Introduction to bioinformatics of cancer: high-throughput sequencing of the genome, epigenome and transcriptome

Valentina Boeva

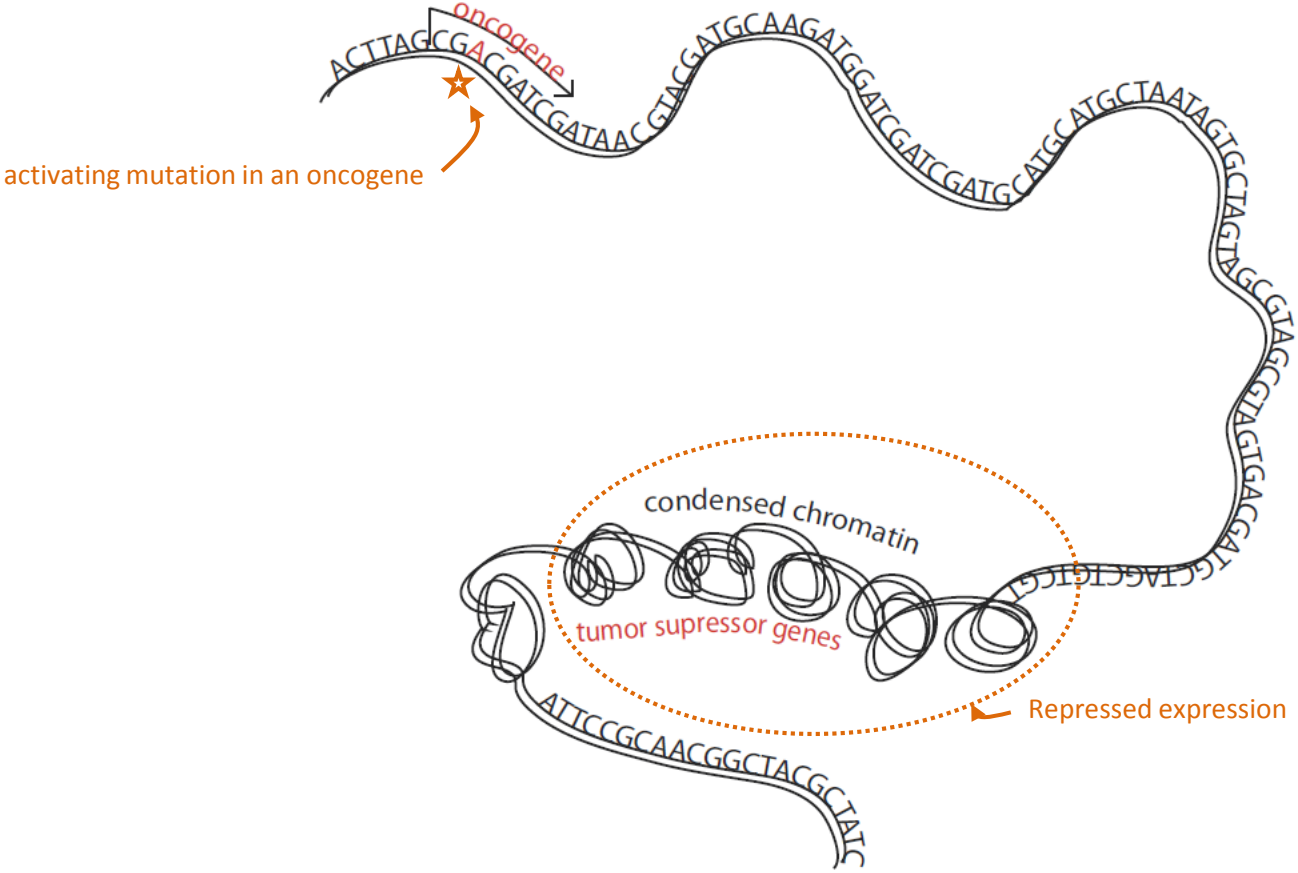
Institut Curie, INSERM U900, Mines ParisTech

institutCurie

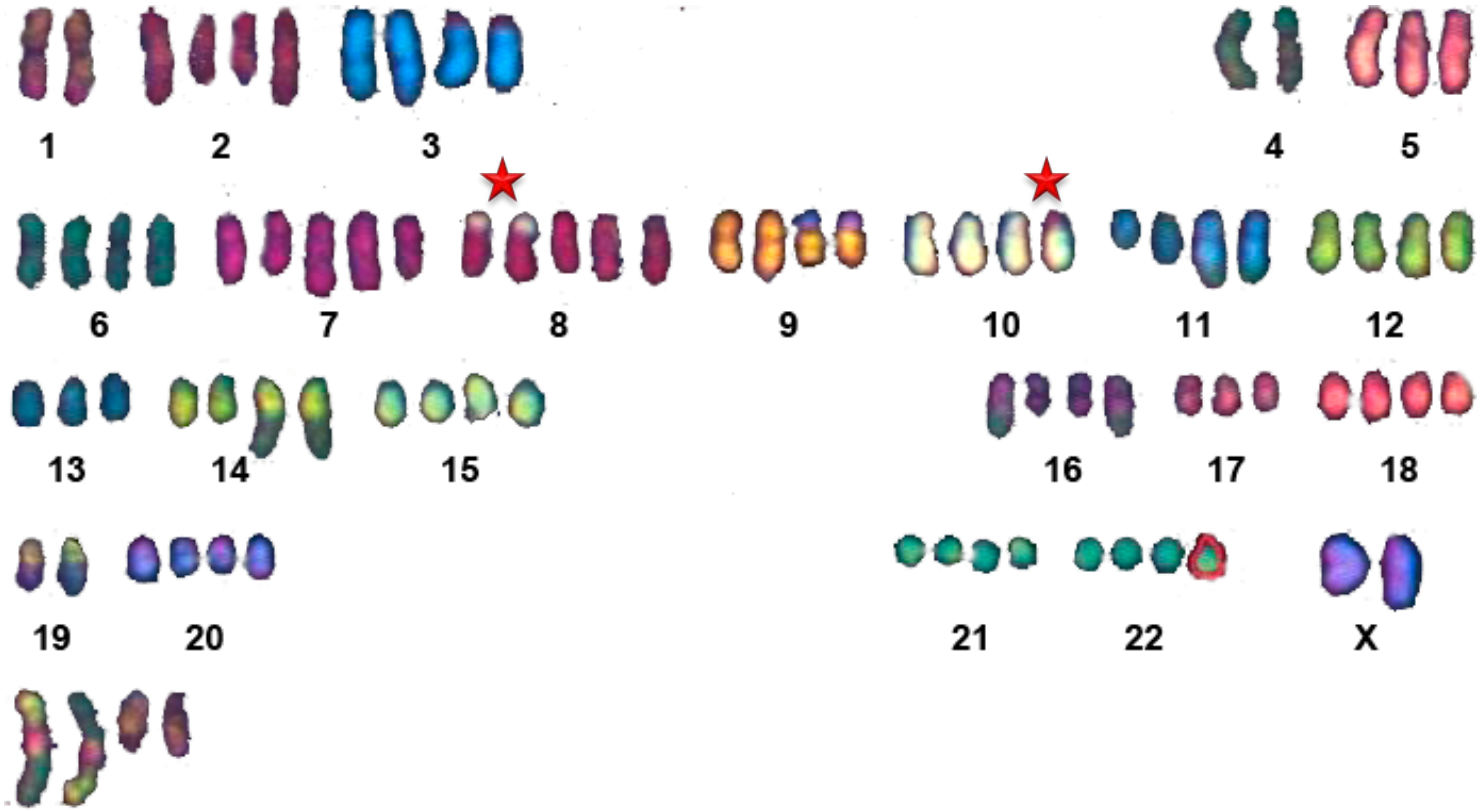
Inserm
Institut national de la santé et de la recherche médicale

MINES PARIS

Cancer cells are characterized by genomic mutations, altered epigenomics and, as a consequence, abnormal gene expression



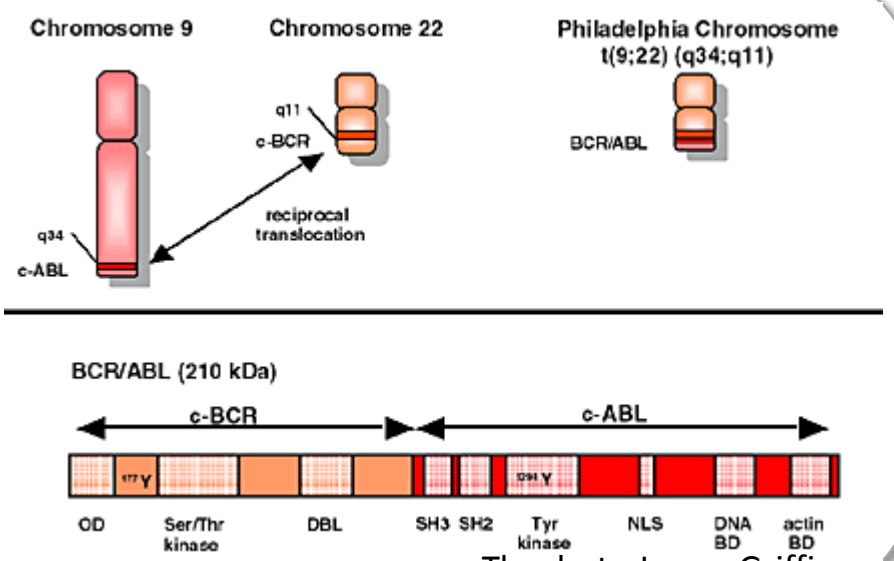
Cancer genomes are often significantly rearranged



A 24 color karyotype of a neuroblastoma cell line

Chromosomal translocations can result in cancer-causing fusion proteins

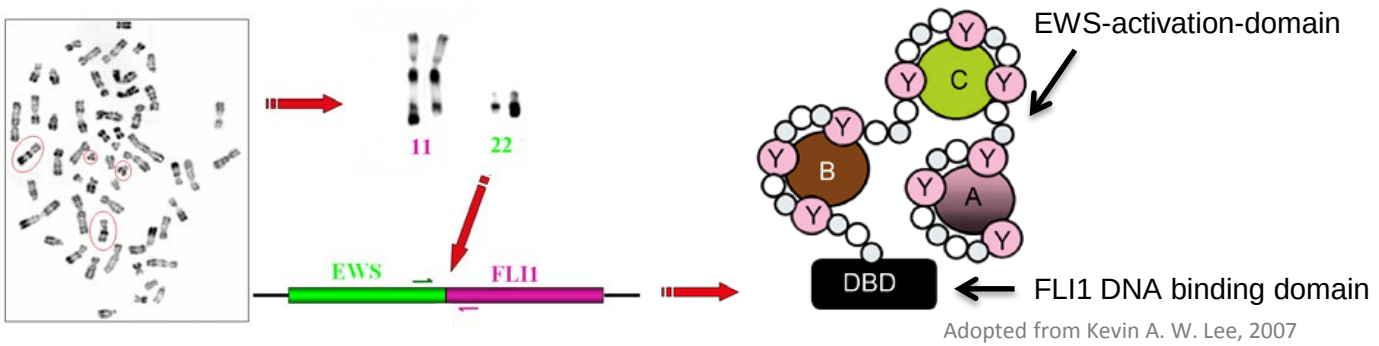
The Philadelphia chromosome (fusion of chr9 and chr22) encodes a new oncogenic protein called BCR/ABL.



Thanks to James Griffin

- BCR/ABL fusion in CML (chronic myelogeneous leukemia)
- BCL1/IGH in multiple myeloma
- EWS/FLI1 in Ewing sarcoma
- etc.

Chromosomal translocations can result in cancer-causing fusion proteins



Adopted from Kevin A. W. Lee, 2007

Fusion of chr11 and chr22 encoding for EWS/FLI1 causes Ewing sarcoma.

- BCR/ABL fusion in CML (chronic myelogeneous leukemia)
- BCL1/IGH in multiple myeloma
- EWS/FLI1 in Ewing sarcoma
- etc.

Translocation is not the only type of structural variant, scientists are interested in.

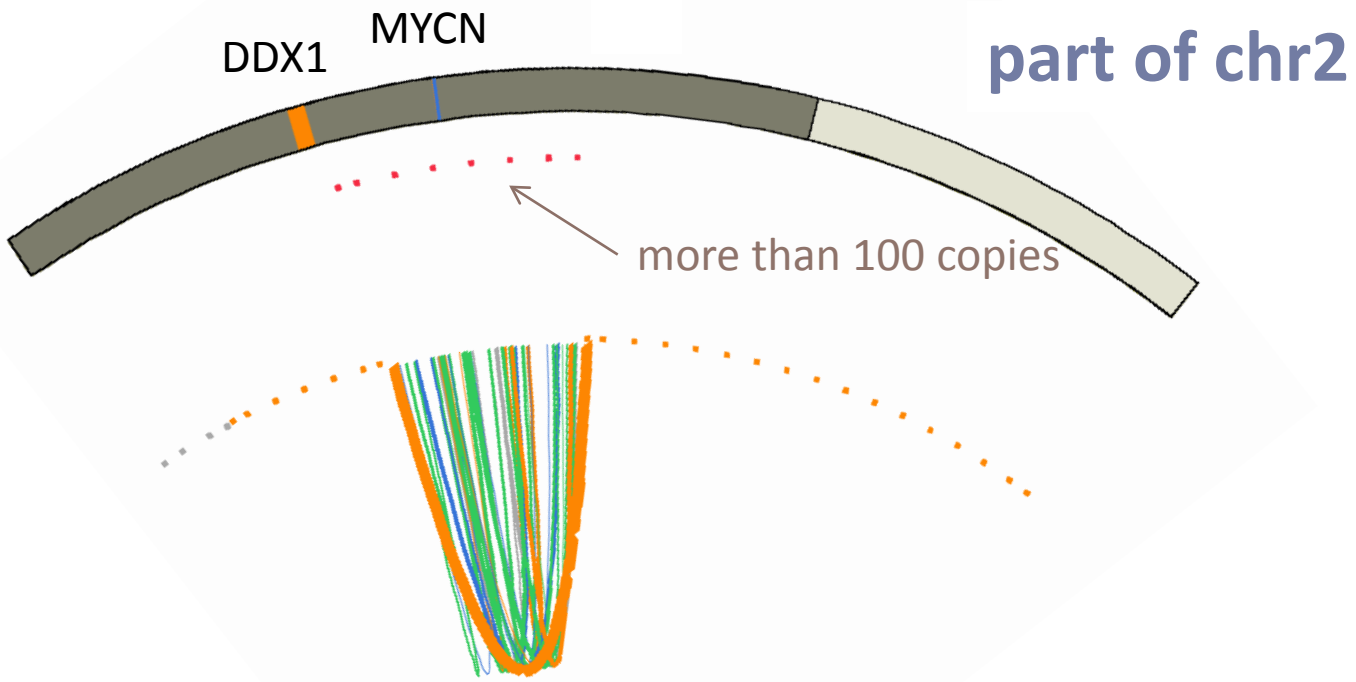
Structural variants:

- Large-scale genomic deletions
- Tandem duplications
- Amplicons
- Insertions
- Inversions

All this can disrupt a tumor suppressor gene or create a functional mutation in a proto-oncogene.

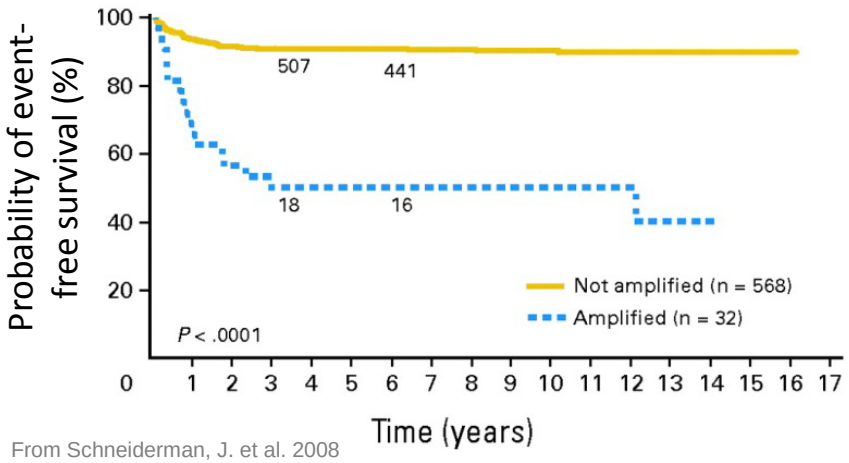
Amplification of an important gene can favor cancer development

- *MYCN* amplification, which occurs in approximately 22% of primary neuroblastomas, is one of the most powerful prognostic factors identified to date. It is significantly associated with advanced-stage disease, rapid tumor progression, and poor prognosis.



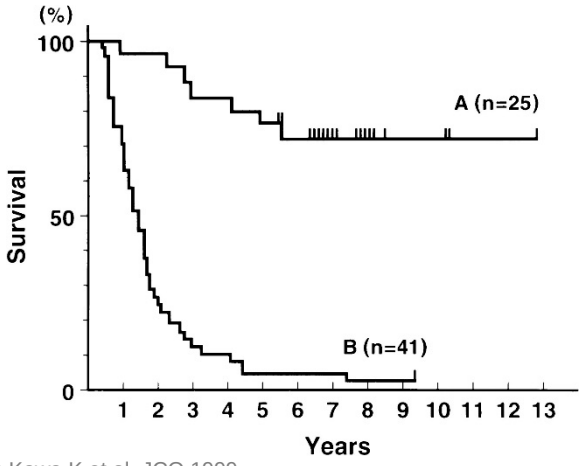
Amplification of an important gene can favor cancer development

- *MYCN* amplification, which occurs in approximately 22% of primary neuroblastomas, is one of the most powerful prognostic factors identified to date. It is significantly associated with advanced-stage disease, rapid tumor progression, and poor prognosis.



From Schneiderman, J. et al. 2008

Kaplan-Meier survival curves for 600 stage A, B, and Ds patients by MYCN status. Event-free survival.

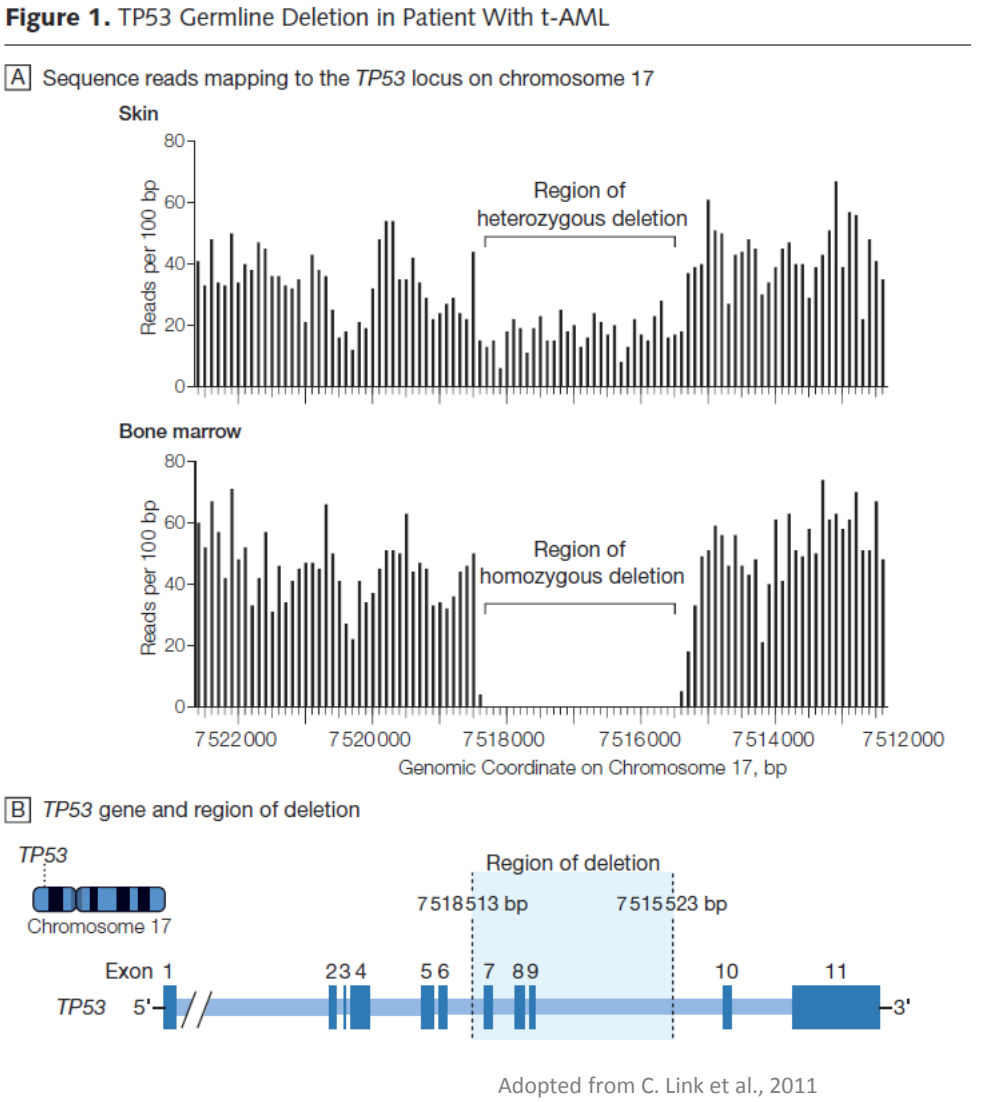


From Kawa K et al. JCO 1999

Overall survival curve for MYCN-amplified neuroblastoma patients relative to treatment after induction chemotherapy.
A, patients who underwent autologous bone marrow transplantation (ABMT)/peripheral-blood stem-cell transplantation (PBSCT) ;
B, patients who did not undergo ABMT/PBSCT.

Deletion in an important gene can favor cancer development

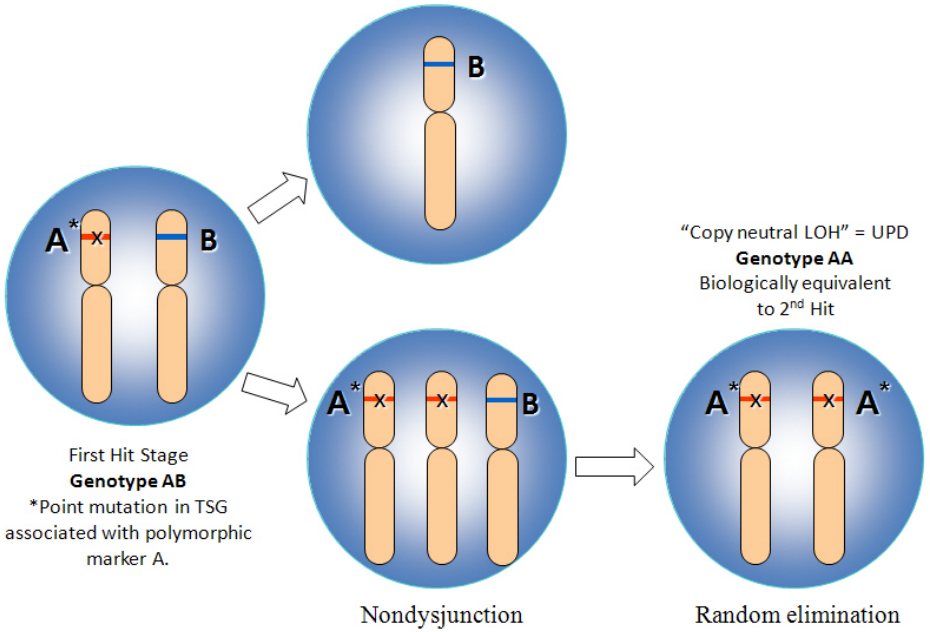
- Patient was treated again breast and ovarian cancer
- She developed therapy-related acute myeloid leukemia (t-AML)
- Whole-genome sequencing revealed a novel, heterozygous 3-kilobase deletion removing exons 7-9 of *TP53* in the patient's normal skin DNA, which was homozygous in the leukemia DNA as a result of acquired uniparental disomy .



Adopted from C. Link et al., 2011

Copy neutral loss of heterozygosity (LOH) or acquired uniparental disomy (UPD) often happens in cancer

Acquired Uniparental Disomy (UPD)



In UPD, a person receives two copies of a chromosome, or part of a chromosome, from one parent and no copies from the other parent.
This acquired homozygosity could lead to development of cancer if the individual inherited a non-functional allele of a tumor suppressor gene.

From Wikipedia

Identification of regions of gain and loss helps to predict the aggressiveness of cancer

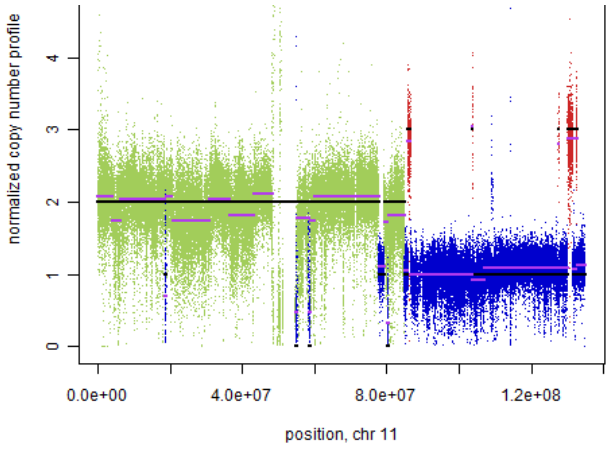
PNAS

High-risk neuroblastoma tumors with 11q-deletion display a poor prognostic, chromosome instability phenotype with later onset

Helena Carén[a], Hanna Kryh[a], Maria Nethander[b], Rose-Marie Sjöberg[a], Catarina Träger[c], Staffan Nilsson[d], Jonas Abrahamsson[e], Per Kogner[c], and Tommy Martinsson[a,1]

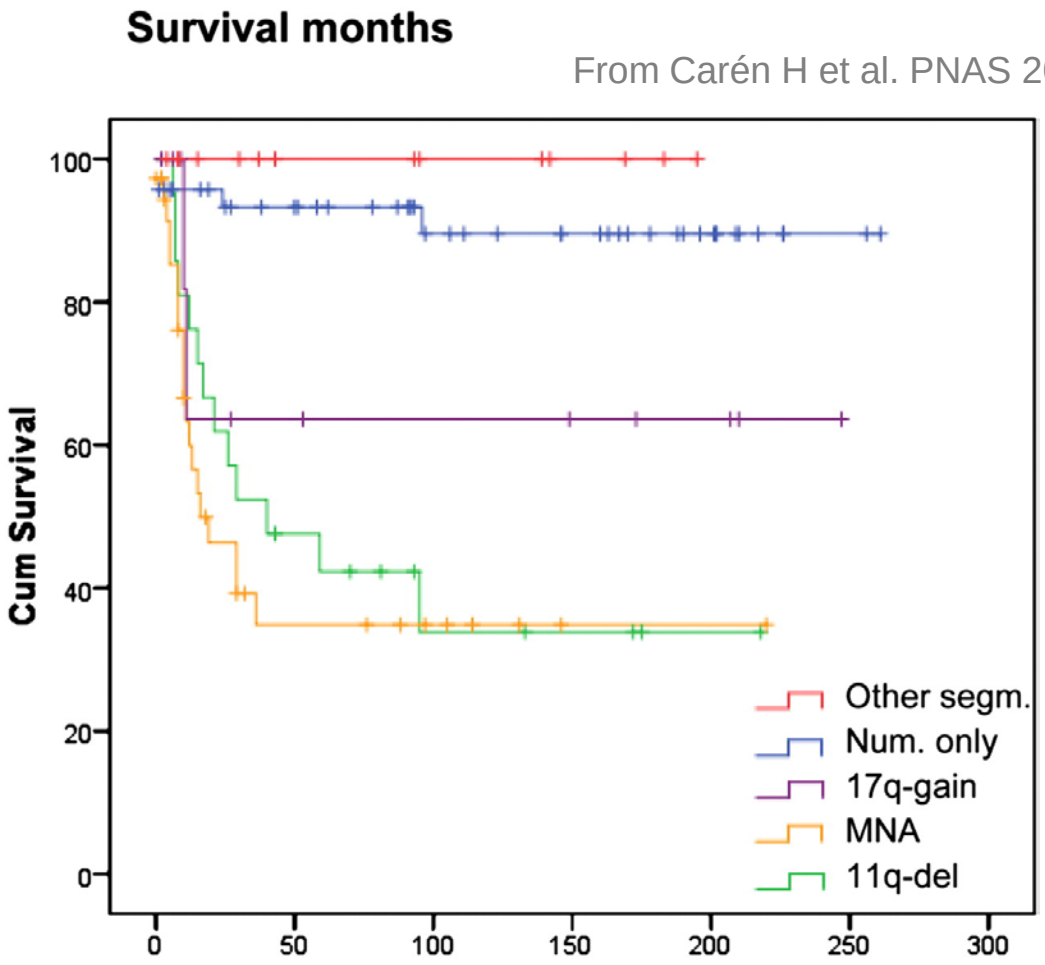
[a]Department of Clinical Genetics, Institute of Biomedicine, University of Gothenburg, Sahlgrenska University Hospital, SE-413 45 Göteborg, Sweden; [b]Genomics Core Facility, University of Gothenburg, SE-405 30 Göteborg, Sweden; [c]Childhood Cancer Research Unit, Department of Women's and Children's Health, Karolinska Institutet, Karolinska Hospital, SE-171 76 Stockholm, Sweden; [d]Department of Mathematical Statistics, Chalmers University of Technology, SE-412 96 Göteborg, Sweden; and [e]Department of Pediatrics, University of Gothenburg, Queen Silvia Children's Hospital, SE-416 85 Göteborg, Sweden

Copy number profile (chr 11) of a metastatic neuroblastoma sample:



Identification of regions of gain and loss helps to predict the aggressiveness of cancer

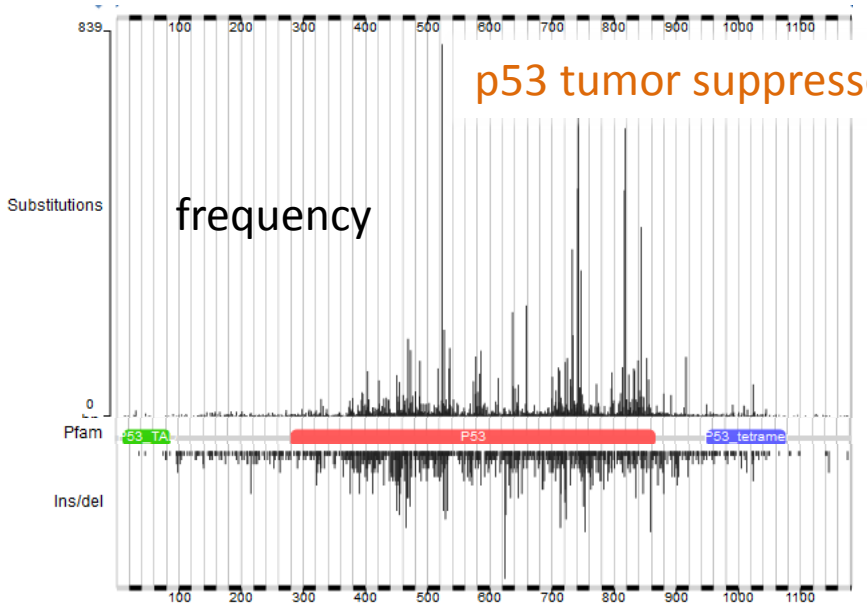
From Carén H et al. PNAS 2010;107:4323-4328

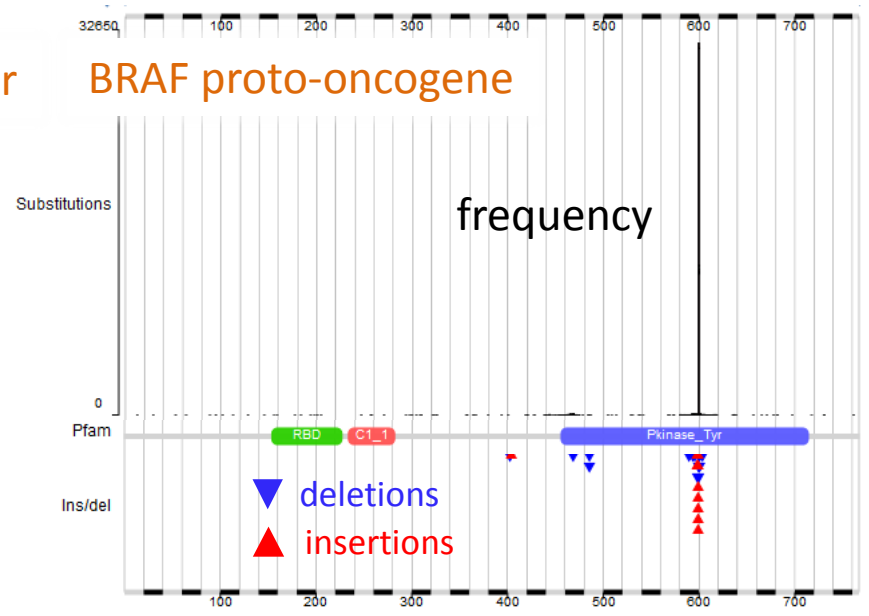


Kaplan-Meier overall survival for patients with tumors with different genomic profiles.

Point mutations can disrupt the function of a tumor suppressor or activate an oncogene

- Single nucleotide variants (SNVs) and short indels (<20bp) can change protein function





- Common activating mutations in proto-oncogenes:

  NMYC, *CCND1, CCNE1, BCR-ABL, BRAF, ALK*

- Common disrupting mutations in tumor suppressor genes:

  TP53 (p53), *BRCA1*, *BRCA2*, *APC*, and *RB1*

COSMIC database provides the list of SNVs and indels in most cancer types

COSMIC
Catalogue of somatic mutations in cancer

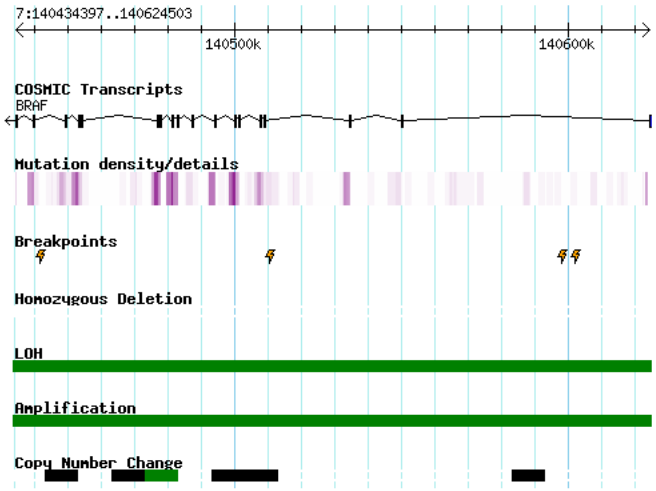
http://cancer.sanger.ac.uk/cosmic/

Search Enter search here...

Home About Download Publications News Contact Help FAQ

Cosmic » Gene » Overview » BRAF

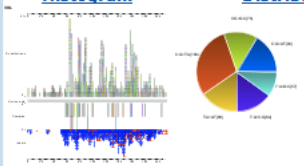
Overview | Genome View | Sequence | Fusions | Studies | References

Gene name	BRAF
Gene Id	COSG2
Synonyms	B-raf 1,B-raf1,BRAF1,MGC126806,MGC138284,RAFB1,CCDS5863.1,P15056,ENSG00000157764
Genomic Summary	

Drug Sensitivity Data:	Click here for BRAF drug sensitivity data Mutations in BRAF are associated with altered sensitivity to the following drug(s) : Inceased sensitivity: PF-562271, CEP-701, FTI-277, 17-AAG, PD-0325901, SB590885, AZD6244, PD-173074, ZM-447439, BIBW2992, Temsirolimus, Metformin, AZD6482, Gefitinib, PLX4720, Nutlin-3a, AZ628, Bortezomib, Embelin, RDEA119, FH535, CI-1040, CHIR-99021, AP-24534, Obatoclax Mesylate
No.of Samples	Total number of unique samples: 168311 Unique samples with mutations: 33263

Mutation Analysis

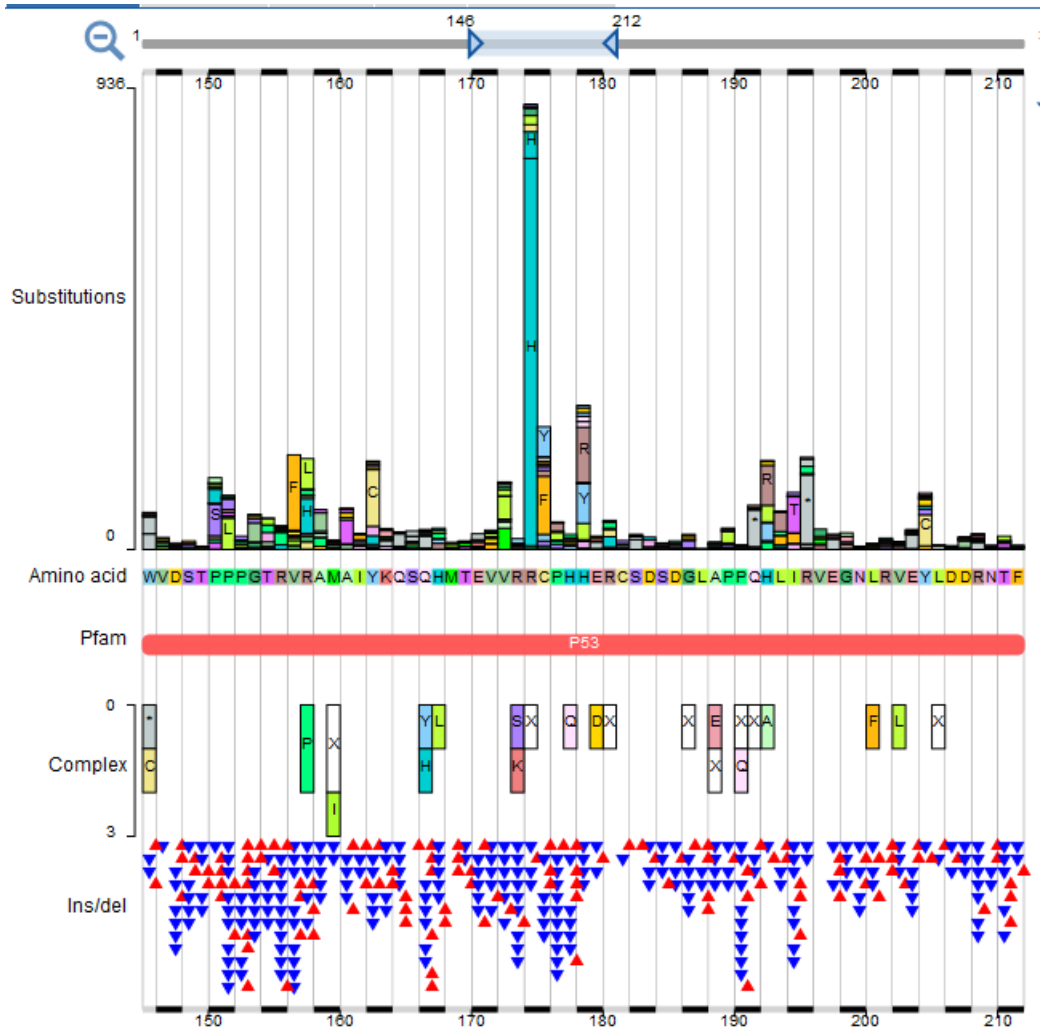
Histogram Distribution



External Links

OMIM:	164757
Transcript:	NM_004333
Ensembl Contig View:	BRAF
UCSC Browser:	BRAF
Copy Number:	CONAN
CCDS:	CCDS5863.1
Pfam:	P15056
Atlas Genetic Oncology:	BRAFID828
HGNC:	1097

Small insertions and deletions (indels) have similar consequences as single nucleotide variants (SNVs)



▼ deletions

▲ insertions

ZOOM on mutations in the P53 domain of the TP53 tumor suppressor gene

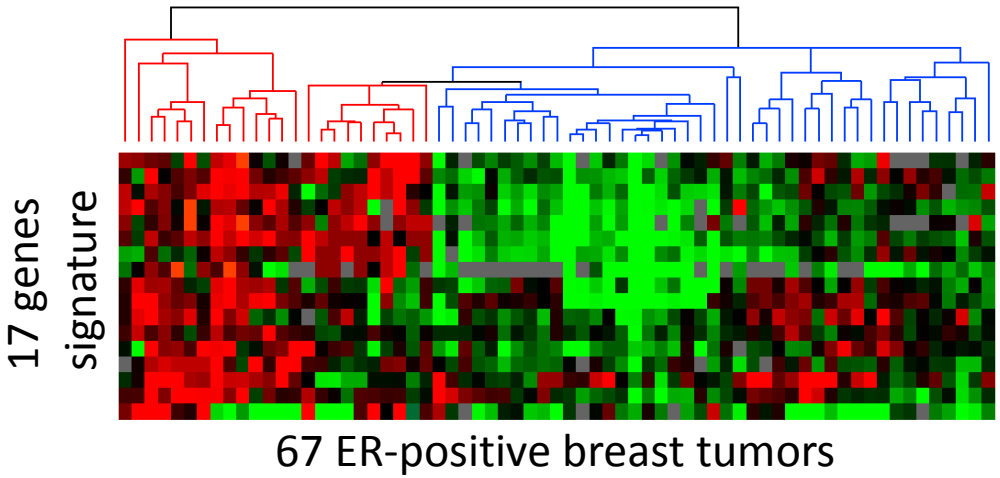
To study genetic profile of a tumor, we need to detect SVs, CNAs, LOH and SNVs.

To study a tumor genome, we need to detect:

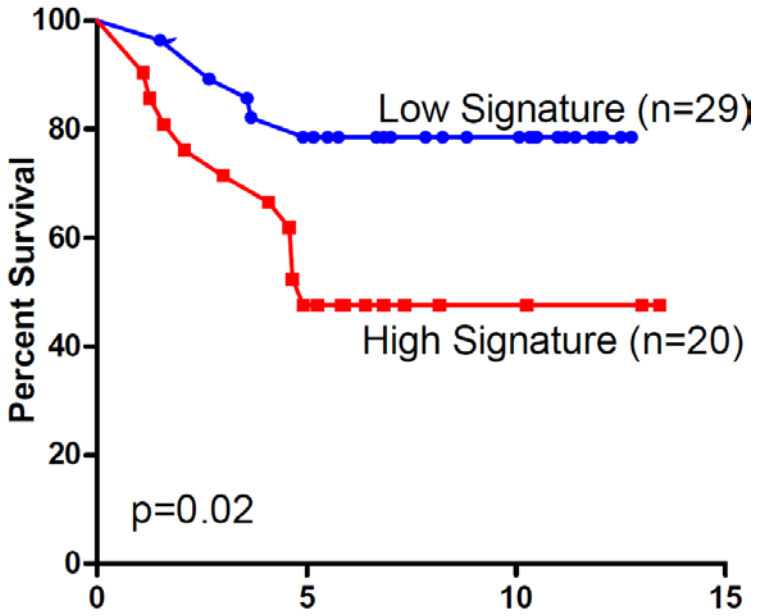
- Structural variants (SVs)
  - large-scale genomic deletions
  - Tandem duplications
  - Amplicons
  - Insertions
  - Inversions
- Copy number alterations (CNAs)
- Loss of heterozygoucity (LOH) regions
- Single nucleotide variants (SNVs) and short indels

Additionally, one can study gene expression and epigenetic profile

- Gene expression signature can distinguish cases with bad and good prognosis



Heat maps and dendrograms for expression of the 14-3-3ζ signature genes.

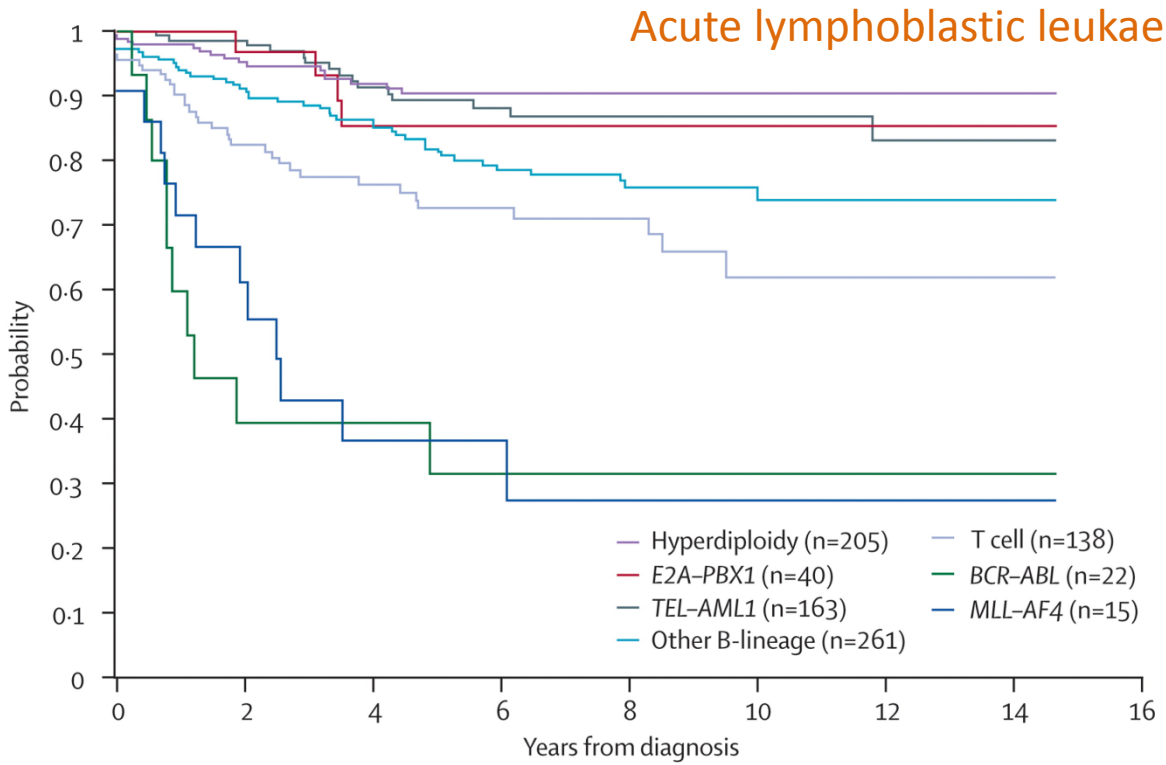


Kaplan-Meier survival curves of patients with ER-positive breast tumors based on expression patterns of the 14-3-3ζ gene signature

from Bergamaschi *et al. Breast Cancer Research* 2011 **13**:R70; data of Sorlie *et al.*

Additionally, one can study gene expression and epigenetic profile

- Expression of chimeric genes can cause tumor development and is often associated with clinical prognosis

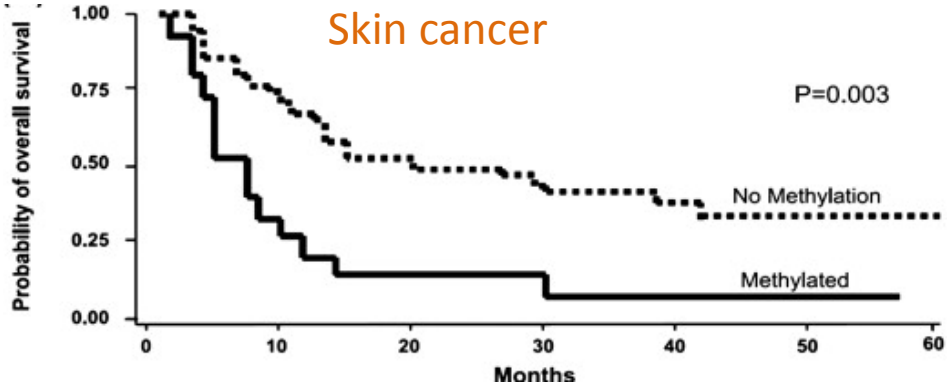


In general, the Philadelphia chromosome, t(4;11) with MLL-AF4 fusion, and hypodiploidy (<44 chromosomes per leukaemic cell) all confer a poor outcome, whereas hyperdiploidy (>50 chromosomes), TEL-AML1 fusion, and trisomy 4, 10, and 17 are associated with favourable prognosis

Additionally, one can study gene expression and epigenetic profile

- DNA methylation and chromatin modifications can be associated with survival

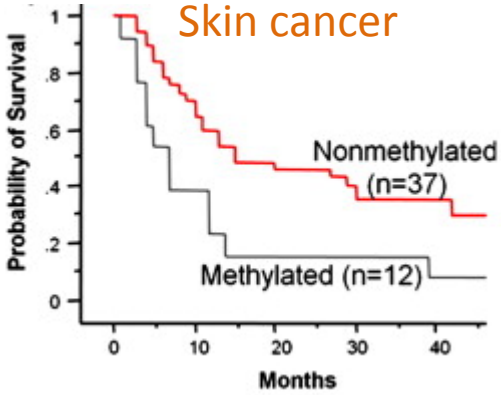
ER-α methylation
Skin cancer



Kaplan–Meier curves showing the correlation of pre-biochemotherapy serum ER-α methylation status with OS (p = 0.003)

From Mori et al, 2006

RASSF1A methylation
Skin cancer

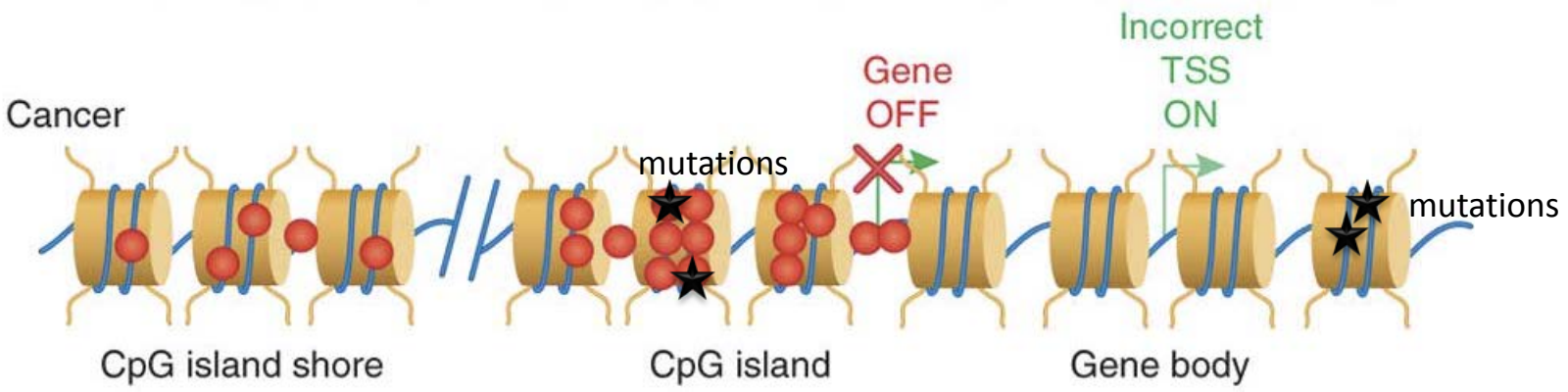


Kaplan–Meier survival curves of biochemotherapy patients: Correlation of pre-BC serum RASSF1A methylation BM with overall survival (p = .013).

From Mori et al, 2005

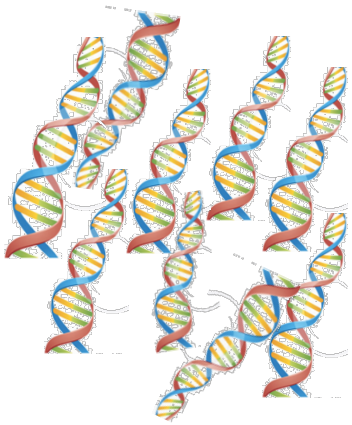
Outline for today's lecture

- Genetic profile of cancer
- Cancer Transcriptome
- Cancer epigenome



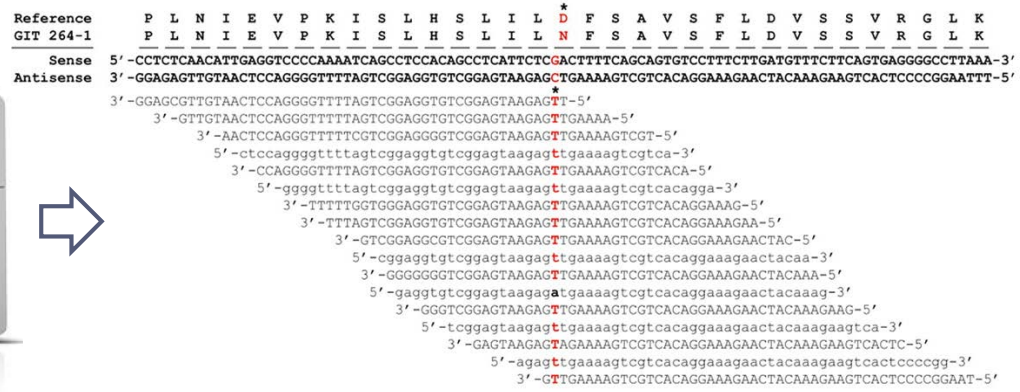
Next generation sequencing is the method of choice to study cancer genome, epigenome and transcriptome

- Genetic profile of cancer – sequence DNA
- Cancer Transcriptome – sequence cDNA
- Cancer epigenome – Bisulfite sequencing & sequence immunoprecipitated DNA



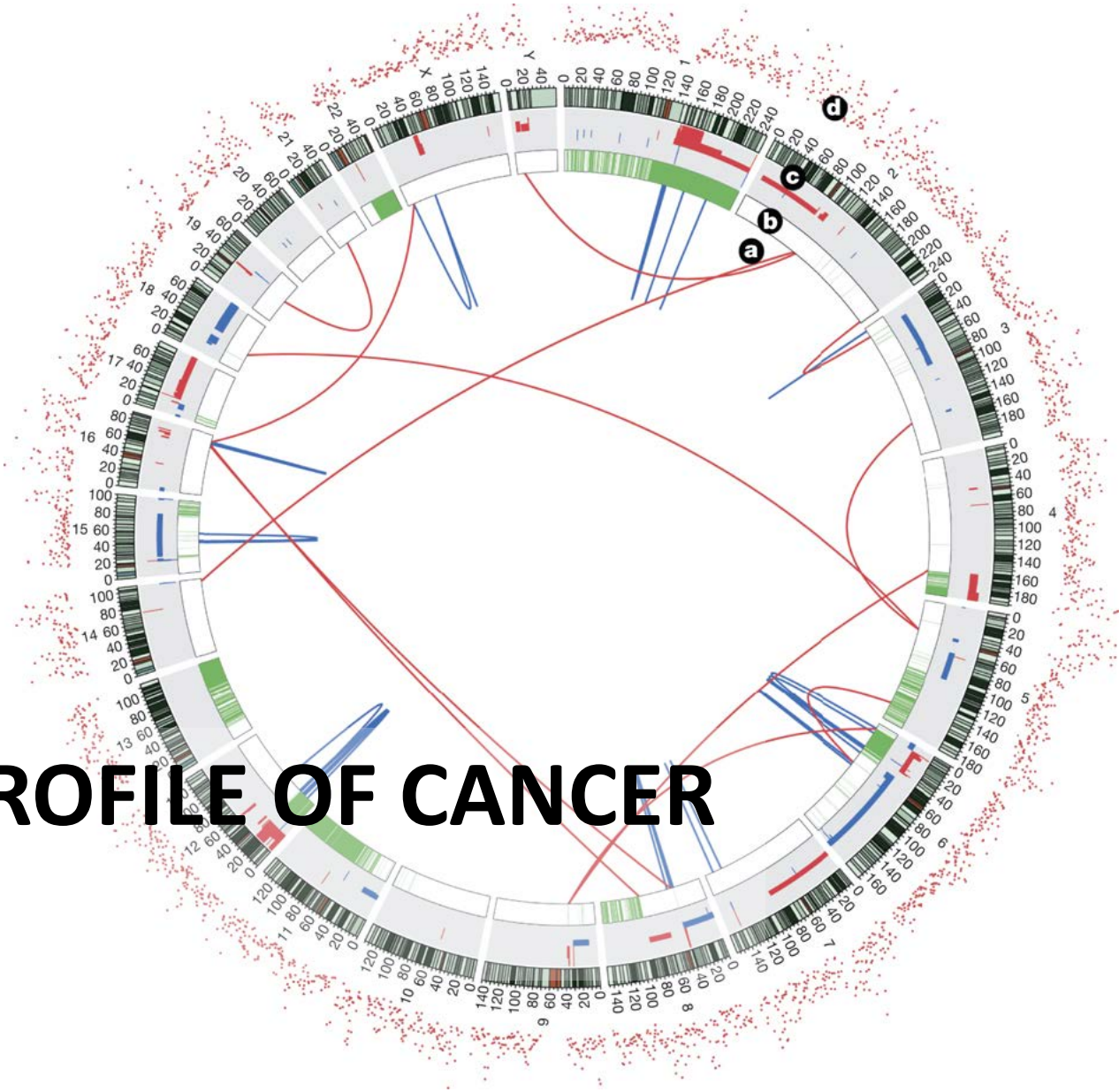
DNA

Sequencer



Reads

GENETIC PROFILE OF CANCER



W Lee *et al. Nature* **465**, 473-477 (2010) doi:10.1038/nature09004

Detection of SNVs, indels, structural variants, copy number changes and LOH has become possible with Next Generation Sequencing (NGS)

- Next Generation sequencing =

  Fast, Accurate Reading of DNA

  - Whole genome
  - Exome sequencing
  - Targeted sequencing

Detection of SNVs, indels, structural variants, copy number changes and LOH has become possible with Next Generation Sequencing (NGS)

- Next Generation sequencing =

  Fast, Accurate Reading of DNA

  - Whole genome
    - Sequencing of the whole cancer genome including intragenic regions and introns
    - Complete information about the genome
  - Exome sequencing
  - Targeted sequencing

Detection of SNVs, indels, structural variants, copy number changes and LOH has become possible with Next Generation Sequencing (NGS)

- Next Generation sequencing =

  Fast, Accurate Reading of DNA

  - Whole genome
  - Exome sequencing
    - Sequencing of exons of ~20000 well characterized genes
    - Complete information about SNVs, indels and copy number changes of the coding part of the genome
  - Targeted sequencing

Detection of SNVs, indels, structural variants, copy number changes and LOH has become possible with Next Generation Sequencing (NGS)

- Next Generation sequencing =

  Fast, Accurate Reading of DNA

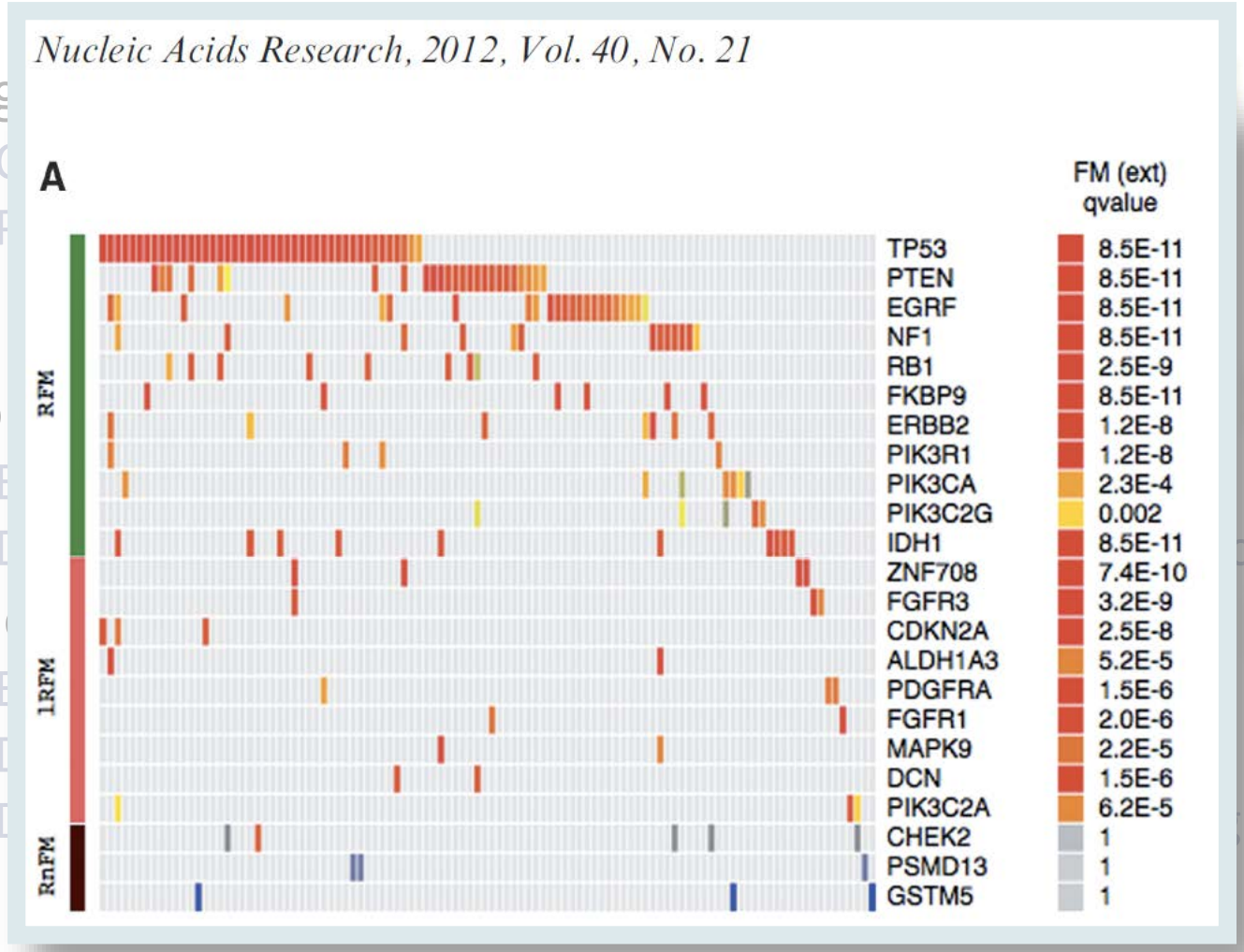
  - Whole genome
  - Exome sequencing
  - Targeted sequencing
    - Complete information about SNVs, indels, copy numbers of a small panel of genes (10-500) actionable in cancer

Shorter the sequenced part lower the price (but less information)

- Targeted sequencing (400 – 1500 Euros) $
  - Clinics
  - Propose targeted therapy, e.g.:
    - Vemurafenib against activating BRAF mutation V600E
    - Erlotinib or Gefitinib against activating EGFR mutations
  - Detection of gene deletions and amplifications $ $
- Exome sequencing (1000-4000 Euros)
  - Biological research (rarely clinics)
  - Detection of driver mutations in different (sub)types of cancer
  - Detection of copy number changes $ $ $
- Whole genome sequencing (2000-10000 Euros)
  - Biological research
  - Detection of driver mutations, rearrangements
  - Detection of copy number changes

Shorter the sequenced part lower the price (but less information)

- Targ
  - o C
  - o F
- Exo
  - o E
  - o D
- Wh
  - o E
  - o D
  - o D



$

cancer

$ $

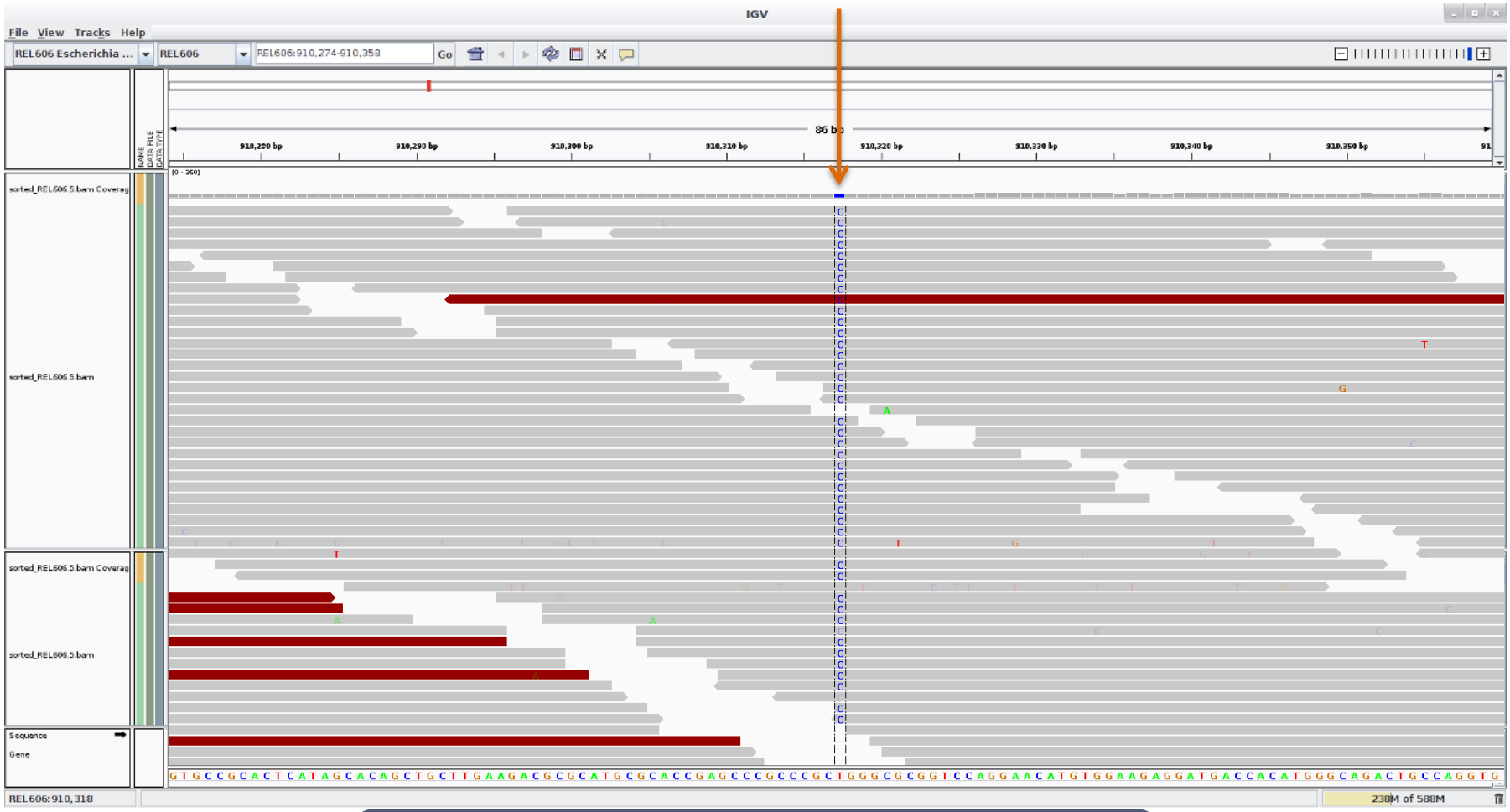
nosis

Shorter the sequenced part lower the price (but less information)

- Targeted sequencing (400 – 1500 Euros) $
  - Clinics
  - Propose targeted therapy, e.g.:
    - Vemurafenib against activating BRAF mutation V600E
    - Erlotinib or Gefitinib against activating EGFR mutations
  - Detection of gene deletions and amplifications $ $
- Exome sequencing (1000-4000 Euros)
  - Biological research (rarely clinics)
  - Detection of driver mutations in different (sub)types of cancer
  - Detection of copy number changes $ $ $
- Whole genome sequencing (2000-10000 Euros)
  - Biological research
  - Detection of driver mutations, rearrangements
  - Detection of copy number changes

High depth of targeted and exome sequencing allows for accurate identification of SNVs and small indels

IGV genome browser



Similar to SNV and indel detection in high-depth whole genome data

Data can be very noisy

IGV genome browser

Homozygous SNVs/SNPs and a deletion

One can detect SNVs/SNPs in the coding regions + some flanks

IGV genome browser

Heterozygous SNV or SNP

Binomial test is commonly used to evaluate the significance of SNP predictions

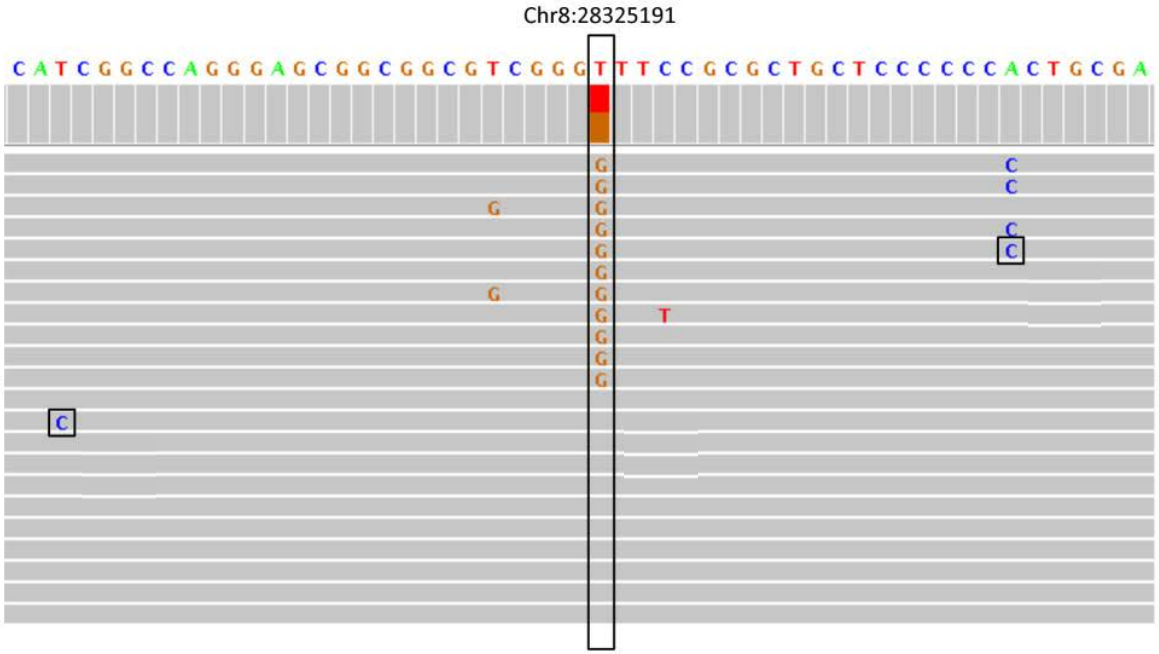
- The basic model is a binomial model for the counts $X_{i,j}$ of nucleotide j at position i in the test :

$$X_{i,j} \sim Bin(ni, p_{i,j})$$

- $H_0 = \{p_{i,j} = \varepsilon, \text{if } j \neq Reference\}$
- $H_1 = \{p_{i,j} > \varepsilon, \text{if } j \neq Reference\}$

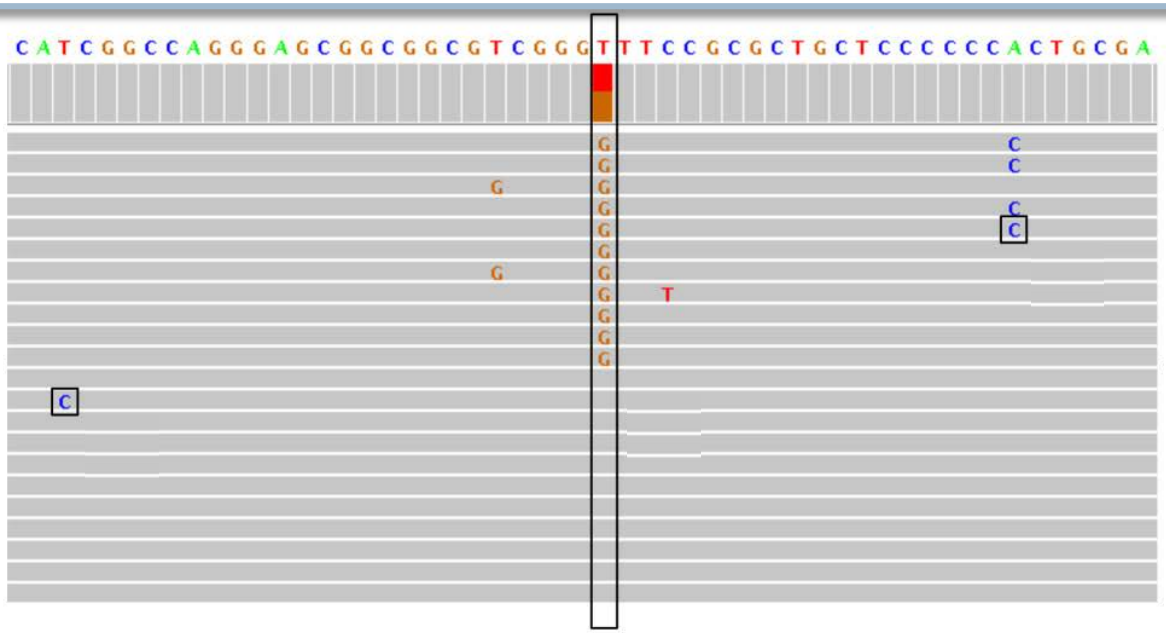
ε is the probability of sequencing error

n_i denotes the coverage



Binomial test is commonly used to evaluate the significance of SNP predictions

What is the probability to get 11 « G » out of 22 if the probability to get « G » (probability of error) is 0.01???



Binomial test is commonly used to evaluate the significance of somatic SNV predictions

- The basic model is a binomial model for the counts $X_{i,j}$ and $Y_{i,j}$ of nucleotide j at position i, in the test and the control experiment, respectively:

$$X_{i,j} \sim Bin(ni, p_{i,j})$$
$$Y_{i,j} \sim Bin(m_i, q_{i,j})$$

  n_i and m_i denote the coverage in the two experiments

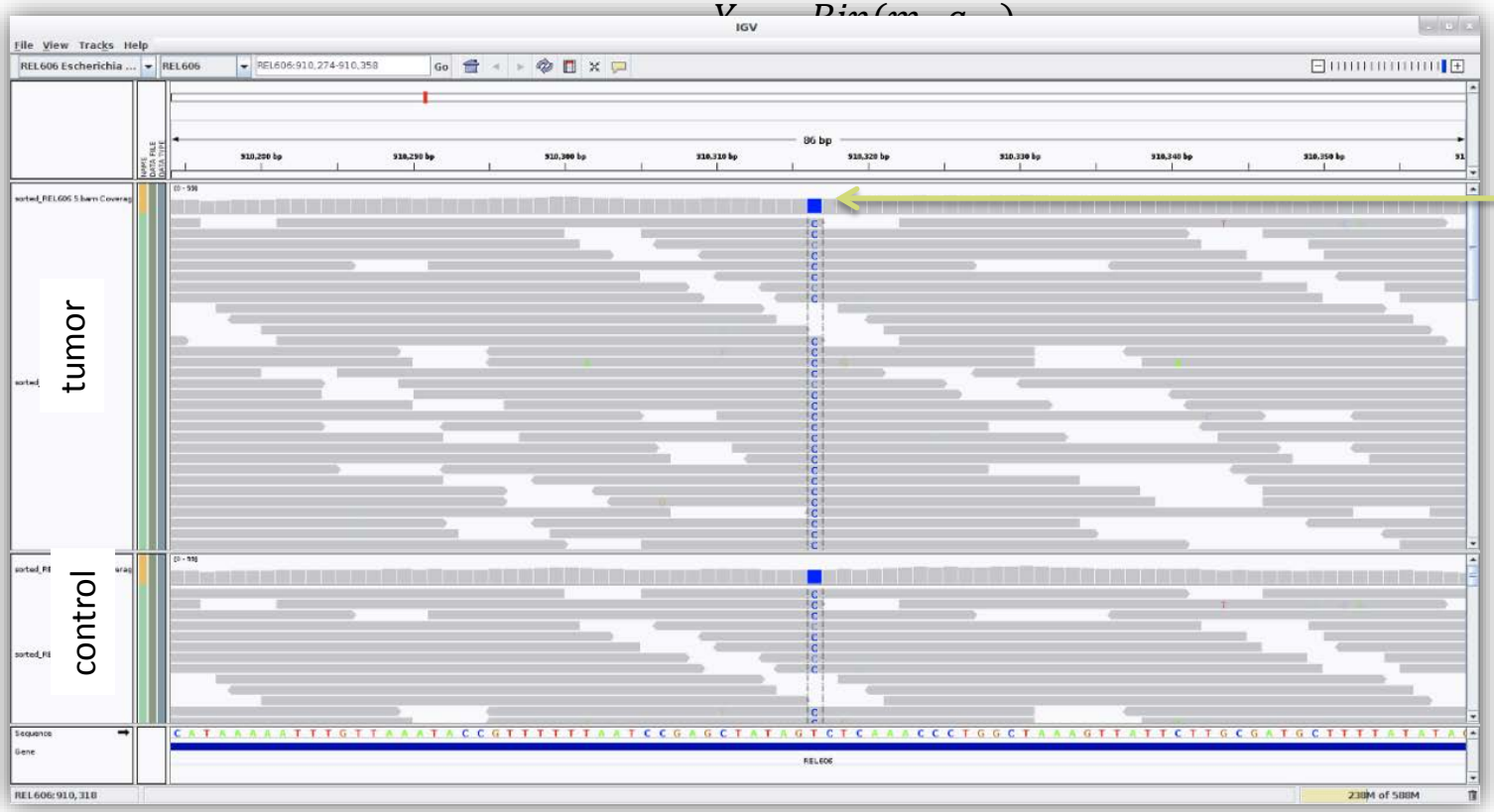
- $H_1 : p_{i,j} > q_{i,j}$
- $H_0 : p_{i,j} = q_{i,j}$
- One can use likelihood ratio test with a χ^2 -distribution (deepSNV)

Binomial test is commonly used to evaluate the significance of somatic SNV predictions

- The basic model is a binomial model for the counts $X_{i,j}$ and $Y_{i,j}$ of nucleotide j at position i, in the test and the control experiment, respectively:

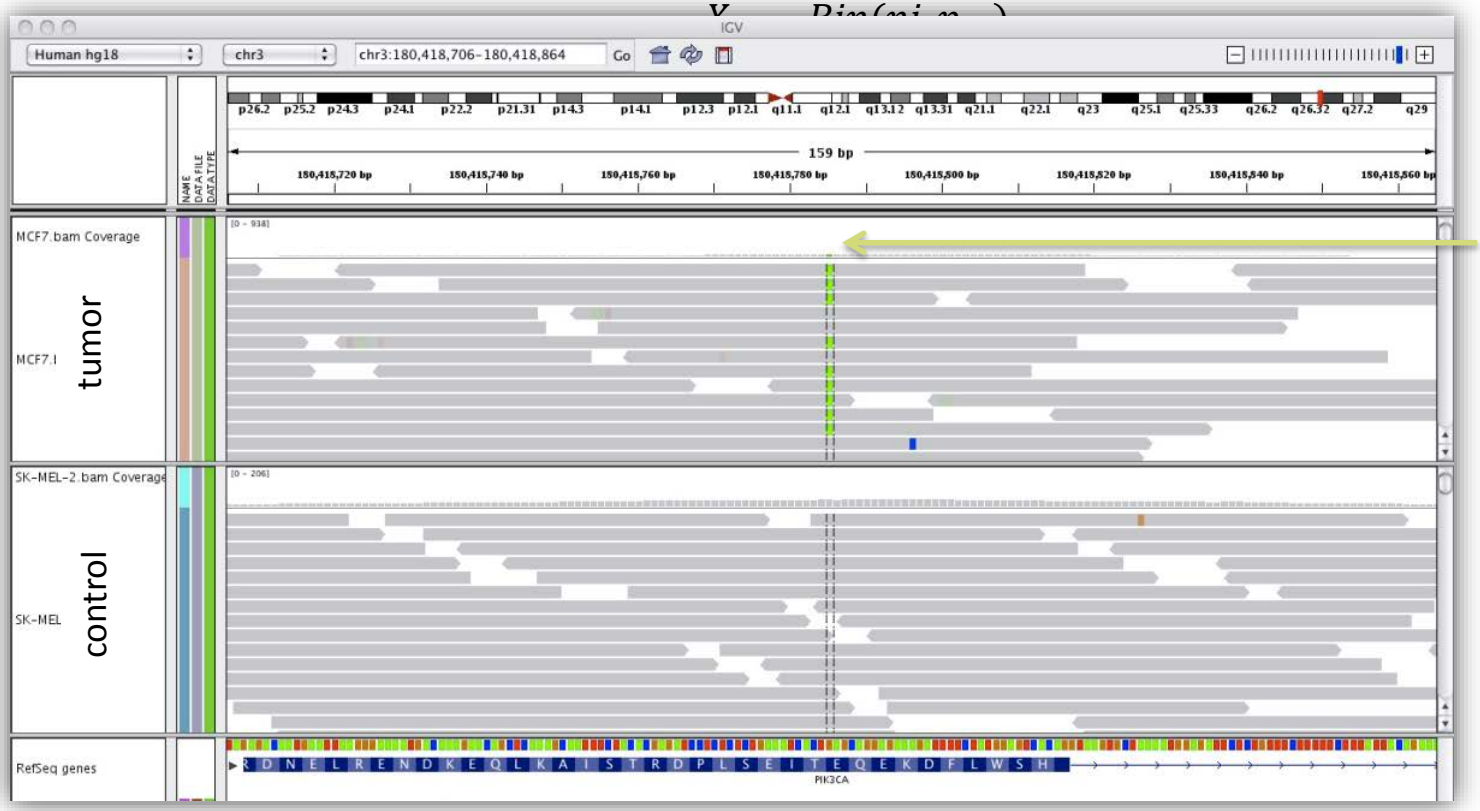
$$X_{i,j} \sim Bin(ni, p_{i,j})$$

$$Y \sim Bin(m, q_{i,j})$$

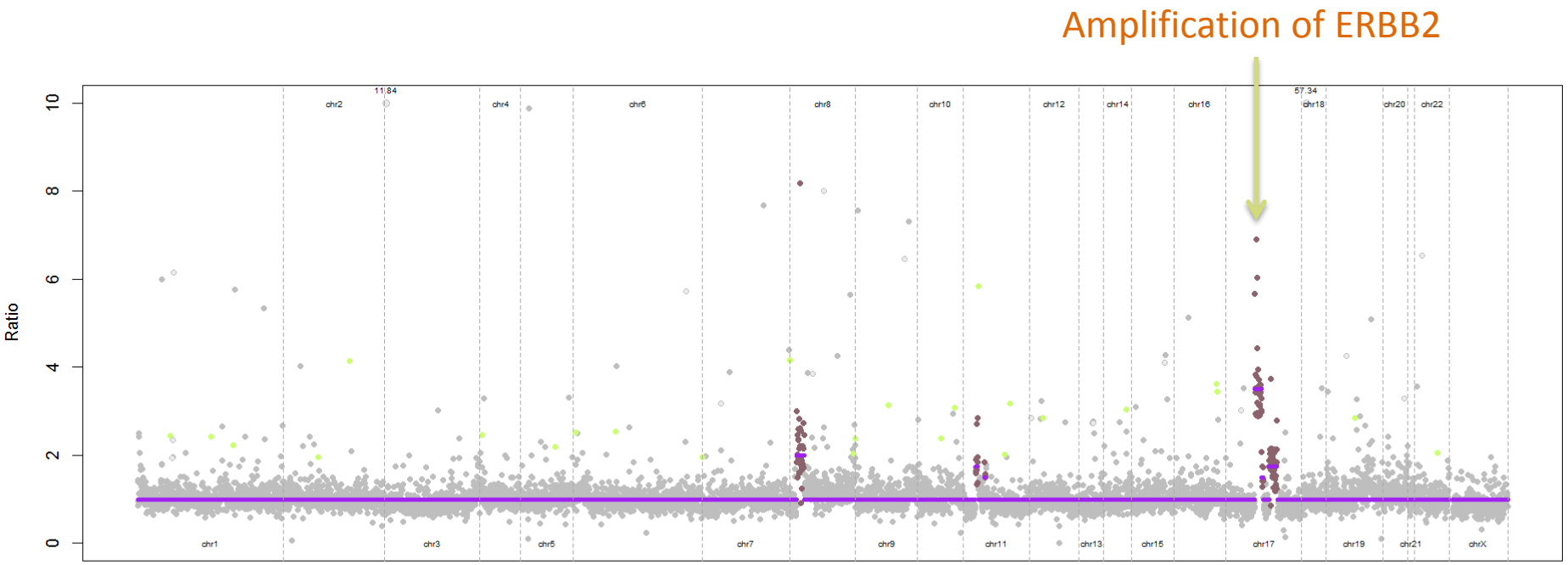


Binomial test is commonly used to evaluate the significance of somatic SNV predictions

- The basic model is a binomial model for the counts $X_{i,j}$ and $Y_{i,j}$ of nucleotide j at position i, in the test and the control experiment, respectively:



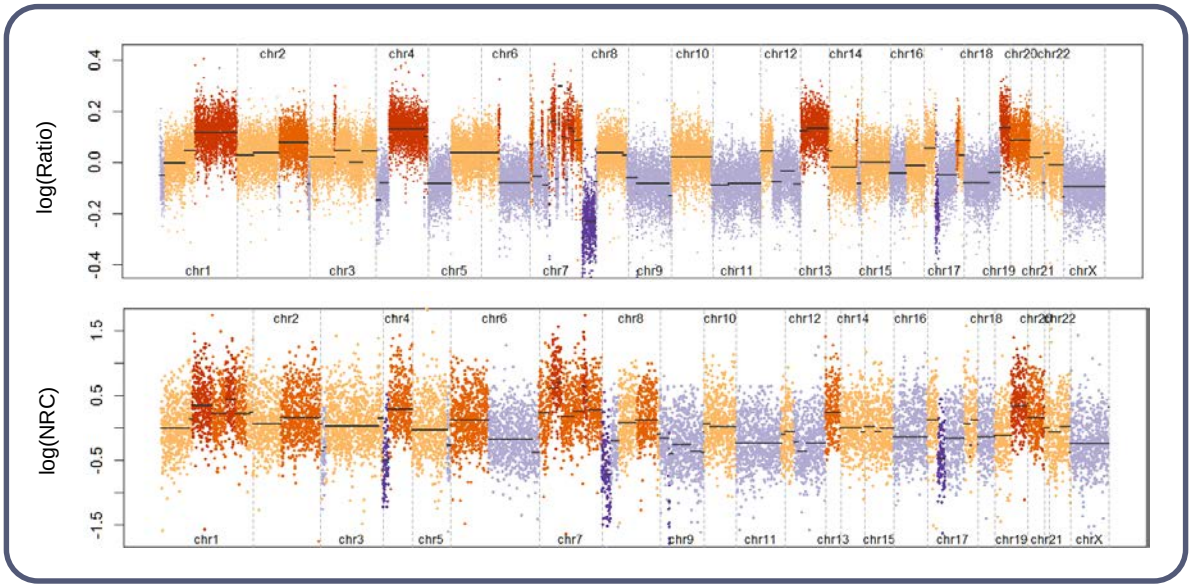
High depth of targeted sequencing may be used to detect amplified or deleted genes from the gene panel of interest



High depth of targeted sequencing may be used to detect amplified or deleted genes from the gene panel of interest

ONCOCNV:

- Multi-factor data normalization enables the detection of copy number aberrations in amplicon sequencing data



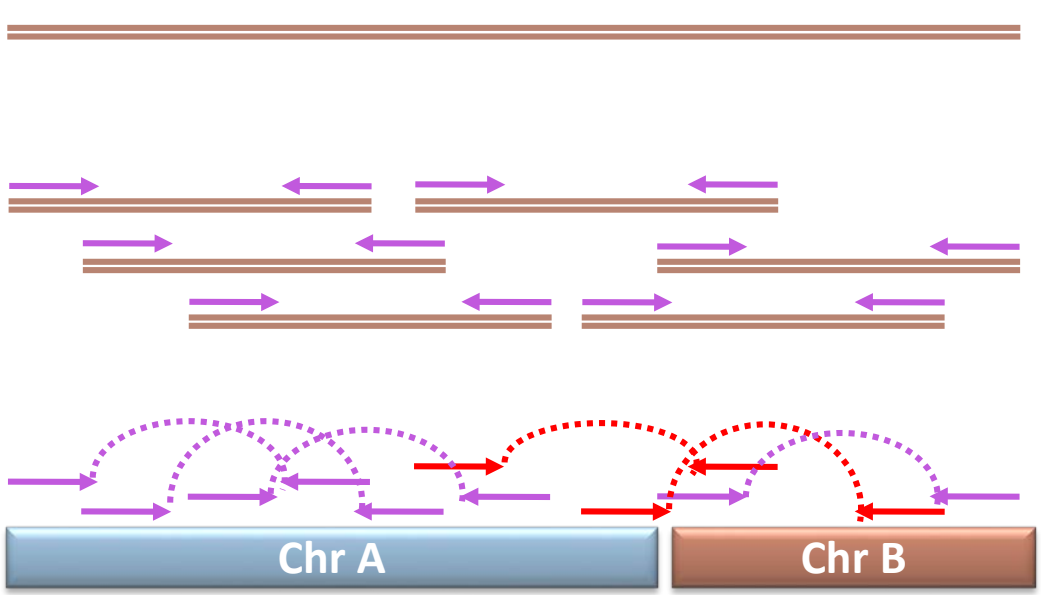
Array CGH profile

Amplicon sequencing profile (ONCOCNV)

Accuracy = 0.89

V. Boeva *et al*, *Bioinformatics*, 2014

In whole genome sequencing and exome seq, paired-end reads are used to detect SNVs, structural and copy number variants



Primary DNA

⇩

Fragment at random and sequence both ends of each DNA fragment

⇩

Map sequenced reads to the reference genome

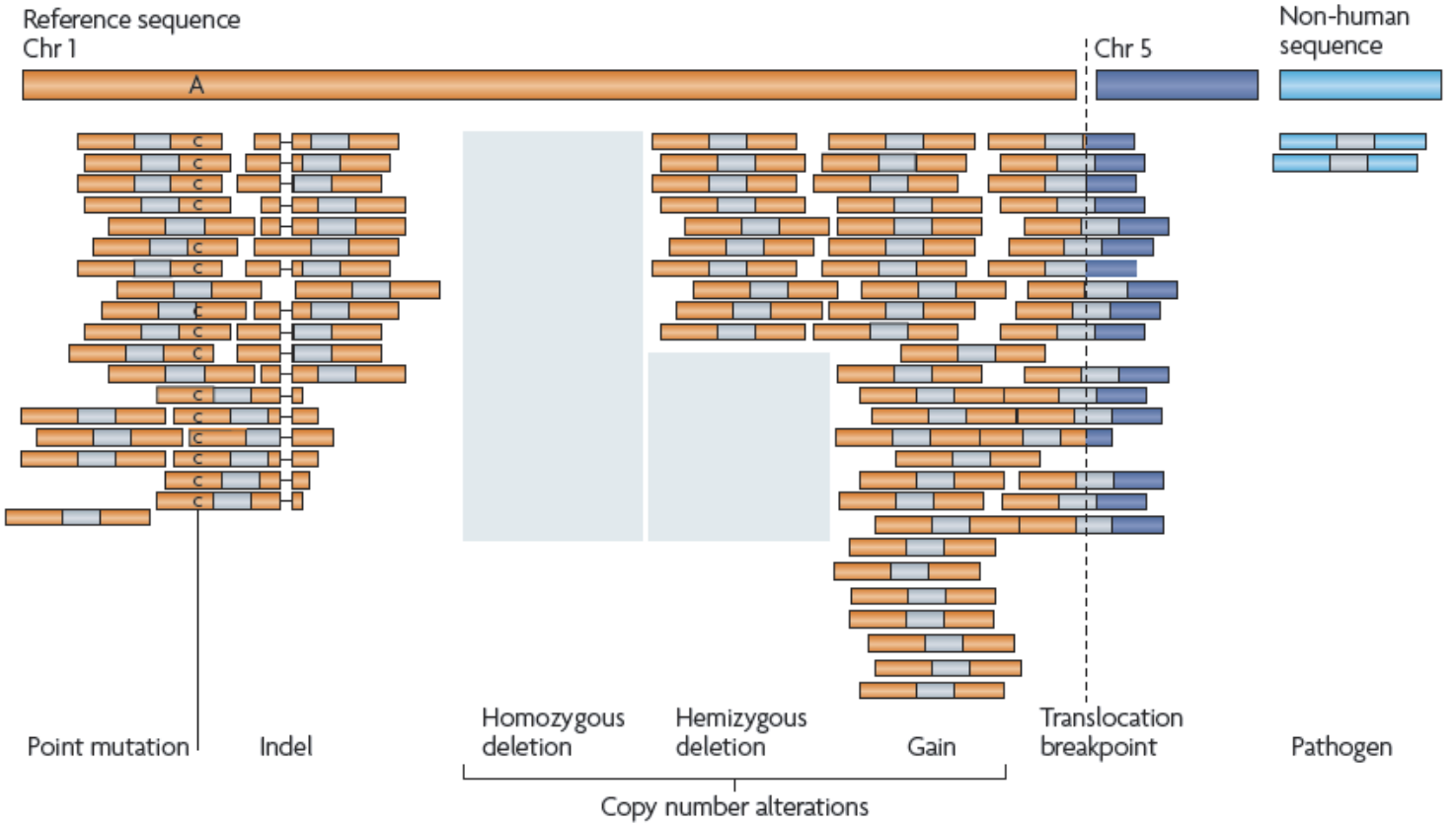
⇙ SNVs and short indels

⇓ Large structural variants (translocations, large indels, inversions, etc.)

⇘ Copy number alterations and LOH

Paired-end (mate-pair) sequencing is used to detect mutations, structural and copy number variants

from E. Martin presentation, 2011

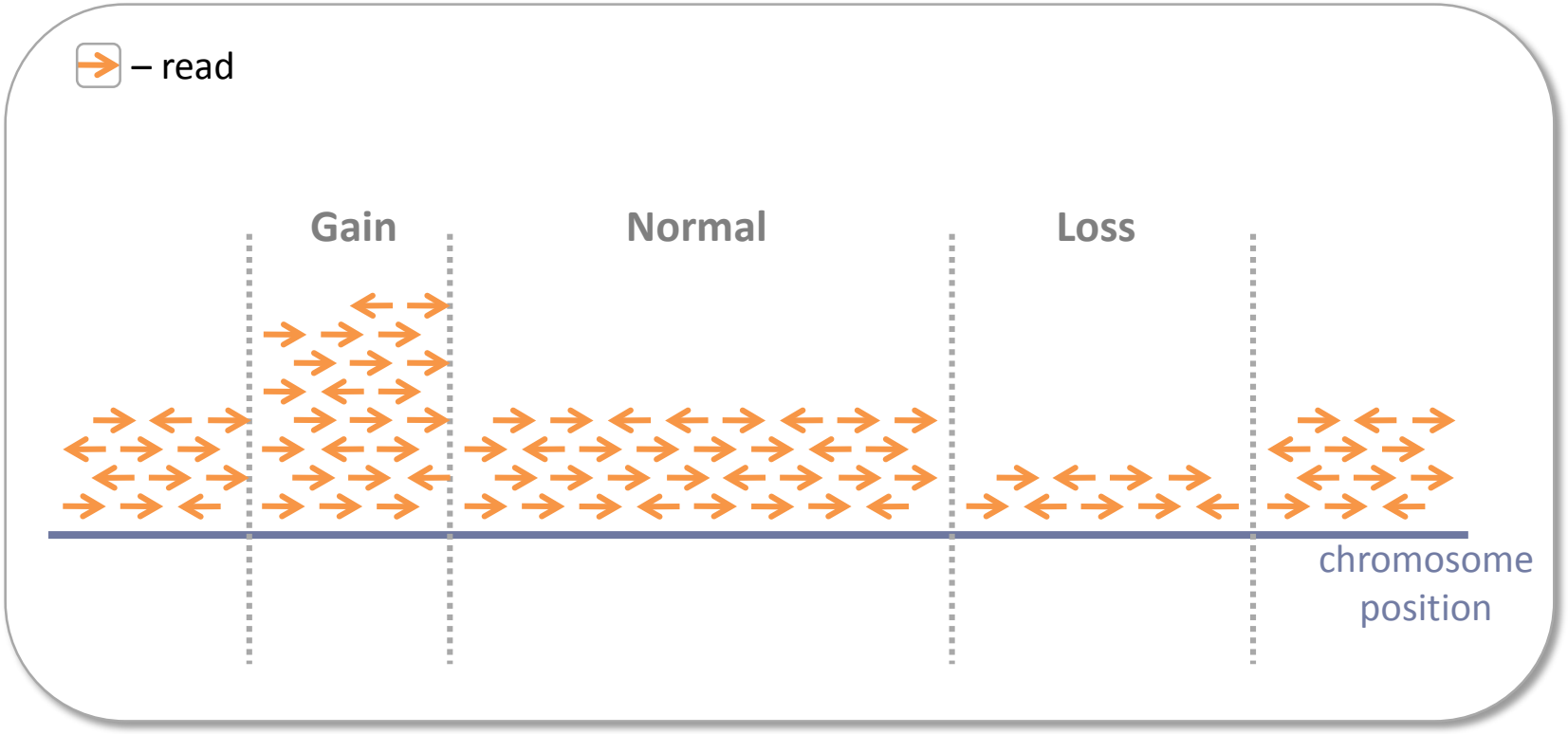


SNVs and short indels

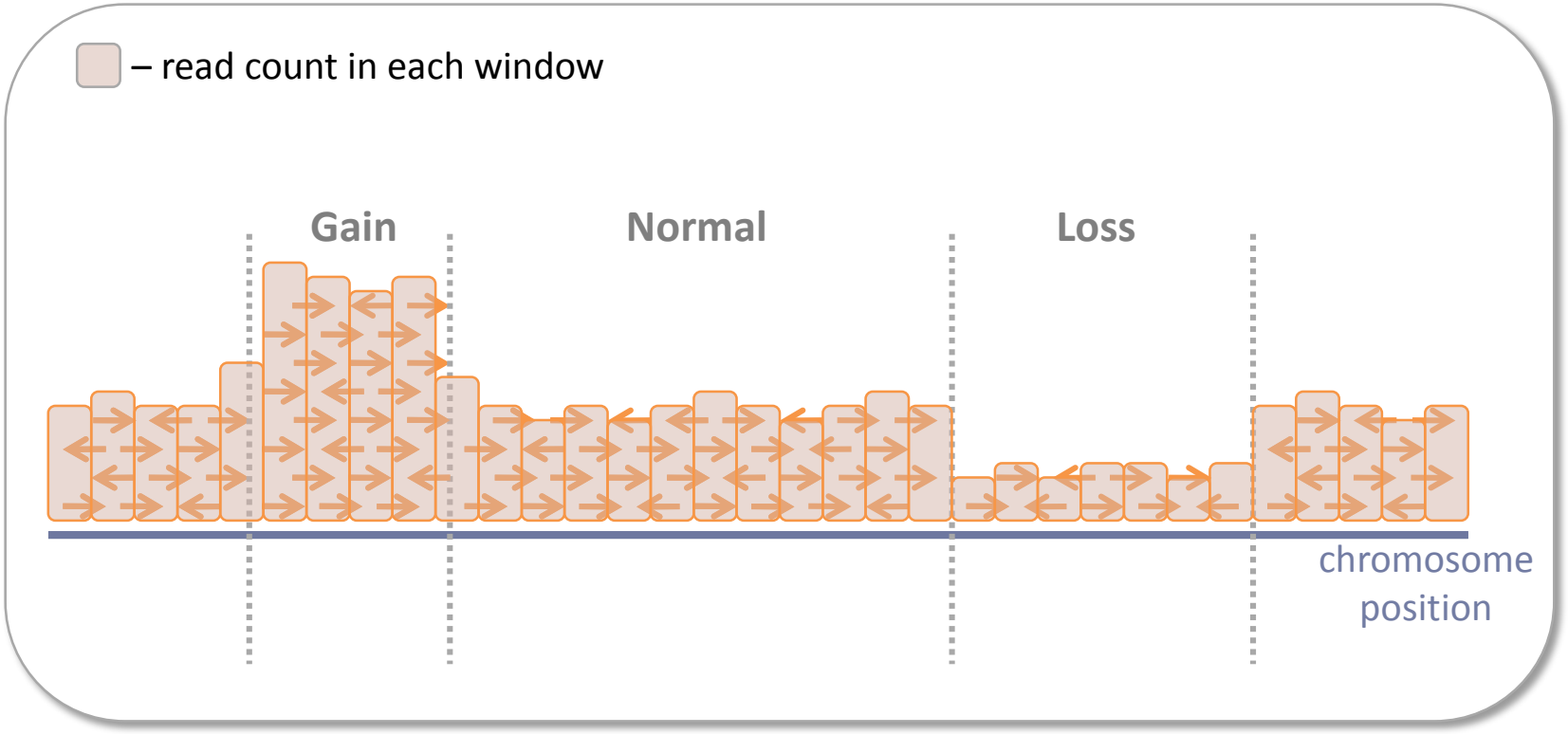
Copy number alterations and LOH

Large structural variants (translocations, large indels, inversions, etc.)

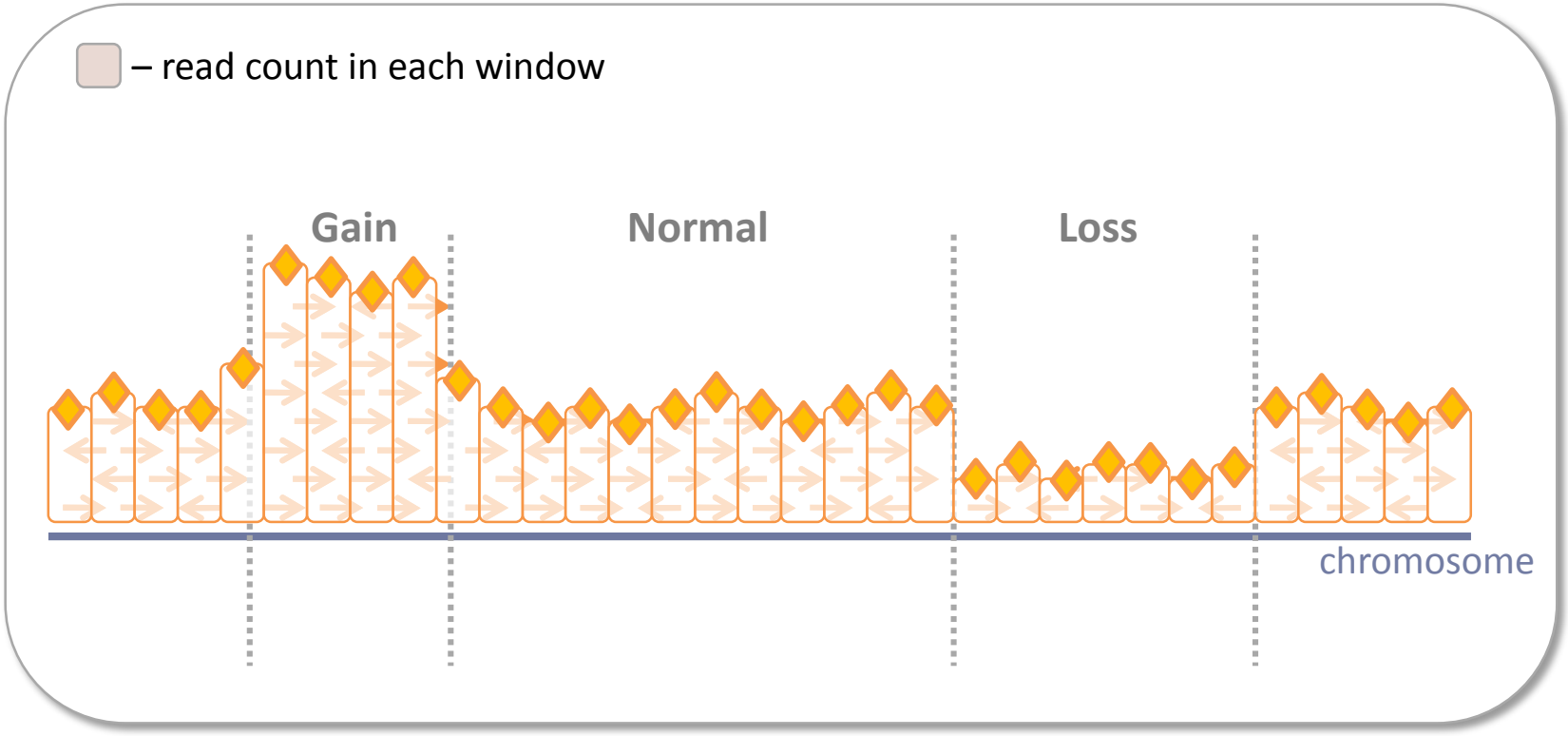
Gains and losses should be detected using depth of coverage



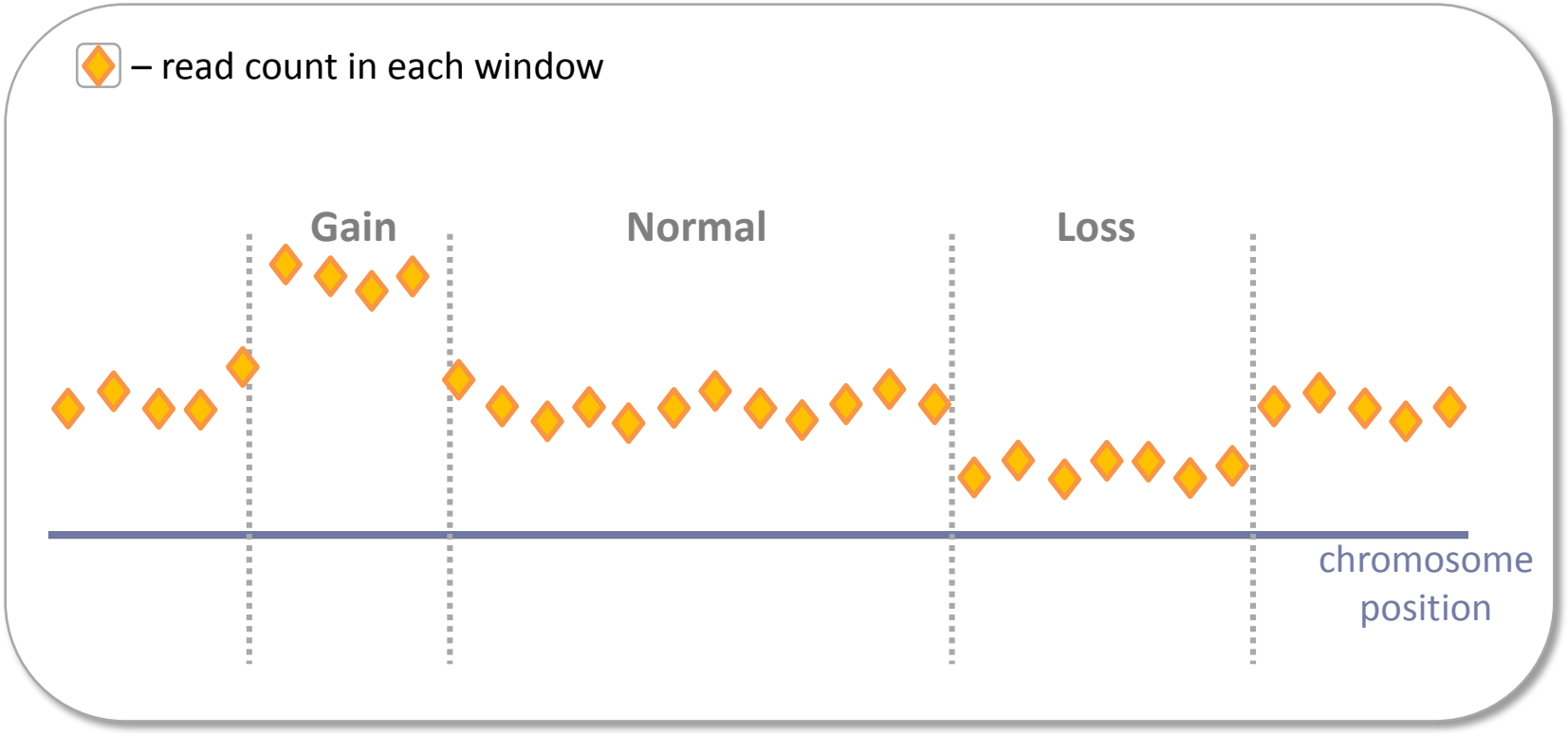
Read count (RC) is calculated in sliding windows



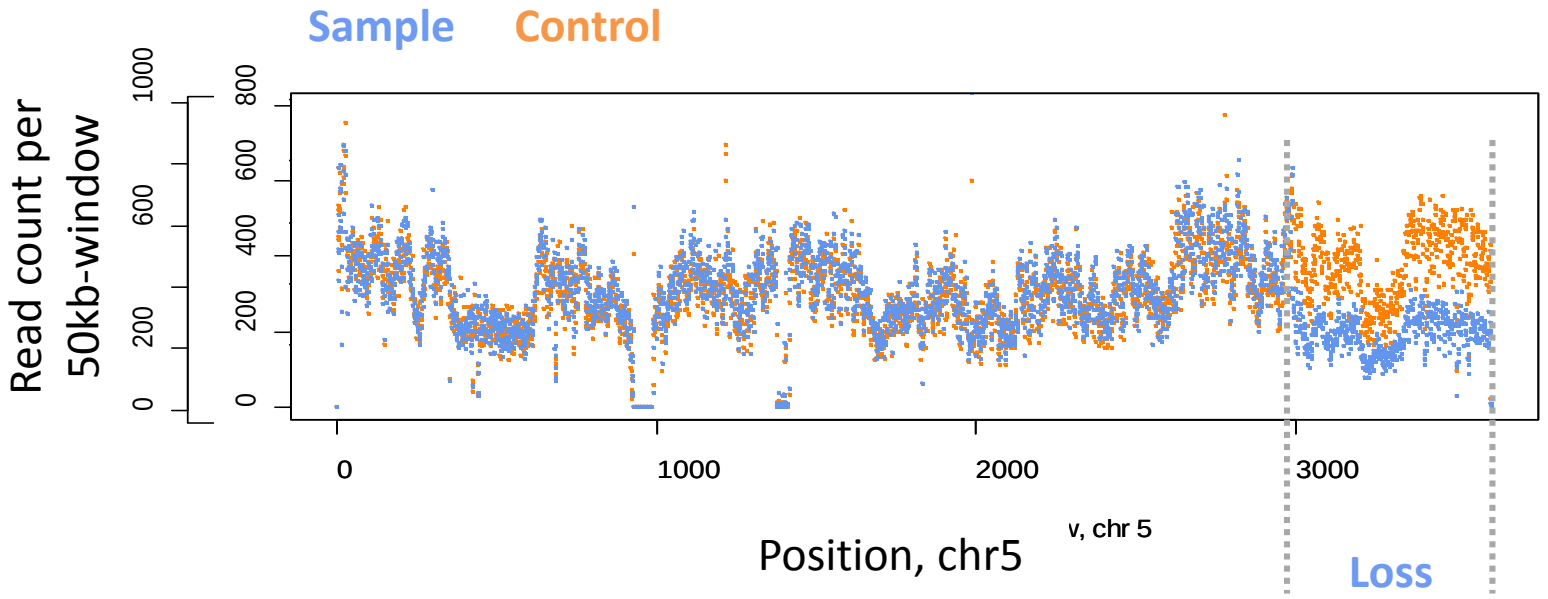
Read count (RC) is calculated in sliding windows



Read count (RC) is calculated in sliding windows



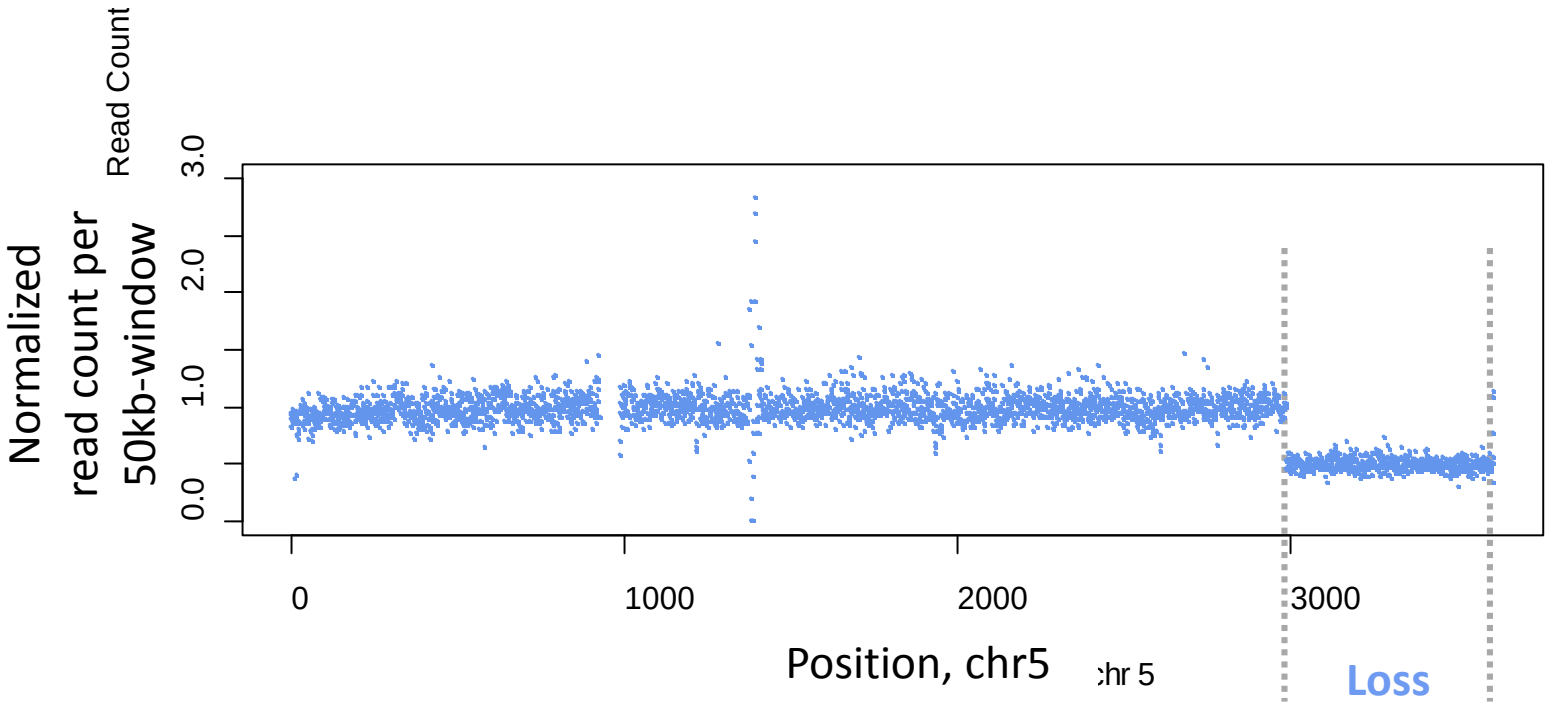
We need to normalize read count per window to get meaningful profiles



If control is available, the problem is easily solved

$$\text{Normalized RC} = \alpha \times \frac{\text{Observed RC}}{\text{RC in control}}$$

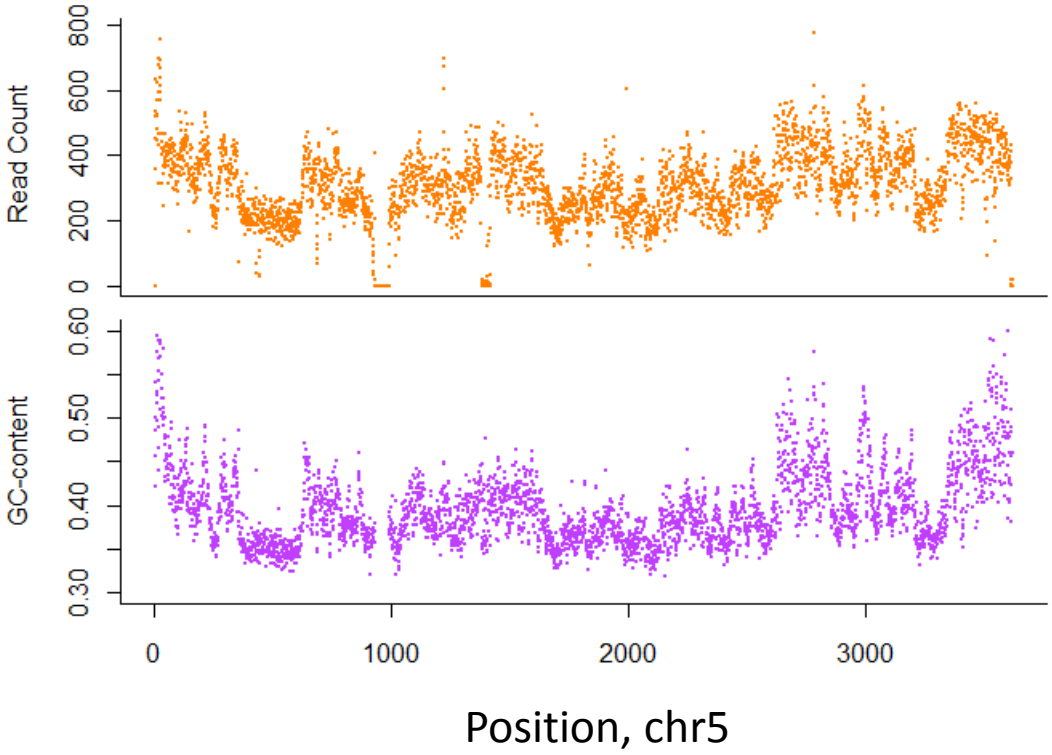
RC = read count



If there is no control dataset, normalization can be done using the GC-content

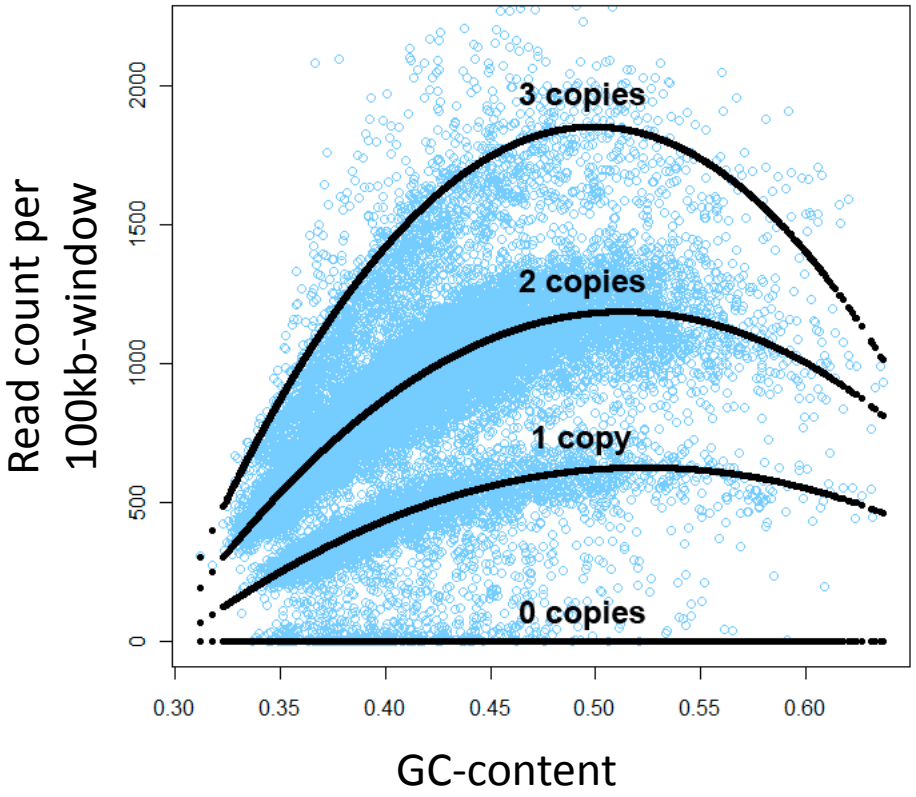
Control

GC-content

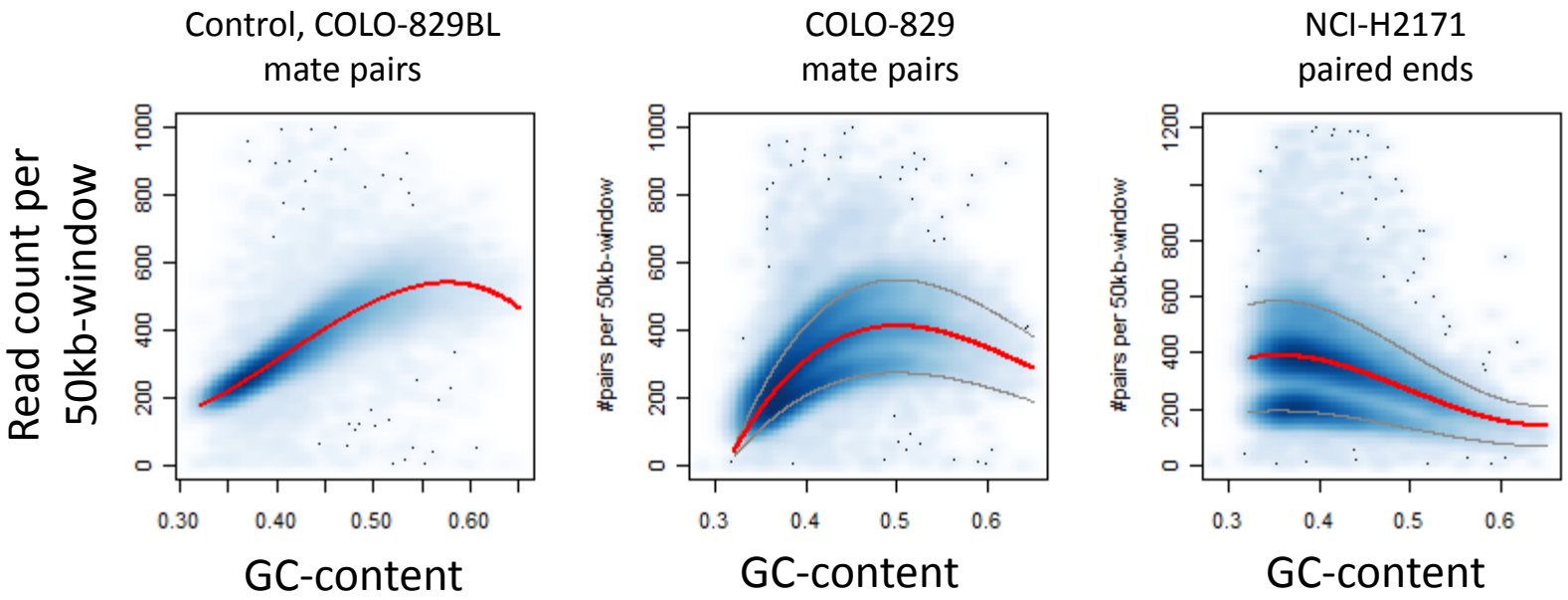


RC can be modeled as a polynomial on GC-content

A scatter plot shows the dependency **RC ~ GC-content**



RC can be modeled as a polynomial on GC-content



– main component

– components corresponding to losses and gains

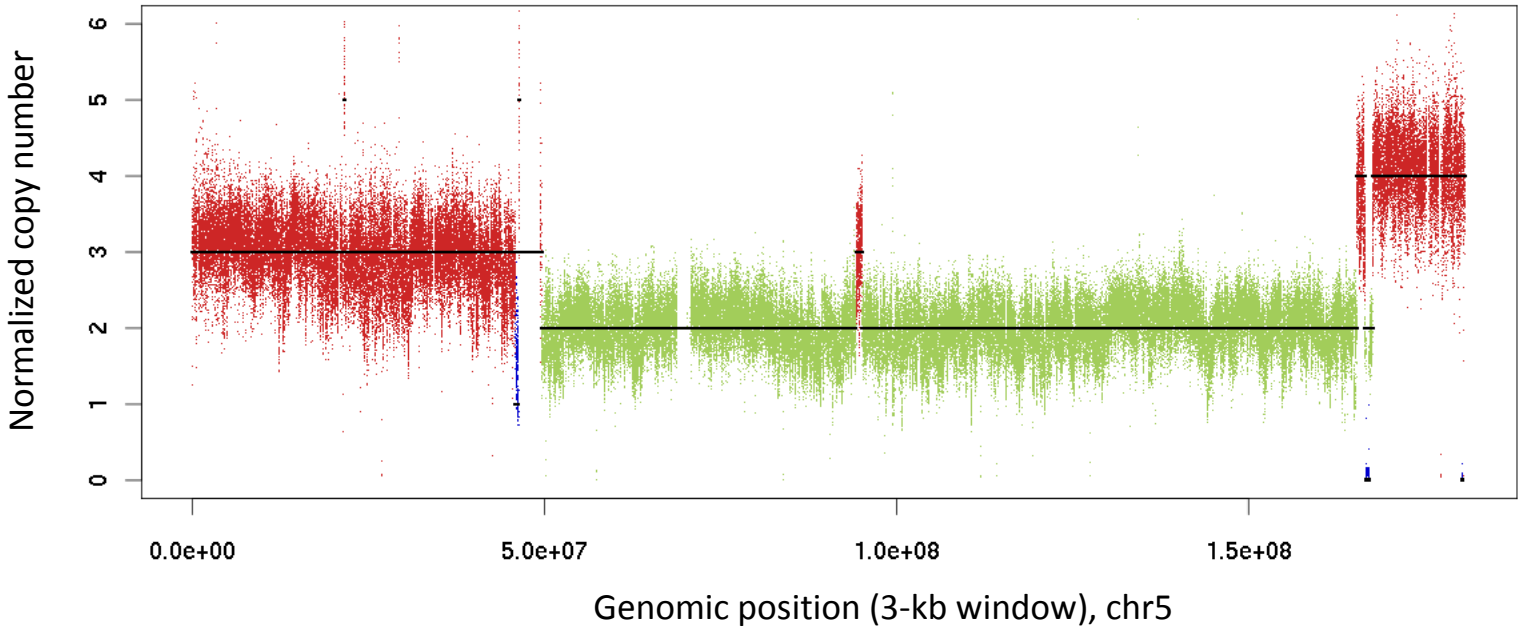
Here RC was modeled as a polynomial of order three on GC-content

The resulting profiles are segmented to detect gains and losses

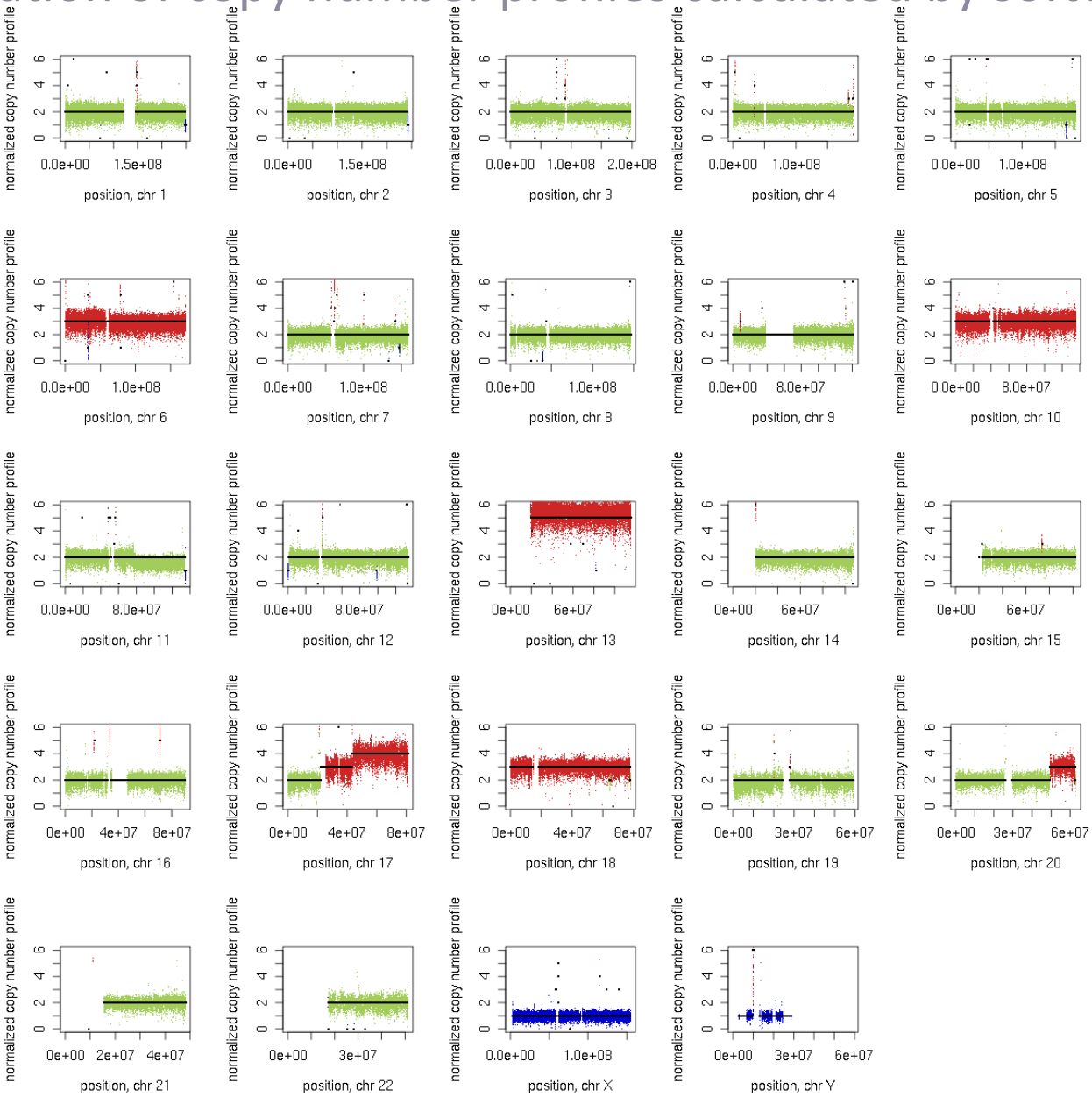
Transformation: $NRC_i = \frac{RC_i}{f(g_i)} \cdot ploidy$

g_i = GC-content in window i
RC_i = is read count in window i,
NRC_i = resulting normalized read count

● – normal copy number ● – loss ● – gain

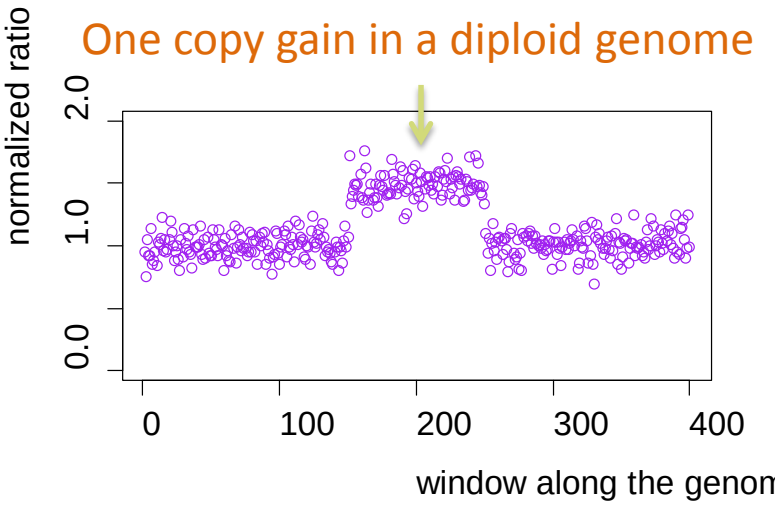


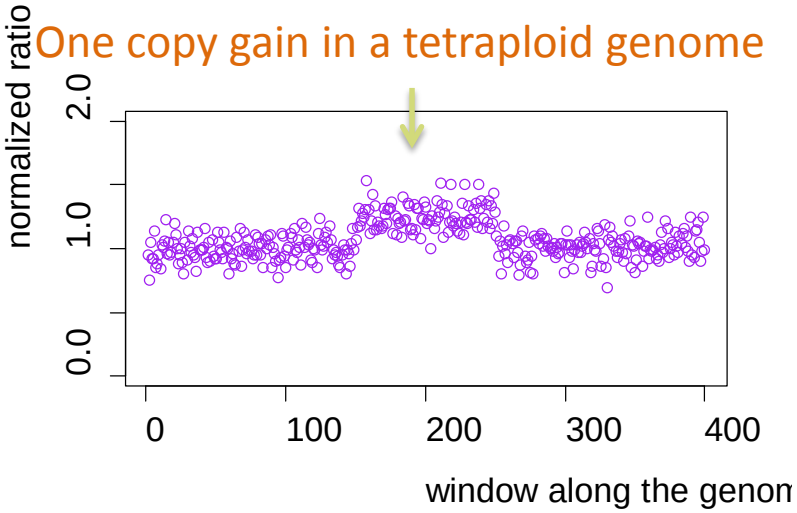
Visualization of copy number profiles calculated by software FREEC



There are 3 problems of genomic profiling

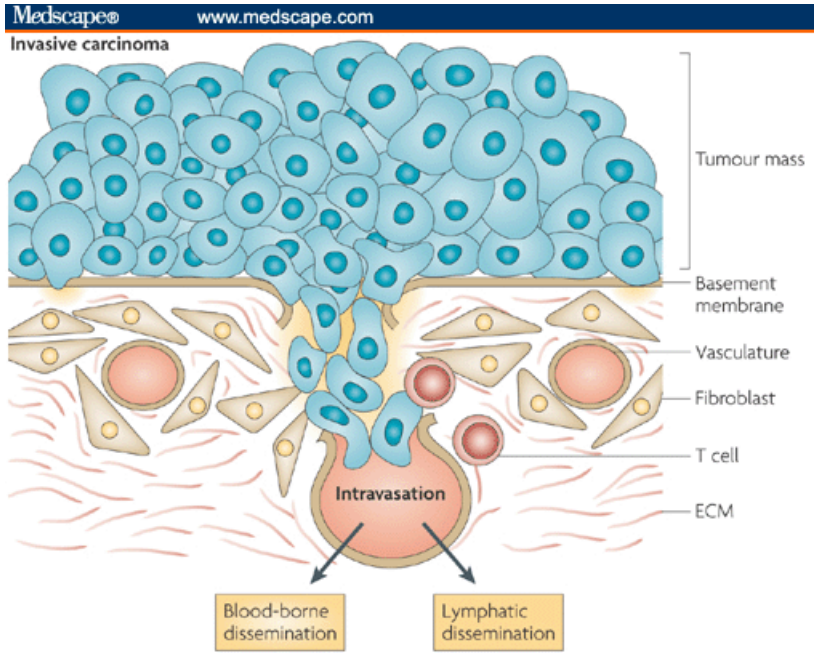
1. Reference point for copy number variation (diploid, triploid, tetraploid genomes)





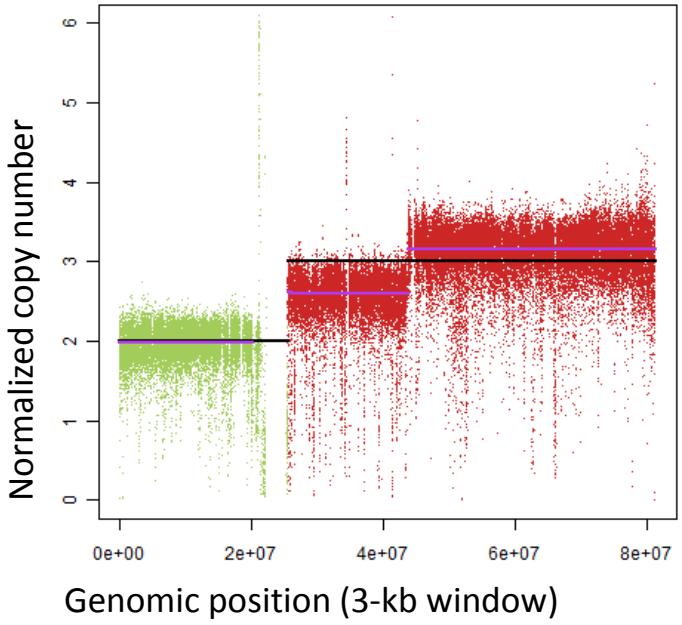
There are 3 problems of genomic profiling

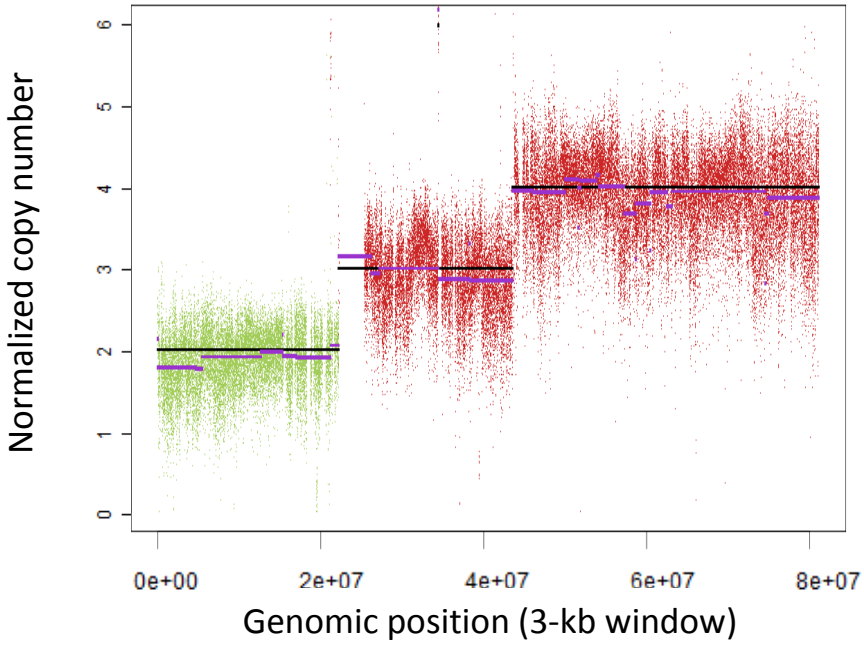
2. Contamination of tumor samples by normal stroma cells



We can evaluate contamination of a tumor sample by normal cells

NRC_i = normalized read count
E_i = expected read count
p = proportion of normal cells
2 = normal ploidy

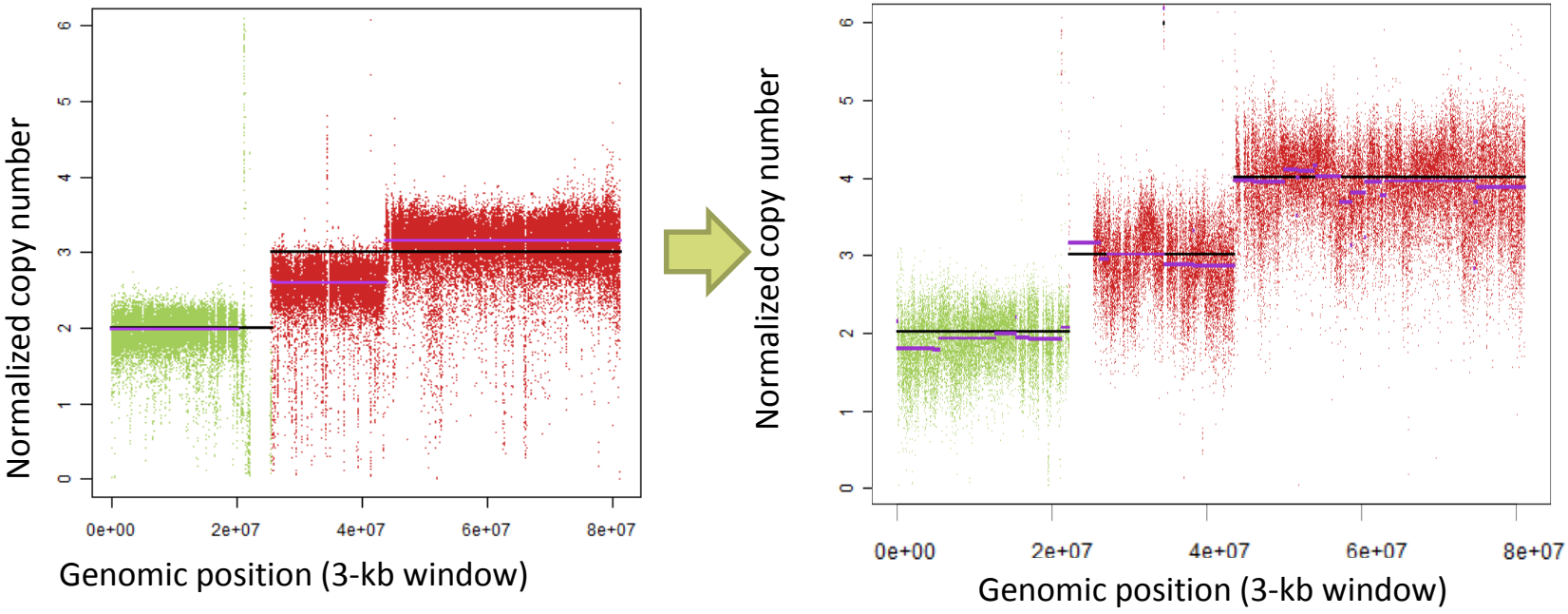




We can evaluate contamination of a tumor sample by normal cells

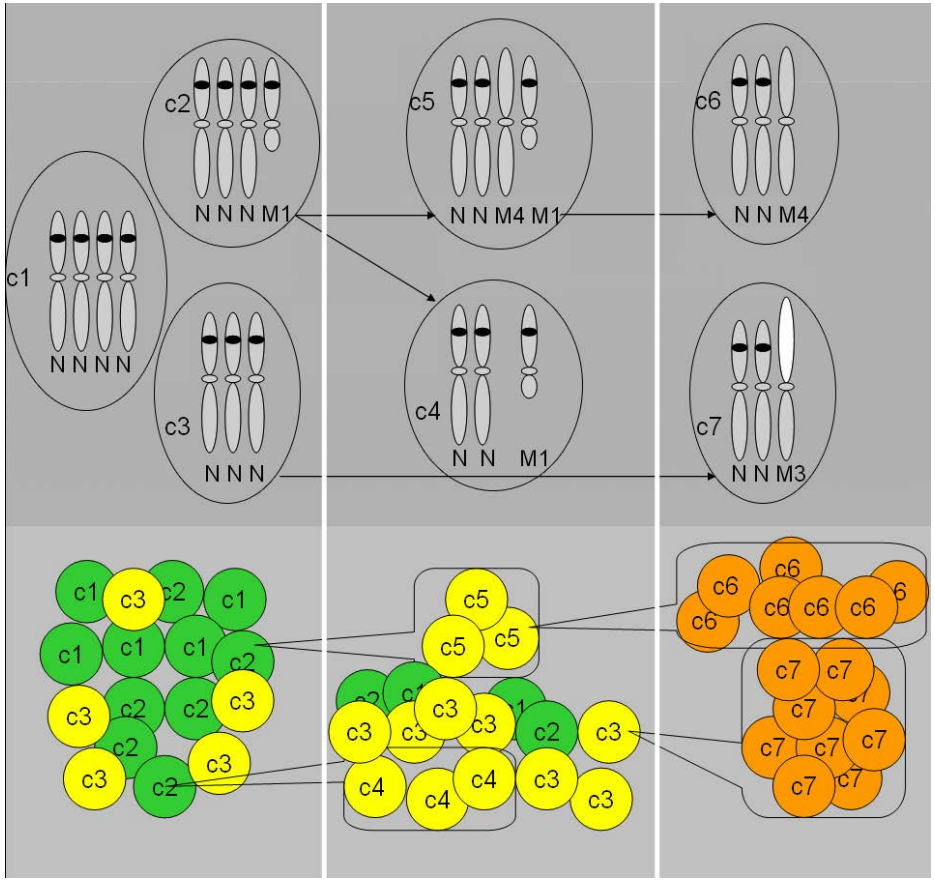
$$Observed\ NRC_i \approx (1-p) \times E_i + p \times 2$$

NRC_i = normalized read count
E_i = expected read count
p = proportion of normal cells
2 = normal ploidy



There are 3 problems of genomic profiling

3. Intra-tumoral heterogeneity

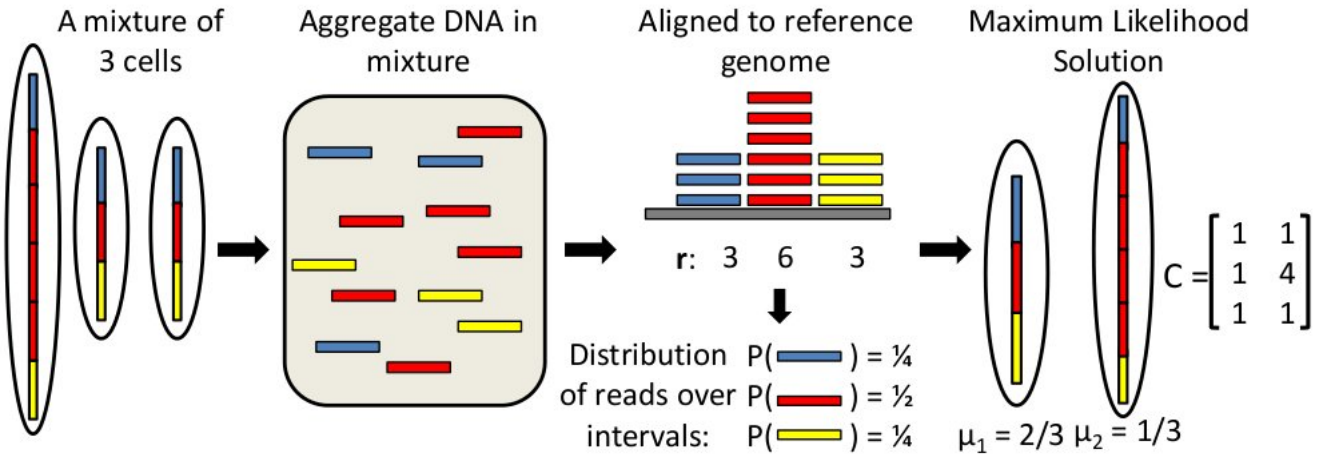


from Kost-Alimova et al, BMC Cancer 2007

There are 3 problems of genomic profiling

3. Intra-tumoral heterogeneity

One solution: Tumor Heterogeneity Analysis (THetA)

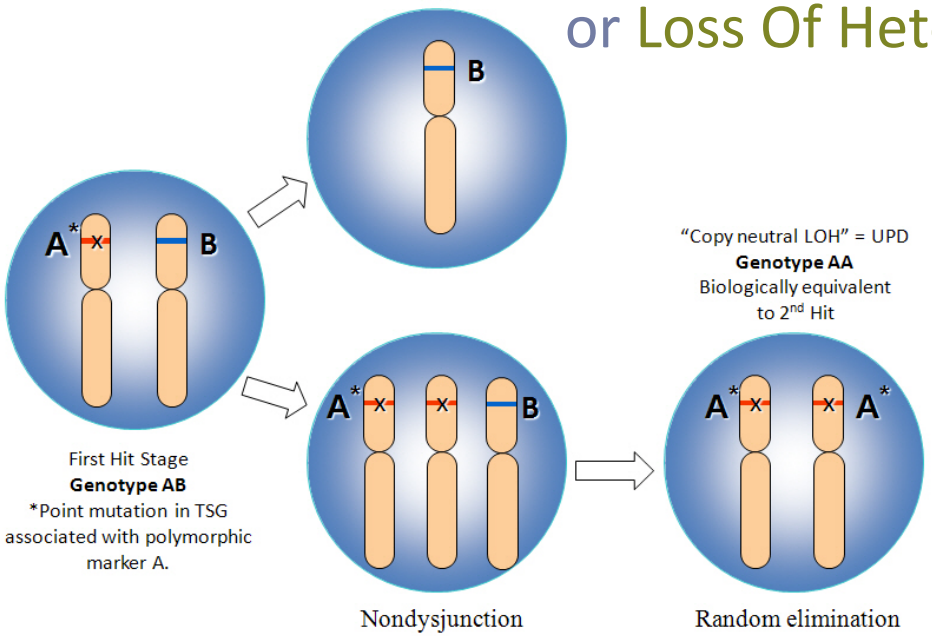


http://compbio.cs.brown.edu/projects/theta/

L. Oesper, A. Mahmoody, and B.J. Raphael. (2013) THetA: Inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. Genome Biology. 14:R80.

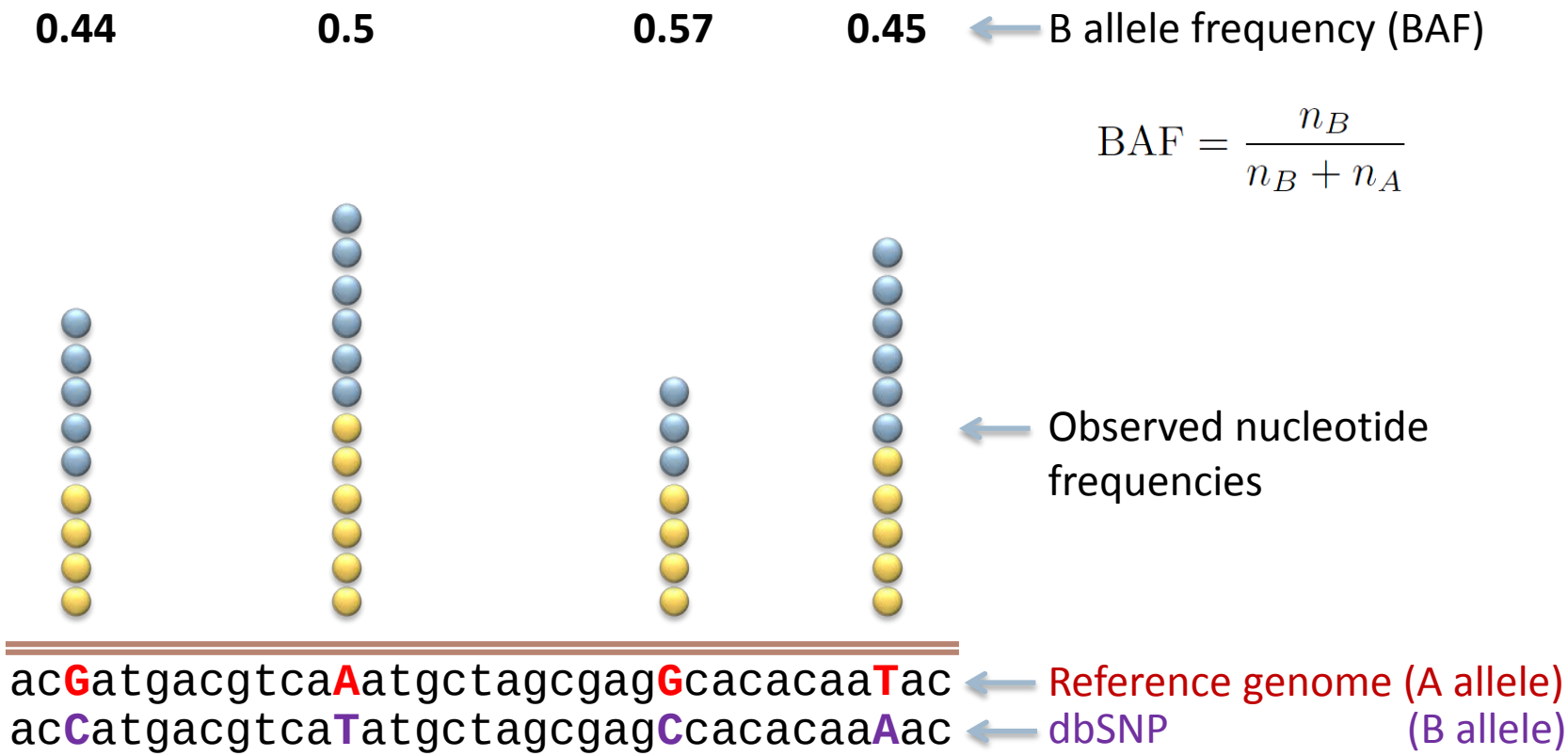
Now we want to detect genotype status (including LOH)

Acquired Uniparental Disomy (UPD)
or Loss Of Heterozygosity (LOH)



We characterize the allelic content *via* the B allele frequency (BAF)

- B allele = alternative variant in dbSNP

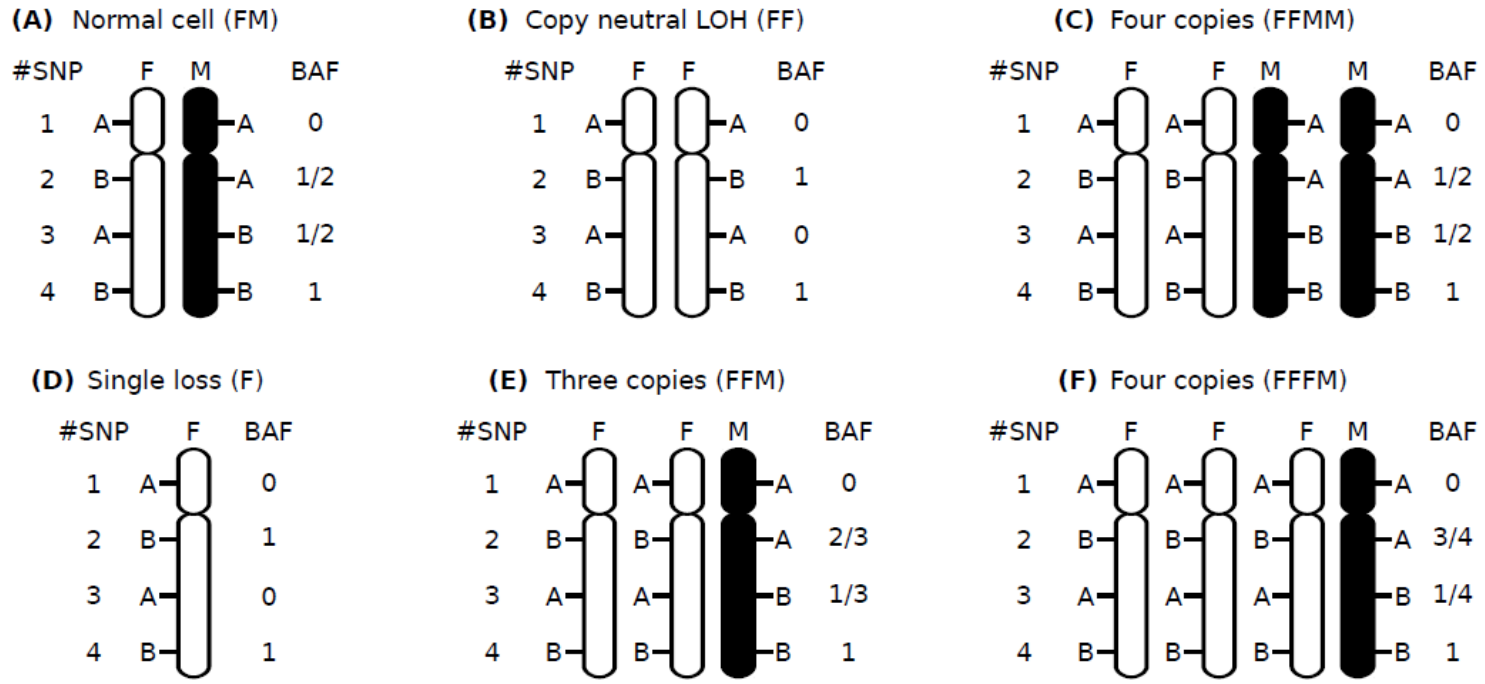


There is a correspondence between copy number and possible BAF

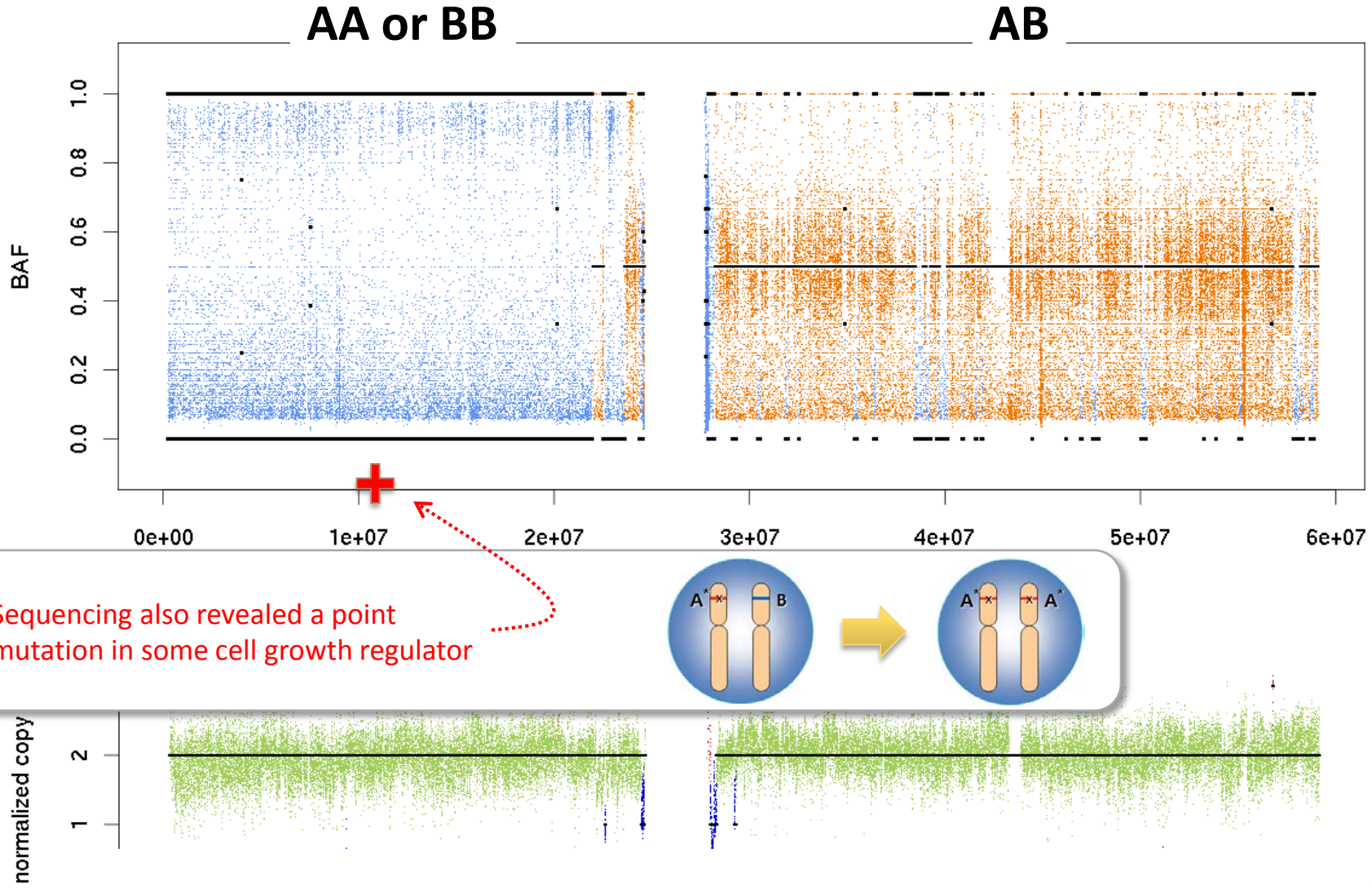
A allele: gtcacccatccctc **c** gtgctggtaatcaga

B allele: gtcacccatccctc **g** gtgctggtaatcaga

$$\mathrm{BAF} = \frac{n_B}{n_B + n_A}$$



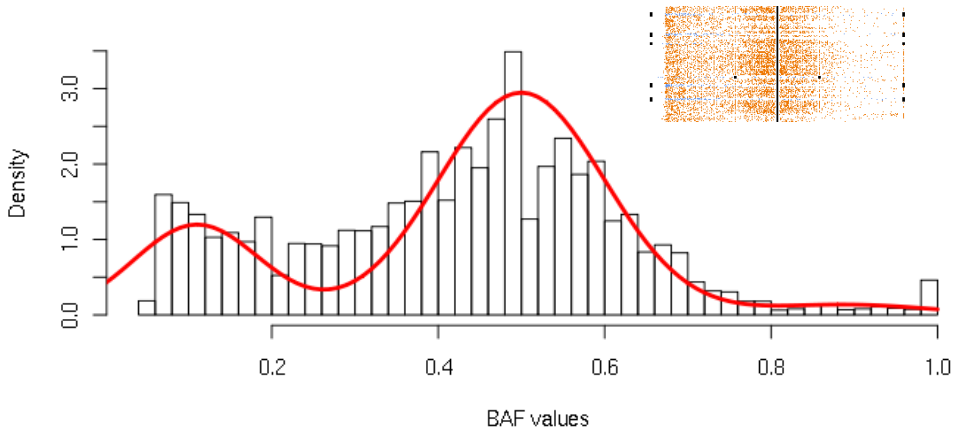
We infer the genotype status of a region from B allele frequency profiles



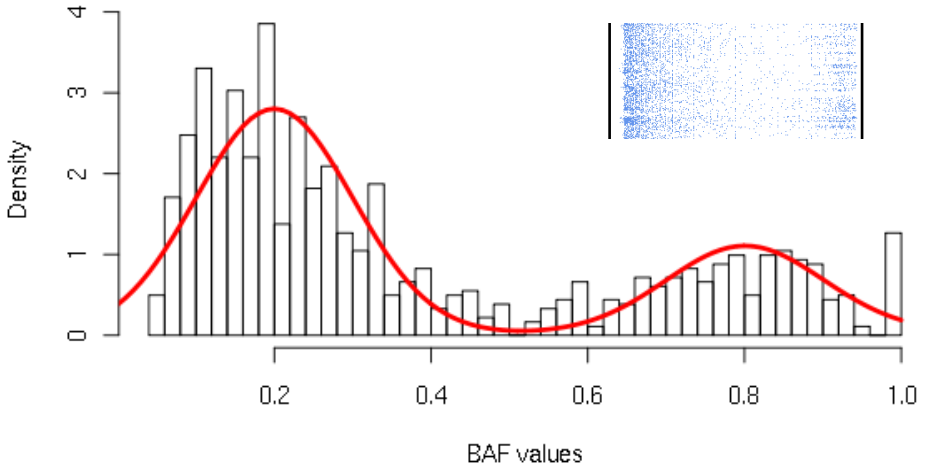
To infer the genotype status of a region from B allele frequency profiles we use Gaussian mixture model (GMM) fit

- We try different fits and choose a fit with the best likelihood

The fit indicates that the genotype = **AB**



The fit indicates that the genotype = **AA/BB** with **40%** contamination by normal ("AB") cells



Fit with 3 modes:

- AA
- AB
- BB

$$\text{mean for } (A)_{N-k}(B)_k = \frac{k(1-c)+c}{N(1-c)+2c}$$

$$\text{mean for } (A)_k(B)_{N-k} = 1 - \frac{k(1-c)+c}{N(1-c)+2c}$$

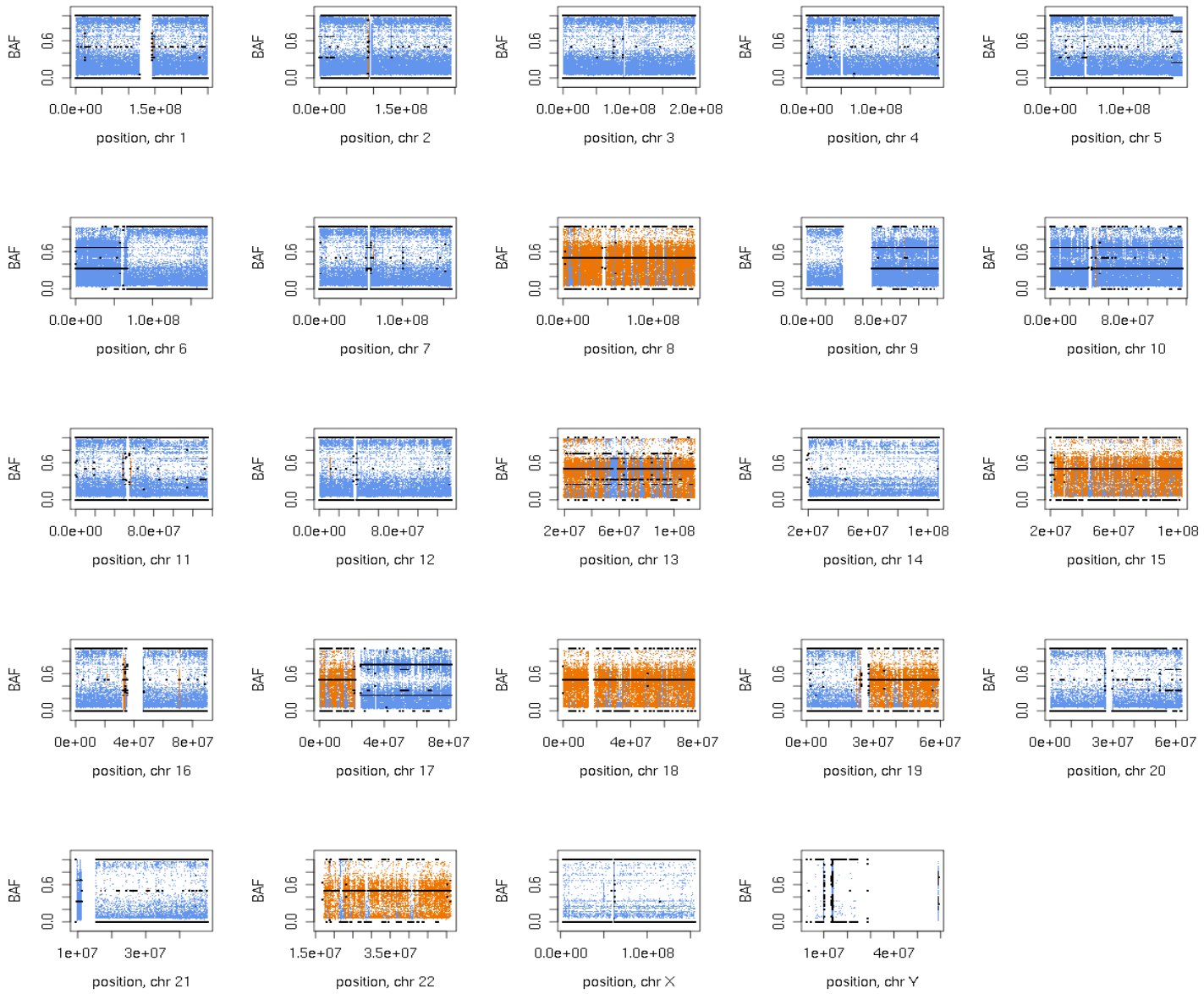
$$\text{means for } (A)_N = \begin{cases} 0.11, & \text{when testing for } k > 0 \\ 0.11 \text{ and } \max\left(0.11, \frac{c}{N(1-c)+2c}\right), & \text{when testing for } k = 0 \end{cases}$$

$$\text{means for } (B)_N = \begin{cases} 0.89, & \text{when testing for } k > 0 \\ 0.89 \text{ and } \min\left(0.89, 1 - \frac{c}{N(1-c)+2c}\right), & \text{when testing for } k = 0 \end{cases}$$

Fit with 4 modes:

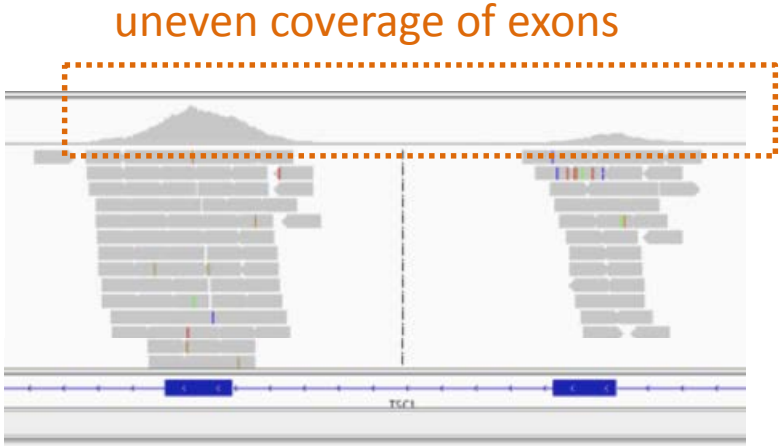
- AA
- BB
- AA*0.6+AB*0.4
- BB*0.6+AB*0.4

Visualization of BAF



Extending Control-FREEC to the exome sequencing data

Exome data:

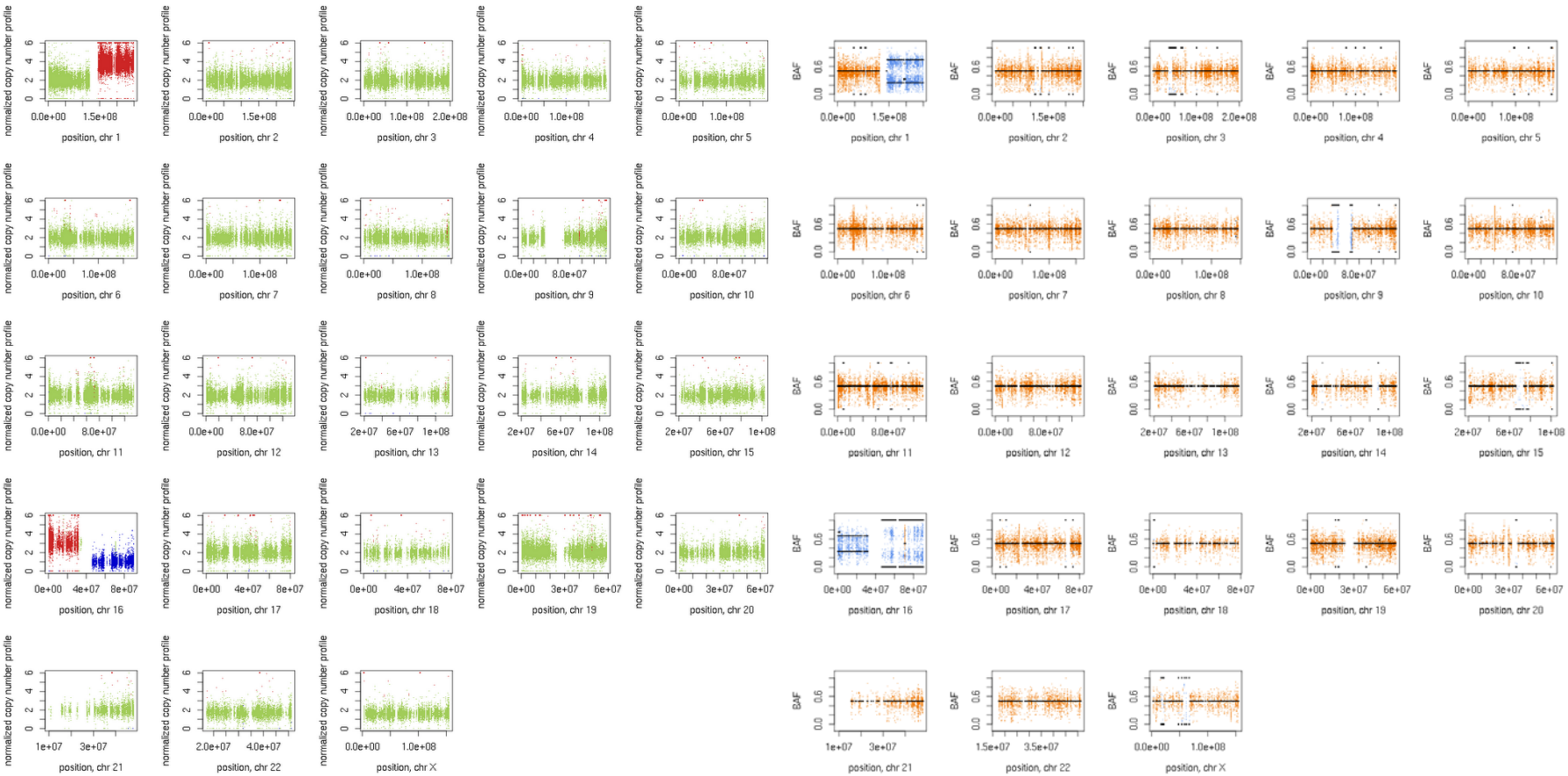


- Capture bias
- GC-content and mappability correction is not enough

Mandatory use of a control sample to normalize read counts

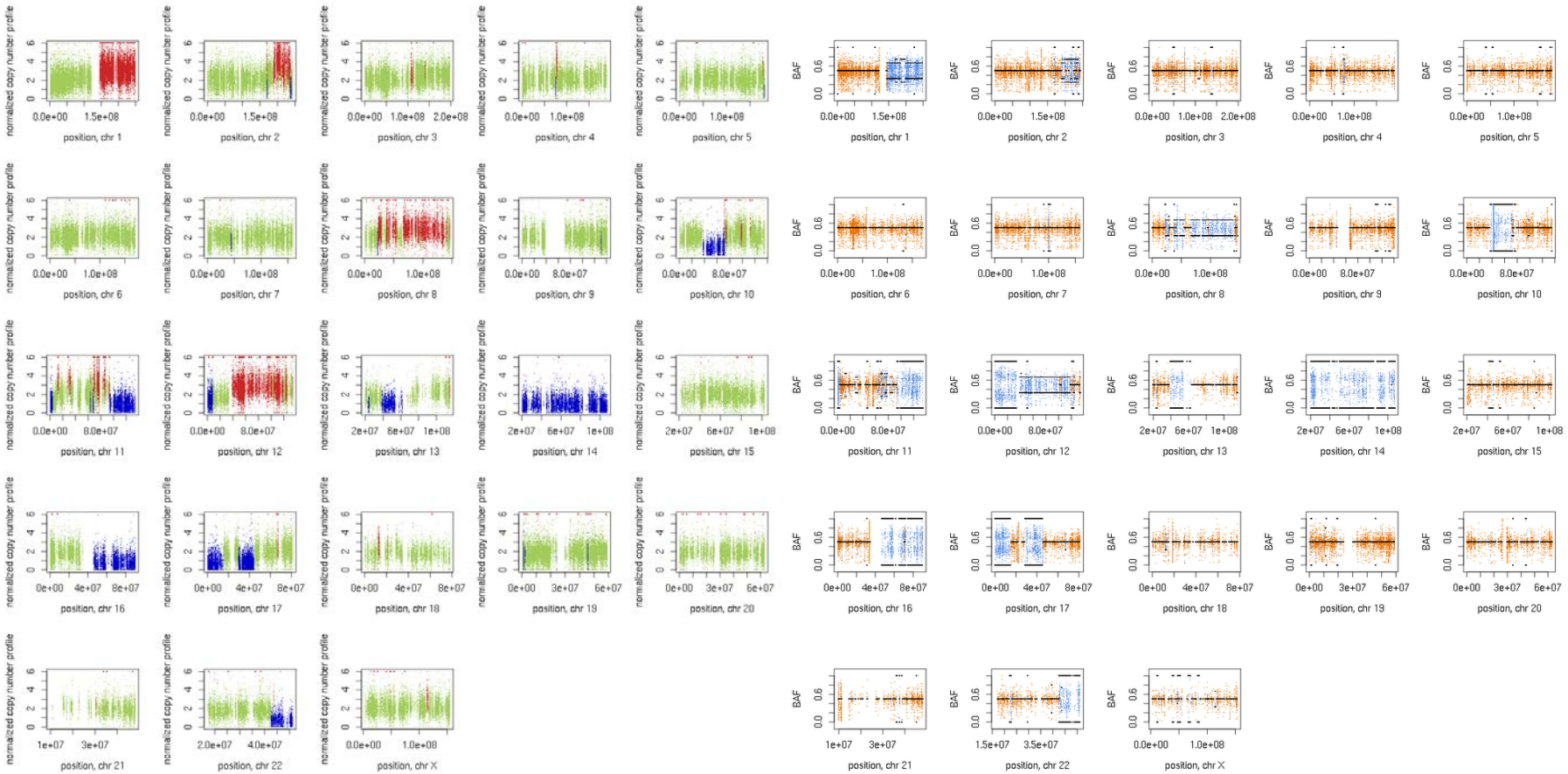
Exome sequencing data may be much more noisy than whole genome sequencing data

Additional bias (capture) => additional noise

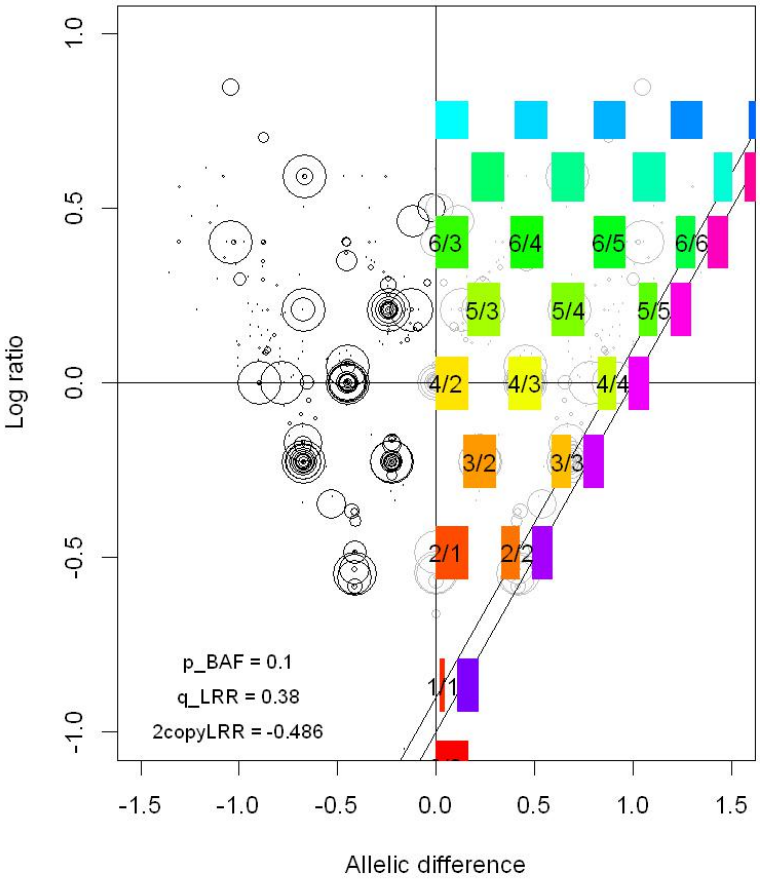


Exome sequencing data may be much more noisy than whole genome sequencing data

Additional bias (capture) => additional noise



Idea: use BAF profiles to infer correct copy numbers



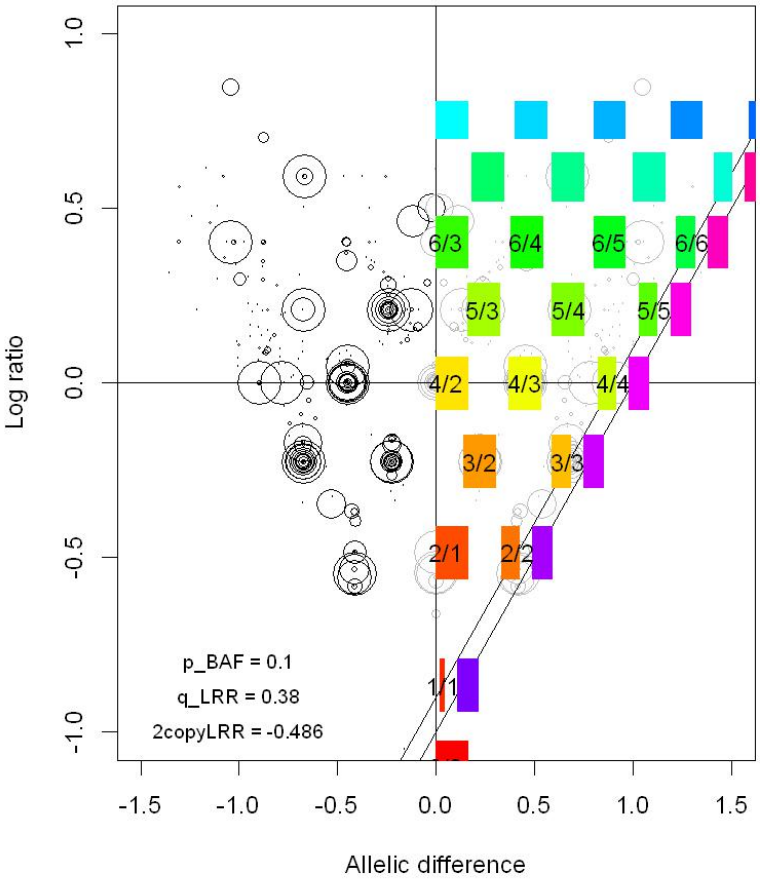
≈

Allelic status	Copy number	BAF*
-	0	NA
A/B	1	0 1
AB	2	0 1 0.5
AA/BB	2	0 1
AAB/ABB	3	0 1 0.33 0.66
AAA/BBB	3	0 1
AABB	4	0 1 0.5
AAAB/ABBB	4	0 1 0.25 0.75
AAAA/BBBB	4	0 1
…	…	…

*No contamination by normal cells

GAP (Popova et al., Genome Biology 2009)

Idea: use BAF profiles to infer correct copy numbers



≈

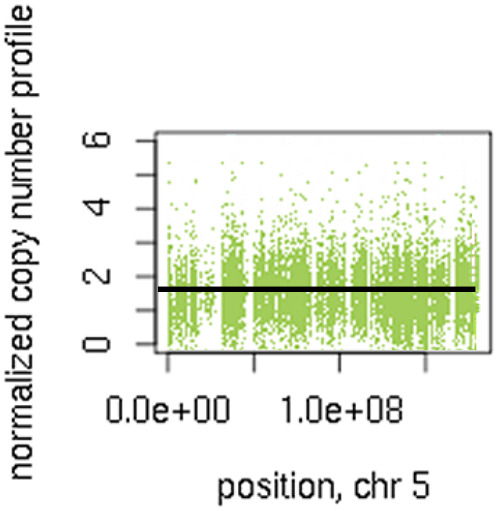
Allelic status	Copy number	BAF**
-	0	NA
A/B	1	0 1 p/(1+p) 1/(1+p)
AB	2	0 1 0.5
AA/BB	2	0 1 p/2 1-p/2
AAB/ABB	3	0 1 1/(1-p) 1-1/(1-p)
AAA/BBB	3	0 1 p/(3-p) 1-p(3-p)
AABB	4	0 1 0.5
AAAB/ABBB	4	0 1 1/(4-2p) 1-1/(4-2p)
AAAA/BBBB	4	0 1 p/(4-2p) 1-p/(4-2p)
…	…	…

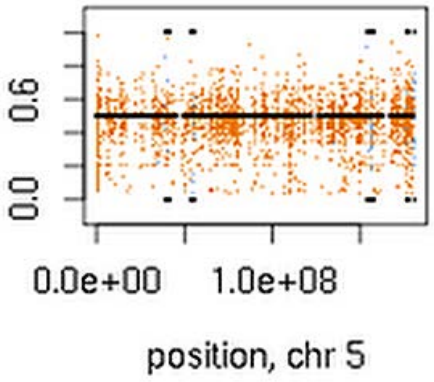
**Contamination p by normal cells

GAP (Popova et al., Genome Biology 2009)

Realization: in case of doubt, get max(logLikelihood) for observed BAF

1 or 2 copies?



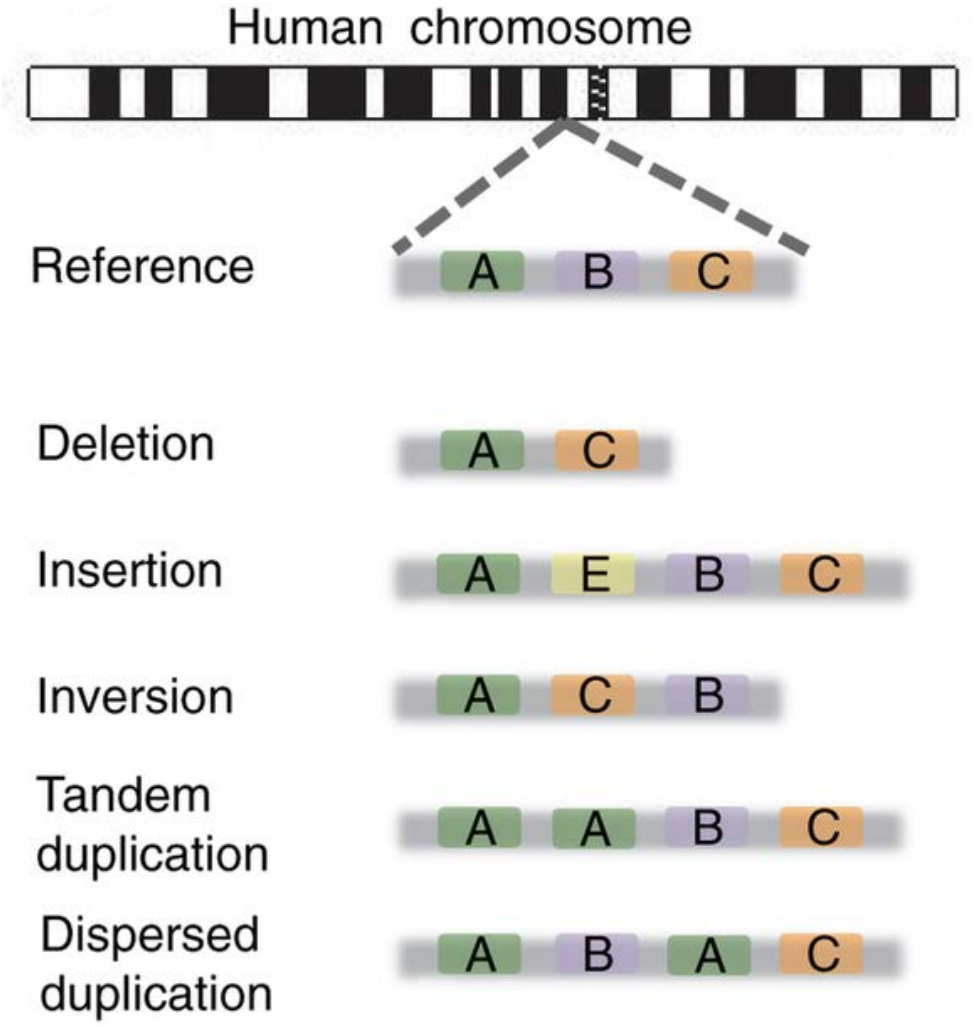


2 copies!

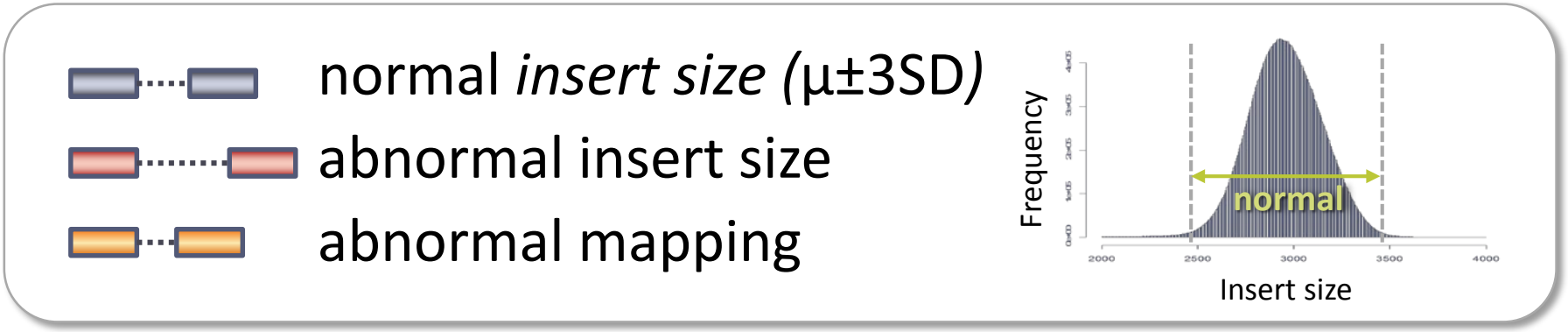
Allelic ratio about ½:

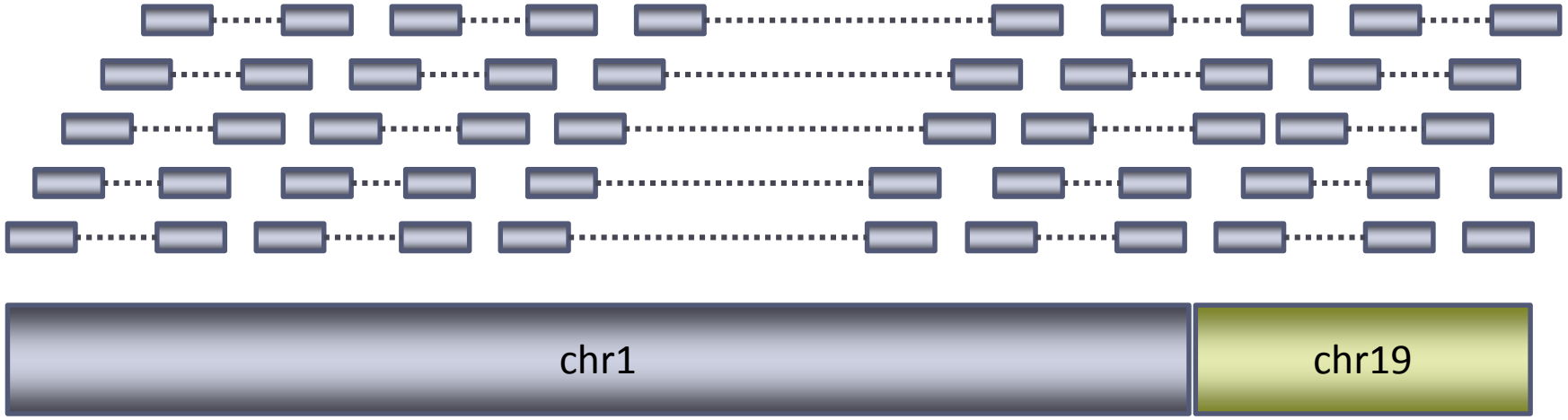
A/B: logLikelihood	-2340
AA/BB: logLikelihood	-3270
AB: logLikeluhood	120

Now we want to detect structural variants – translocations, insertions, inversions and so on

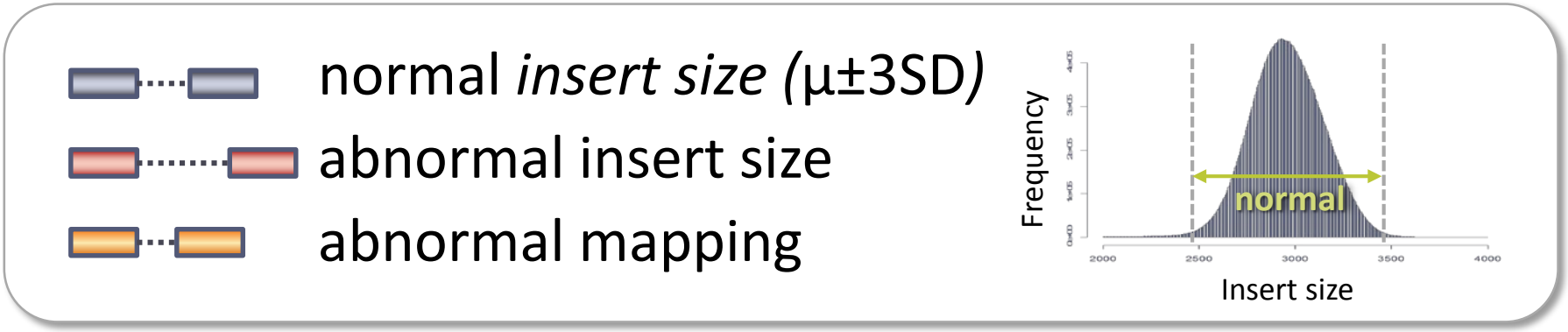


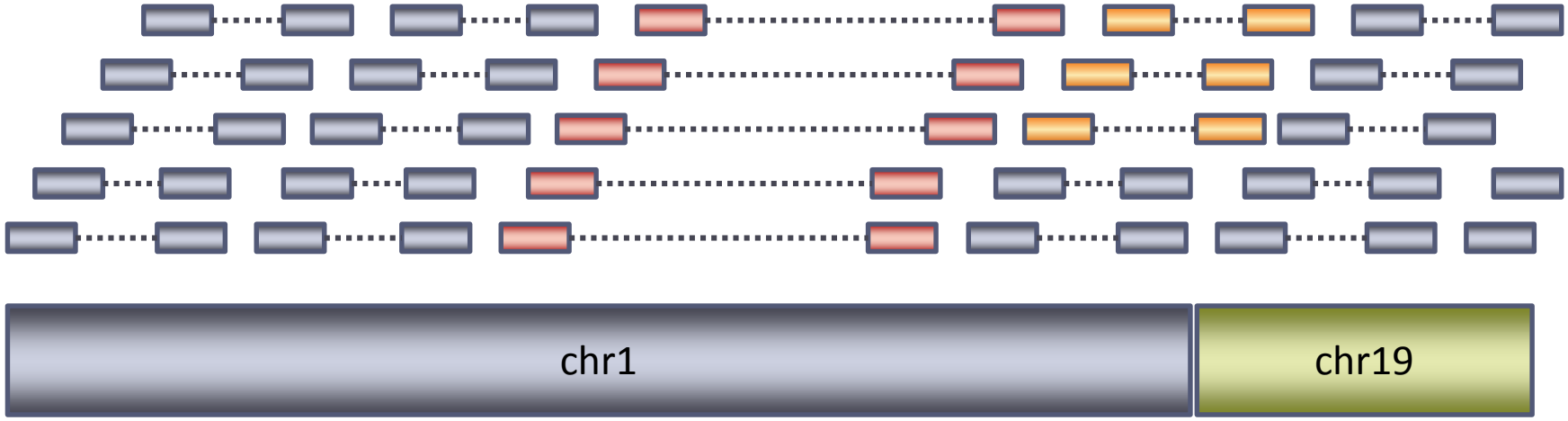
To detect structural variants we should separate “normally” mapped read pairs from “abnormal” pairs



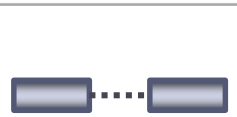


To detect structural variants we should separate “normally” mapped read pairs from “abnormal” pairs





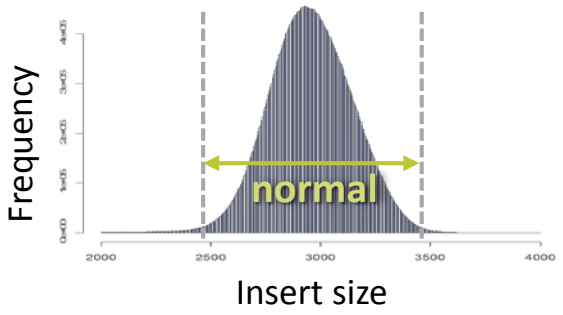
Ne need to cluster “abnormal” pairs and analyze these clusters

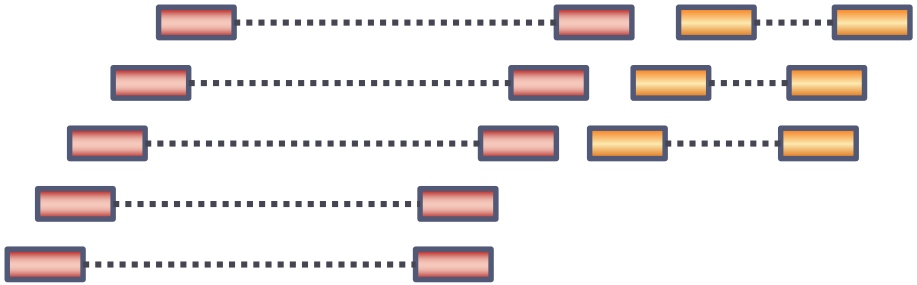


normal *insert size (*µ±3SD*)*

abnormal insert size

abnormal mapping



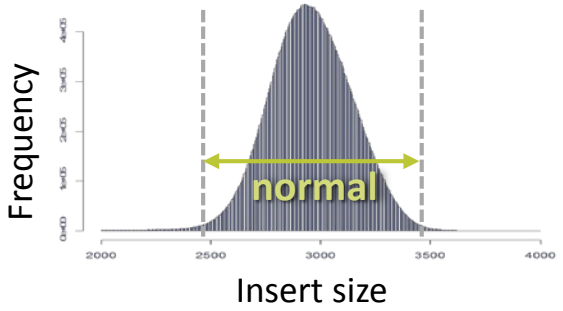


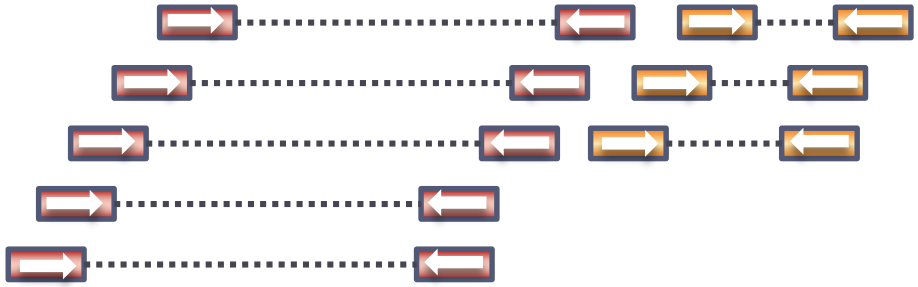
Ne need to cluster “abnormal” pairs and analyze these clusters

normal *insert size (*µ±3SD*)*

abnormal insert size

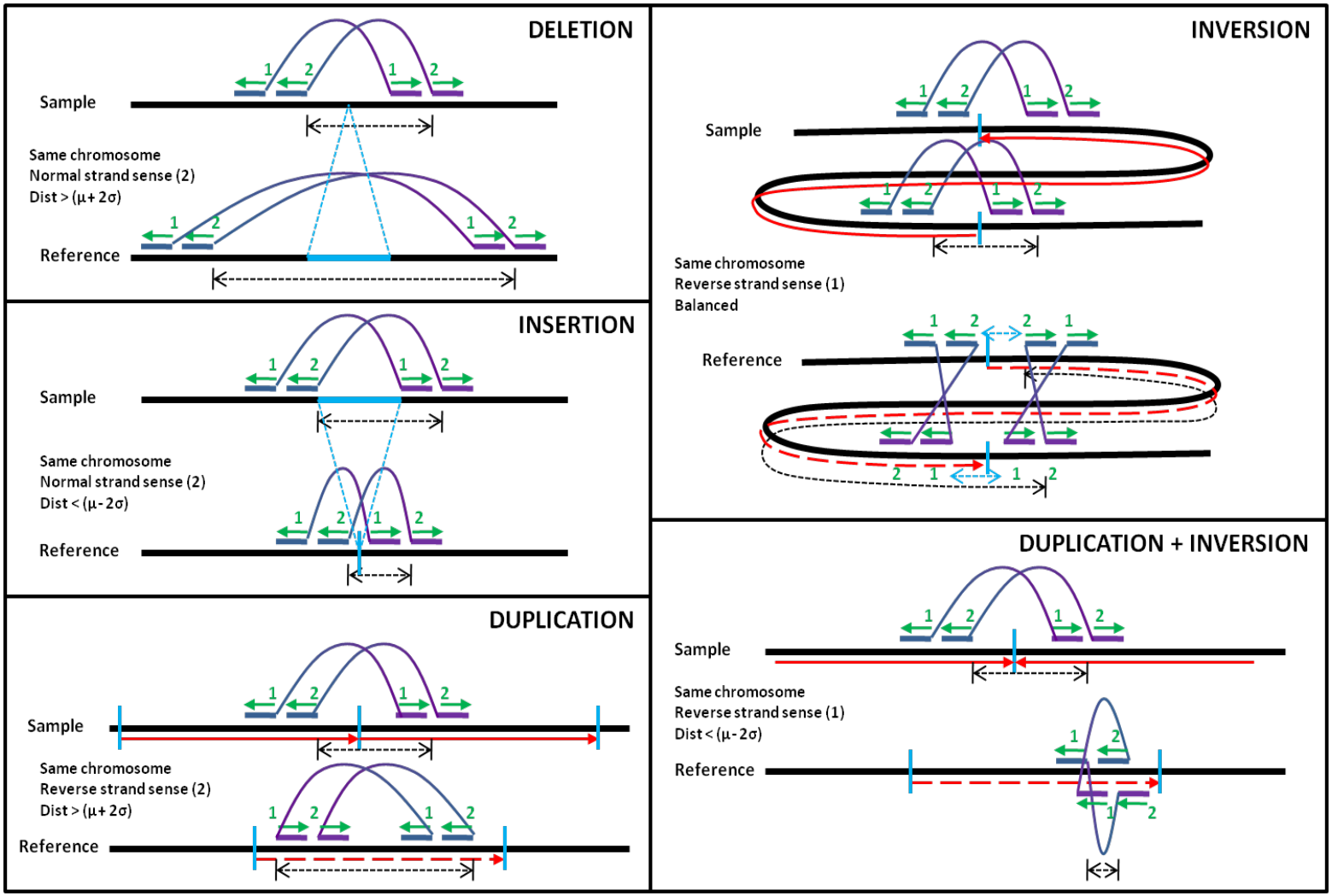
abnormal mapping



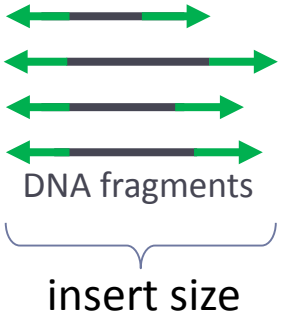


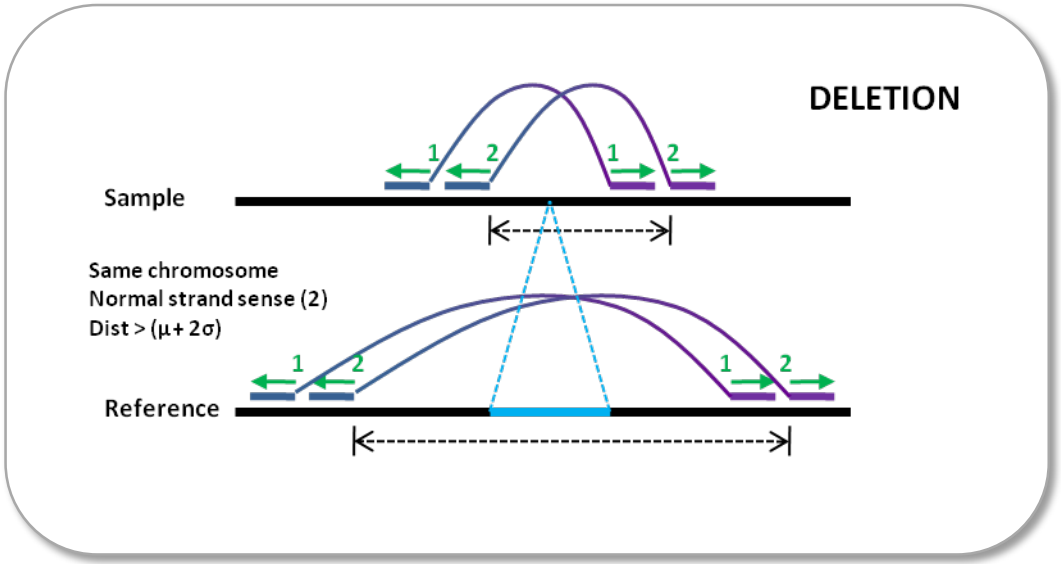
Signature of abnormally mapped reads let us know type of an SV

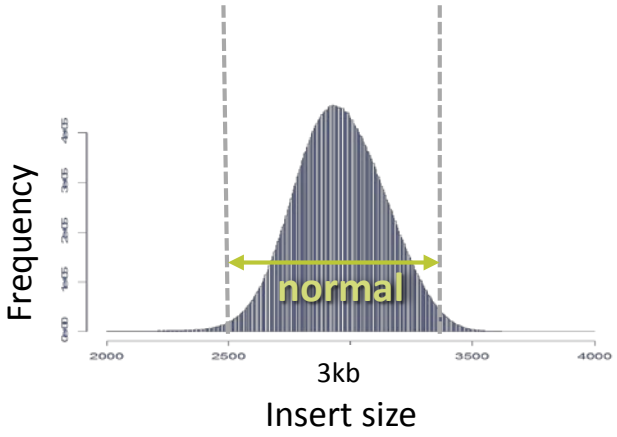
Intra-chromosomal SVs

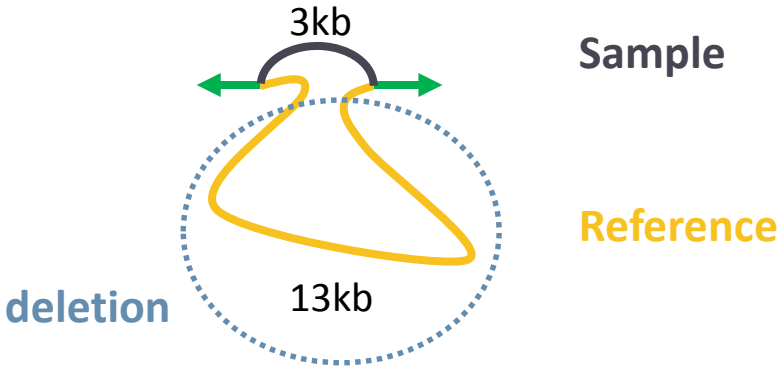


Normal orientation of reads in a cluster but a larger insert size mean "deletion"

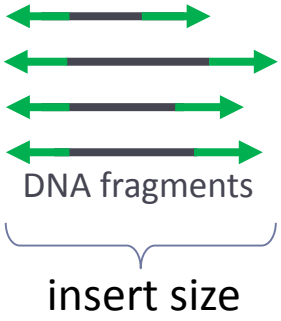


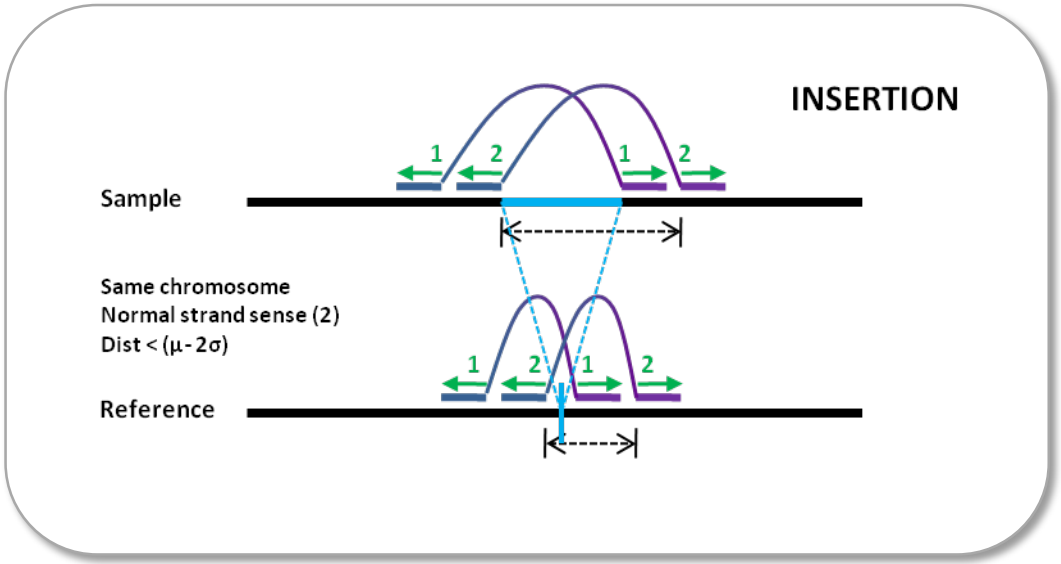


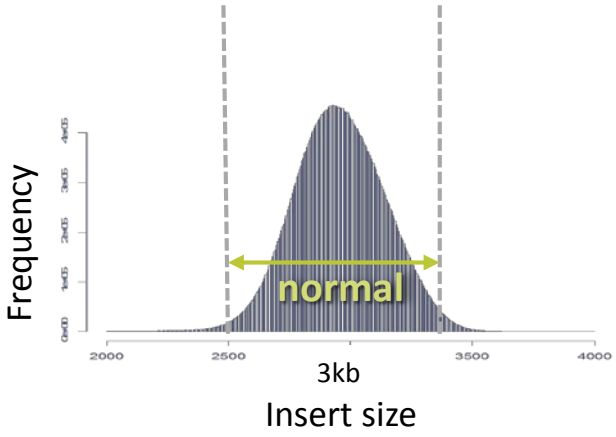


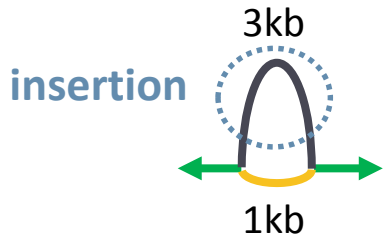


Normal orientation of reads in a cluster but a smaller insert size mean "insertion"







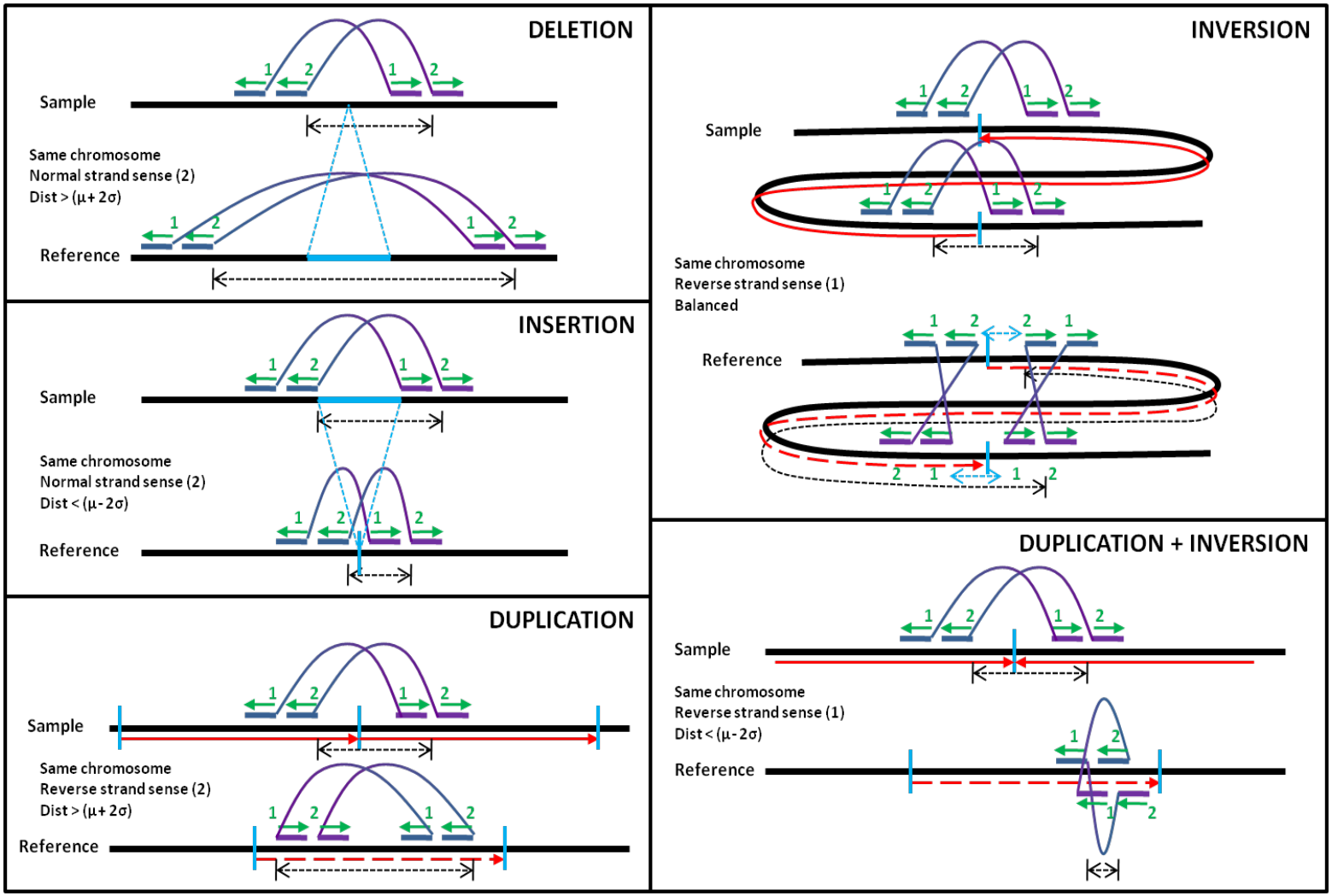


Sample

Reference

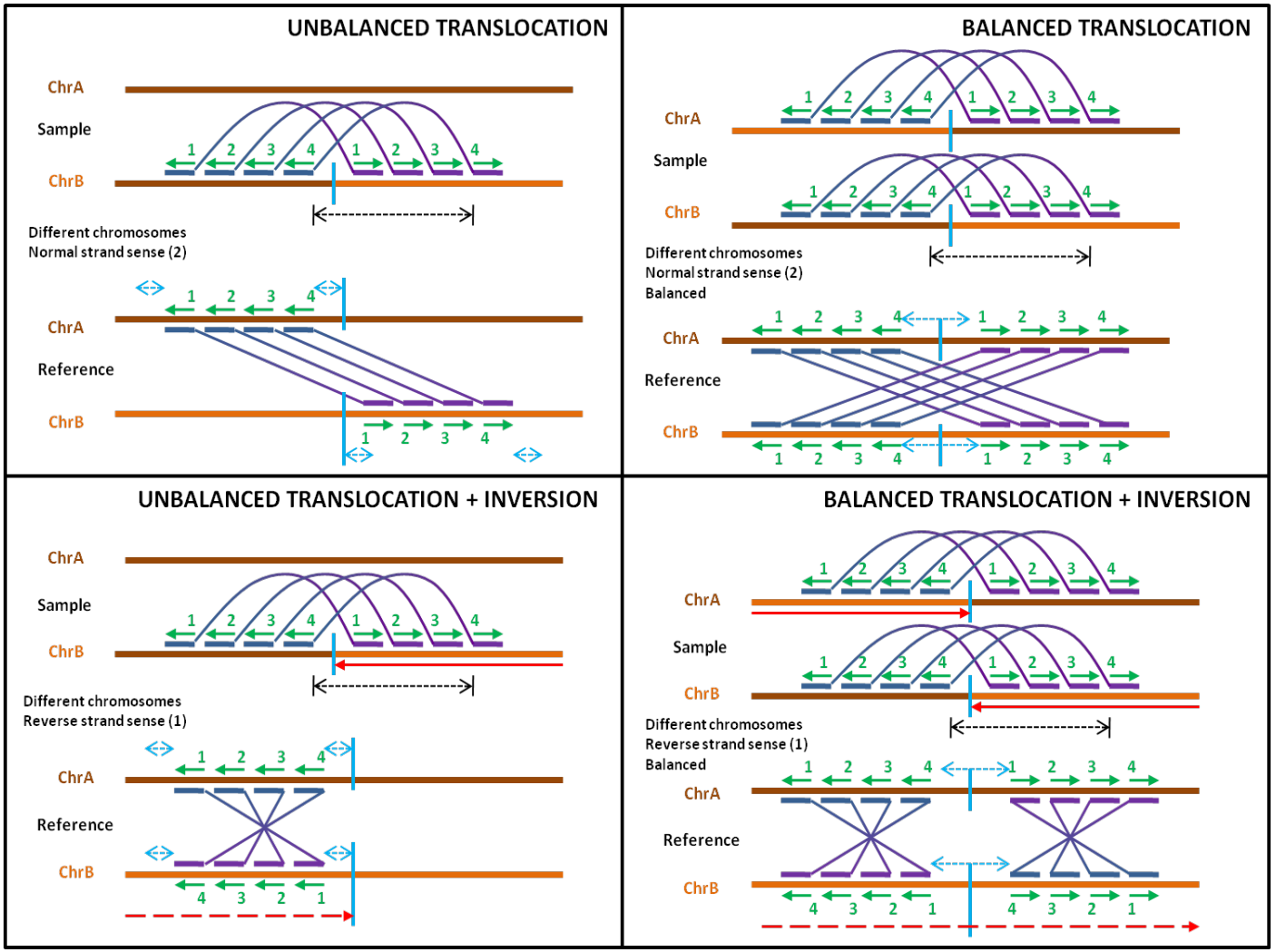
Signature of abnormally mapped reads let us know type of an SV

Intra-chromosomal SVs



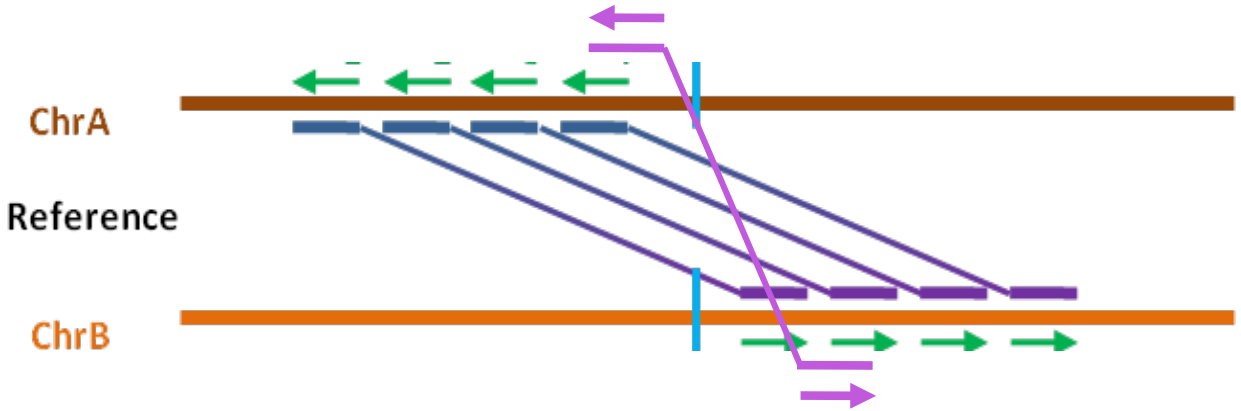
SVDetect identify structural variants (SVs) using clusters of abnormally mapped reads

Inter-chromosomal SVs



Some advantages of our method SVDetect

- Predict >10 types of structural variants (SVs) using the pair-end mapping signature
- Annotate pairs **inconsistent** with the main signature of the predicted SV

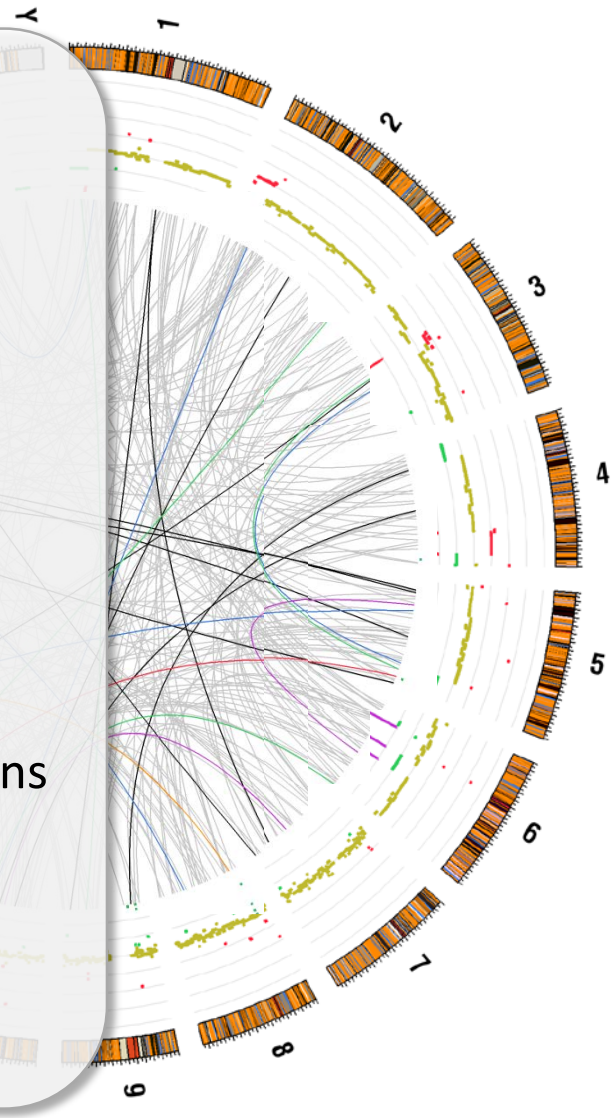


- Annotate SVs predicted in different samples

We need to apply filters to remove noise from predictions of SVs

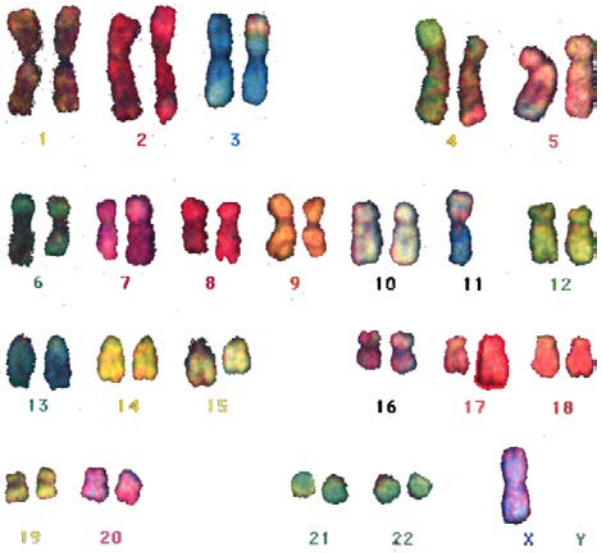
Filters:

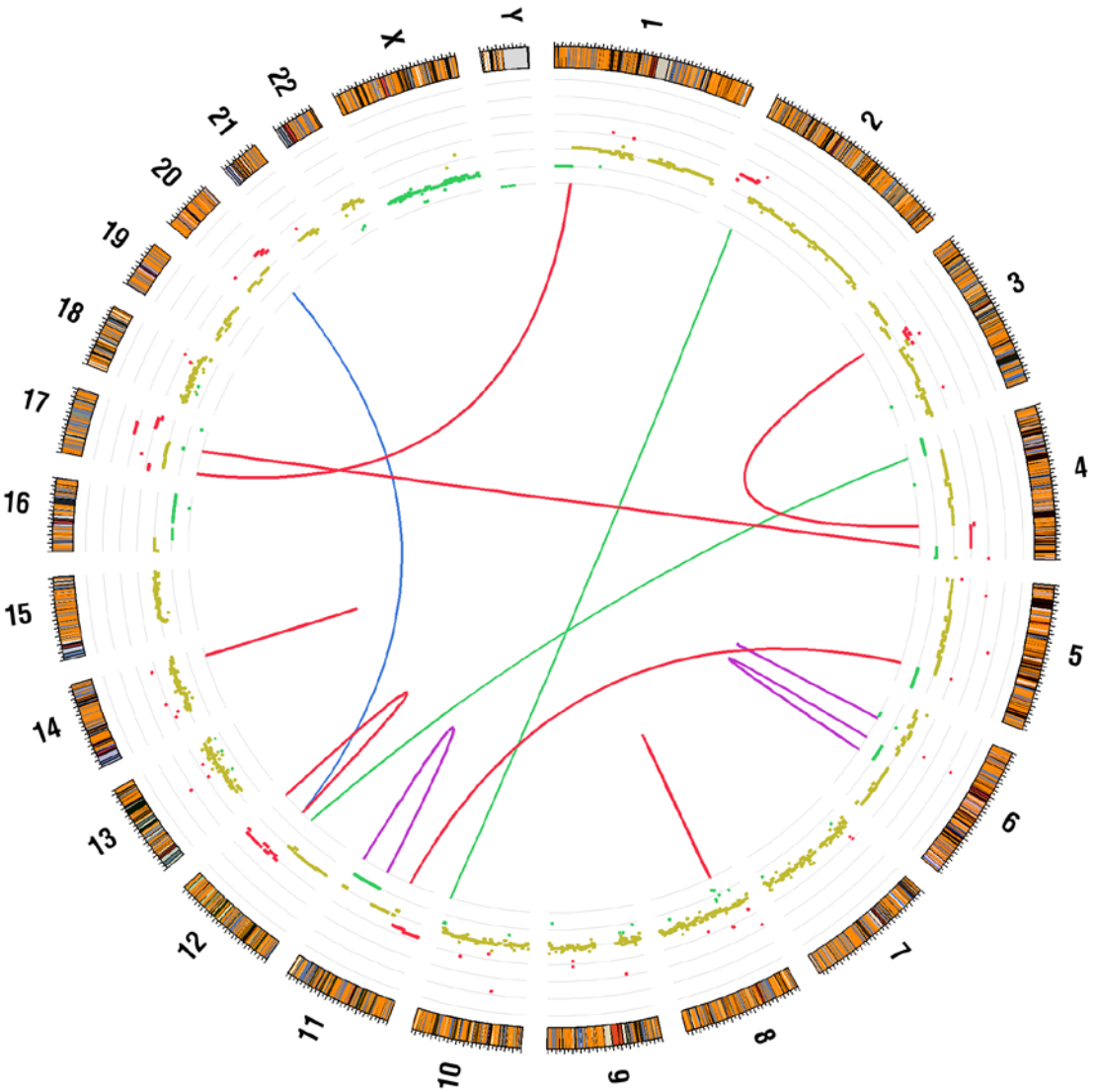
- Number of pairs/link
- Proportion of consistent pairs
- Links close to poly-N non-assembled telomeric/centromeric regions
- Links falling in satellite sequences
- Links falling in segmental duplication regions
- Links with too large or too short clusters
- Links found in the normal sample



Circos © representation of predicted SVs

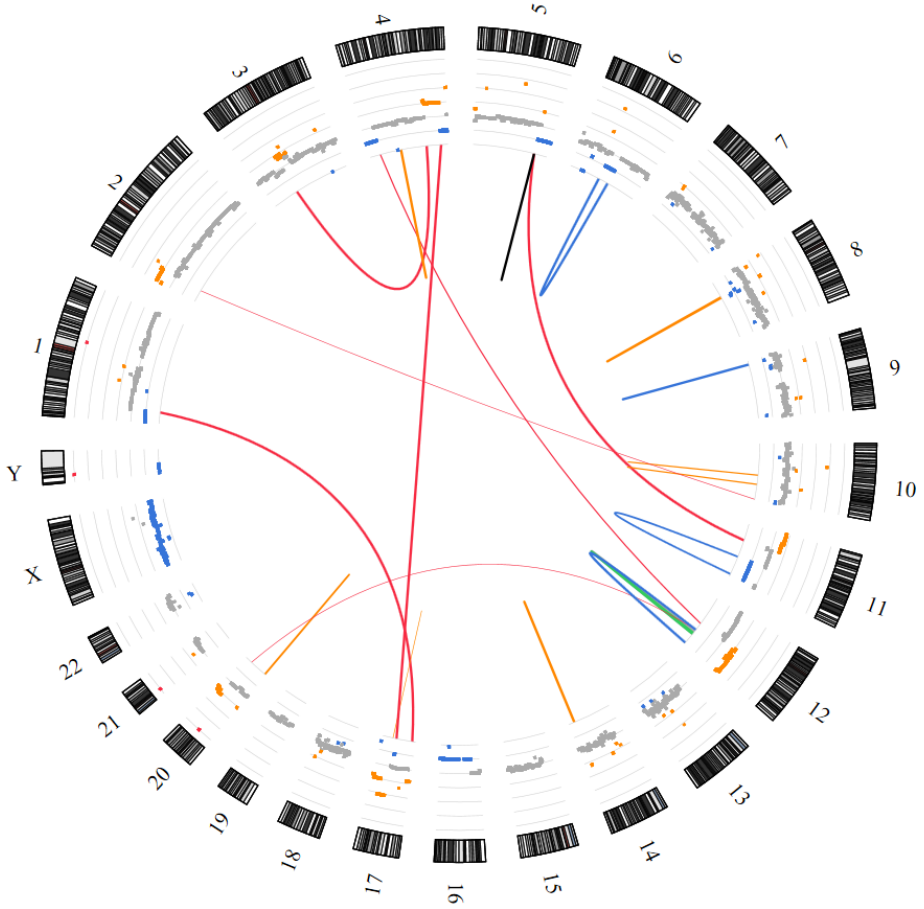
We can complement the information about copy number changes with the predictions of structural variants



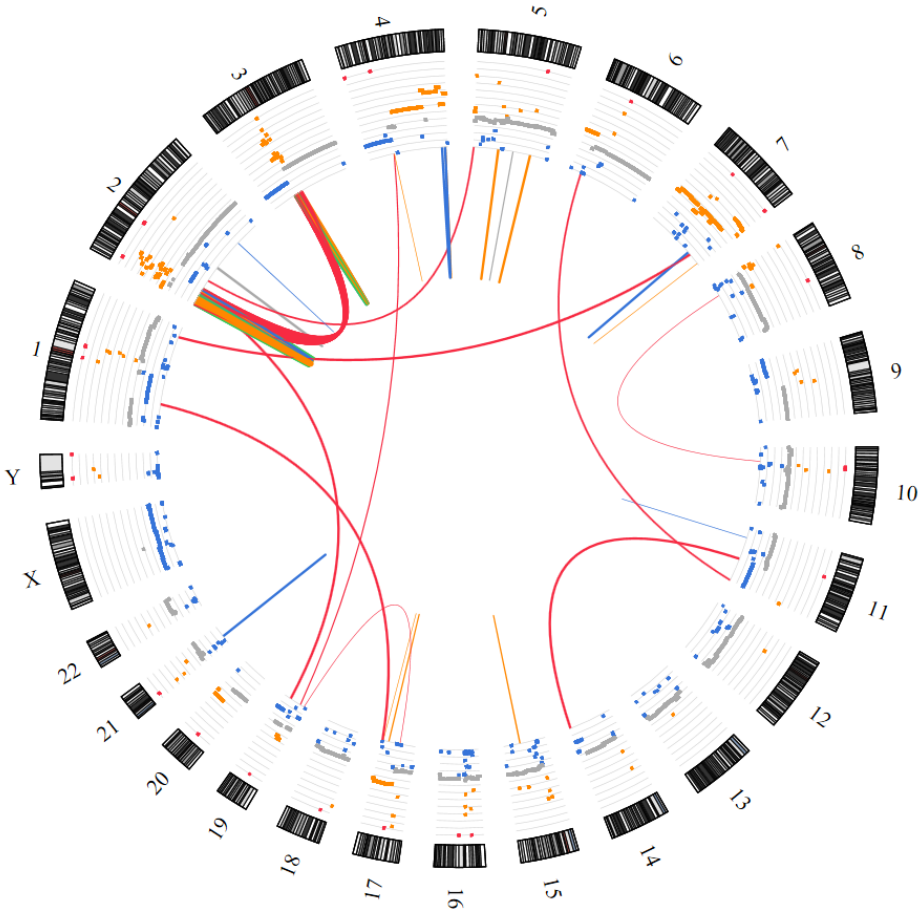


Circos representation of SVs predicted by SVDetect confirmed by the CNAs identified by FREEC

Using FREEC and SVDetect we studied several neuroblastoma samples

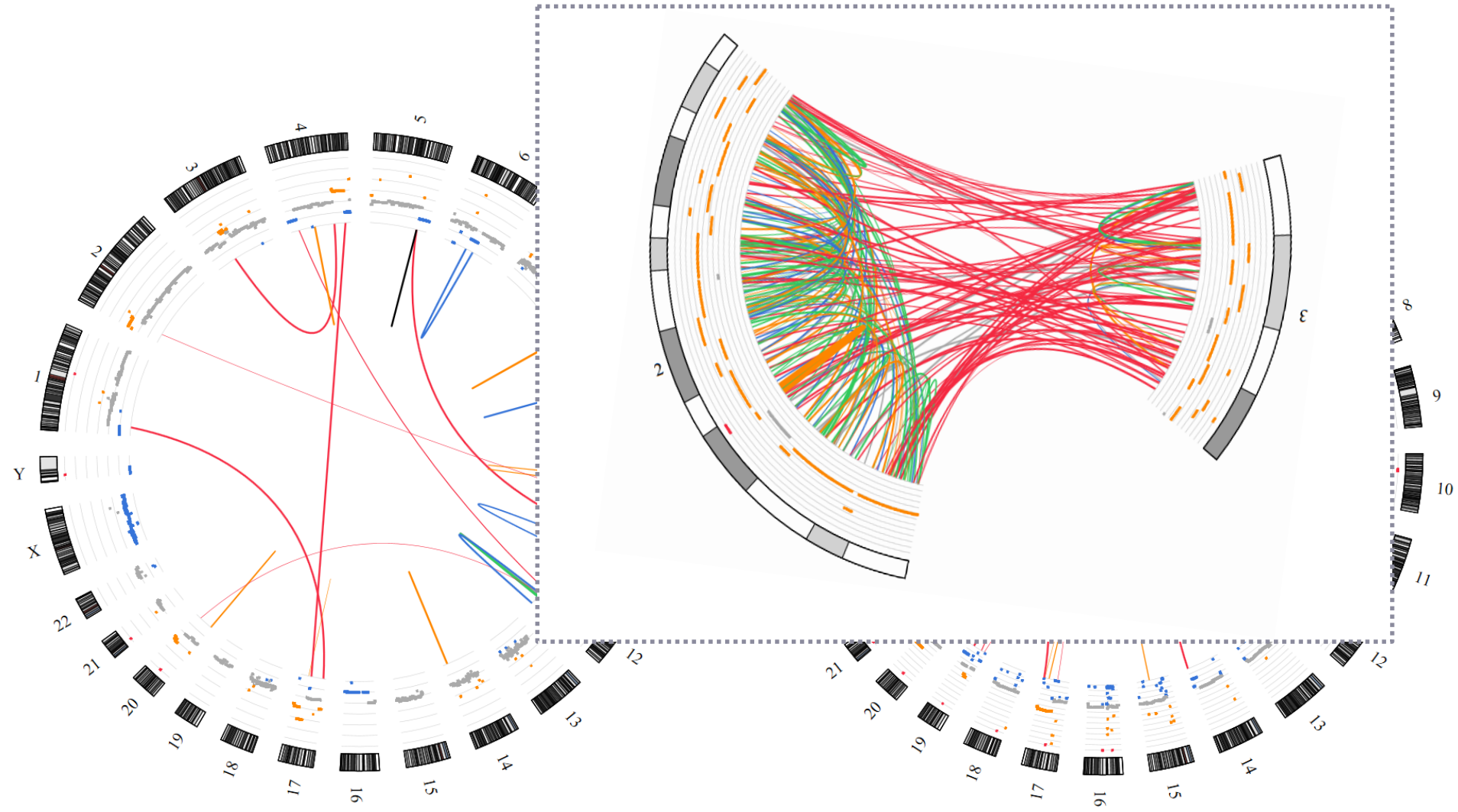


CLB-GA

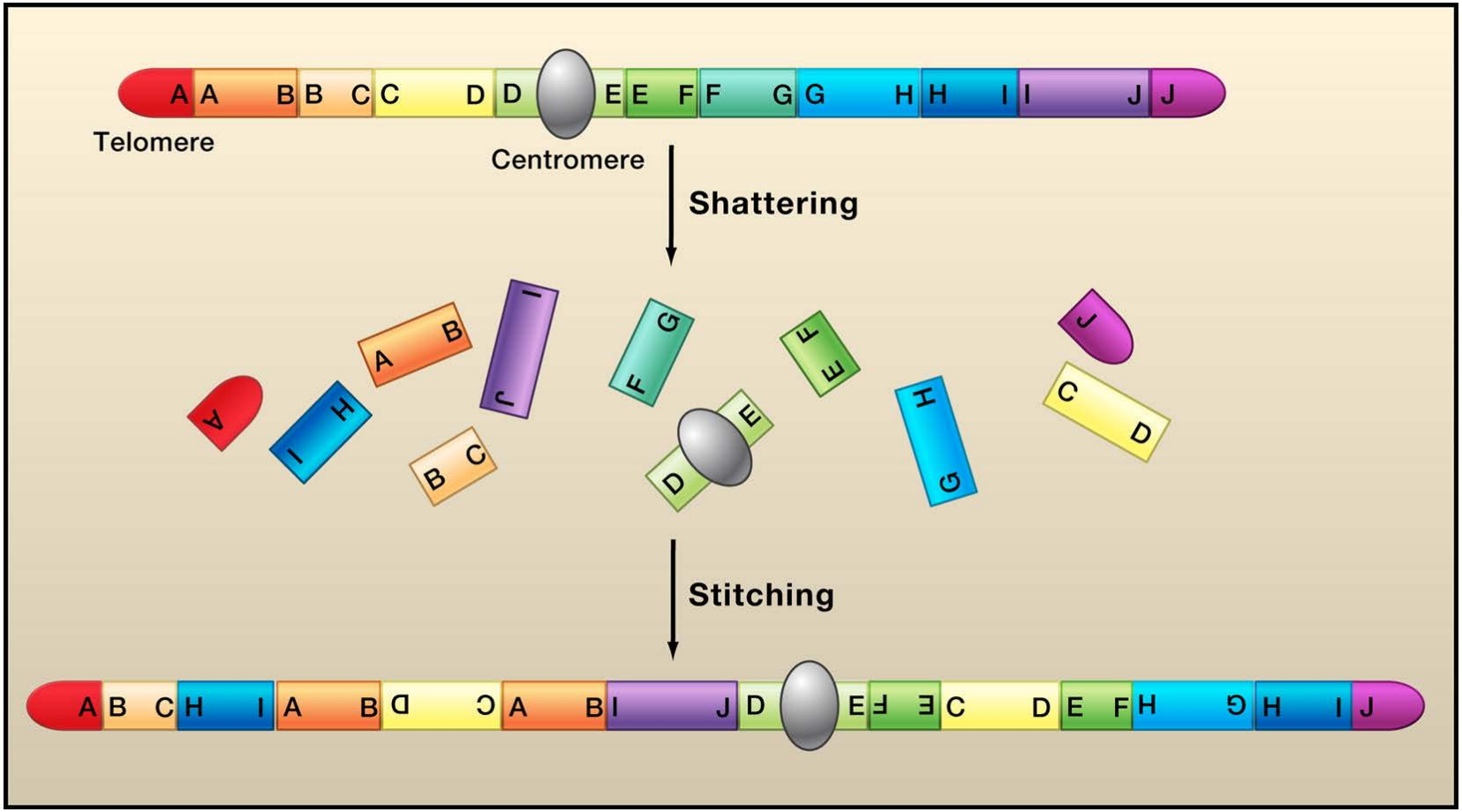


CLB-RE

We observed chromothripsis in neuroblastoma

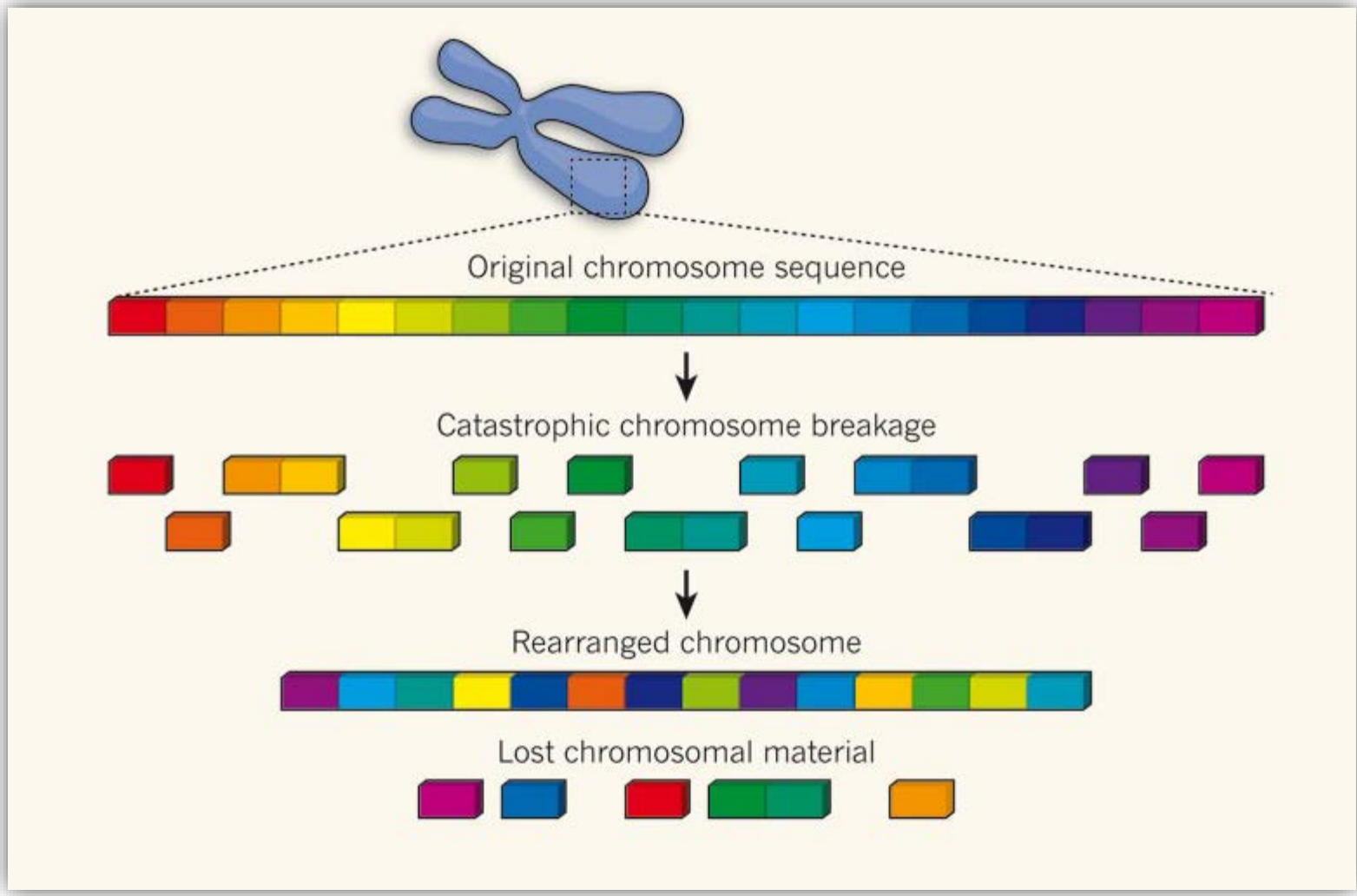


Chromothripsis – a cataclysmic event in which a single chromosome is fragmented and then reassembled.



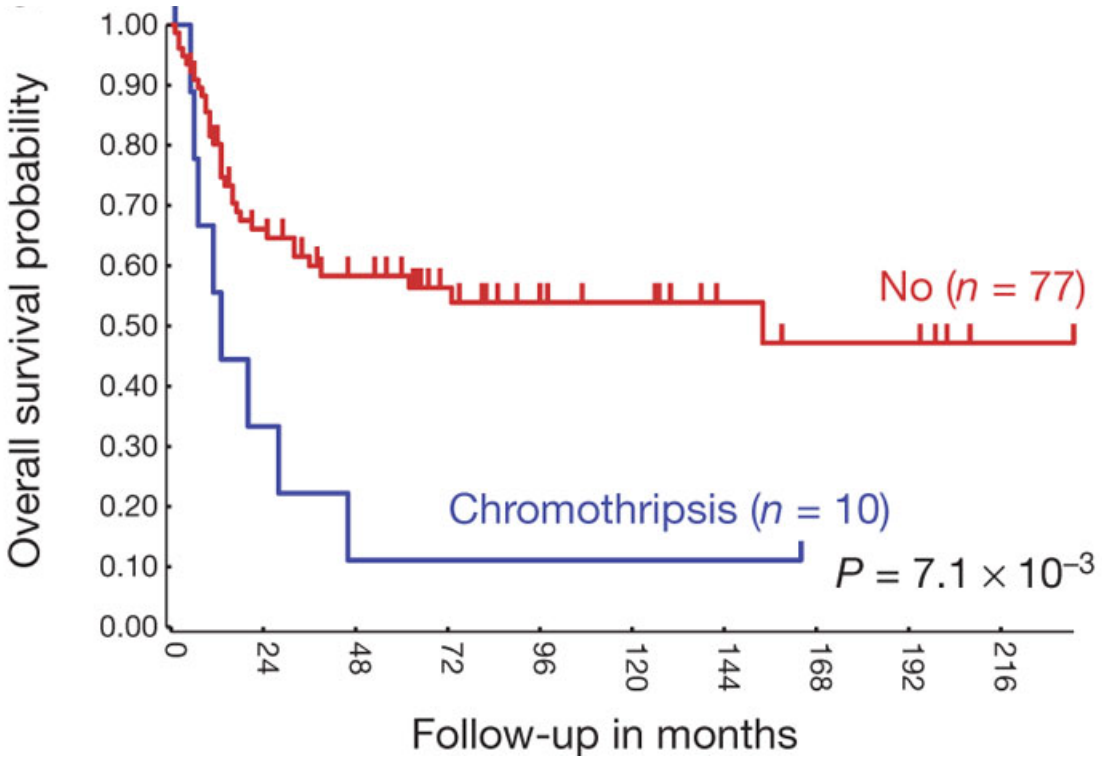
from Meyerson & Pellman, Cell, 2011

Chromothripsis usually results in a loss of chromosomal material



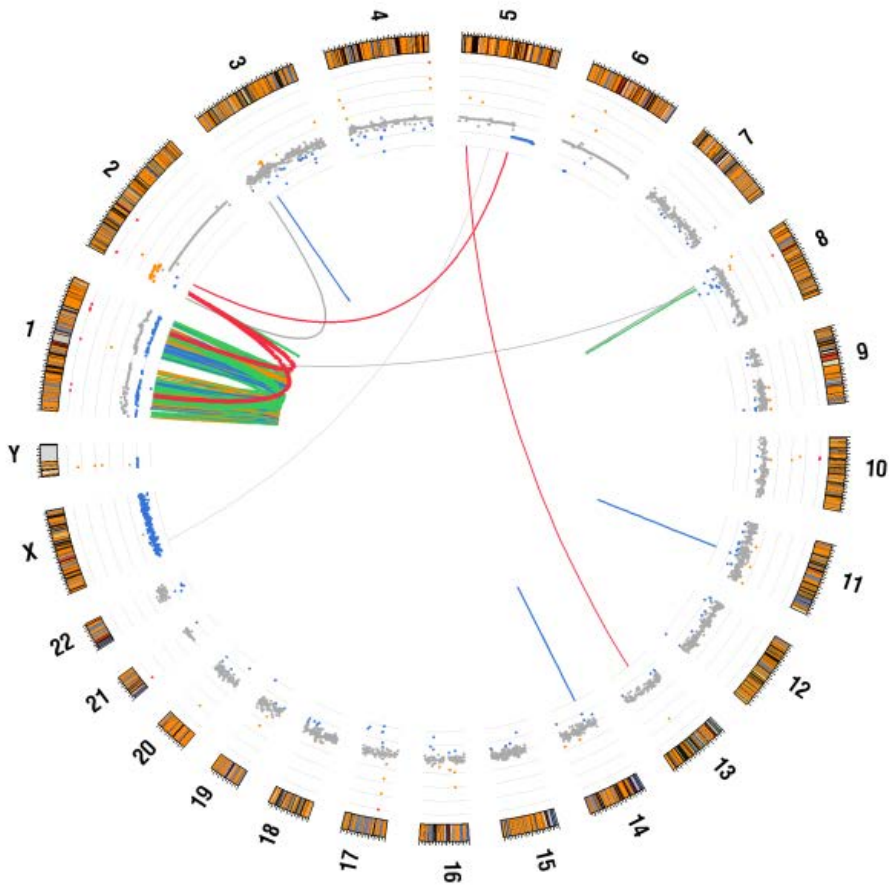
from Tubio & Estivill, Nature, 2011

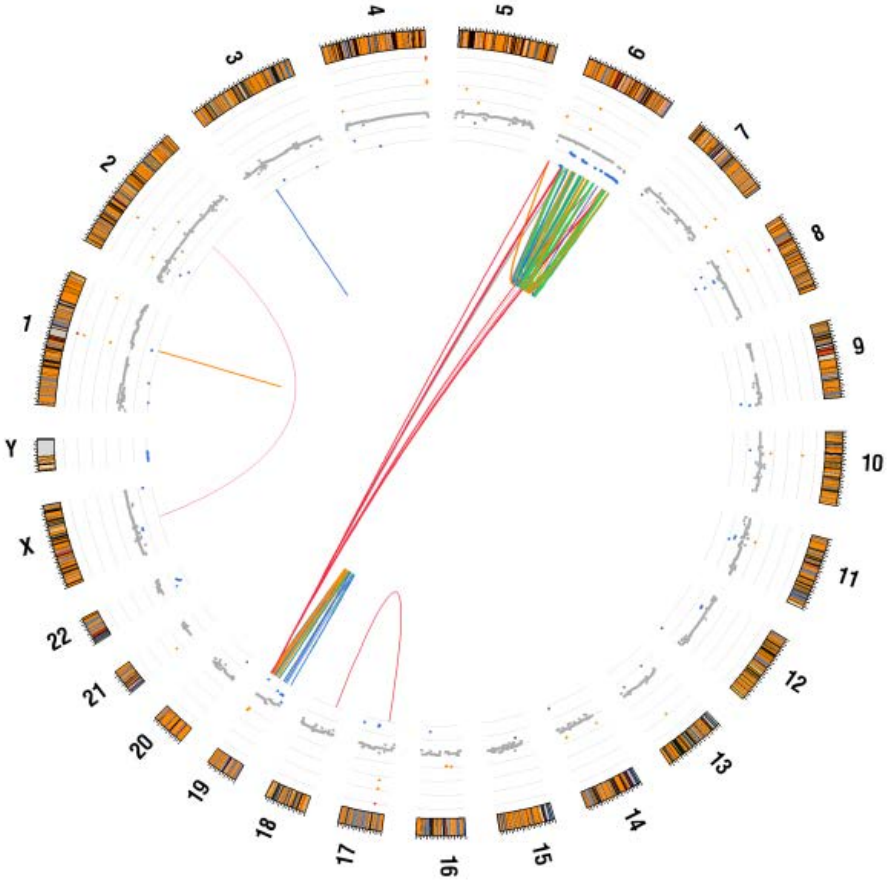
The neuroblastoma with chromothripsis were associated with a poor prognosis (log-rank test $P = 7.1 \times 10^{-3}$)



from JJ Molenaar et al. Nature (2012)

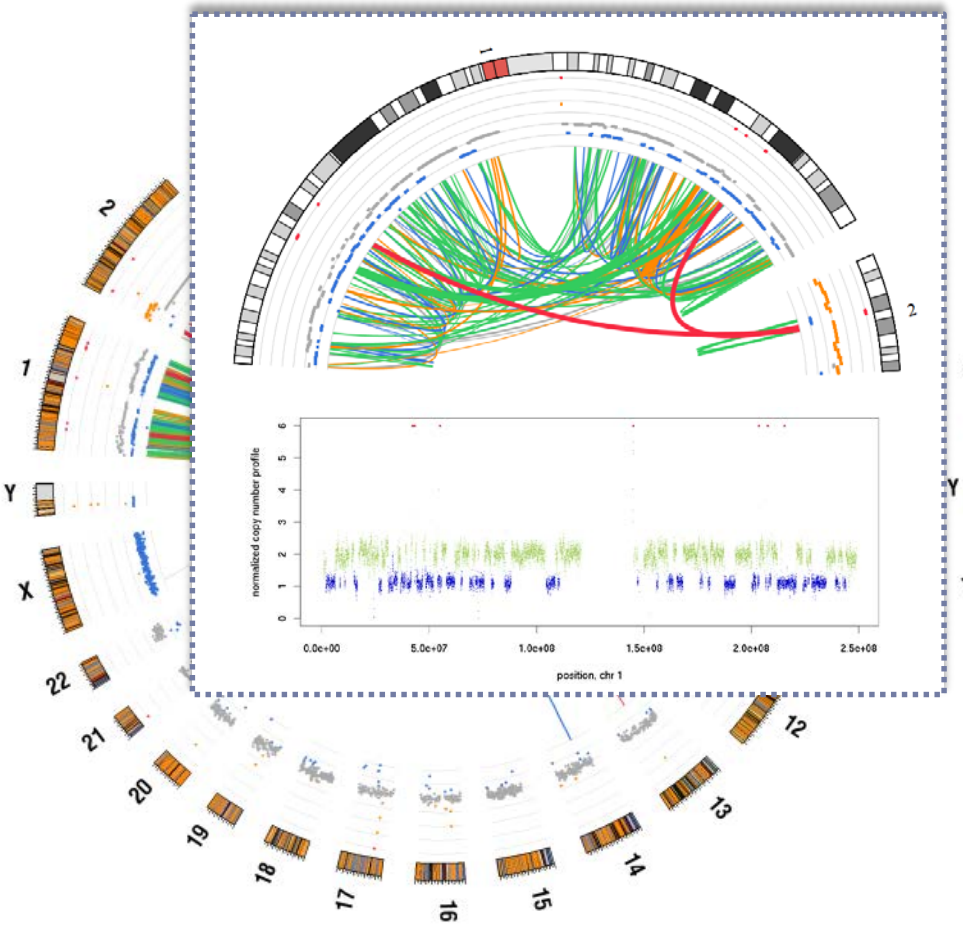
Examples of “classical” chromothripsis cases in neuroblastoma

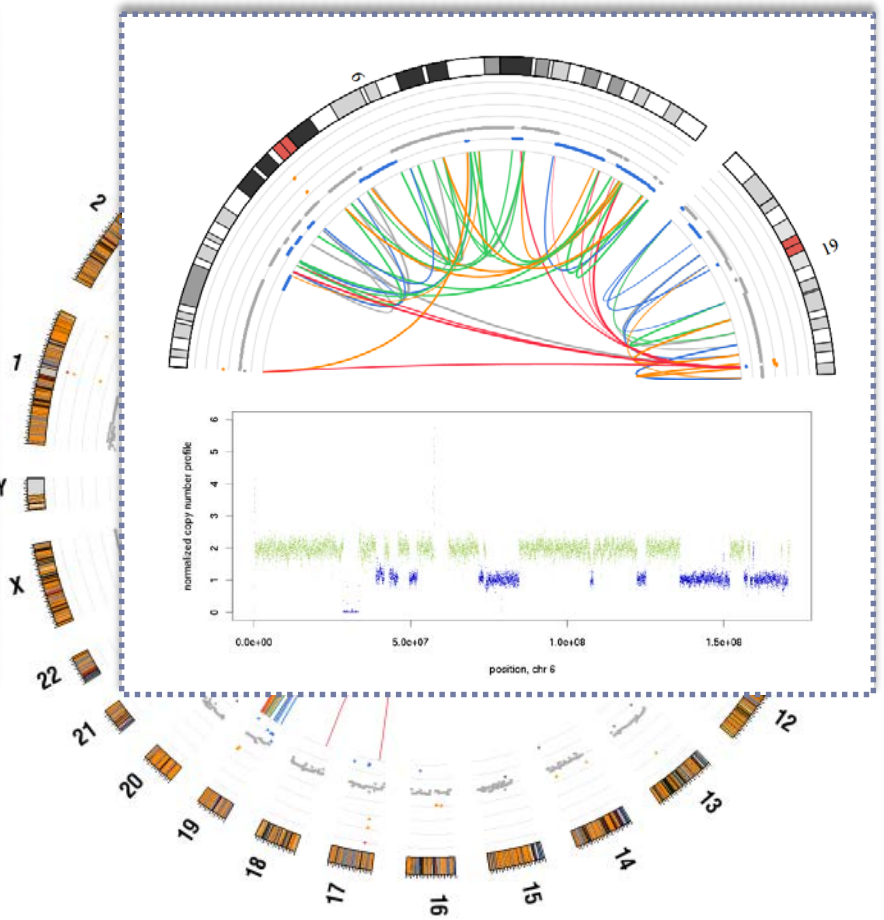




Boeva et al., PLOS One, 2013.

Examples of “classical” chromothripsis cases in neuroblastoma



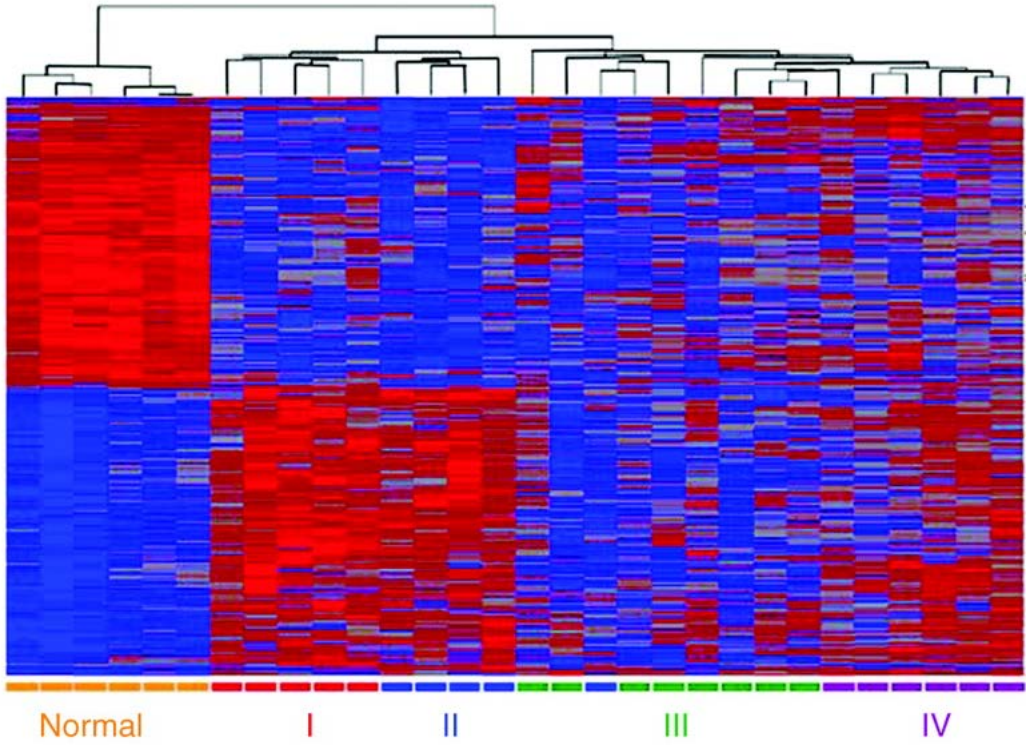


Boeva et al., PLOS One, 2013.

Summary

- **Single nucleotide variants** can be detected using frequencies of non-reference nucleotides in sequencing data
- **Copy number alterations** can be detected using **Read Count per Window or per target region**
  - Control data set is optional (we can use GC-content to normalize the read count) in whole genome sequencing
  - We can adjust for a possible contamination by normal cells
- **Structural variant** can be detected using **“abnormal” Paired-end Mappings**
  - There are >10 types of Paired-end Mapping Signature, each of them corresponds to a proper type of structural variant
  - To eliminate false predictions, we can filter out links falling satellites, segmental duplications, links with inconsistent pairs, etc.
  - To focus on somatic events, we need to filter out events detected in the normal sample (constitutional DNA)

Asian Gastric Cancer, Kim et al., 2012

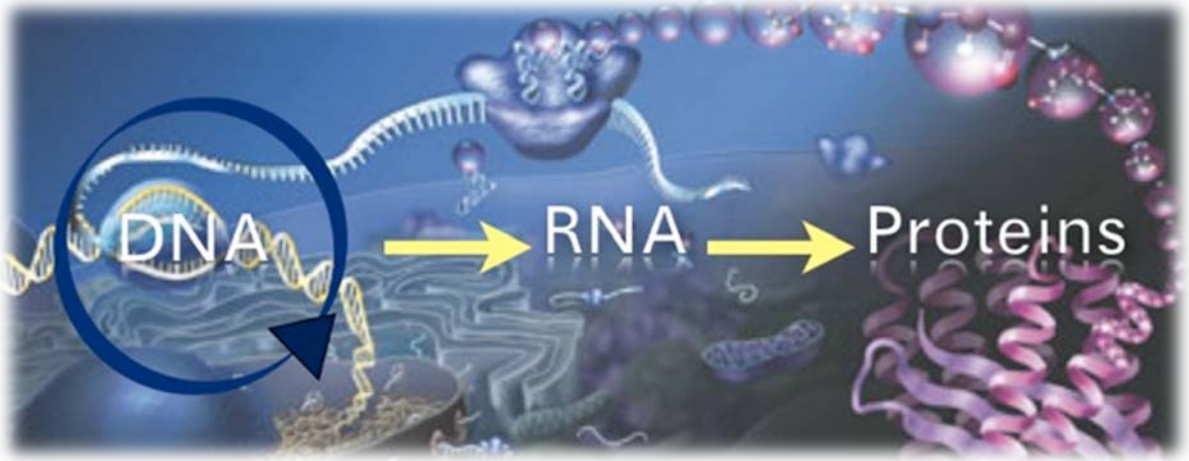


CANCER TRANSCRIPTOME

3 things that characterize a cancer transcriptome

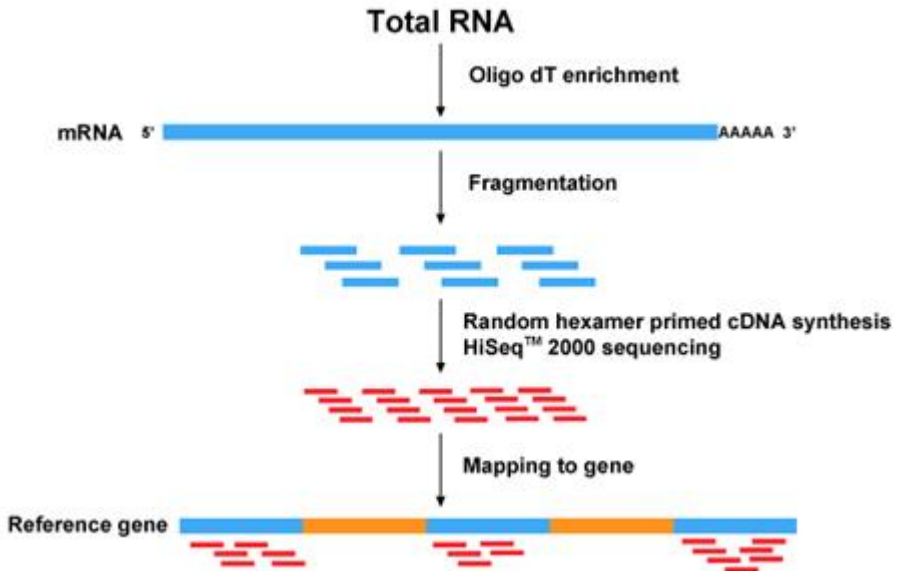
Transcriptome =

- Gene expression ⇐ Expression of cell cycle genes, oncogenes and tumor suppressors
- Splicing
- Abnormal transcripts ⇐ Expression of chimeric and truncated genes leading to formation of proteins with abnormal function



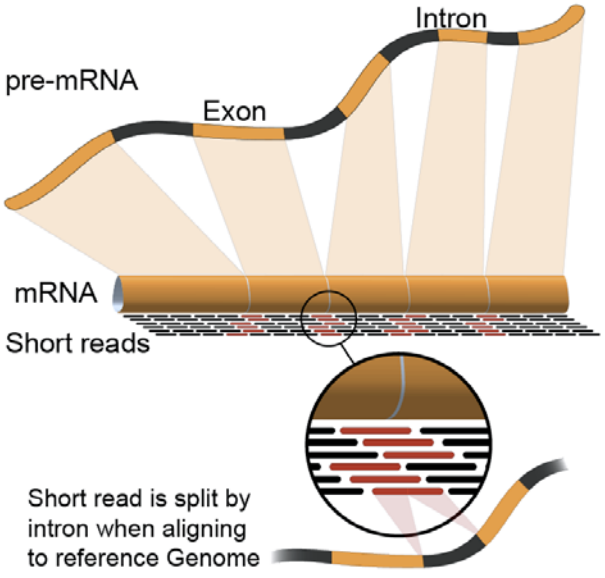
Cancer transcriptome can be fully characterized with RNA-seq

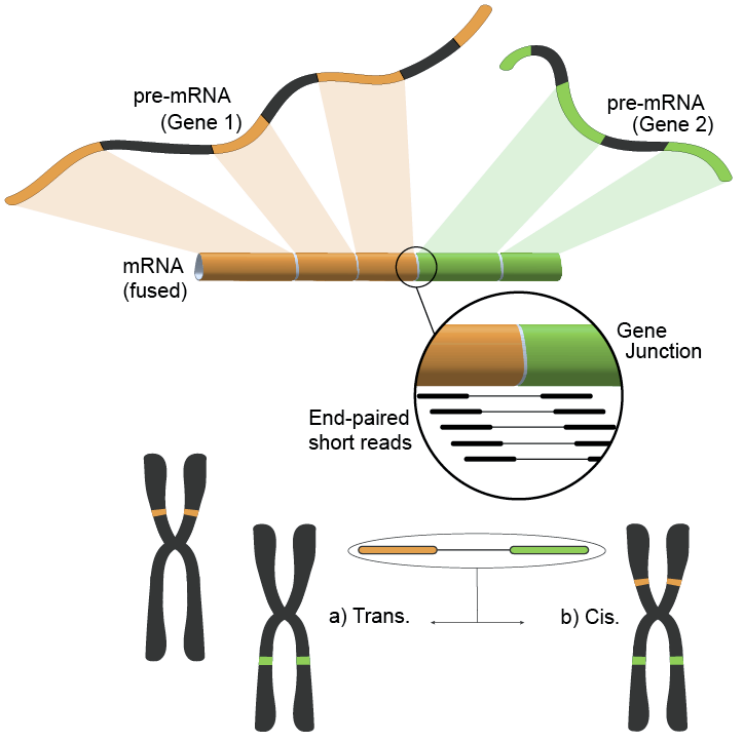
- RNA-seq consists in sequencing cDNA (complementary to mRNA)
- More reads = more expression



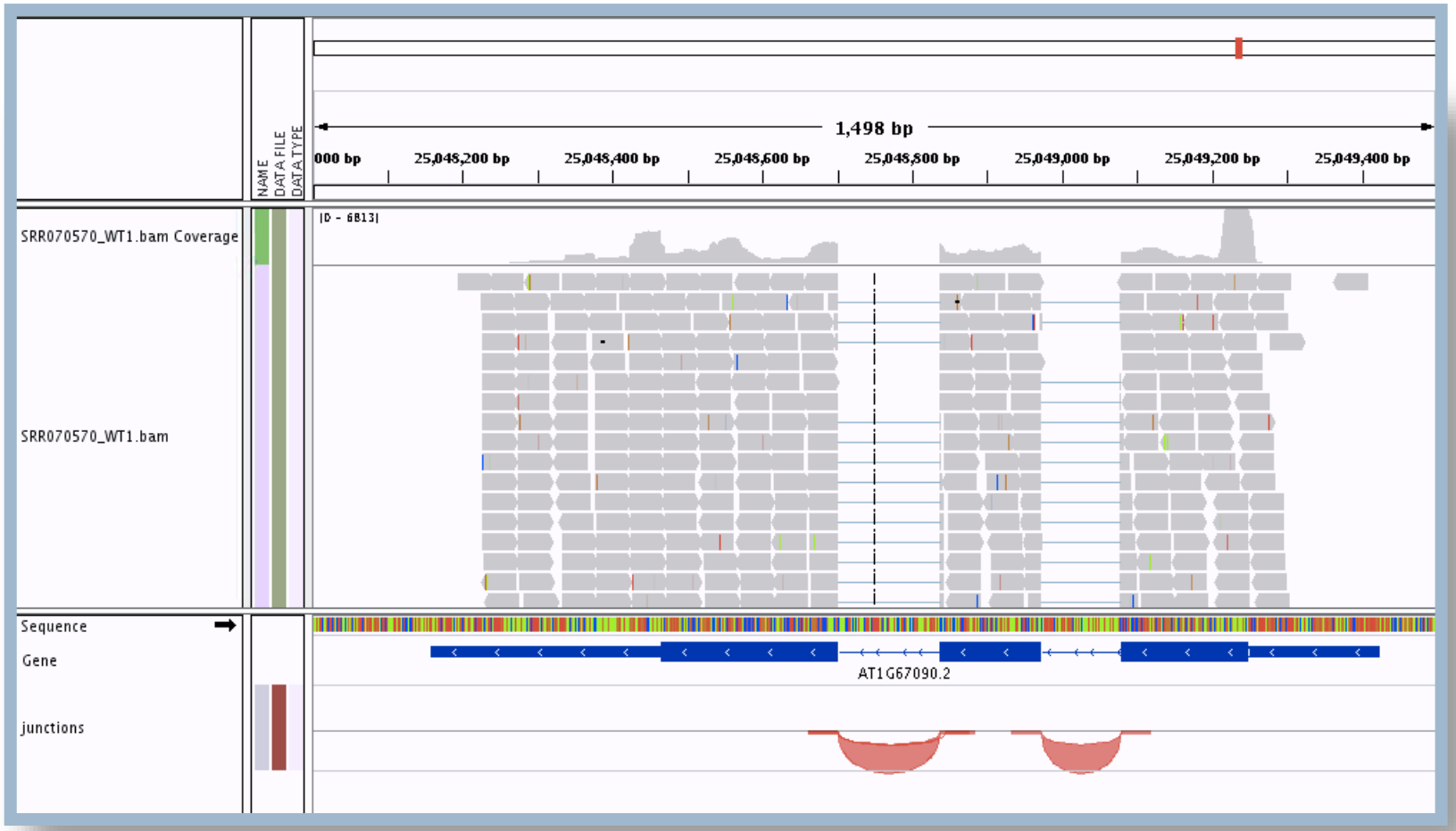
Cancer transcriptome can be fully characterized with RNA-seq

- RNA-seq allows us to identify splicing and fusion genes





Visualization of RNA-seq data with a genome browser (IGV)

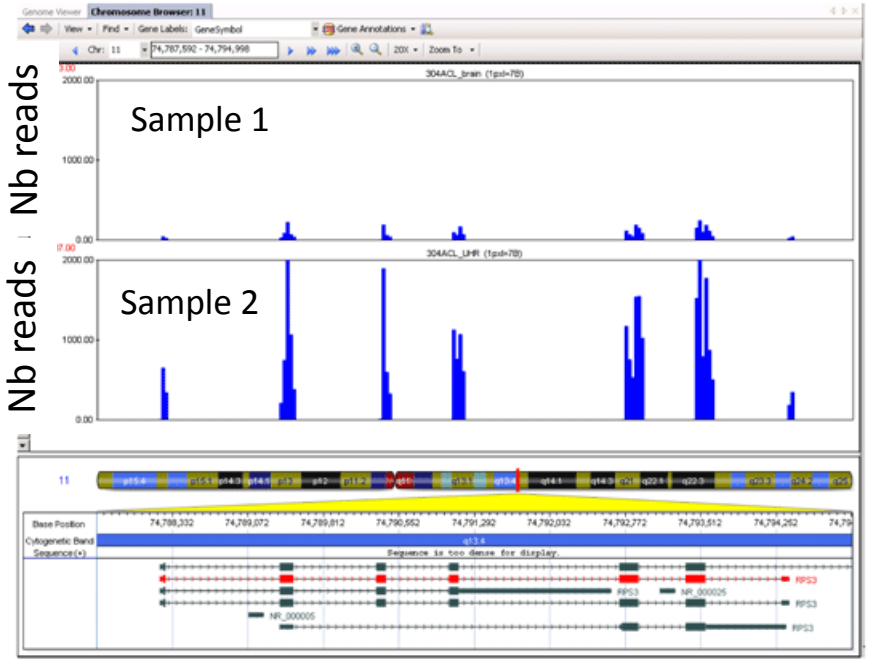


RPKM measure to quantify transcript abundance

- RPKM = Reads Per Kilobase per Million mapped reads to make samples comparable
  - Normalization for “Long genes ⇨ more reads”
  - and for “more reads in the library ⇨ higher the coverage”

RPKM measure to quantify transcript abundance

TISSUE-SPECIFIC EXPRESSION DETECTED WITH RNA-SEQ



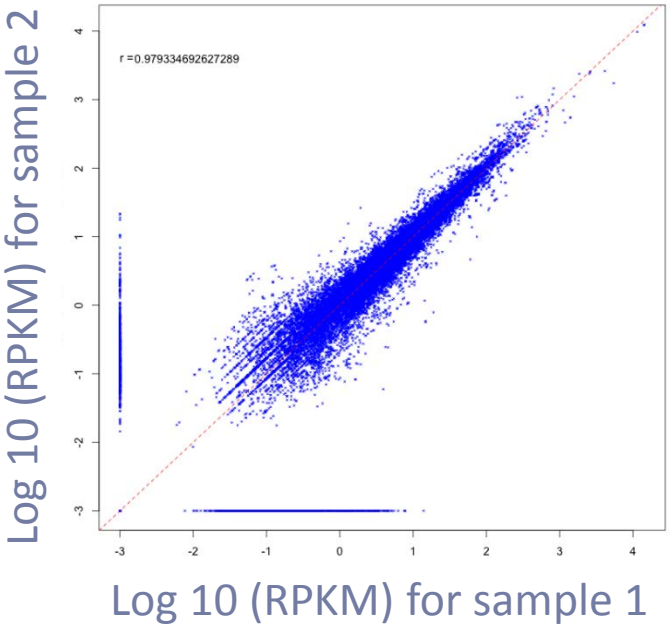
The quantity of individual reads are indicated at each genomic location (y-axis). Expressed exons are clearly seen as peaks, and are consistent with RefSeq annotation (bottom). Sample-specific expression is quantifiable by comparing results from different samples. The brain sample (top) exhibited 3,115 reads, whereas UHR sample (middle) exhibited 31,109 reads, indicating a ten-fold higher level of expression.

Sample 1 does not necessarily have a lower expression!!!

To be sure, we need to correct for the library size (i.e. total number of reads for each sample)

RPKM measure to quantify transcript abundance

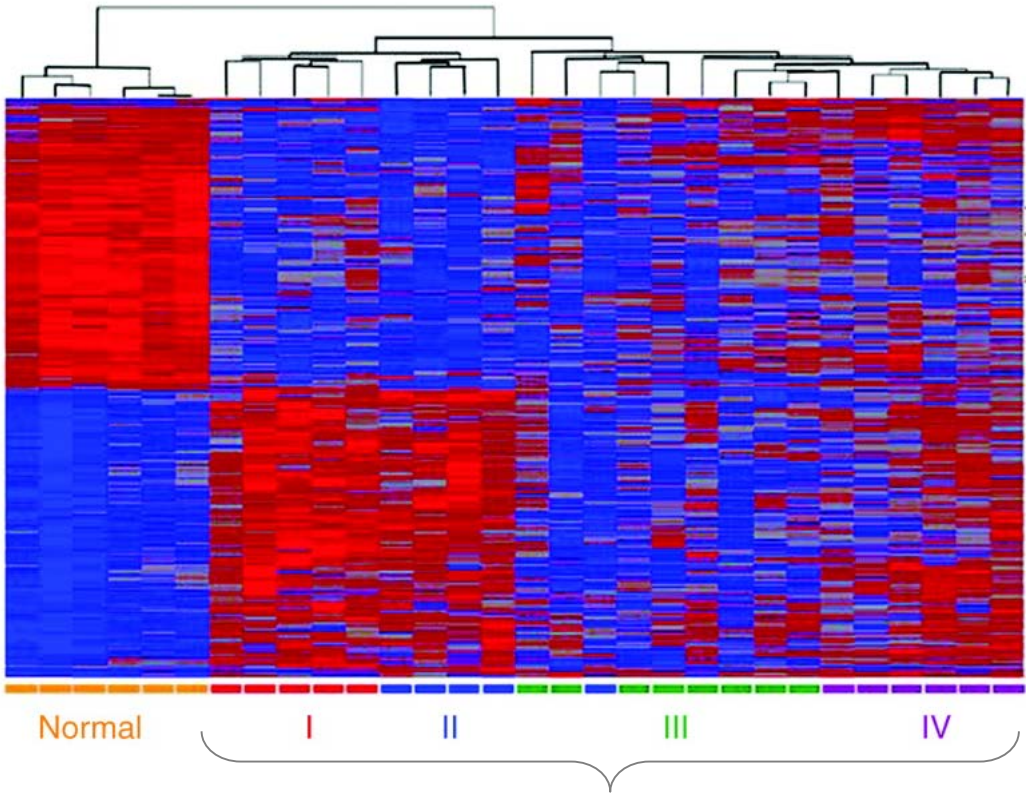
- RPKM = Reads Per Kilobase per Million mapped reads to make samples comparable
  - Normalization for “Long genes ⇨ more reads”
  - and for “more reads in the library ⇨ higher the coverage”



After RPKM normalization, we can compare expression of different genes in different samples

Gene expression profiles can be used to characterize tumor types

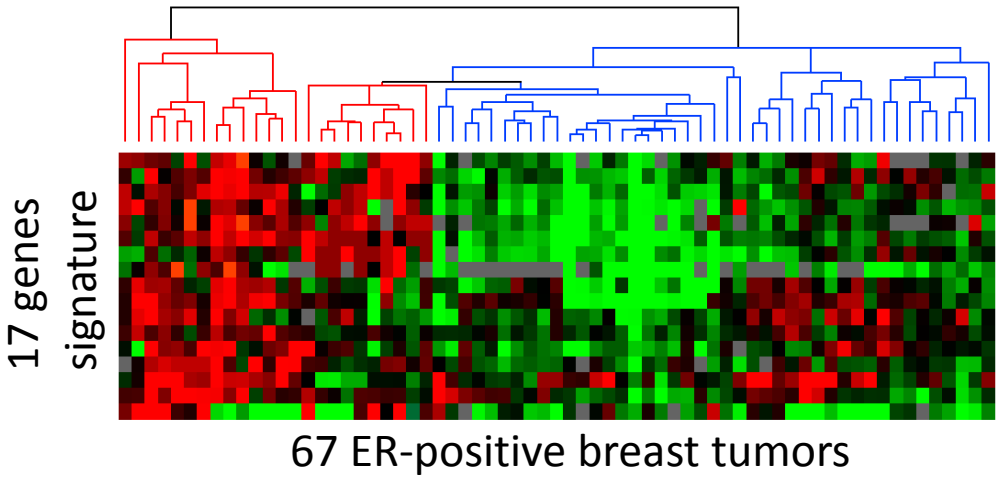
Unsupervised gene expression clustering of Asian Gastric Cancer samples



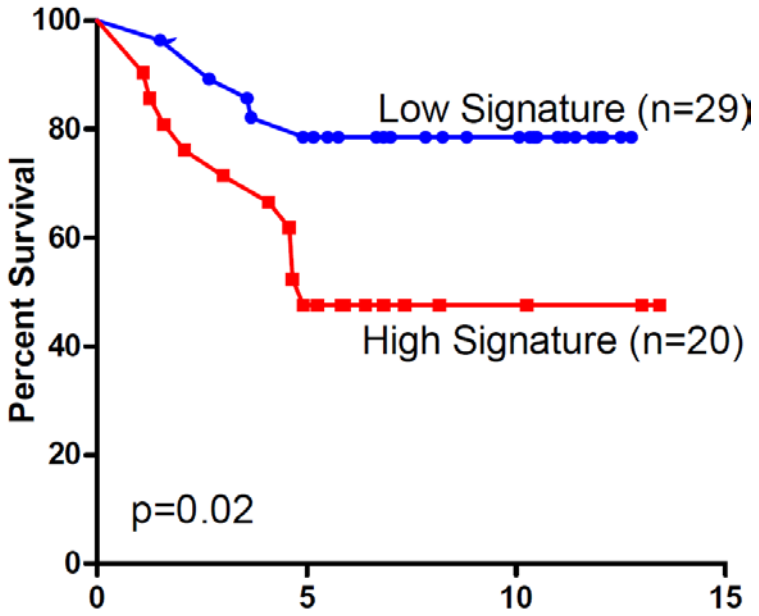
From Kim et al., 2012

Gene expression profiles can be used to characterize tumor aggressiveness

- Gene expression signature can distinguish cases with bad and good prognosis



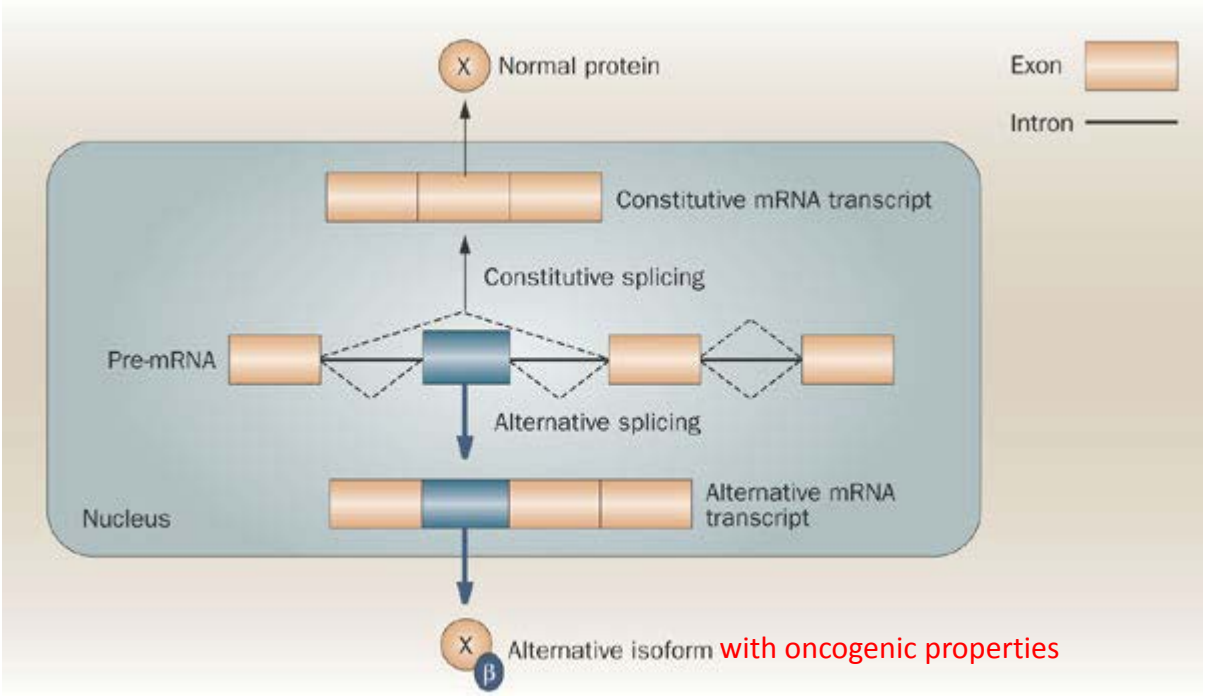
Heat maps and dendrograms for expression of the 14-3-3ζ signature genes.



Kaplan-Meier survival curves of patients with ER-positive breast tumors based on expression patterns of the 14-3-3ζ gene signature

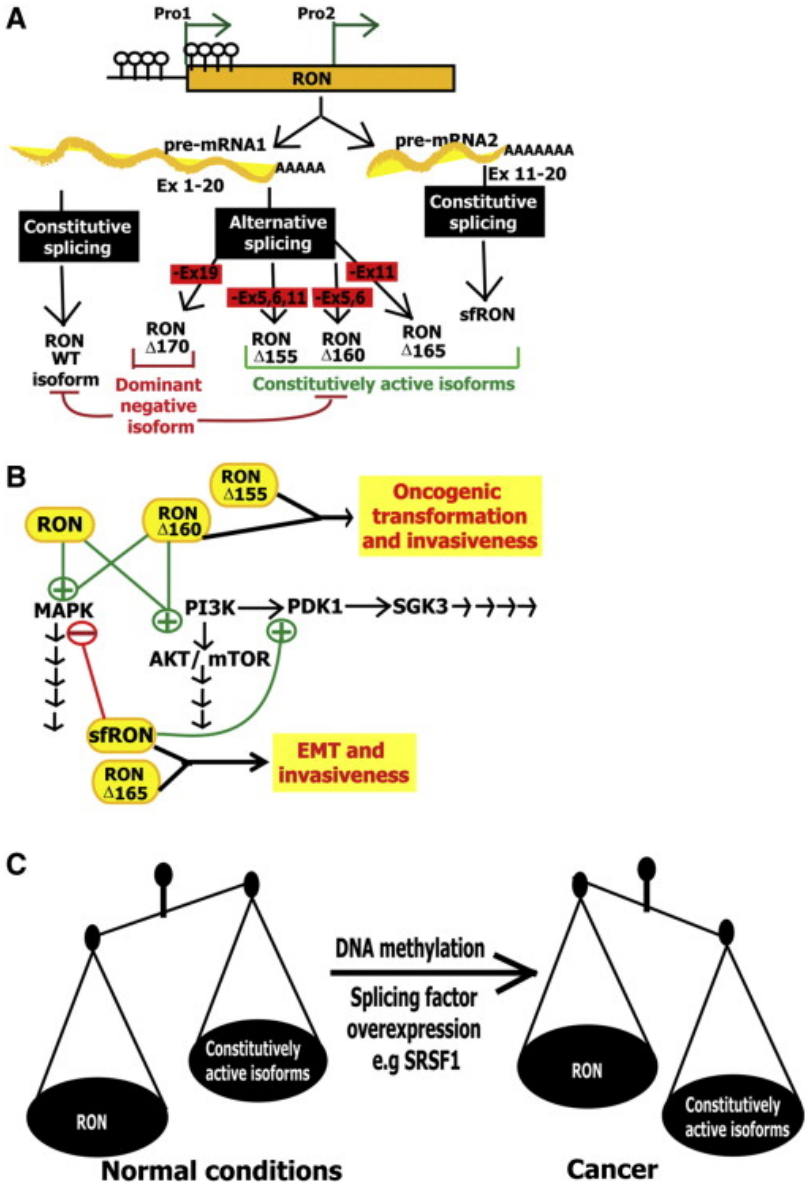
from Bergamaschi *et al. Breast Cancer Research* 2011 **13**:R70; data of Sorlie *et al.*

Cancer transcriptome often shows alternative splicing



from Rajan et al, 2009

Role of alternative promoters and alternative splicing in regulating receptor tyrosine kinase expression and functions is exemplified by RON (Pal et al, 2012)



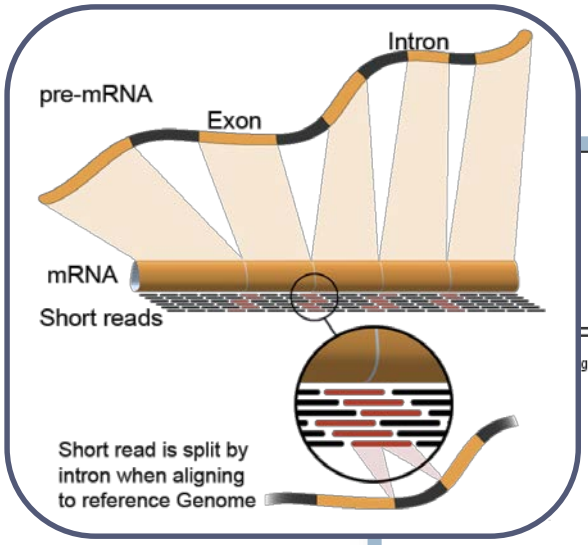
A. pre-mRNA1 undergoes alternative splicing to produce a variety of *RON* isoforms that include the kinase-deficient, dominant negative form-RONΔ170, and the constitutively active *RON* isoforms RONΔ165, RONΔ160, and RONΔ155.

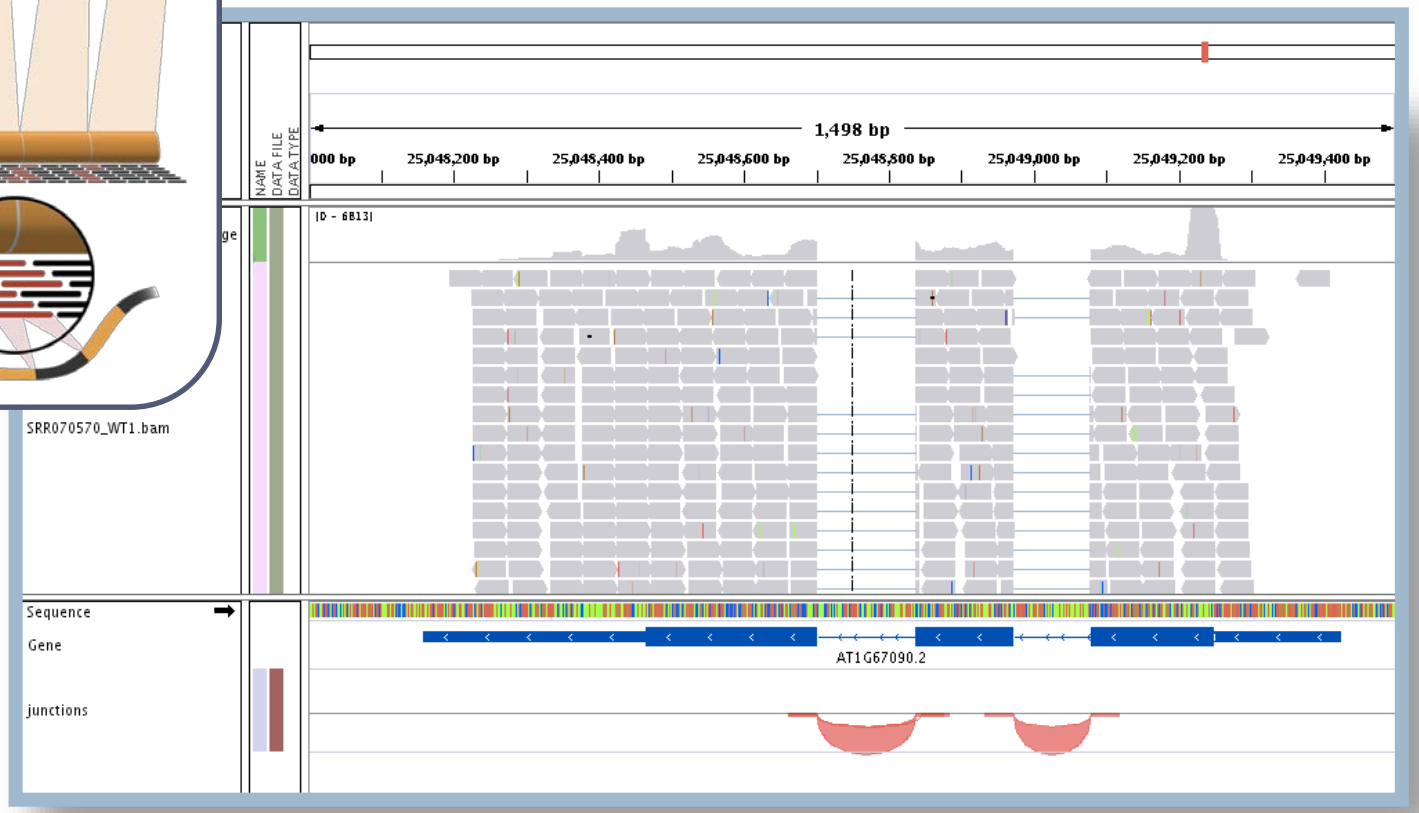
B. *RON* isoforms have different biological activities. While WT isoform-RON requires the binding of its ligand MSP to activate downstream MAPK, PI3K, AKT, and mTOR pathways, RONΔ160 activates these pathways in a ligand independent manner. On the other hand, the sfRON inhibits MAPK pathway and activates PI3K pathway and downstream PDK1 and SGK3 segment. The *RON* isoforms also differ in their impact on tumorigenesis, i.e., RONΔ160 and RONΔ155 have both transforming and metastatic activity, while RONΔ165 and sfRON can only induce EMT and promote invasiveness.

C. Various isoforms have been co-detected in normal cells, with WT RON being the major isoform; however in benign and metastatic cancers, this balance is switched such that constitutively active RON-isoforms (specific isoform varies with cancer types) represent the majority of *RON*'s protein in the cancer cells.

Detection of splicing (expressed isoforms)

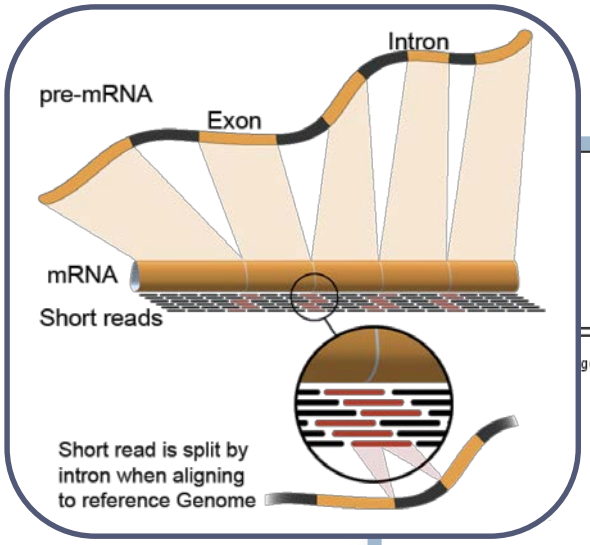
- Detection of splicing using “spliced” read





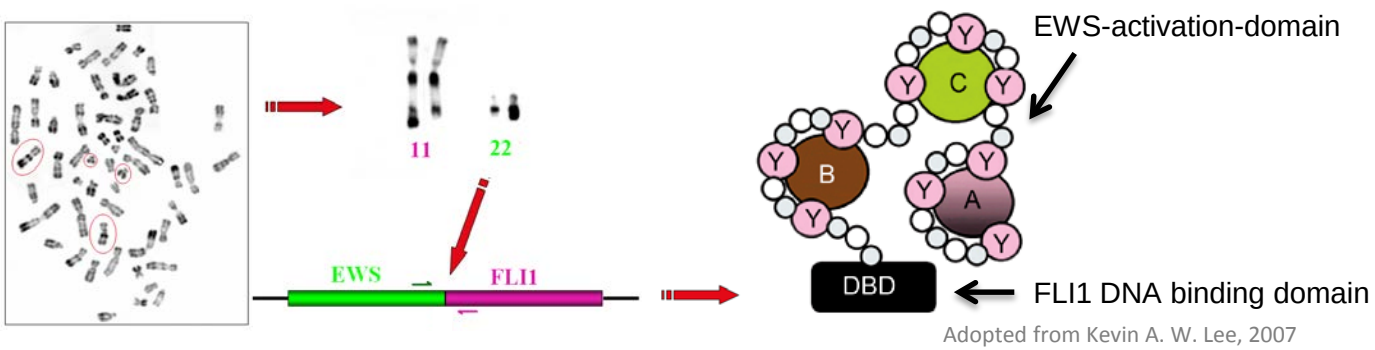
Detection of splicing (expressed isoforms)

- Detection of splicing us



Expressed chimeric transcripts can produce proteins with abnormal function

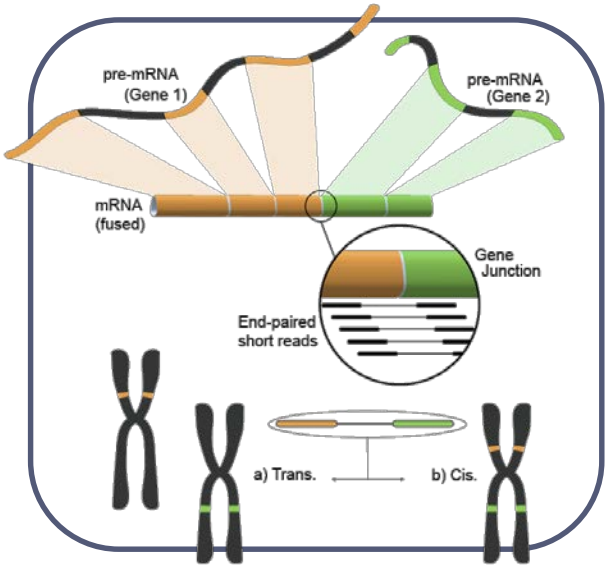
- In cancer, about 20% of chimeric genes give rise to chimeric proteins
- Chimeric proteins can disrupt normal cell functioning

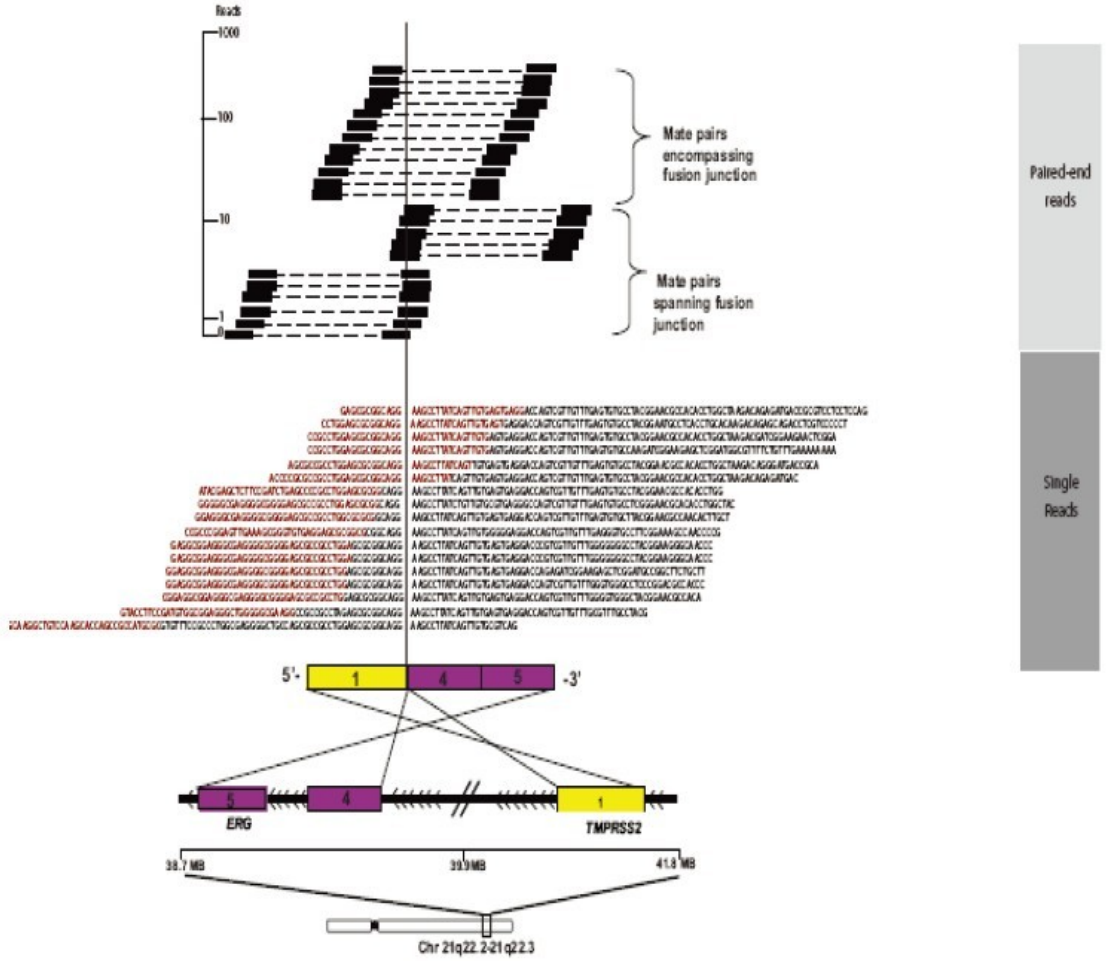


Adopted from Kevin A. W. Lee, 2007

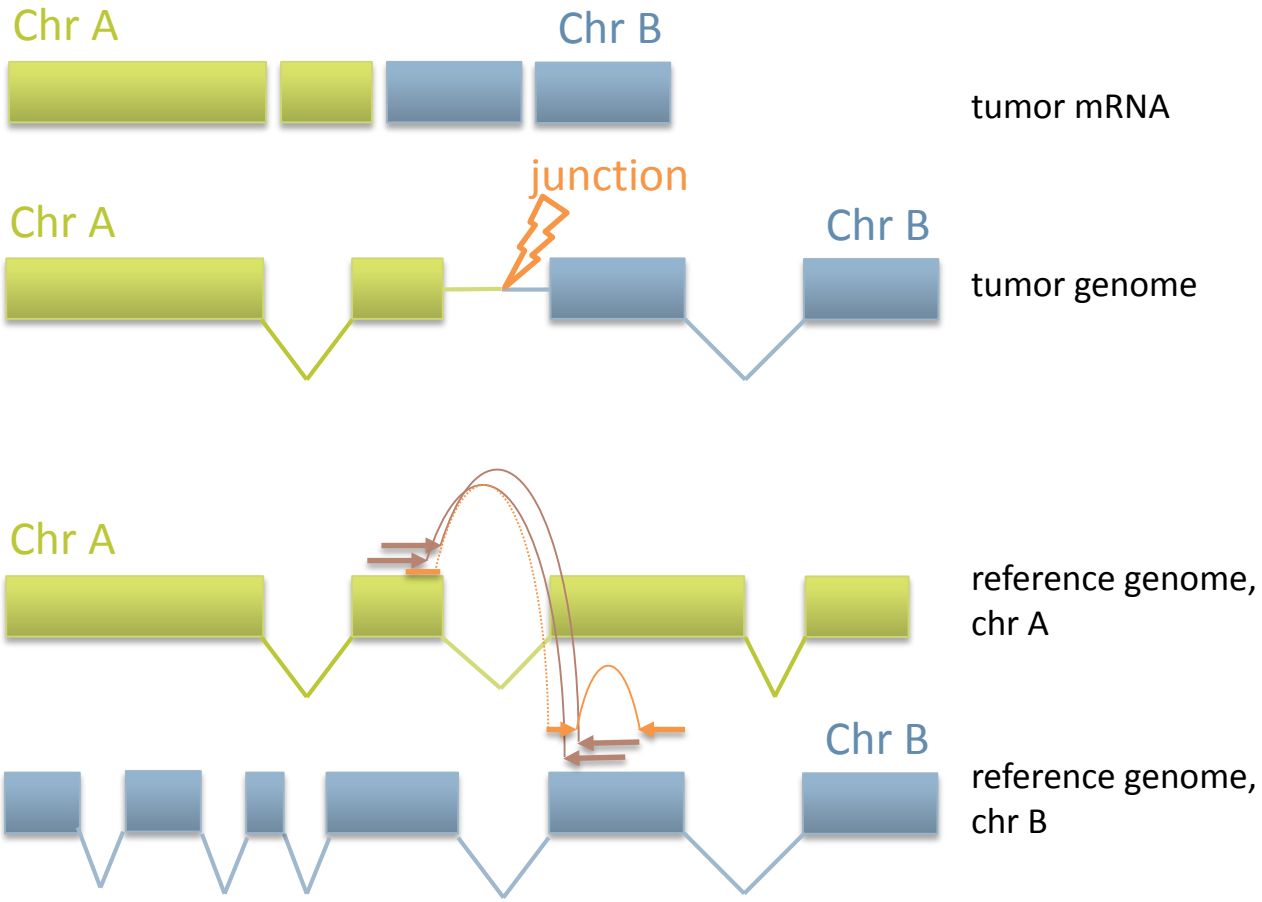
Fusion of chr11 and chr22 encoding for EWS/FLI1 causes Ewing sarcoma.

Detection of expressed chimeric transcripts is done using paired-end data and “spliced” alignment of reads





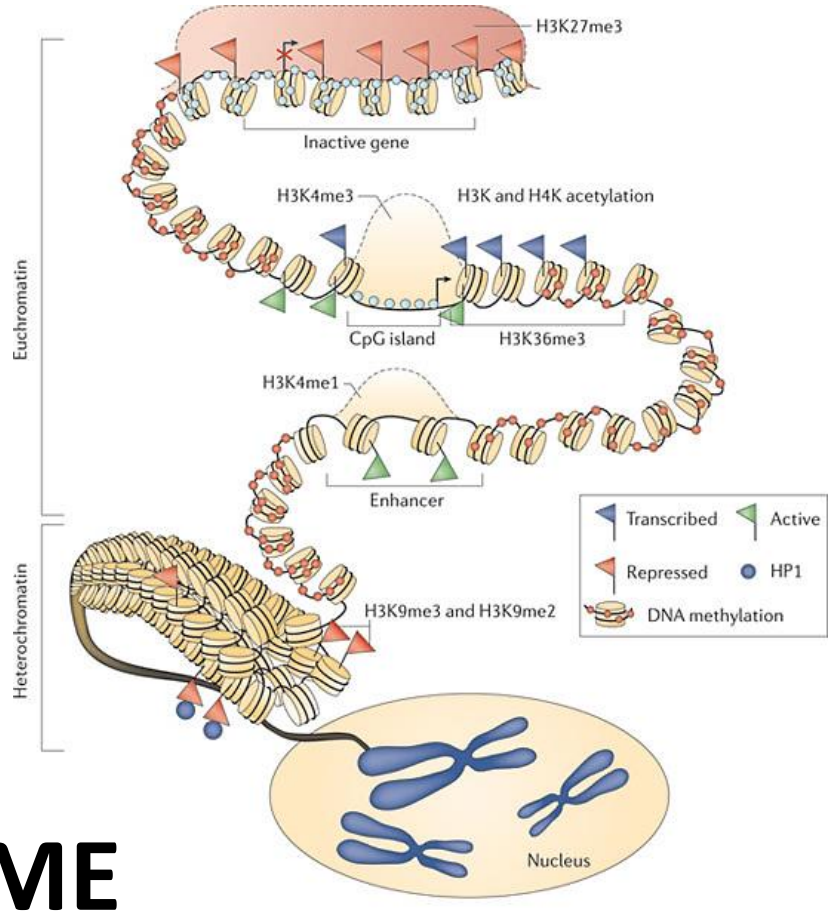
We are confident in a chimera or another translocation if it is confirmed both by abnormal paired-end read mapping and “spliced” reads



Summary

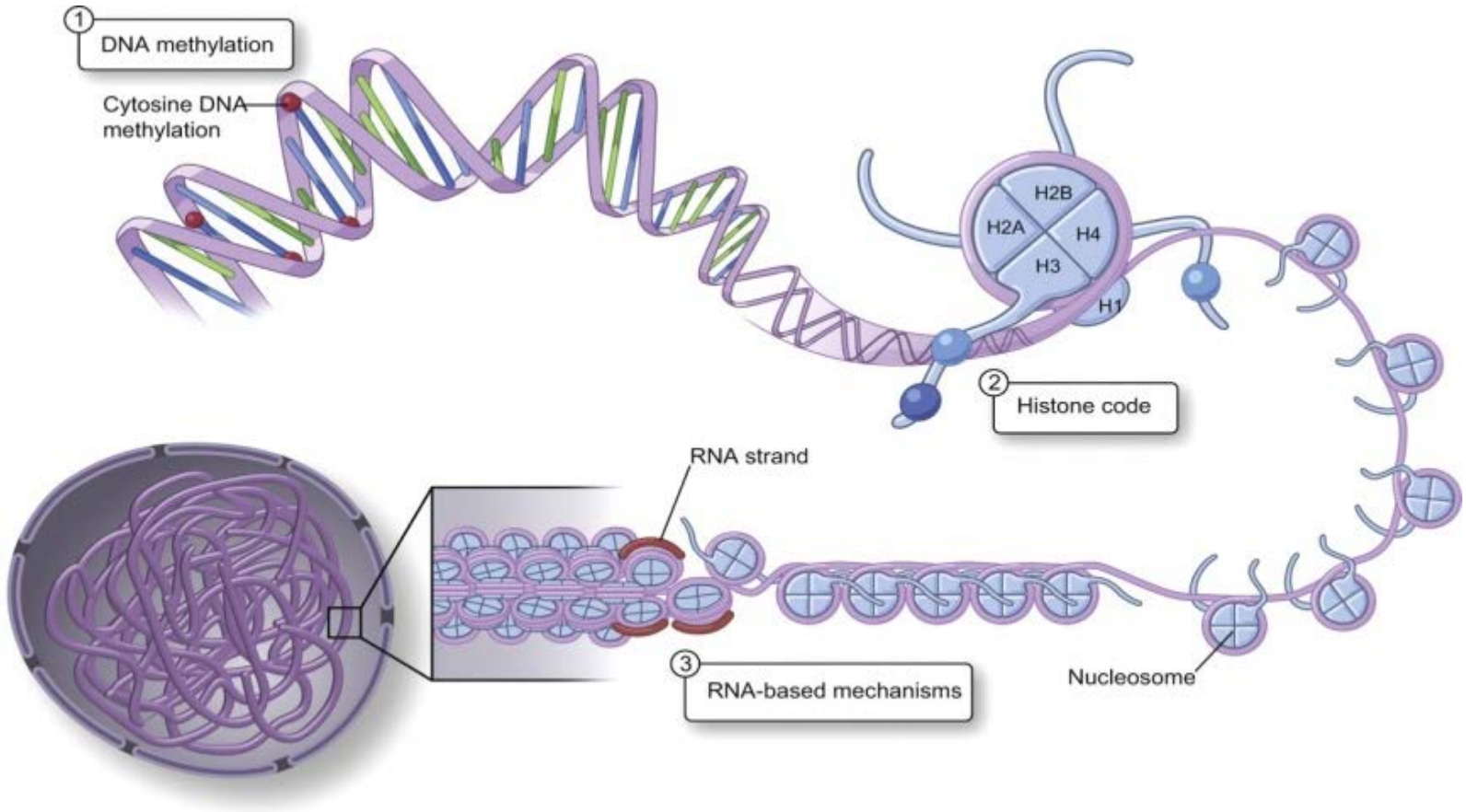
- **RNA-seq** can be used to detect
  - Gene expression
  - Splicing
  - Abnormal transcripts
- **Normalized read counts (RPKM)** should be analyzed to quantify gene expression
- **Paired-end read mapping** and **spliced read mapping** should be analyzed to detect splicing and expressed fusion transcripts

CANCER EPIGENOME



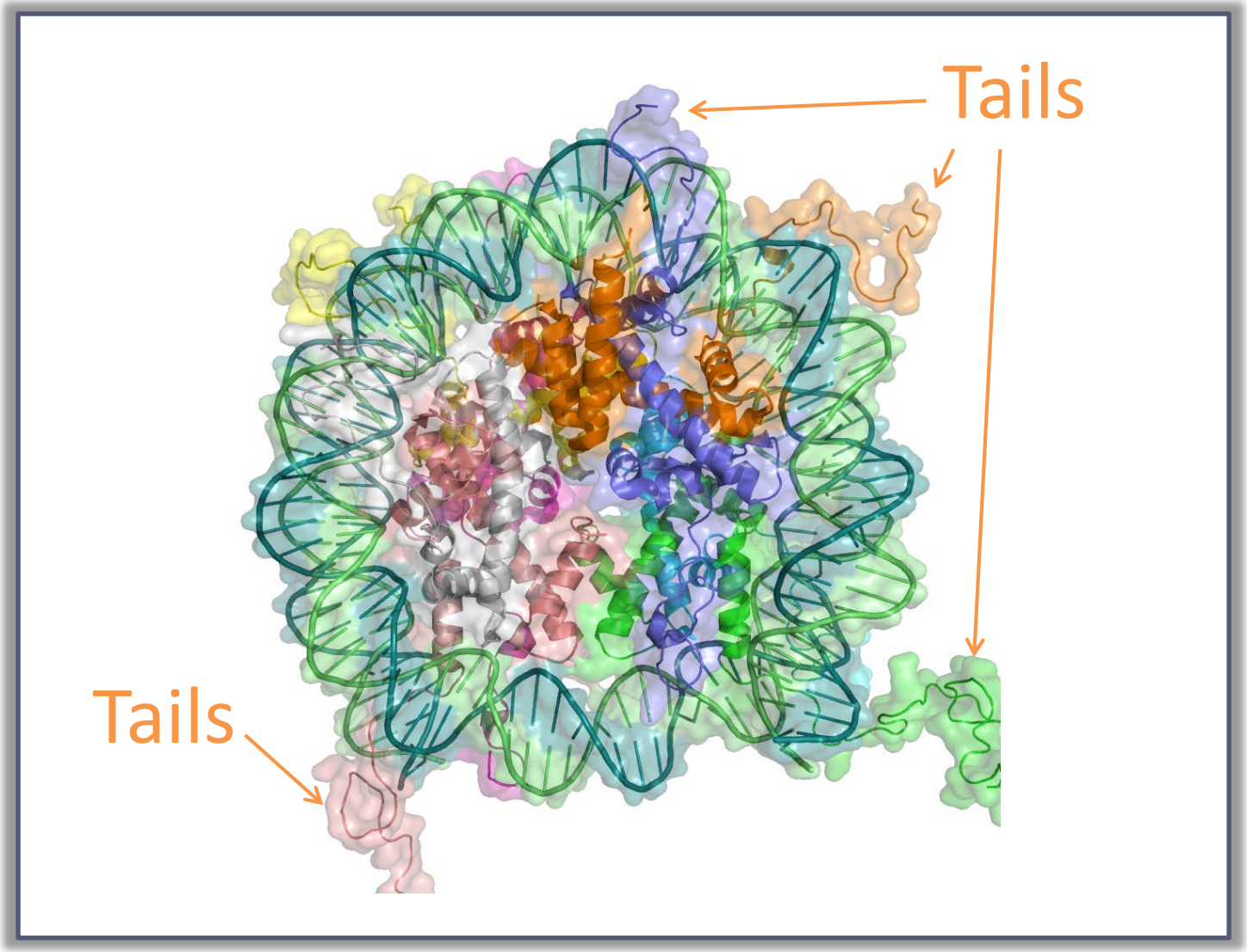
Nature Reviews | Cancer

DNA is bound around nucleosomes composed of histones (called chromatin)

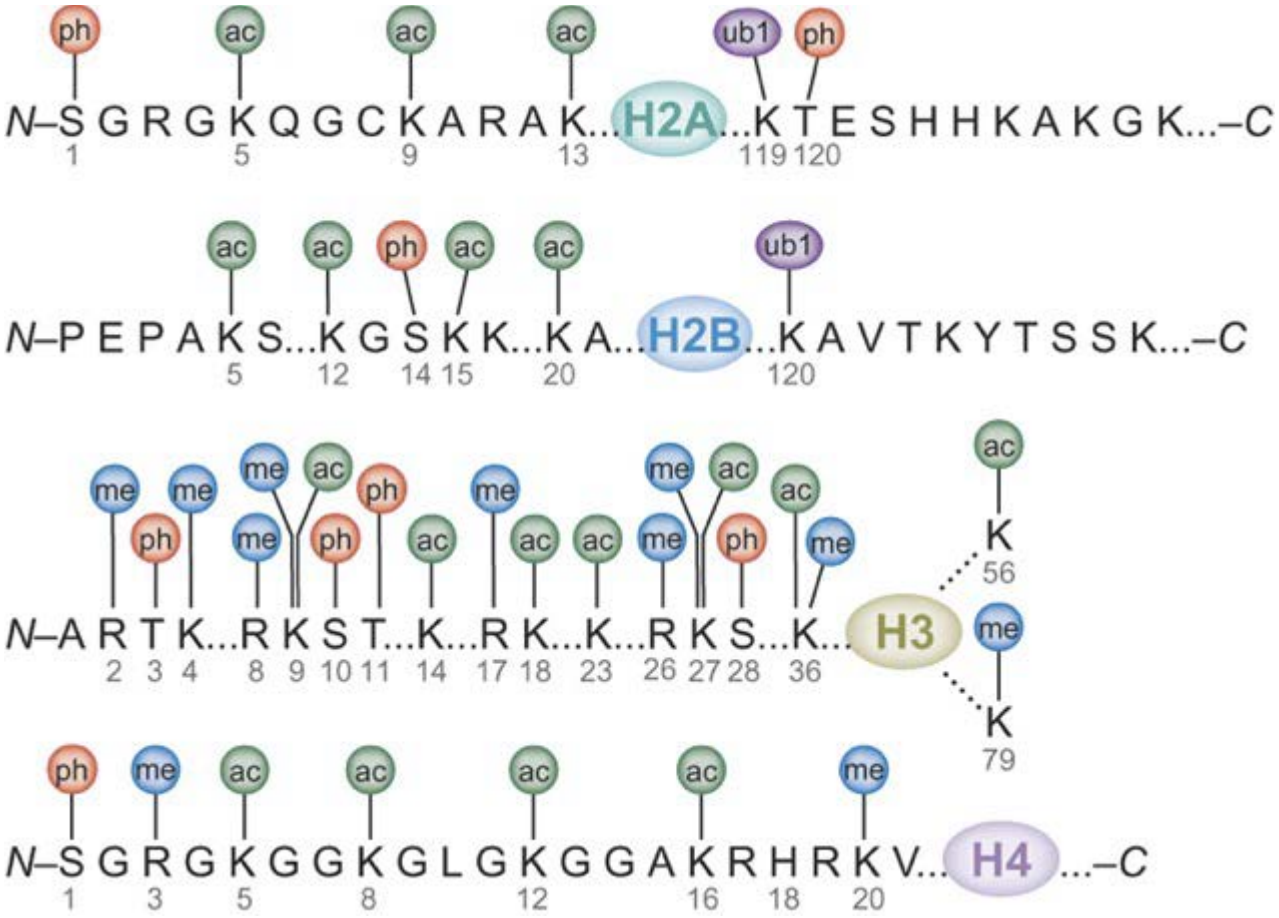


from Yan, M. S. et al. J Appl Physiol 2010

Histones have tails that can be covalently modified

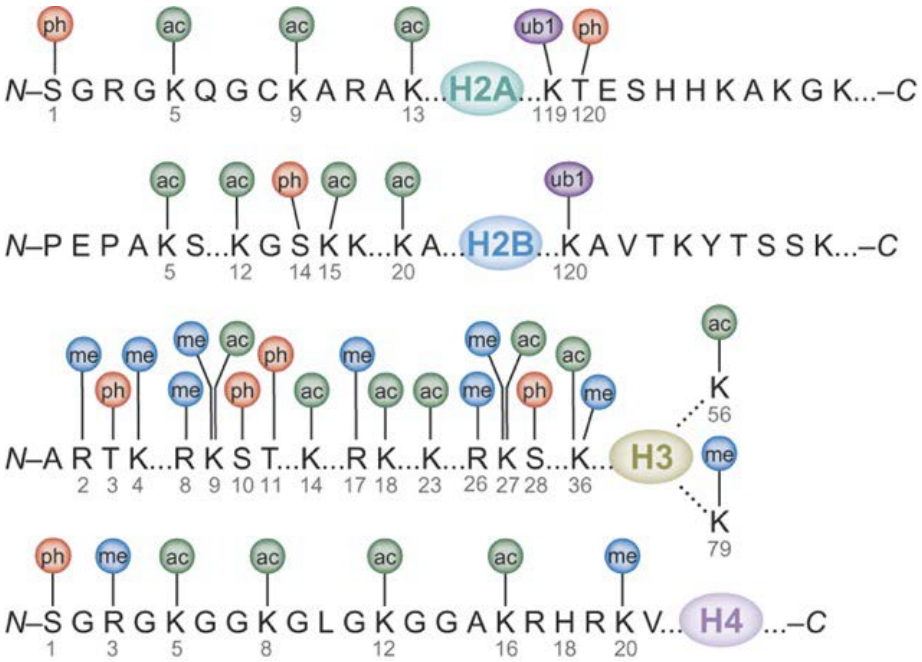


There are many well-known histone modifications (covalent modifications of histone tails)

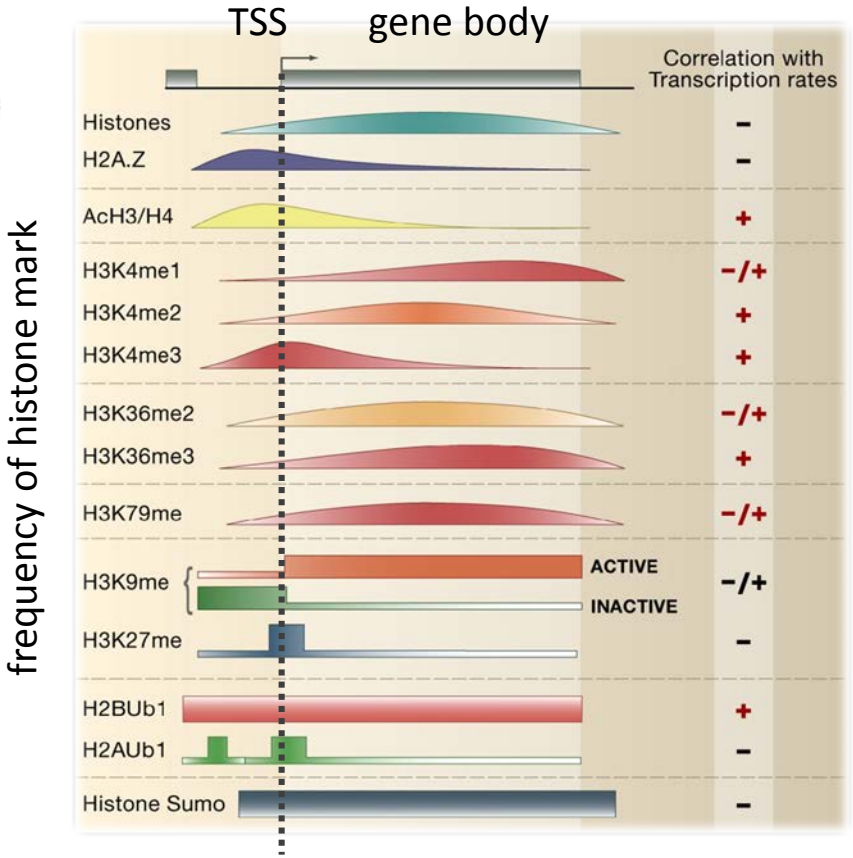


Bhaumik et al, Nat Str & Mol Bio, 2007

Epigenetic marks are associated with gene expression



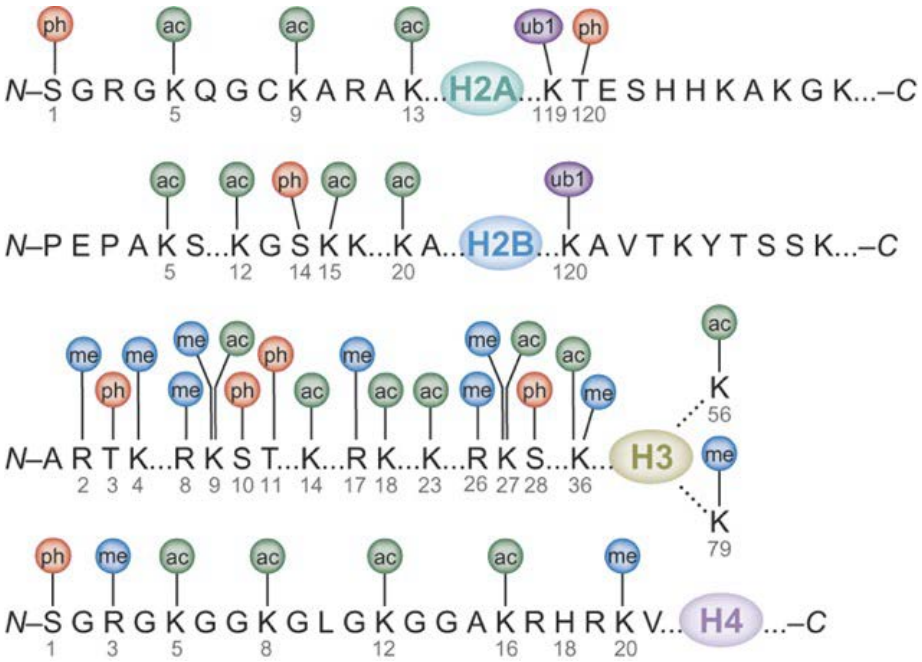
from Bhaumik et al, *Nat Str & Mol Bio*, 2007



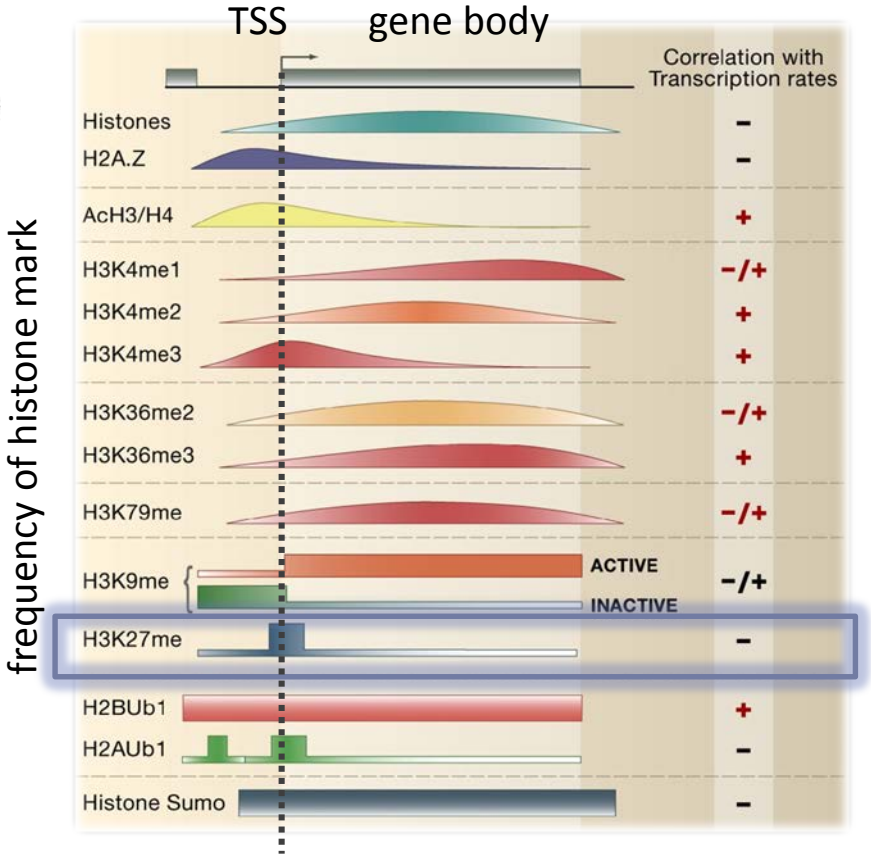
from Li et al., 2007. *Cell*

TSS = transcription start site

Epigenetic marks are associated with gene expression



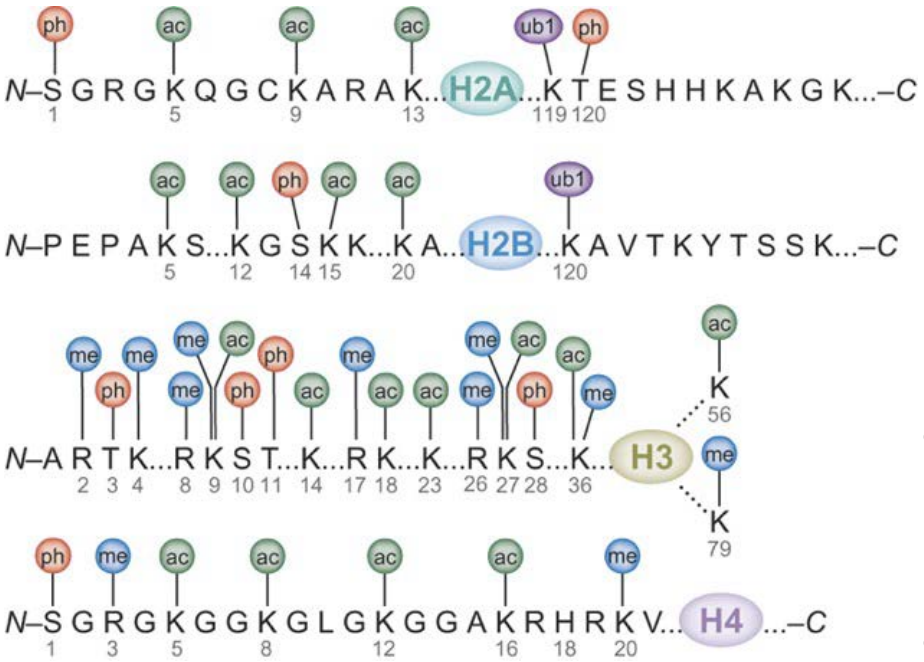
from Bhaumik et al, *Nat Str & Mol Bio*, 2007



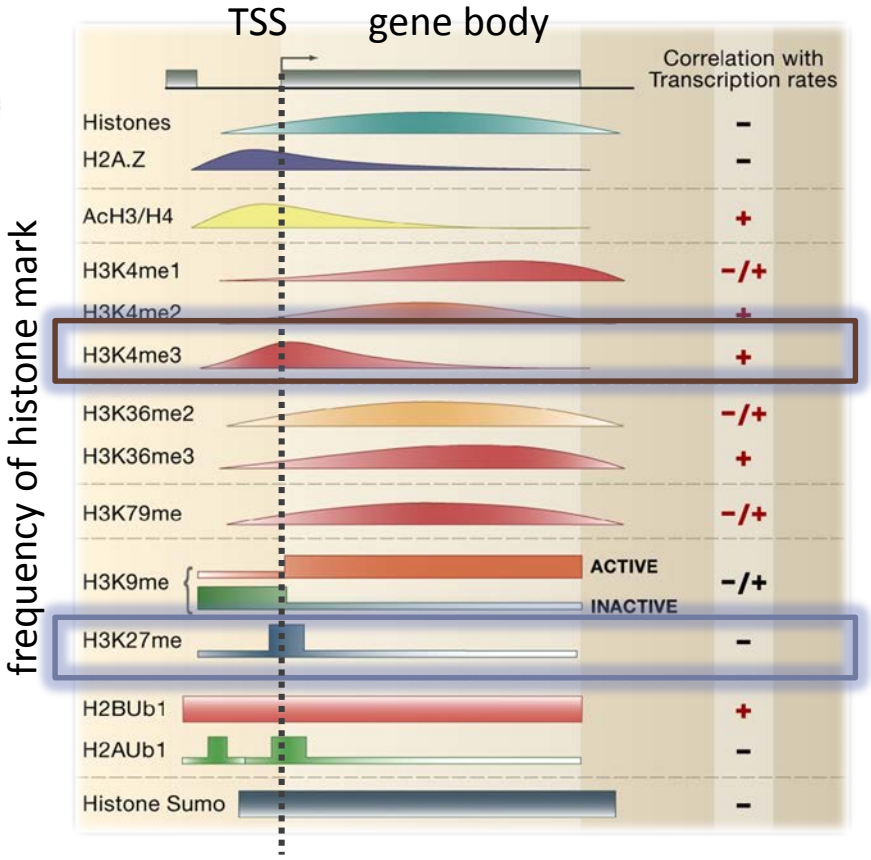
from Li et al., 2007. *Cell*

TSS = transcription start site

Epigenetic marks are associated with gene expression



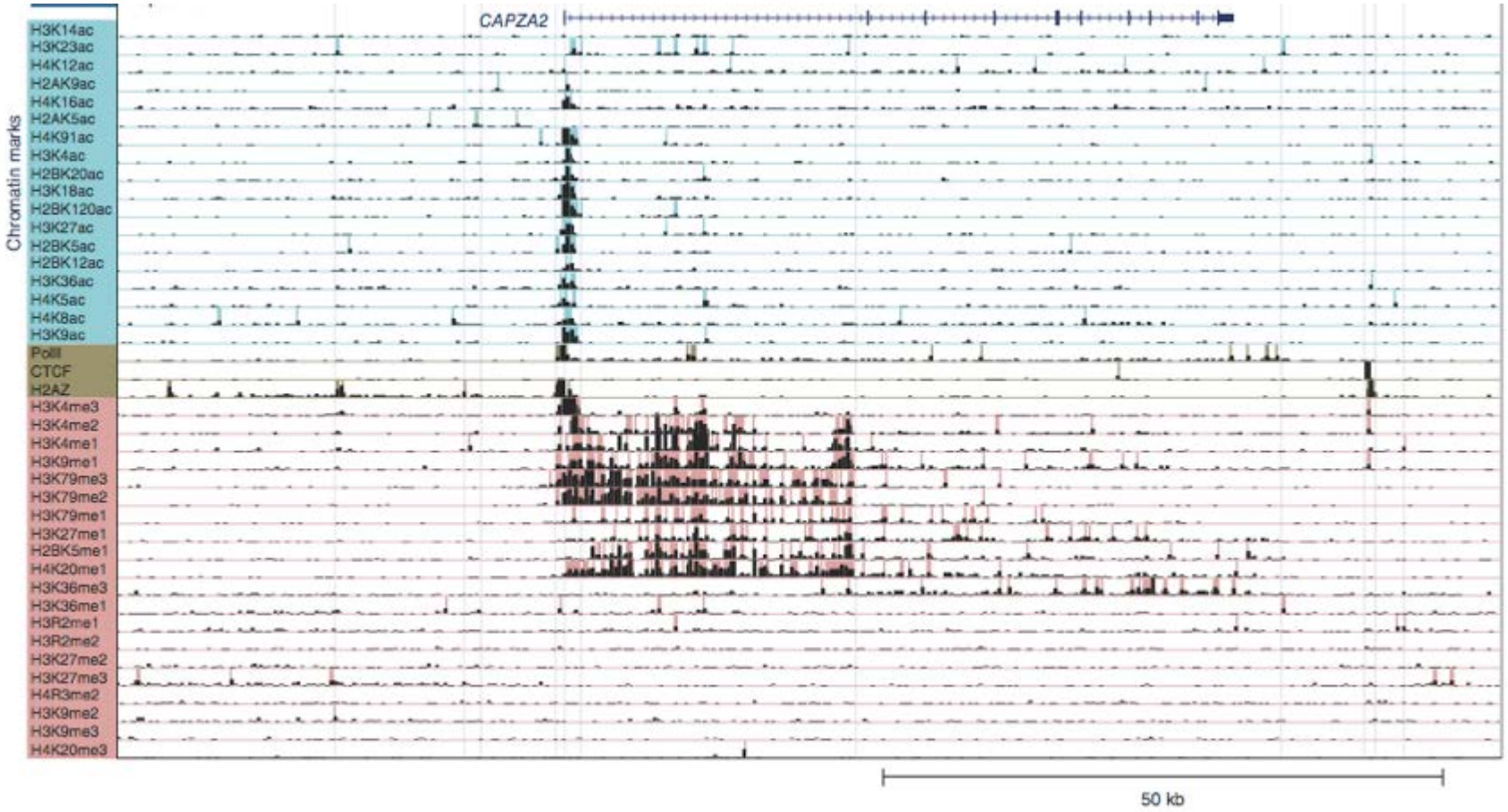
from Bhaumik et al, *Nat Str & Mol Bio*, 2007



from Li et al., 2007. *Cell*

TSS = transcription start site

Visualization of chromatin marks measured by ChIP-Seq



Histone modifying proteins are often mutated or deleted in cancer

- Epigenome-modifying gene mutations in human cancer

Gene	Cancer	Frequency or stage of cancer	Frequency of mutation (*N*)	Effect
Histone variants				
HIST1H1B	Colorectal cancer	Common	4% (24)	
HIST1H1C	Non-Hodgkin's lymphoma	Common	7% (127)	
H3F3A	Paediatric glioblastoma	Rare aggressive paediatric, high grade	36% (90)	Prevents PTMs on H3K27 or H3K36
	Diffuse intrinsic pontine glioma	Rare aggressive paediatric	60% (50)	Prevents PTMs on H3K27
HIST1H3B	Diffuse intrinsic pontine glioma	Rare aggressive paediatric	18% (50)	Prevents PTMs on H3K27
DNA methyltransferases				
DNMT1	Colorectal cancer		2% (29)	Mutation
DNMT3A	AML	Stage M4	13.6% (66)	
		Stage M5	20.5% (112)	
	AML	Common	22.1% (281)	
DNA demethylases				
TET2	BCR-ABL-negative myeloproliferative neoplasms	Rare form	13% (239)	
	CMML	Common form	50% (68)	
	MDS	Rare	26% (102)	
IDH1	Anaplastic astrocytoma	Rare	73% (52)	
	Diffuse astrocytoma	Rare	90% (30)	
	AML	Common	6.2% (385)	
IDH2	AML	Common	8.6% (385)	
Histone acetyltransferases				
EP300 (which encodes p300)	Pancreas adenocarcinoma	Common form	8% (24)	Mutation
	DLBCL	Common form	10% (134)	Mutation
	Follicular lymphoma	Uncommon form	8.7% (46)	Mutation
	Head and neck squamous cell cancer	Common	11% (74)	Mutation
	Transitional cell carcinoma (bladder)	Common	13% (97)	Mutation
CREBBP (which encodes CBP)	Ovary	Common	3% (75)	Inactivated
	Breast adenocarcinoma	Common	8% (183)	Gain of copy
	Lung cancer	Common	5.3% (95)	Mutation
	DLBCL	Common form	22.4% (134)	Mutation
	DLBCL	Common form	18% (111)	Mutation
	Follicular lymphoma	Uncommon form	32.6% (46)	Mutation
	Relapsed ALL		18.3% (71)	Mutation
	Transitional cell carcinoma (bladder)	Common	13% (97)	Mutation
ELP4	Breast adenocarcinoma	Common	4% (183)	Gain of copy
Histone deacetylases				
HDAC4	Breast adenocarcinoma	Common	4% (24)	Mutation
HDAC9	Prostate adenocarcinoma	Common	42.9% (7)	Mutation
Histone methyltransferases				
SETD2	Renal clear cell carcinoma	Common	3% (407)	Mutation
MEN1	Pancreas neuroendocrine cancer	Rare	44% (68)	Mutation
	Parathyroid cancer	Rare	35% (185)	Mutation
MLL	Squamous cell lung cancer	Rare	3% (63)	Mutation
	Transitional cell carcinoma (bladder)	Common	7% (97)	Mutation
	Mixed lineage leukaemia	Common	100% (definition)	Fusion

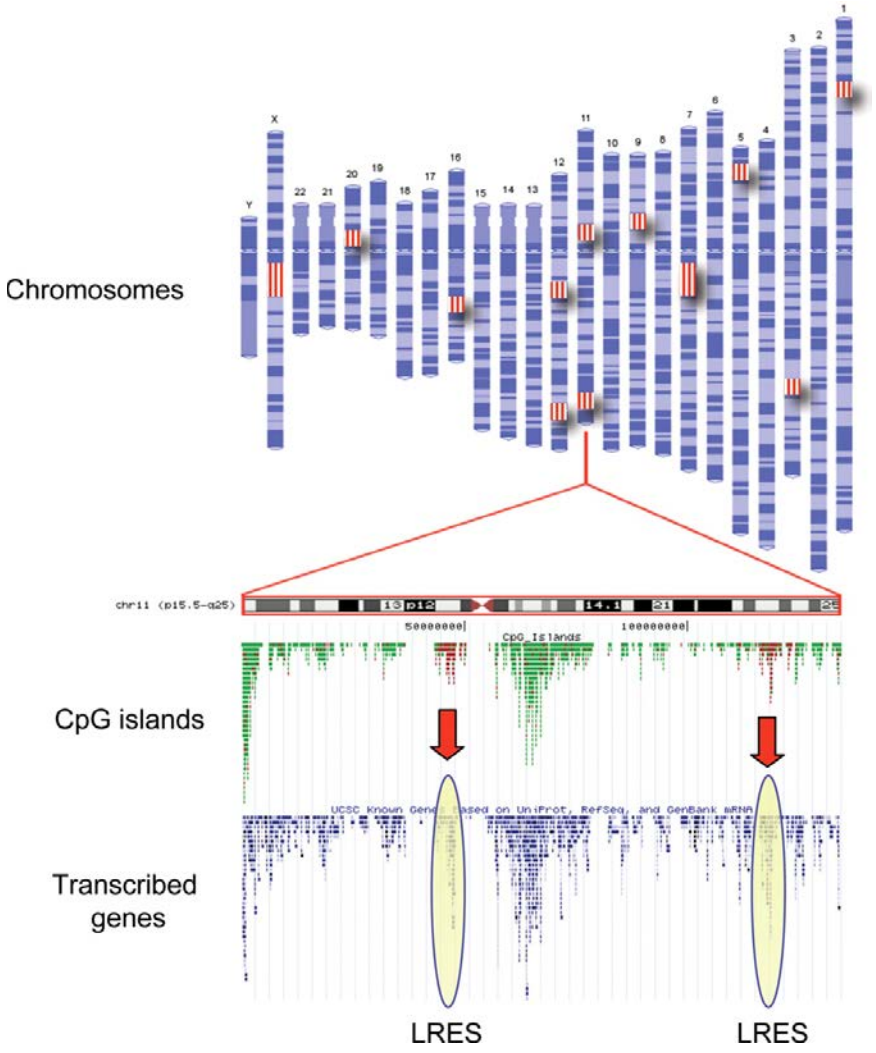
Gene	Cancer	Frequency or stage of cancer	Frequency of mutation (*N*)	Effect
Histone methyltransferases (cont.)				
MLL2	Renal clear cell carcinoma	Common	4% (407)	Mutation
	Childhood medulloblastoma	Rare	8.7% (92)	Mutation
	Childhood medulloblastoma	Rare	13.6% (88)	Mutation
	DLBCL	Common form	32% (37)	Mutation
	DLBCL	Common form	22.8% (92)	Mutation
	Follicular lymphoma	Uncommon form	89% (35)	Mutation
	Head and neck squamous cell cancer	Common	11% (74)	Mutation
MLL3	Childhood medulloblastoma	Rare	3.4% (88)	Mutation
	Transitional cell carcinoma (bladder)	Common	5% (97)	Mutation
	Colorectal cancer	Common	20.8% (24)	
EZH2	Non-Hodgkin's lymphoma	Common	7.8% (661)	Mutations
	DLBCL	Common form	5.6% (107)	Mutation
	MDS and MPNs	Rare	12% (219)	Mutations
	Myelofibrosis	Rare	13% (30)	Mutations
	Follicular lymphoma	Uncommon form	12% (221)	Mutations
Histone demethylases				
KDM5C (also known as *JARID1C*)	Renal clear cell carcinoma	Common	3% (407)	Mutation
KDM6A (also known as *UTX*)	Transitional cell carcinoma (bladder)	Common	20% (97)	Mutation
	Childhood medulloblastoma	Rare	3.2% (92)	Mutation
KDM2B	DLBCL	Uncommon form	7.4% (54)	Mutation
Chromatin remodelling factors				
ARID1A	Pancreas adenocarcinoma	Common	8% (24)	Mutation
	Ovarian clear cell carcinoma	Rare	57% (42)	Mutation
	Ovarian clear cell carcinoma	Rare	46% (119)	Mutation
	Endometrial cancer	Common	30% (33)	Mutation
	Transitional cell carcinoma (bladder)	Common	13% (97)	Mutation
	Hepatocellular carcinoma	Common	16.8% (125)	Mutation
	Colorectal adenocarcinoma	Hypermutated	37% (30)	Mutation
		Non-hypermutated	5% (165)	
ARID1B	Breast adenocarcinoma	Common	5% (100)	Mutation
ARID2	Hepatocellular carcinoma	Common	5.6% (125)	Mutation
	Melanoma	Common	9% (121)	Nonsense mutation
CHD1	Prostate adenocarcinoma	Common	42.9% (7)	Mutation
CHD5	Prostate adenocarcinoma	Common	42.9% (7)	Mutation
PBRM1	Clear cell renal carcinoma	Common	41% (227)	Mutation
ATRX	Pancreas neuroendocrine cancer	Rare	25% (68)	Mutation
DAXX	Pancreas neuroendocrine cancer	Rare	17.6% (68)	Mutation
SMARCD1	Breast adenocarcinoma	Common	4% (100)	Mutation
SMARCB1 (also known as *SNF5* and *INI1*)	Malignant rhabdoid cancer	Rare	100% (29)	Loss of copy or mutation
SMARCA4	Childhood medulloblastoma	Rare	4.3% (92)	Mutation

from Timp and Feinberg, 2013

Regional epigenetic changes are frequent in cancer

Epigenetic changes in cancer are not always focal, but can be global encompassing large chromosomal regions, resulting in Long Range Epigenetic Silencing (LRES).

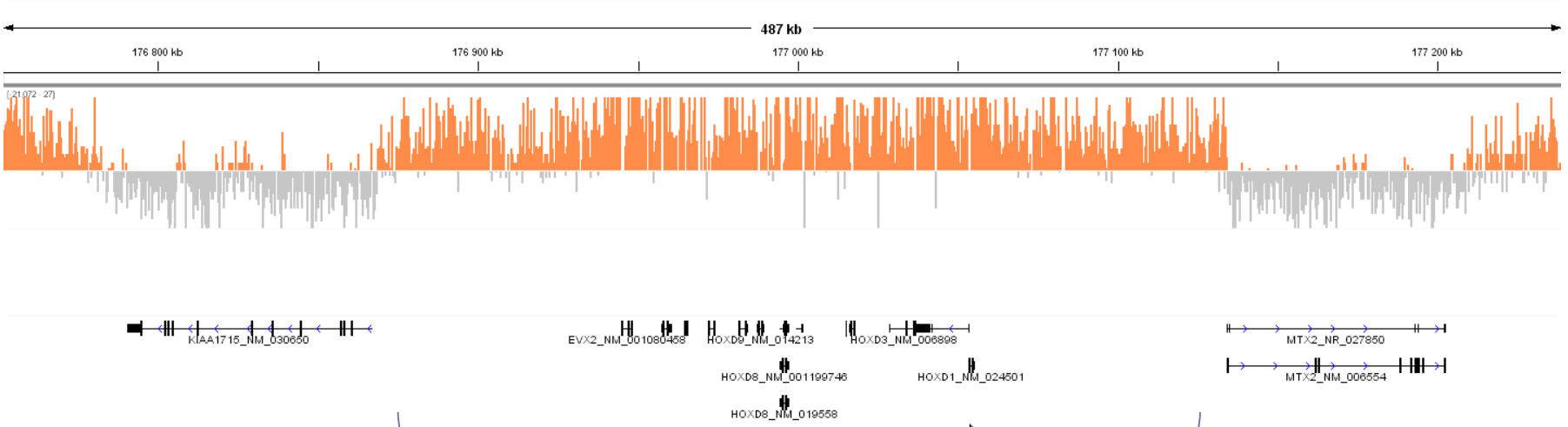
A hypothetical view of LRES in cancer



from Clark S *J Hum. Mol. Genet*. 2007;**16**:R88-R95

Example of silencing of HOXD gene cluster in bladder cancer by epigenetic mechanisms

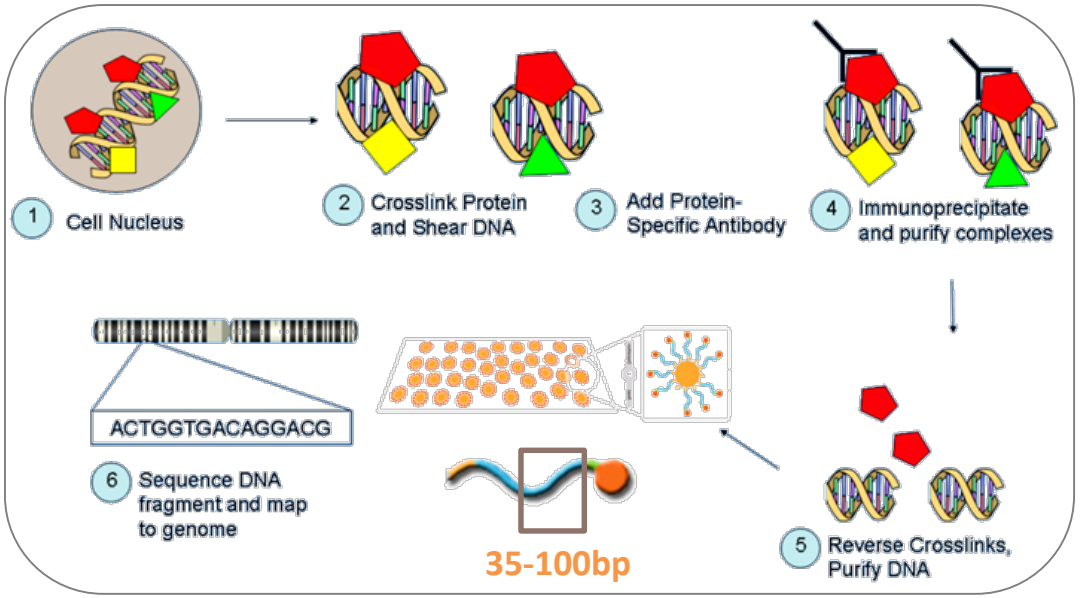
Enrichment in repressive histone mark H3K27me3



Cluster of HOXD genes repressed by epigenetic mechanisms

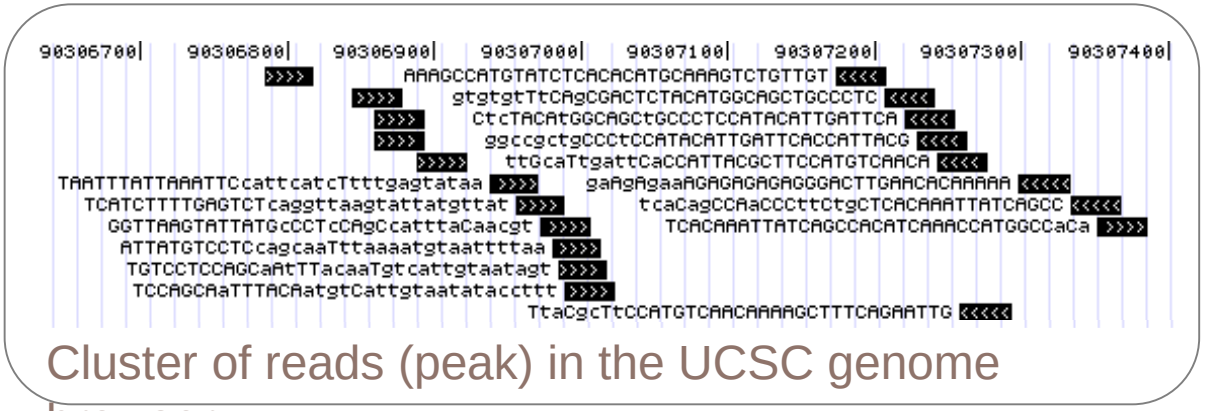
ChIP-seq technique can provide information about medication of histone tails

Mains steps of ChIP-Seq technique:



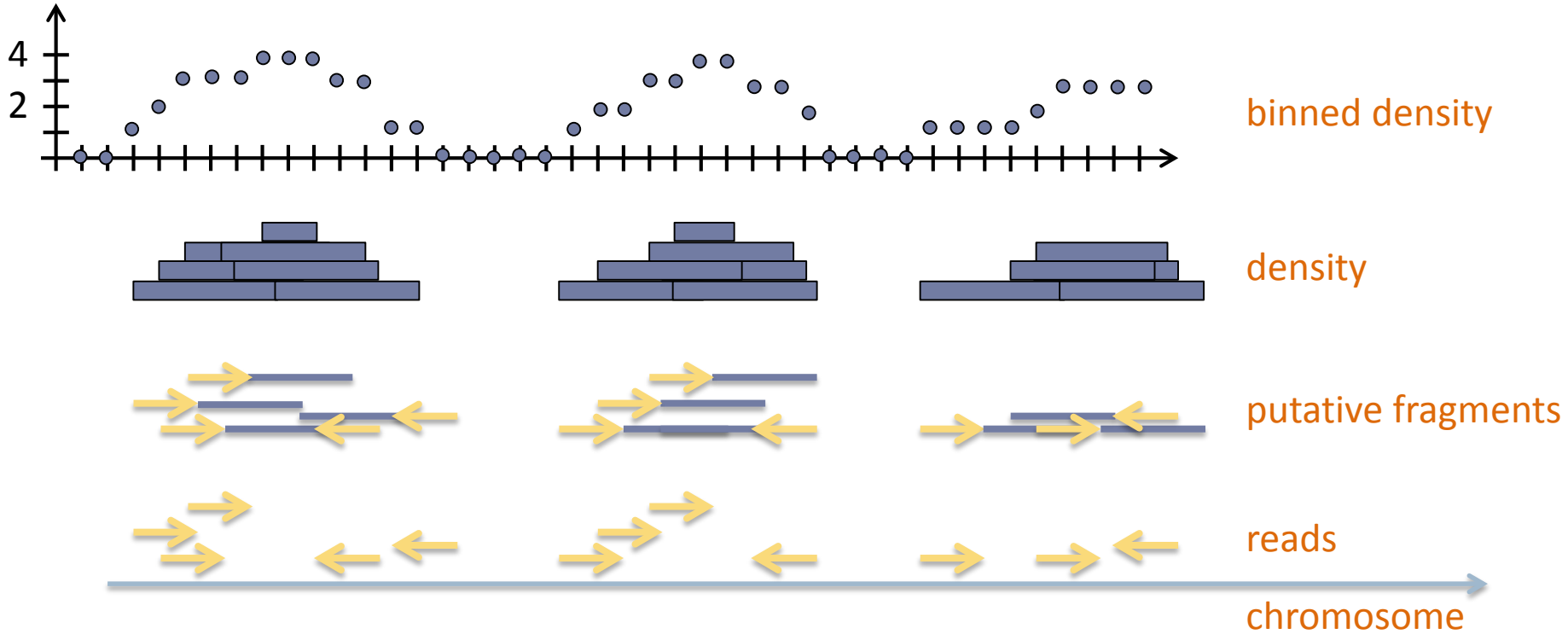
ChIP-seq = chromatin immunoprecipitation + sequencing

+ Control (e.g., input DNA)



Cluster of reads (peak) in the UCSC genome browser

Analysis of ChIP-seq data: density profile calculation and peak calling

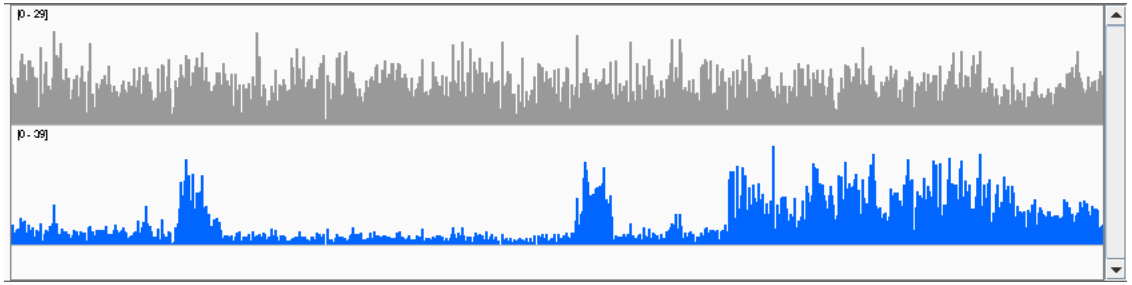


We calculate the density both for the ChIP and control sample

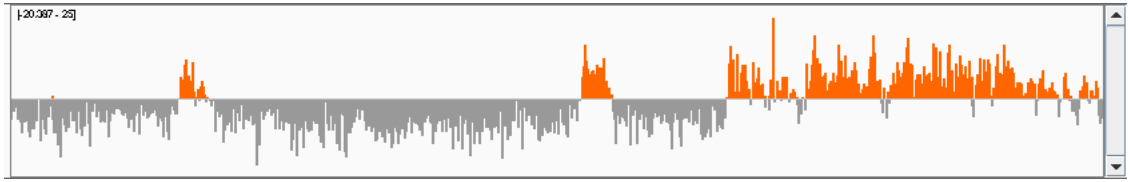
We can “subtract” density for a control sample from the ChIP sample density

Control (Input)

H3K27me3



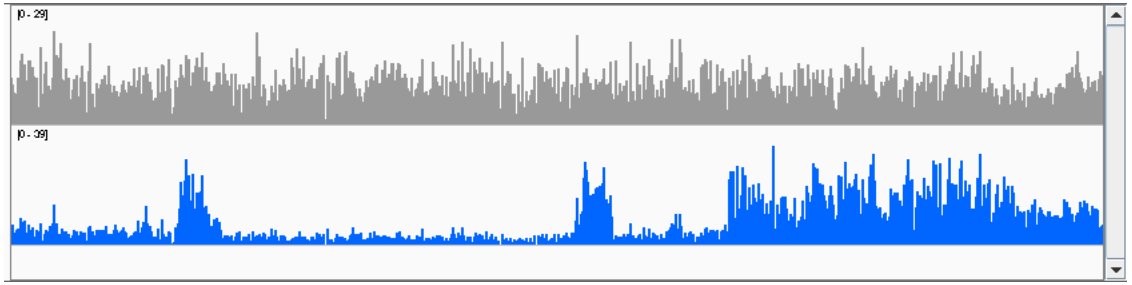
Normalized profile



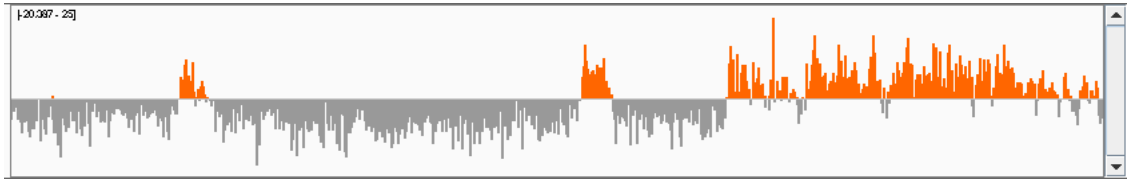
We can “subtract” density for a control sample from the ChIP sample density

Control (Input)

H3K27me3

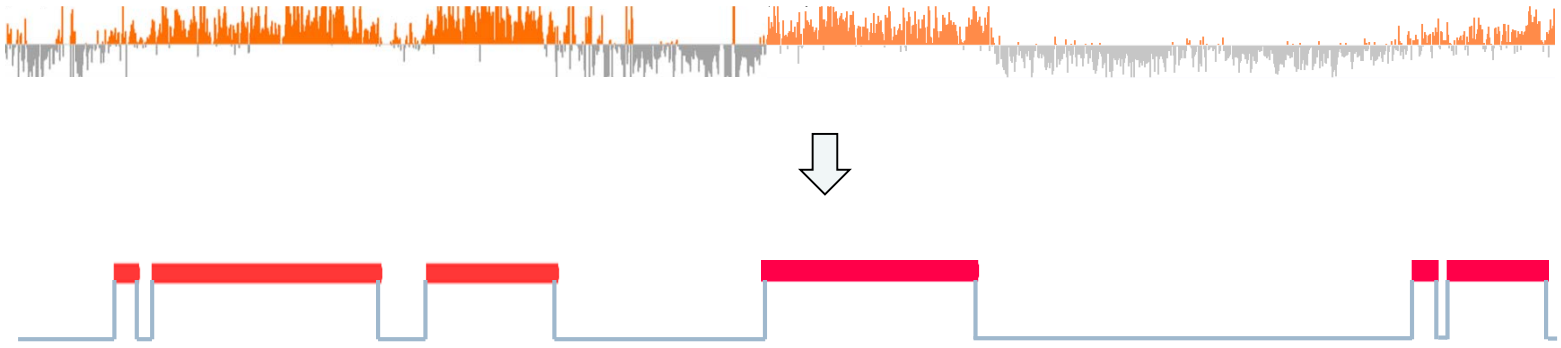


Normalized profile



We can detect enrichment regions using a number of techniques such as Hidden Markov Models, clustering and so on

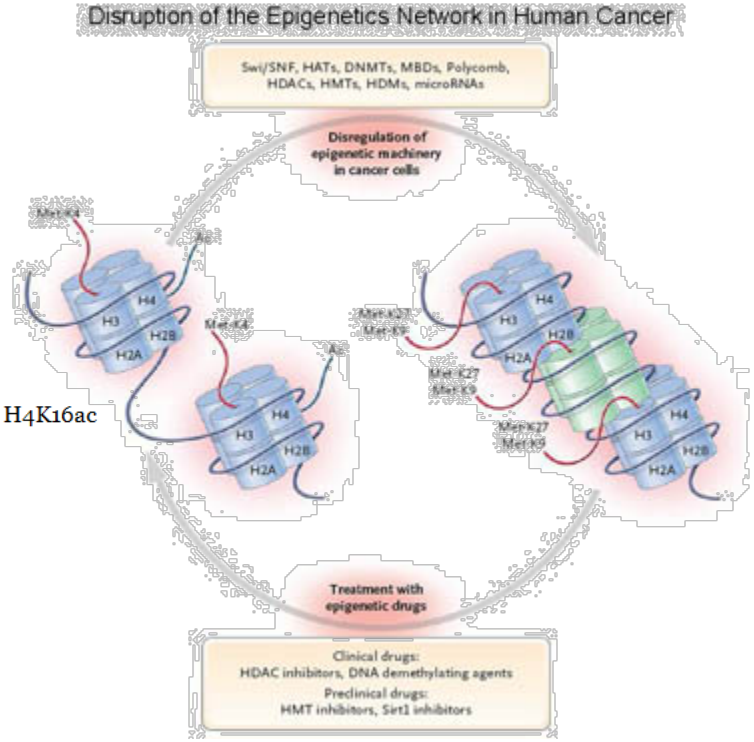
Normalized density ⇨ "Yes-No" Signal



Histone modification patterns predict prognosis in multiple cancers

Study	Cancer type	Histone Modifications
Song *et al.* 2012	Lung	H3K9ac, H3K9me3, H4K16ac
Barlési *et al.* 2007	Lung	H3K4me2, H3K9ac
Van Den Broeck *et al.* 2008	Lung	H4K5ac, H4K8ac, H4K12ac, H4K16ac, H4K20me3
Seligson *et al.* 2009	Lung	H3K4me2, H3K18ac
Seligson *et al.* 2005	Prostate	H3K4me2, H3K18ac
Ellinger *et al.* 2010b	Prostate	H3K4me1, H3K9me2, H3K9me3, H3Ac, H4Ac
Behbahani *et al.* 2012	Prostate	H4K20me1, H4K20me2
Bianco-Miotto *et al.* 2010	Prostate	H3K4me2, H3K18ac
Ellinger *et al.* 2012	Prostate	H3K27me3
Elsheikh *et al.* 2009	Breast	H3K18ac, H4K12ac, H3K4me2, H4K20me3, H4R3me2, H4K16ac
Leszinski *et al.* 2012	Breast	H3K9me3, H4K20me3
Müller-Tidow *et al.* 2010	Leukemia	H3K9me3
Park *et al.* 2008	Stomach	H3K9me3
Zhang *et al.* 2009	Stomach	H3K27me3
Tzao *et al.* 2009	Esophagus	H3K18ac, H4R3me2,H3K27me3
I *et al.* 2010	Esophagus	H3K18ac, H4R3me2
Ellinger *et al.* 2010a	Kindey	H3K4me1, H3K4me2, H3K4me3
Rogenhofer *et al.* 2012a	Kidney	H3K9me1
Rogenhofer *et al.* 2012b	Kidney	H3K27me1, H3K27me2, H3K27me3
He *et al.* 2012	Liver	H3K4me3
Cai *et al.* 2011	Liver	H3K27me3
Manuyakorn *et al.* 2010	Pancreas	H3K4me2, H3K9me2, H3K18ac

from Chervona et al., 2012



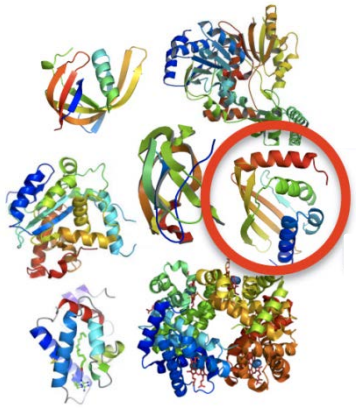
from Esteller, Nat Rev Genet, 2007

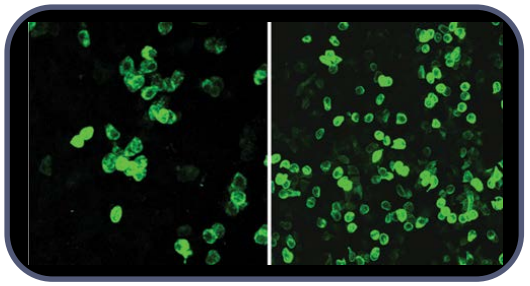
Summary

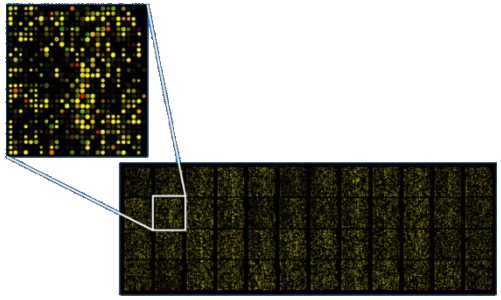
- In cancer **histone modification** profiles are often **distorted**
- The **ChIP-seq** technique can be used to detect histone modifications in cancer genomes
  - = Combination of chromatin immunoprecipitation and sequencing

Topics related to cancer bioinformatics I did not mention:

- Proteomics
  - Mass spectrometry
  - Reverse phase protein array (RPPA)
- Image analysis
- Microarray analysis (since 1995)
  - SNP arrays (copy number)
  - Gene expression arrays
  - ChIP-on-chip (DNA-protein binding)
- Small RNA (miRNA) profiling
- Data integration methods
- Network analysis







Overall Summary

- DNA-sequencing (NGS) can provide lots of information about:
  - Tumor genome structure
  - Transcriptome
  - Epigenome
- Each application has different analysis methods
- Results of this analysis are used both in clinics and research