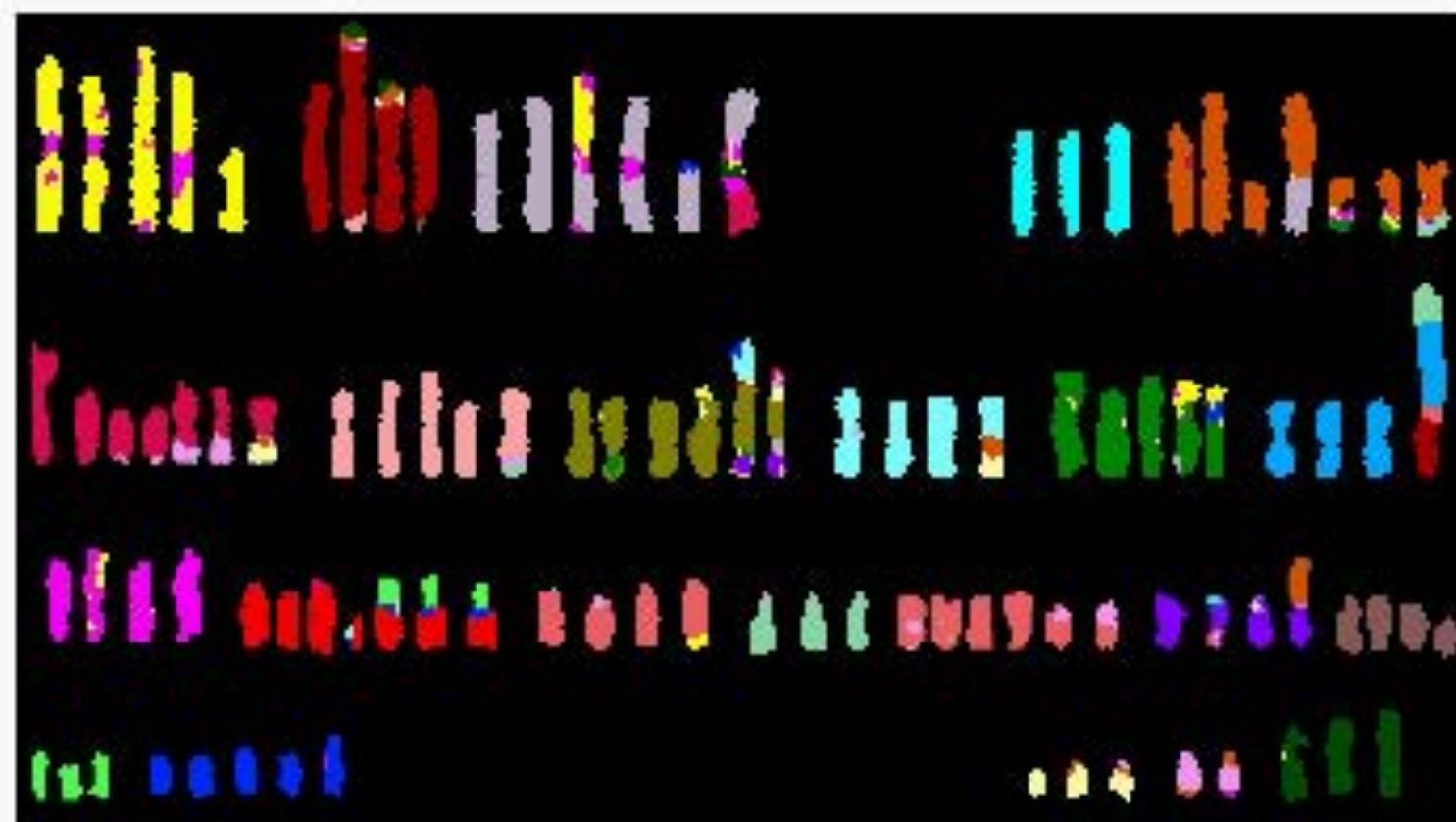
Structural Variants



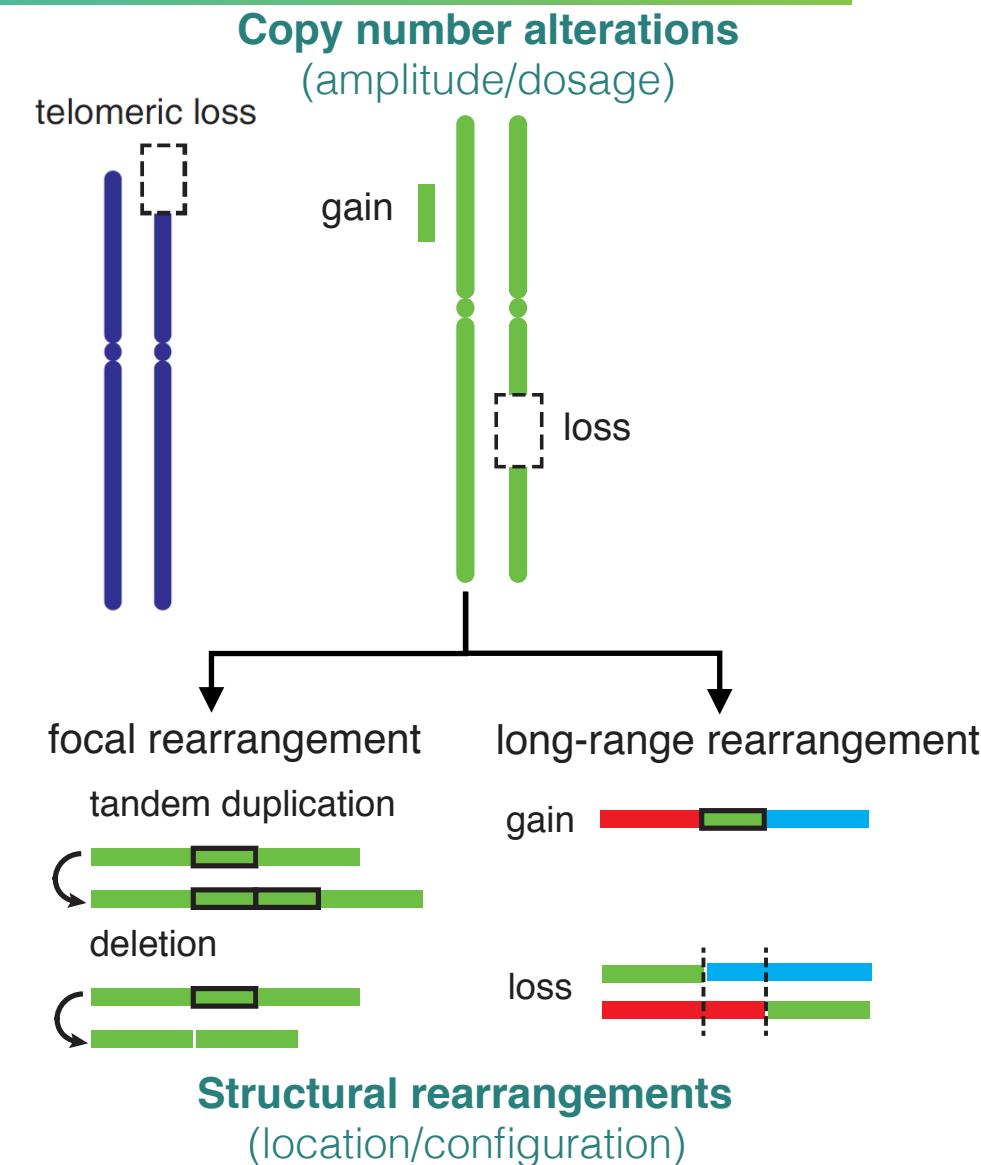
4. Structural Rearrangement Analysis in Cancer Genomes

- Structural variant types predicted from sequencing analysis
- Complex genomic structural rearrangement patterns
- Brief overview of software tools

Abnormal chromosomal rearrangements are prevalent in cancer

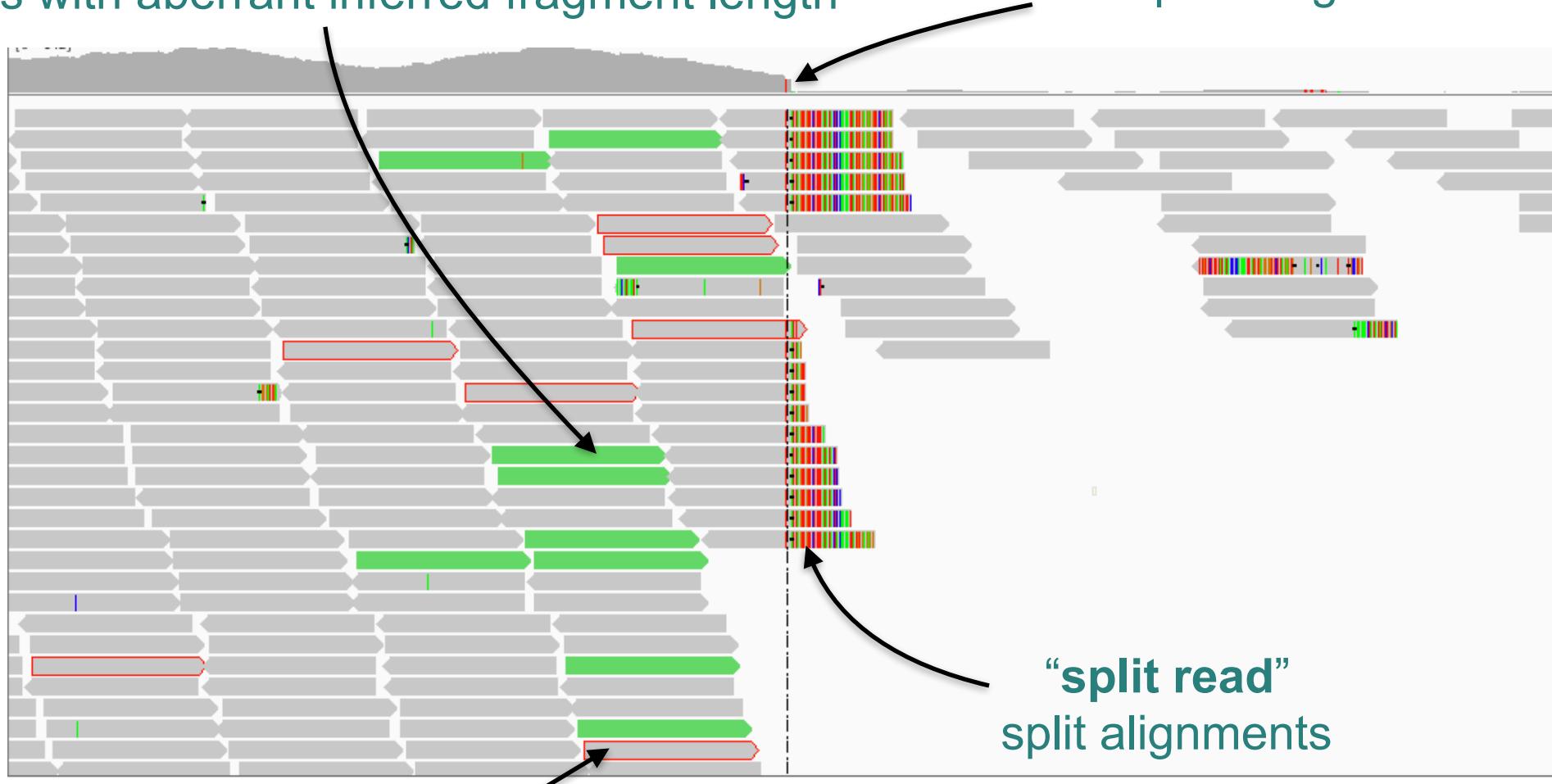


David Huntsman, BC Cancer Agency



Structural Variants: Sequence Features

“discordant read pair”
read pairs with aberrant inferred fragment length



“copy number change”
abrupt change in read coverage

“split read”
split alignments

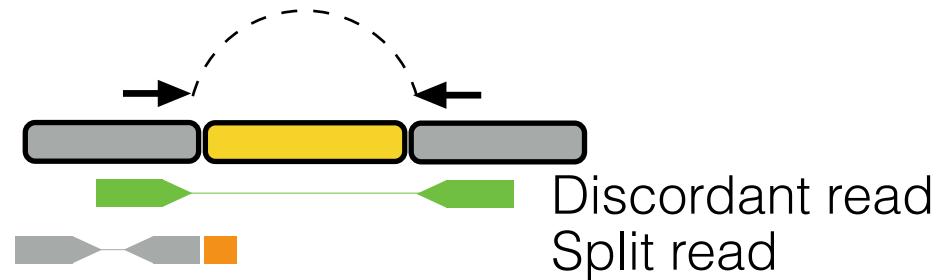
Simple Structural Variants: Deletion & Tandem Duplications

Deletion

Sample

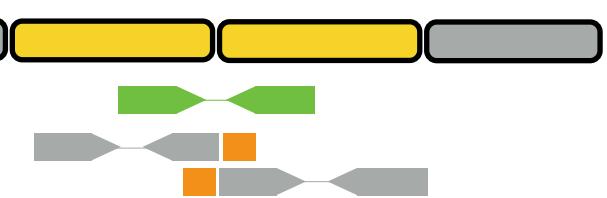


Reference

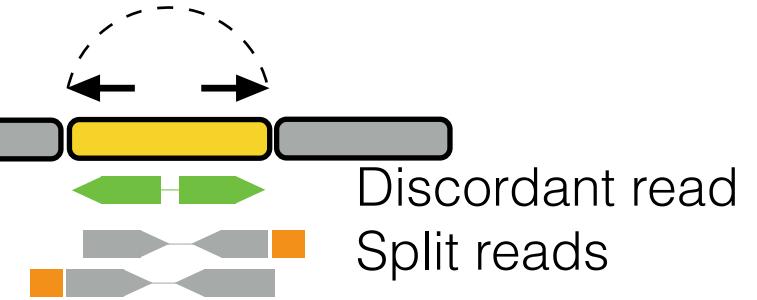


Tandem Duplication

Sample

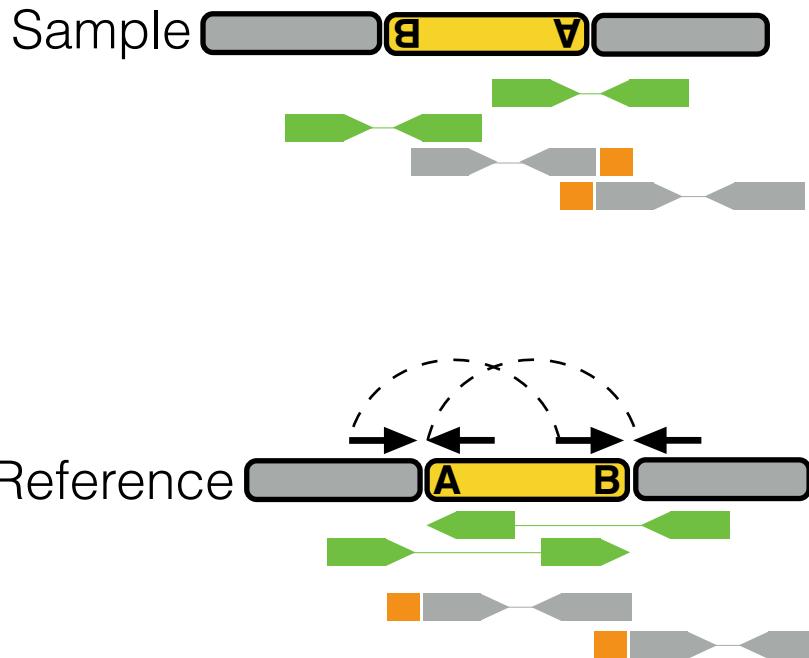


Reference

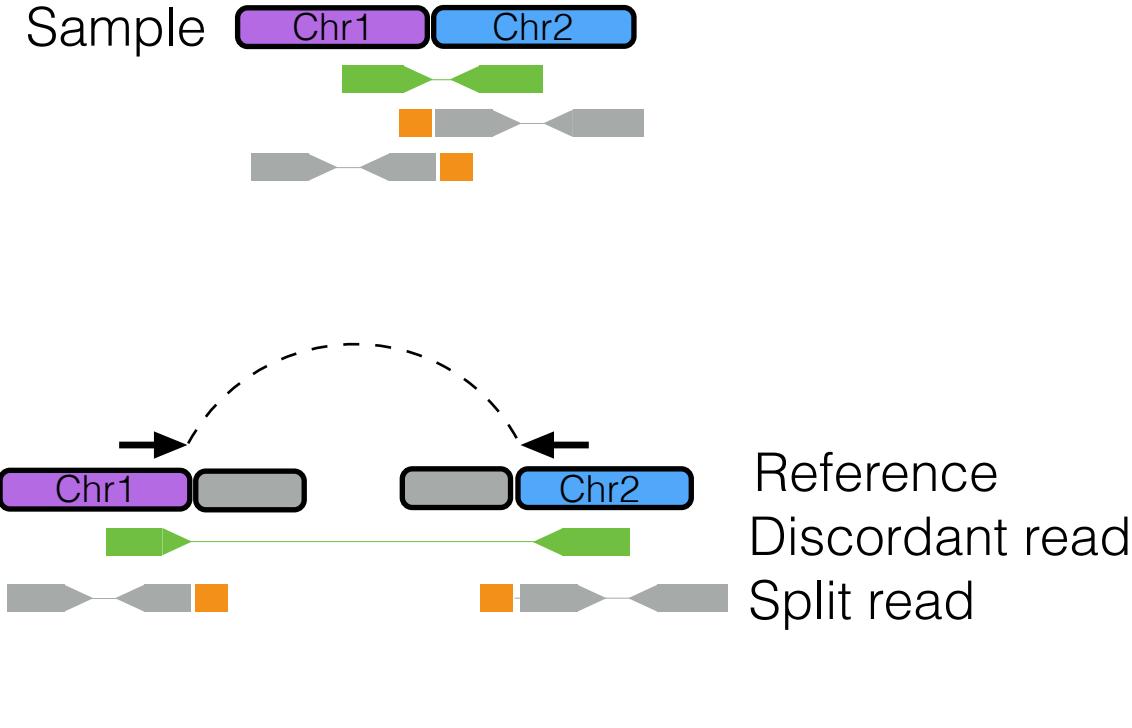


Simple Structural Variants: Inversions & Translocations

Inversion

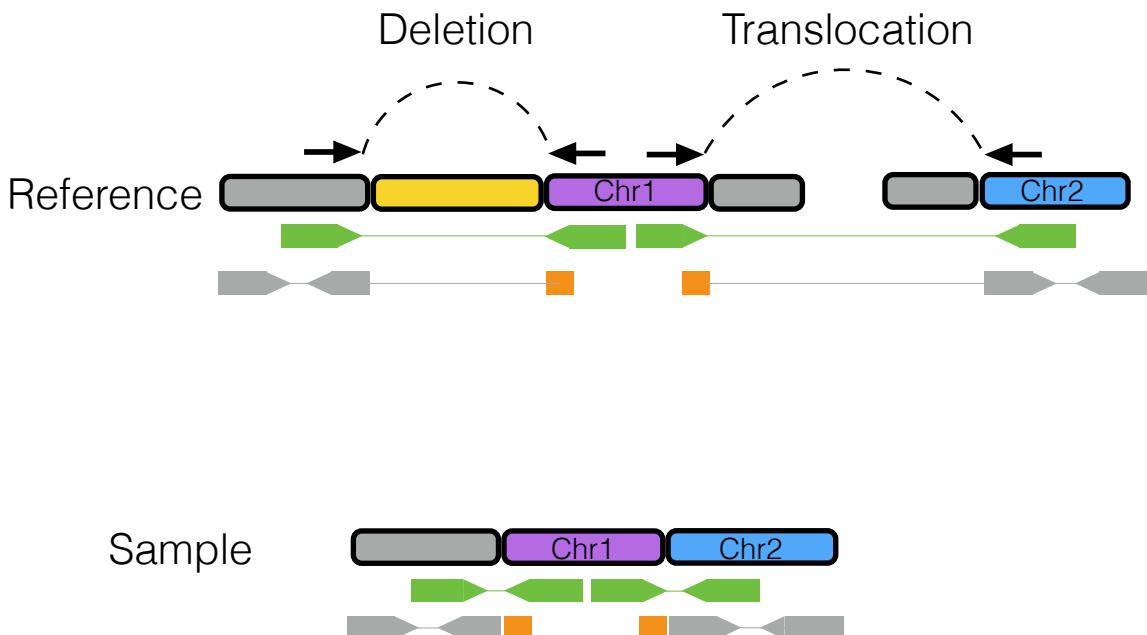


Translocation

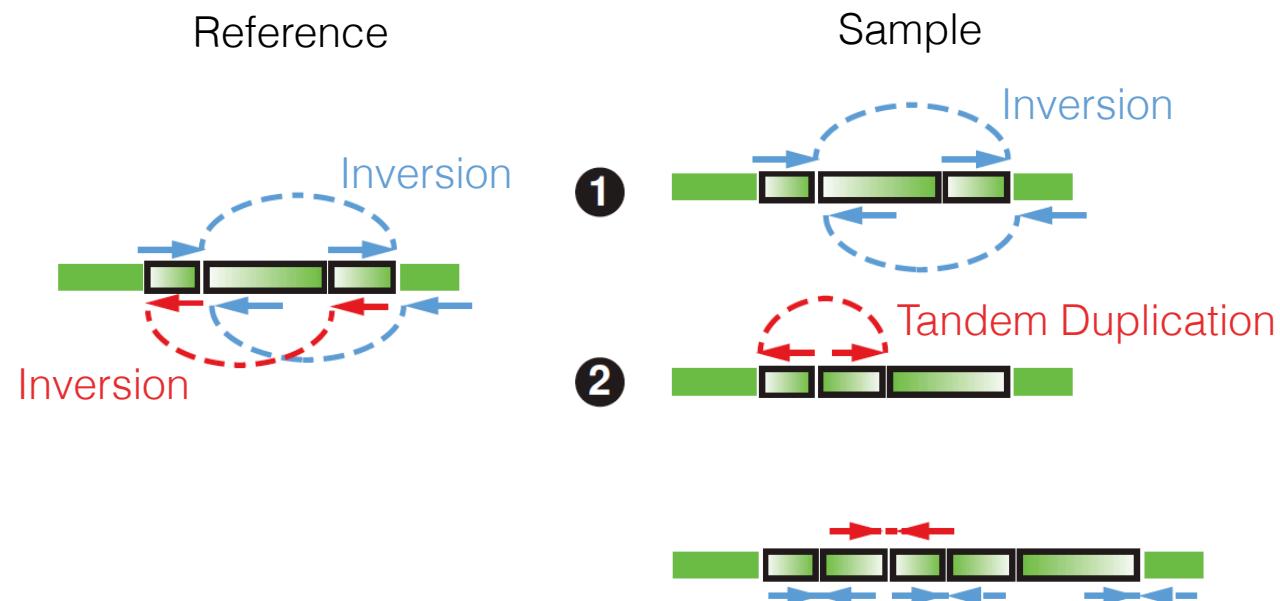


Complex Structural Variants of 2+ more events

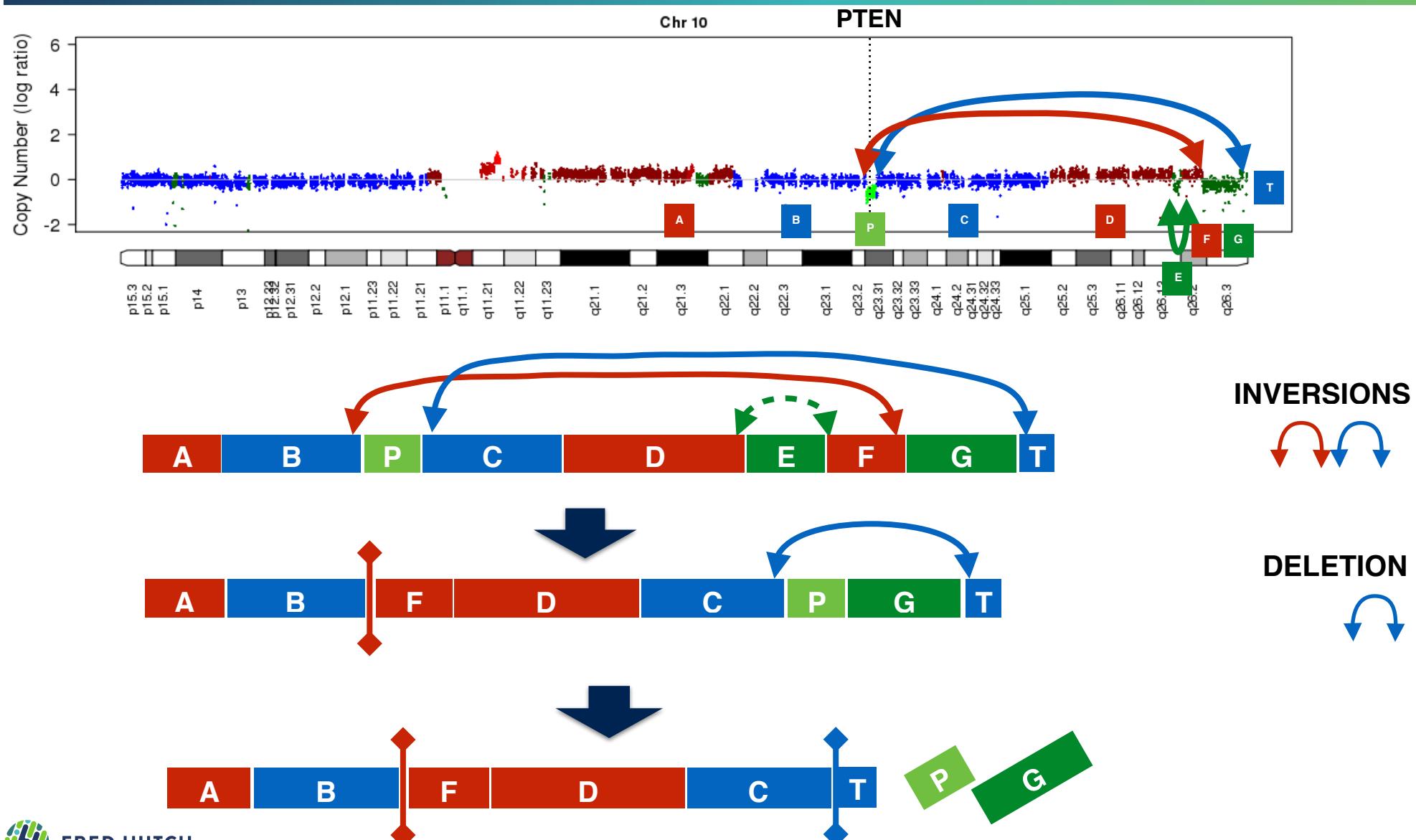
Complex Event (non-overlapping)



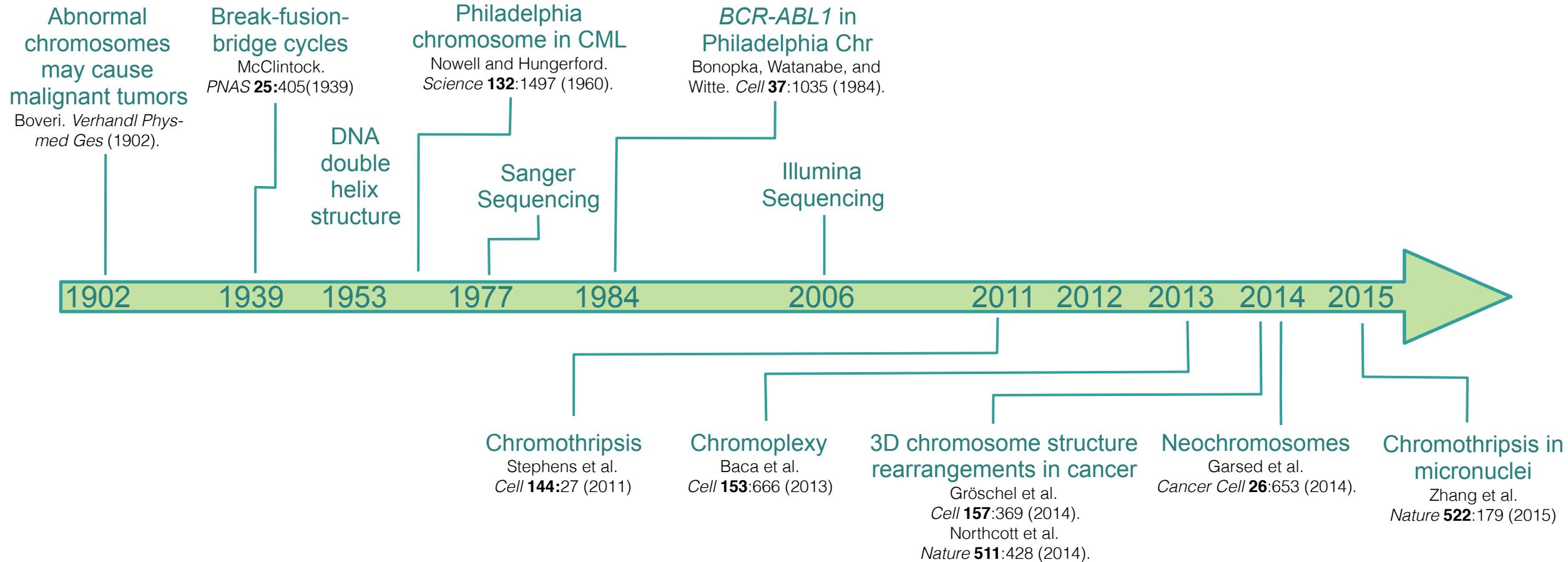
Complex Event (overlapping)



Complex Structural Variant: Example of PTEN deletion

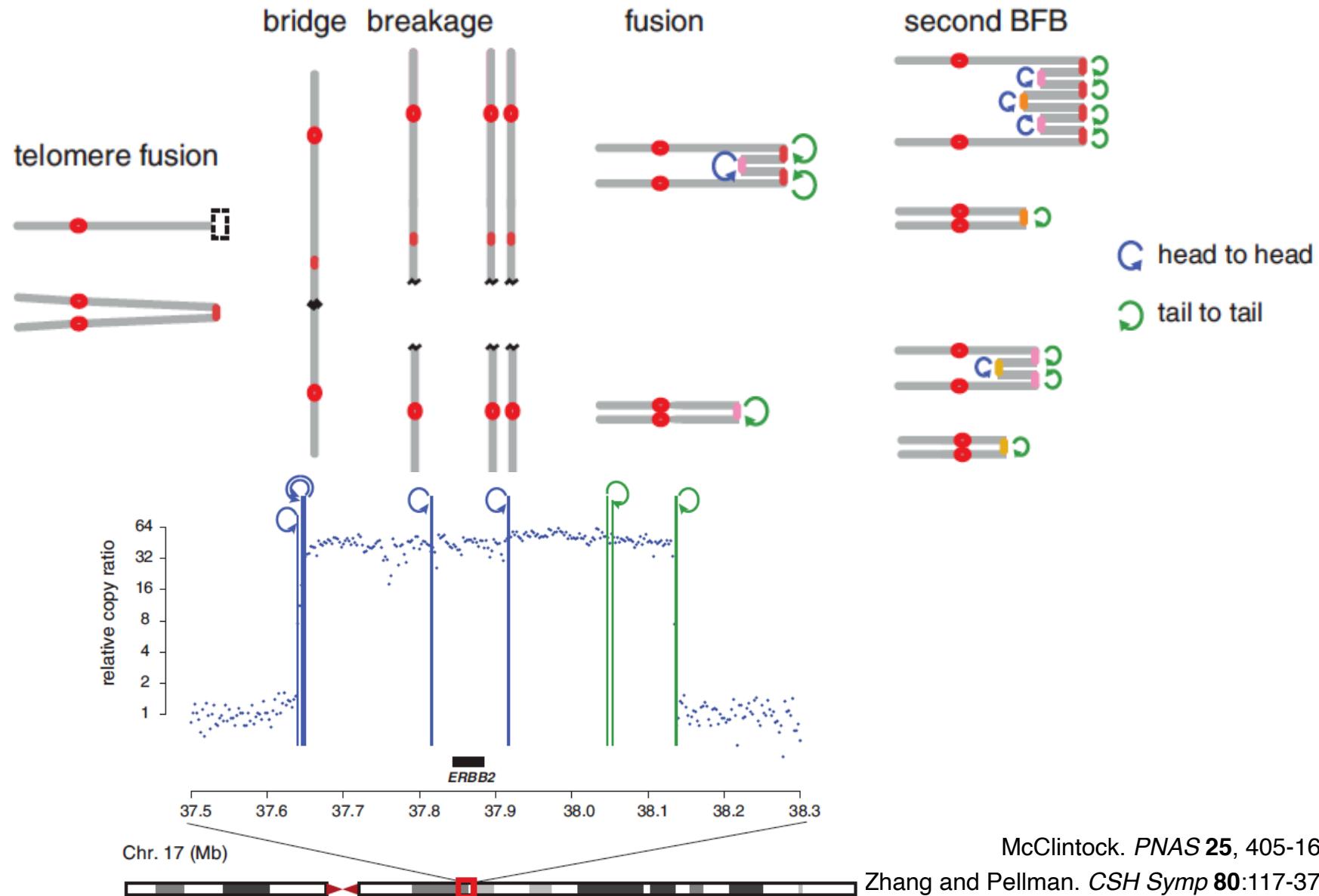


Brief History of Genome Rearrangement Discoveries in Cancer



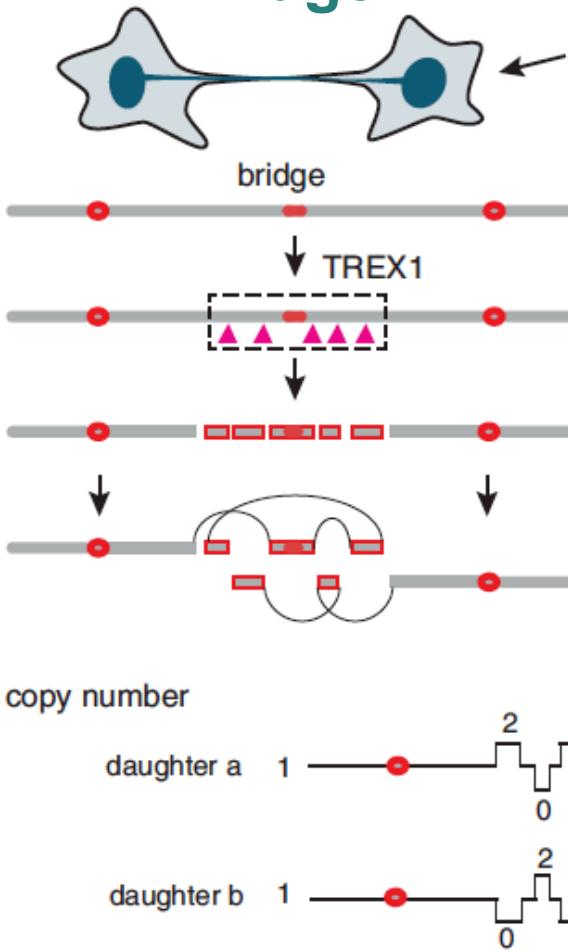
Complex Cancer Genome Rearrangement Patterns

Breakage-Fusion-Bridge (BFB) Cycles

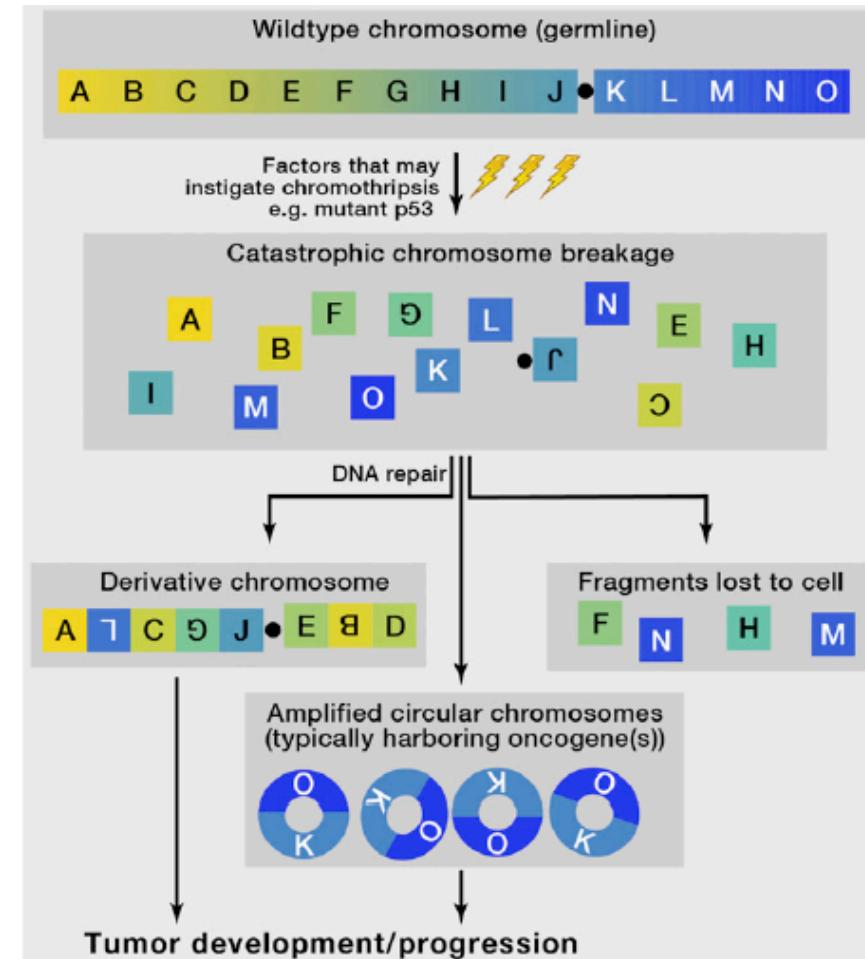
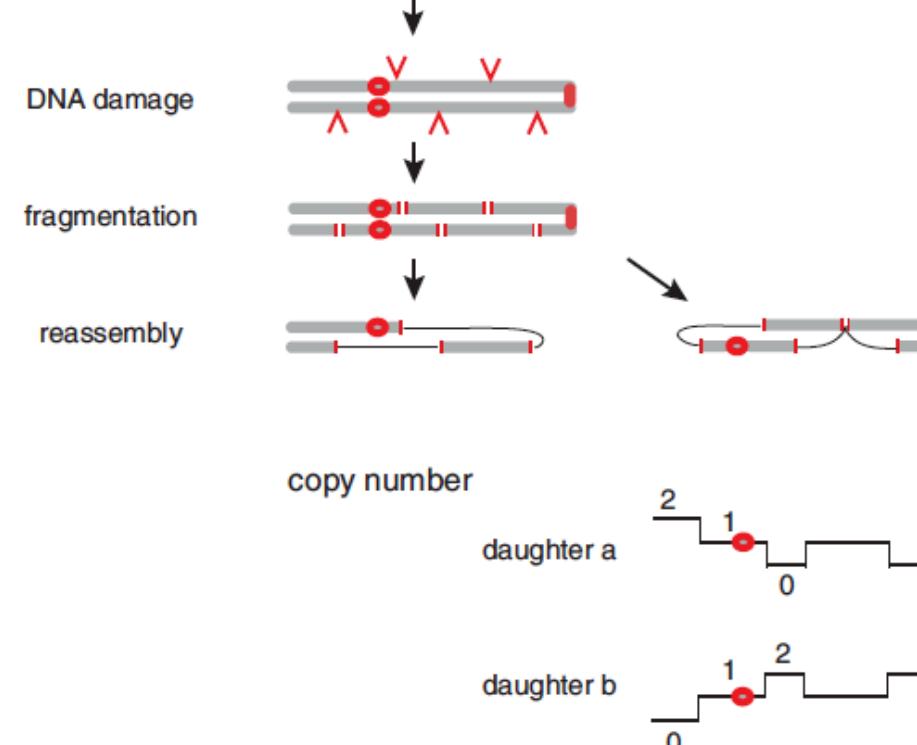


Chromothripsy: Catastrophic DNA shattering

Chromosome Bridge



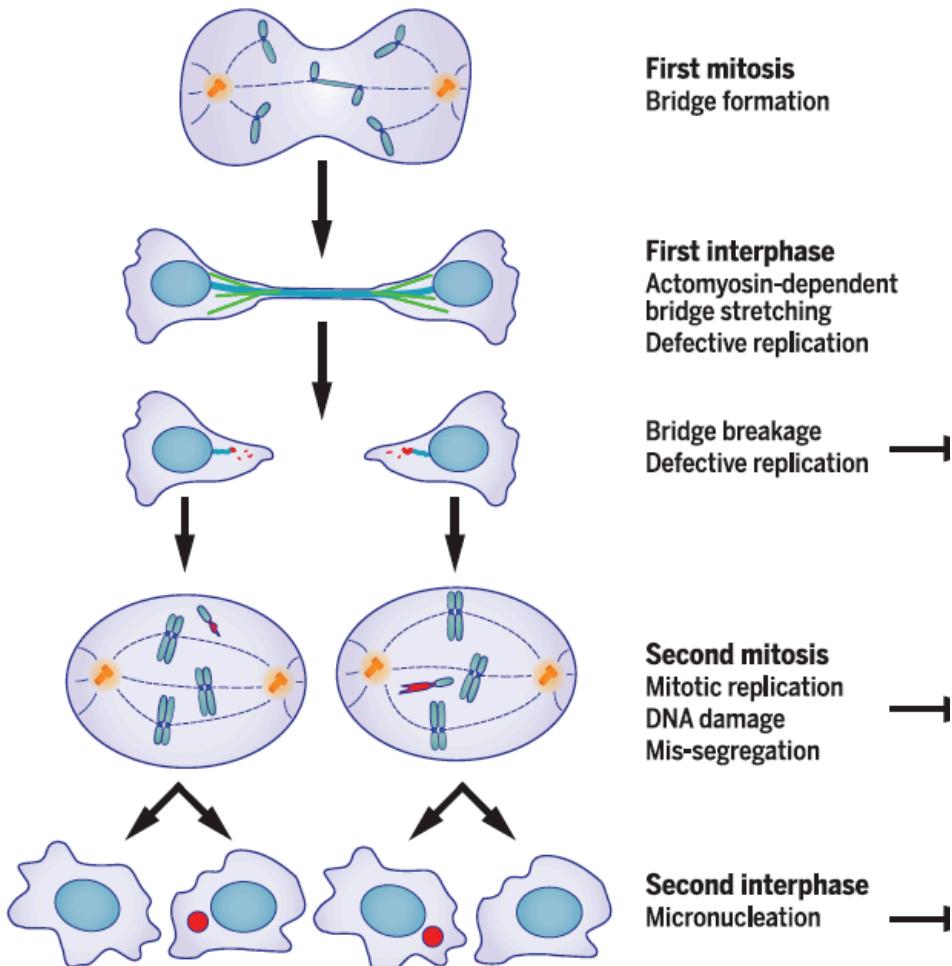
Micronuclei



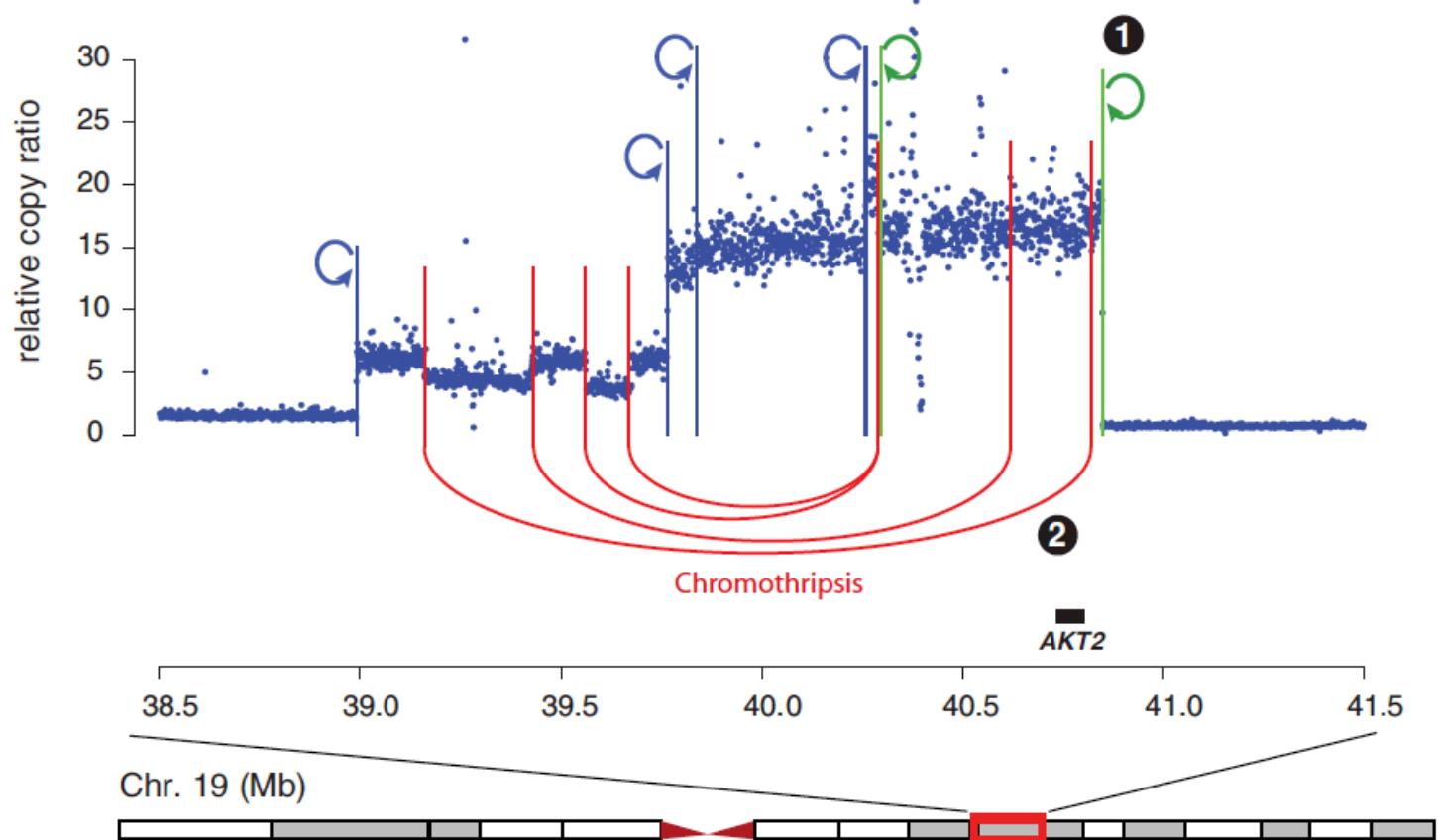
Stephens et al. *Cell* 144:27-40 (2011)

Korbel and Campbell. *Cell* 152:1226-36 (2013)

Concurrent Breakage-Fusion-Bridge & Chromothripsis

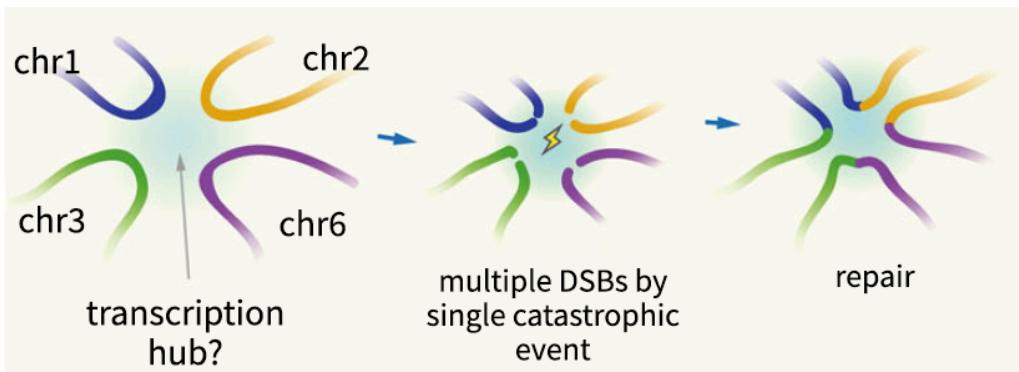


Umbreit et al. *Science* **368**:282 (2020)

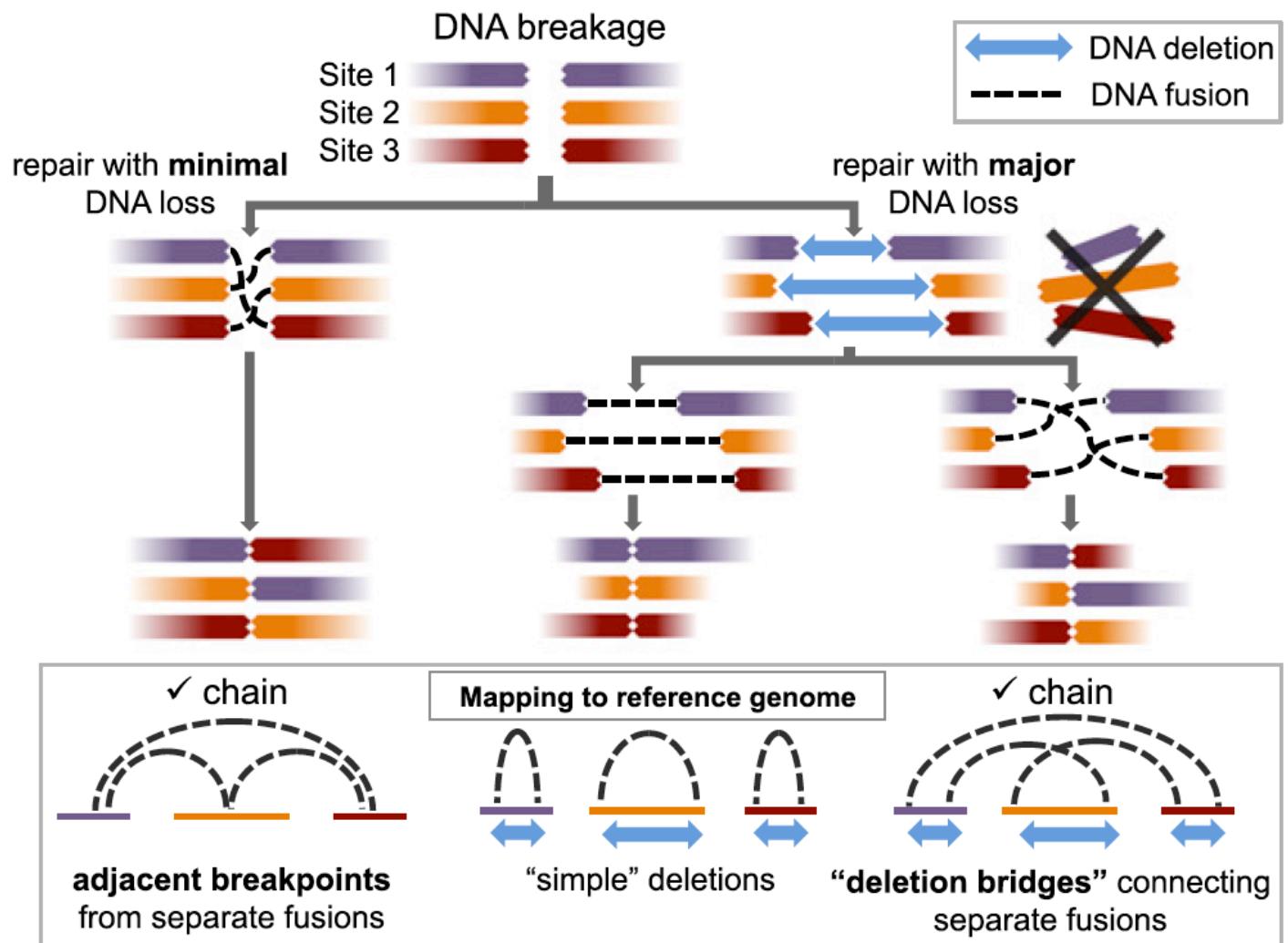


Zhang and Pellman. *CSH Symp* **80**:117-37 (2016)

Chromoplexy: Inter-dependent disruption of DNA within close spatial proximity

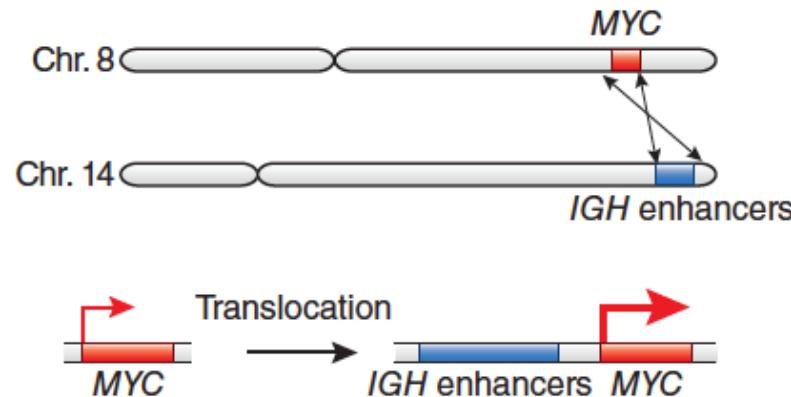


Yi and Ju. *Expt. Mol. Med.* 50:98 (2014).



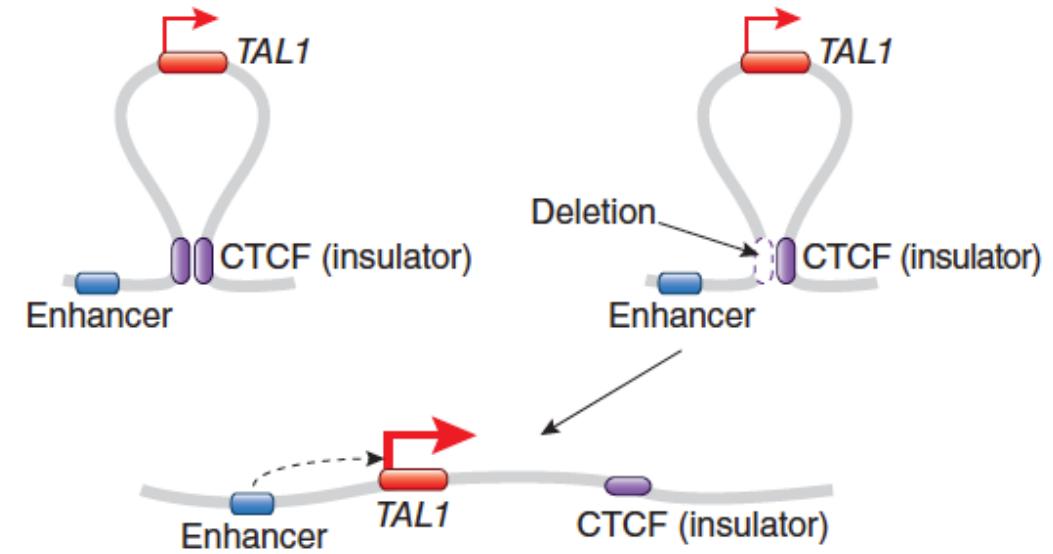
Alterations of oncogene regulation and genome topology

Translocation



Battey et al. *Cell* **34**:779-87 (1983).

Enhancer Hijacking



Beroukhim, Zhang, Meyerson. *Nat Genet* **49**:5-6 (2017).

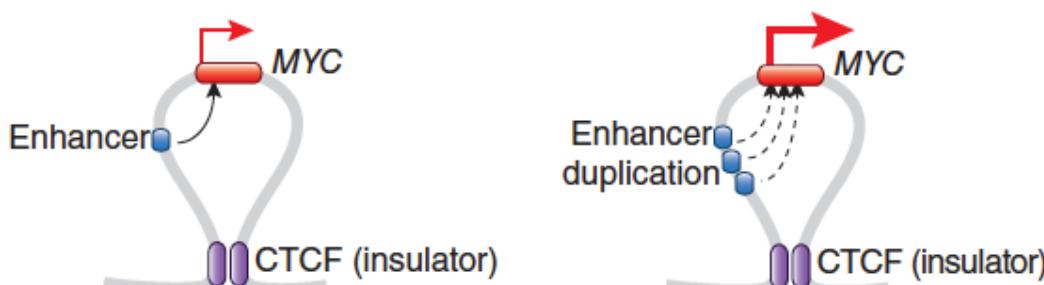
Gröschel et al. *Cell* **157**:369-81 (2014).

Northcott et al. *Nature* **511**:428-34 (2014).

Hnisz et al. *Science* **351**:1454–58 (2016).

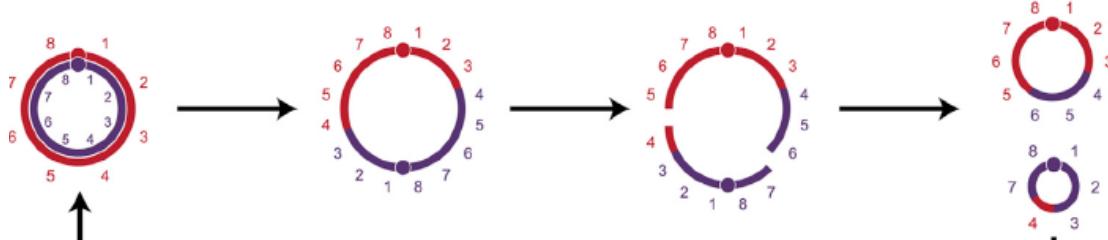
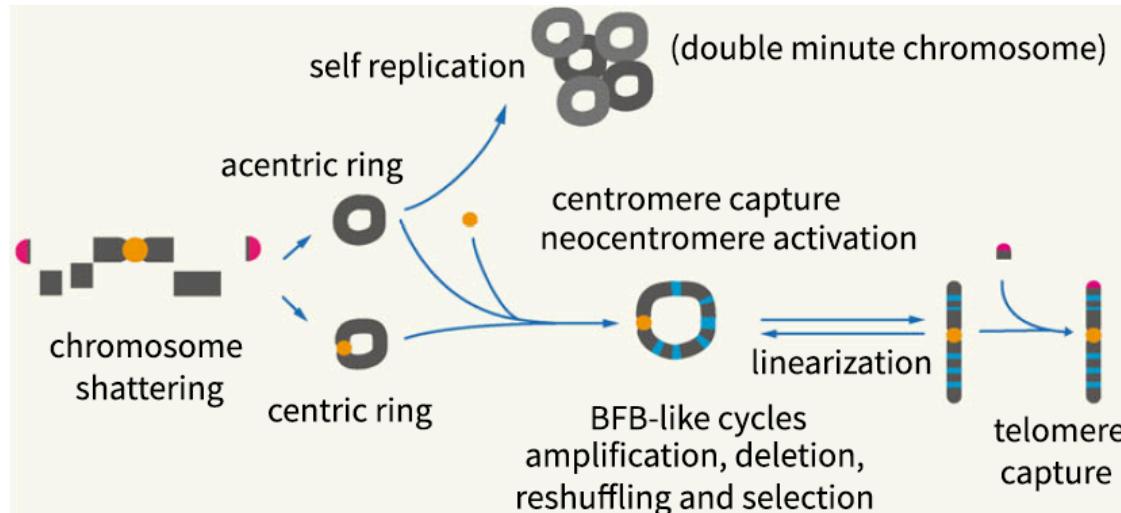
Weischenfeldt et al. *Nat Genet* **49**:65-74 (2017).

Duplication of Enhancer



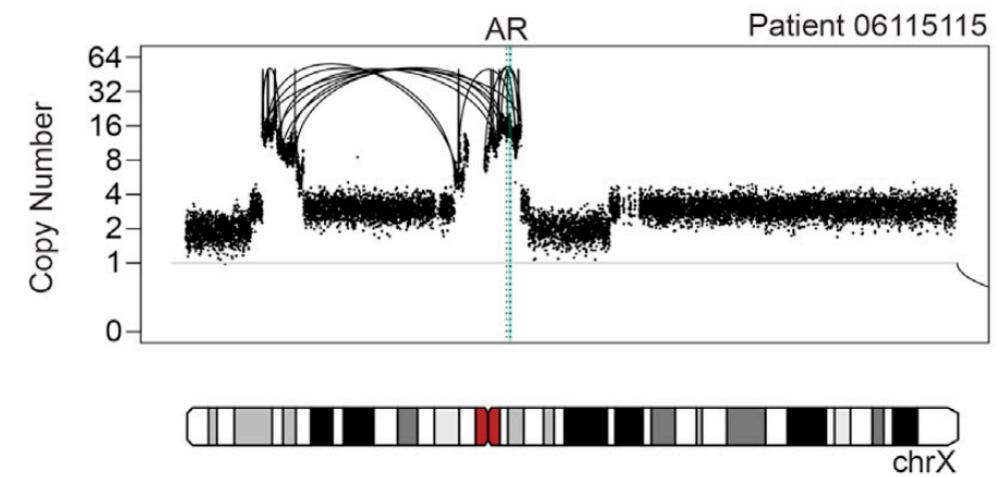
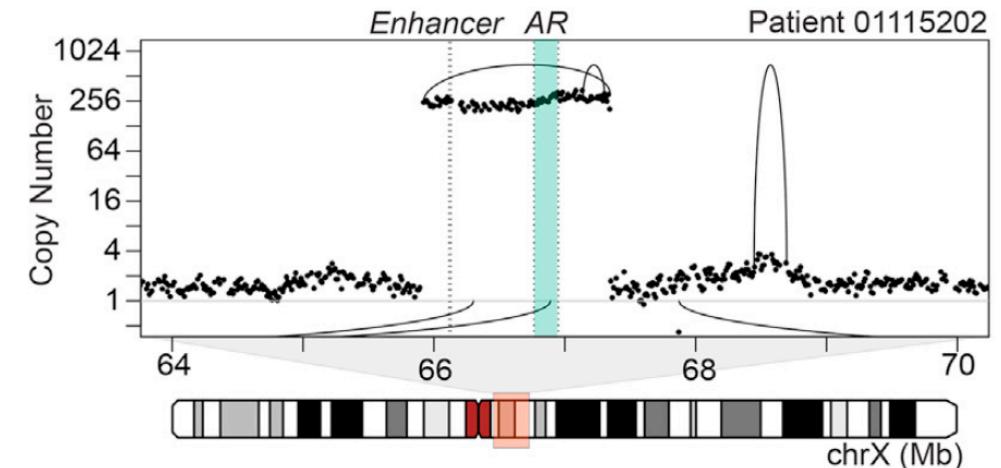
Zhang et al. *Nat Genet* **48**:176-82 (2016).

Extra-Chromosomal DNA: Double Minutes & Neo-chromosomes



Garsed et al. *Cancer Cell* **26**:653-67 (2014).

Double Minute



Neo-Chromosomes

Viswanathan*, Ha*, Hoff*, et al. *Cell* **174**:433-447 (2018)

Structural Variation Tools for Cancer Genome Analysis

Popular SV Methods for Cancer Genomes

SV Breakpoint Methods	Discordant Reads	Split Reads	Assembly	Software	References
DELLY	✓	✓		https://github.com/dellytools/delly	Rausch et al. Genome Biol (2012)
LUMPY	✓	✓		https://github.com/arq5x/lumpy-sv	Layer et al. Genome Biol (2014)
GRIDSS	✓	✓	✓	https://github.com/PapenfussLab/gridss	Cameron et al. Genome Res (2017)
SVABA	✓	✓	✓	https://github.com/walaj/svaba	Wala et al. Genome Res (2018)
BRASS	✓	✓	✓	https://github.com/cancerit/BRASS	Sanger Pipeline

Over 70 tools!

Complex Rearrangements	Methods	References
Chromothripsis	ShatterSeek ShatterProof	Cortés-Ciriano et al. Nat Genet (2020) Govind et al. BMC Bioinf (2014)
Chromoplexy	ChainFinder	Baca et al. Cell (2013)
Extra-chromosomal DNA	AmpliconArchitect	Deshpande et al. Nat Commun (2019)
SV clusters/footprints	ClusterSV GRIDSS	Li et al. Nature (2020) Cameron et al. Genome Res (2017)