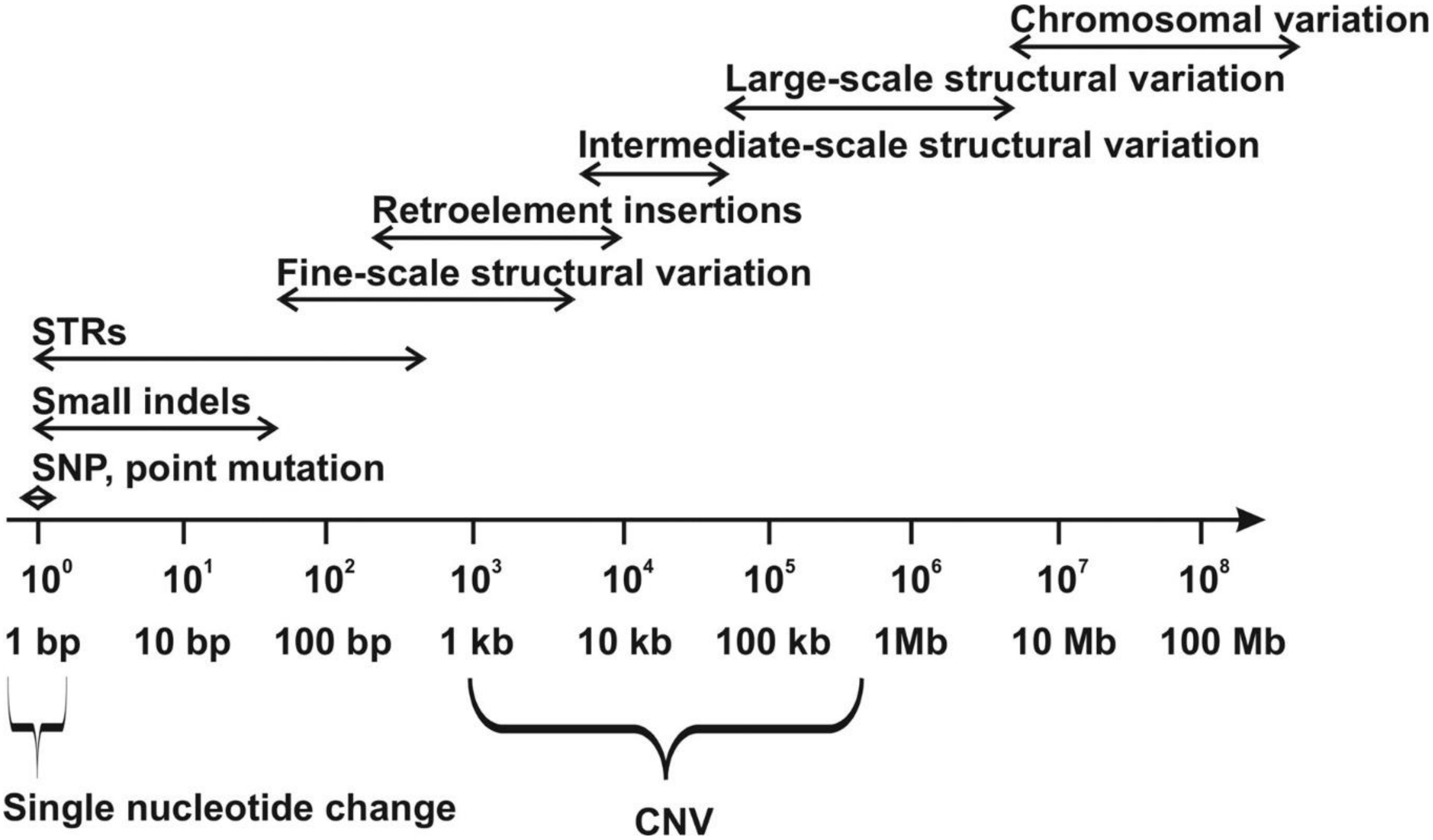
Detection of structural variants in human diseases

2019 Dragon Star Bioinformatics Course (Day 3)

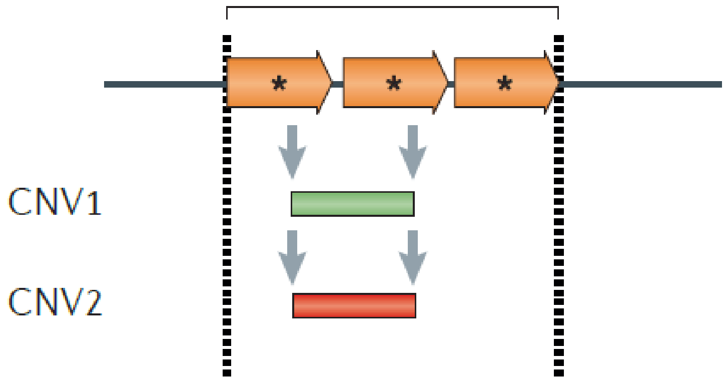
Human genetic variation



Pollex et al, Circulation. 2007

Mechanisms underlying structural variant formation

- **Recurrent structural variants:**
  - Share the same size and genomic content in unrelated individuals
  - Often caused by **NAHR** (Nonallelic homologous recombination--Nonallelic pairing of paralogous sequences and crossover leading to deletions, duplications and inversions)
- The breakpoints map within long, highly identical, flanking interspersed paralogous repeats, which mostly consist of segmental duplications(SDs)



Carvalho, C.M. and J.R. Lupski. *Nat Rev Genet, 2016.*

SVs and repeat sequences

- Approximately 50% of the human genome consists of repeat sequences.
- Different types of repeat sequences:
  - Mobile elements such as *Alu-processed pseudogenes*
  - Simple sequence repeats
  - Tandemly repeated sequences
  - Low-copy repeats (LCRs) such as SDs.
- SDs
  - Computationally defined as segments of DNA that contain ≥90% of sequence identity and ≥1 kb in length in the reference haploid genome
  - Constitute approximately 4–5% of the human genome.

Carvalho, C.M. and J.R. Lupski*. Nat Rev Genet, 2016.*

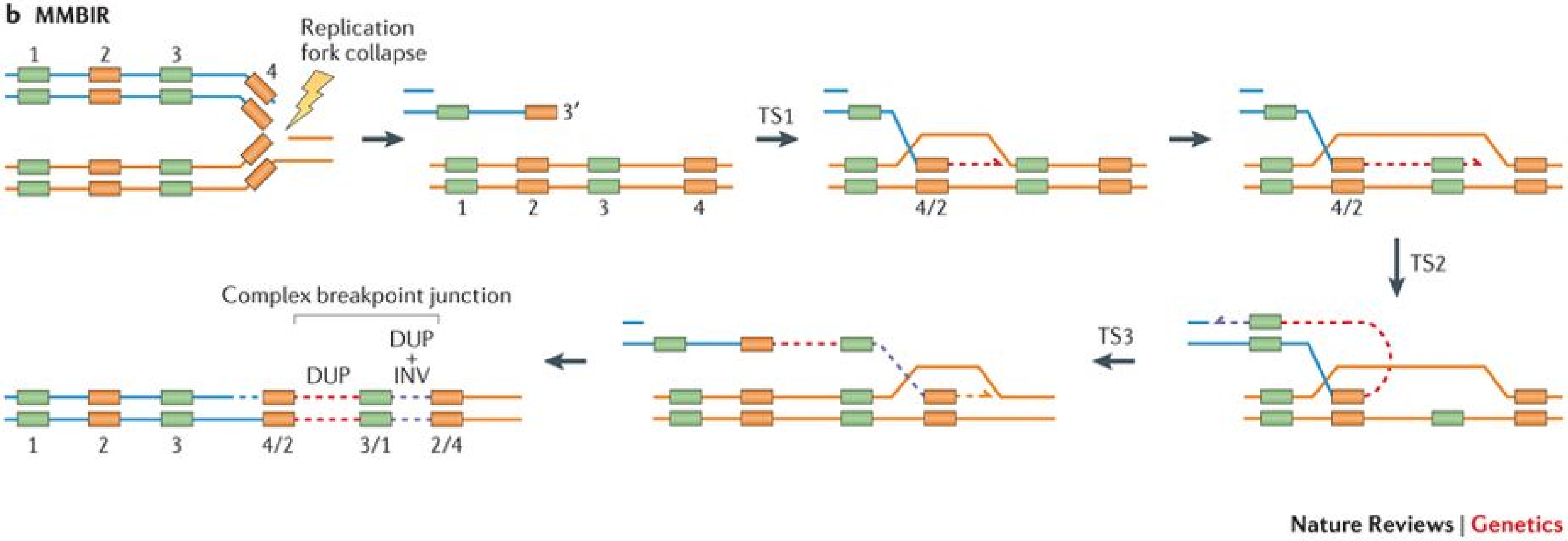
Mechanisms underlying structural variant formation

- **Nonrecurrent rearrangements:**
  - Have a unique size and genomic content at a given locus in unrelated individuals.
- Typical mechanisms:
  - NHEJ: Non-homologous end joining
  - MMEJ: Microhomology-mediated end joining
  - FoSTeS/MMBIR: microhomology-mediated break-induced replication
  - SRS: Smaller complex rearrangements caused by serial replication slippage

Escaramís G, et al. Briefings in Functional Genomics, 2015

Mechanisms underlying structural variant formation

- Microhomology-mediated break-induced replication (MMBIR)

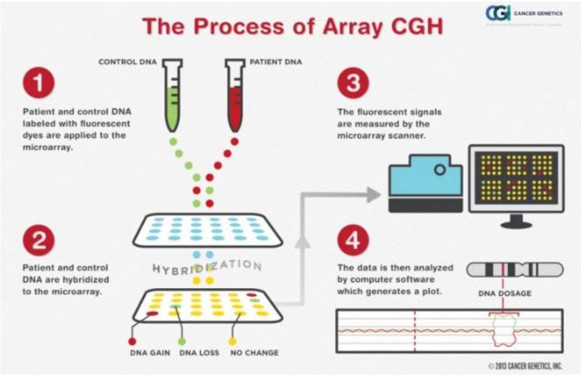

b MMBIR
1
2
3
4
Replication fork collapse
3′
1
2
3
4
TS1
4/2
4/2
TS2
TS3
Complex breakpoint junction
DUP
DUP + INV
1
2
3
4/2
3/1
2/4
Nature Reviews | Genetics

Carvalho, C.M. and J.R. Lupski. *Nat Rev Genet,* 2016.

Technologies for CNV Detection

- Karyotyping and cytogenetic analysis
- Array comparative genomic hybridization (array CGH)
- SNP microarrays (the same arrays used in GWAS)
- Next-generation sequencing (NGS) and long-read sequencing

Array CGH



The Process of Array CGH

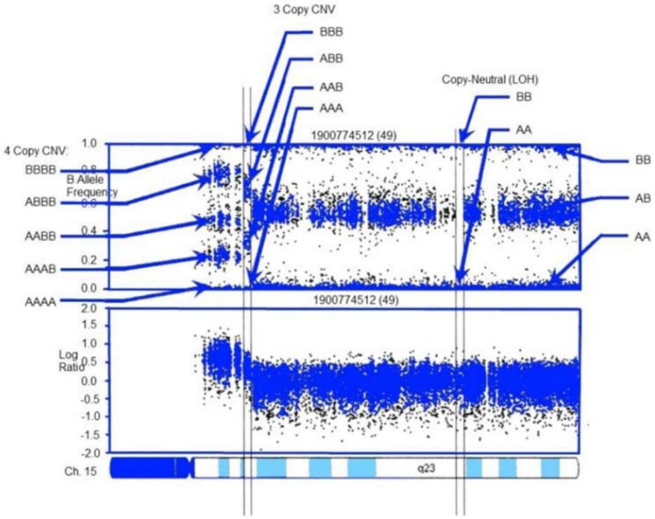
1 Patient and control DNA labeled with fluorescent dyes are applied to the microarray.

2 Patient and control DNA are hybridized to the microarray.

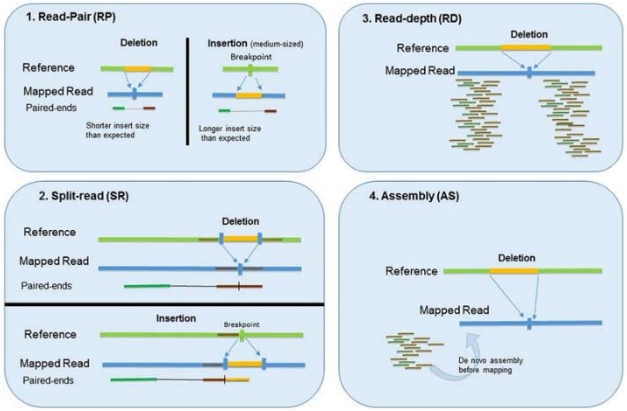
3 The fluorescent signals are measured by the microarray scanner.

4 The data is then analyzed by computer software which generates a plot.

SNP array



Next-generation sequencing

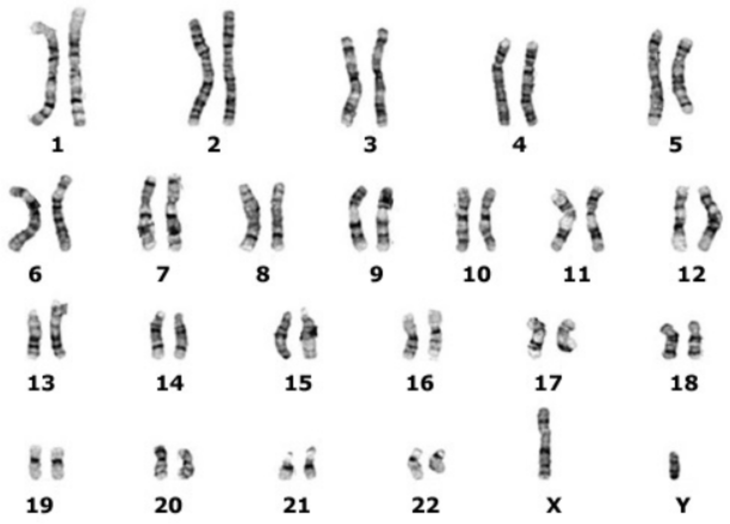


Commonly used cytogenetic techniques

- Giemsa staining
- Fluorescent in situ hybridization (FISH)
- Comparative genomic hybridization (CGH)
- Spectral karyotyping (SKY)

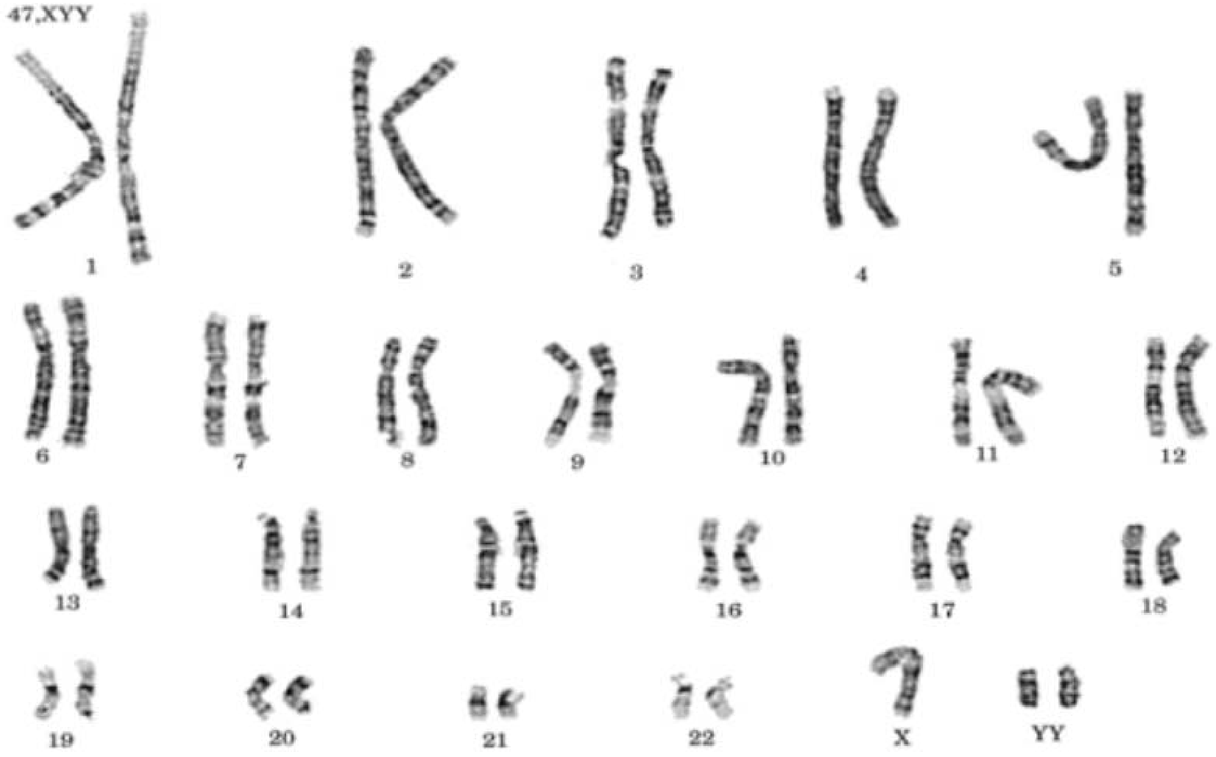
Cytogenetic techniques

- Giemsa banding (G-banding).
- The metaphase chromosomes are treated with trypsin (to digest proteins in the chromosomes) and stained with Giemsa stain.
- Dark bands are AT-rich and have less genes.
- Light bands are GC-rich DNA and are more transcriptionally active.



Karyogram of human male using Giemsa staining

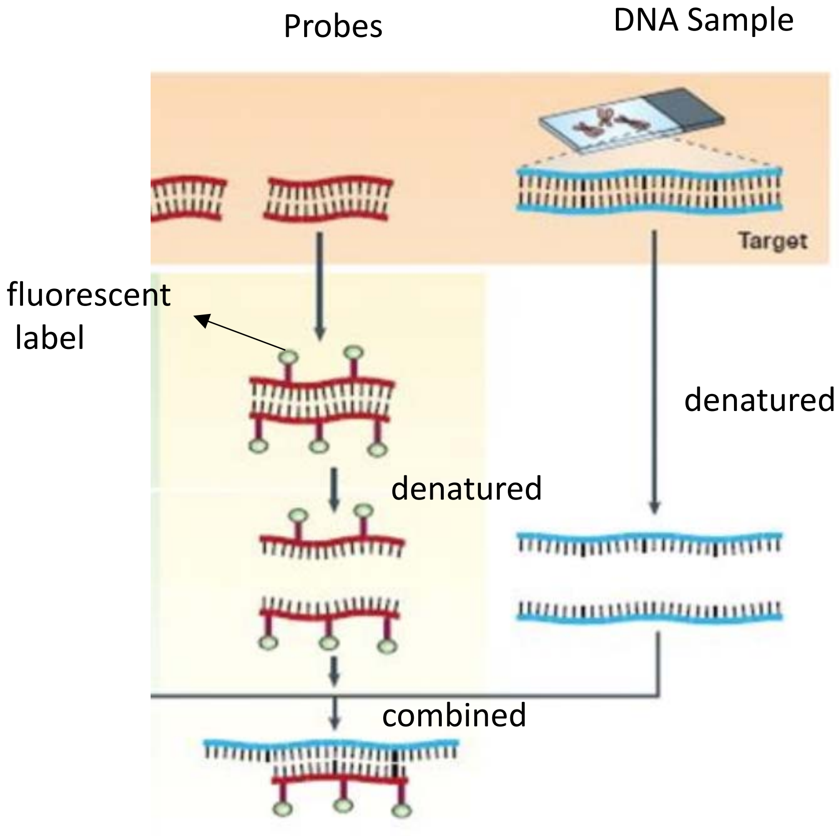
Detection of chromosomal abnormalities by cytogenetic techniques



Polipalli, J Clin Diagn Res. 2016

Cytogenetic techniques

- Fluorescent in situ hybridization (FISH).
  - FISH uses fluorescent probes that bind to specific chromosomal regions where there is a high degree of sequence complementarity.
- Fluorescence microscopy can be used to visualize and evaluate the signals.


Probes

DNA Sample

fluorescent label

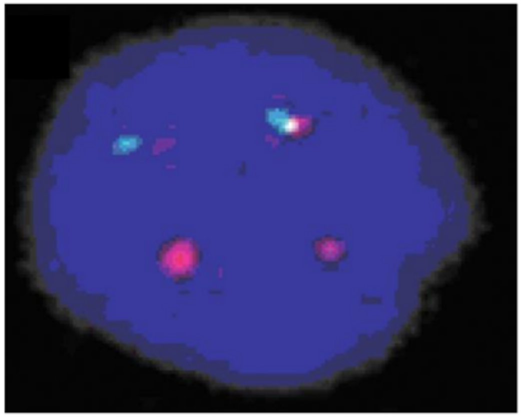
denatured

denatured

combined

Bioscience Horizons: The International Journal of Student Research, Volume 3, Issue 1, 1 March 2010, Pages 85–95,

Detection of SVs using cytogenetic techniques (FISH)



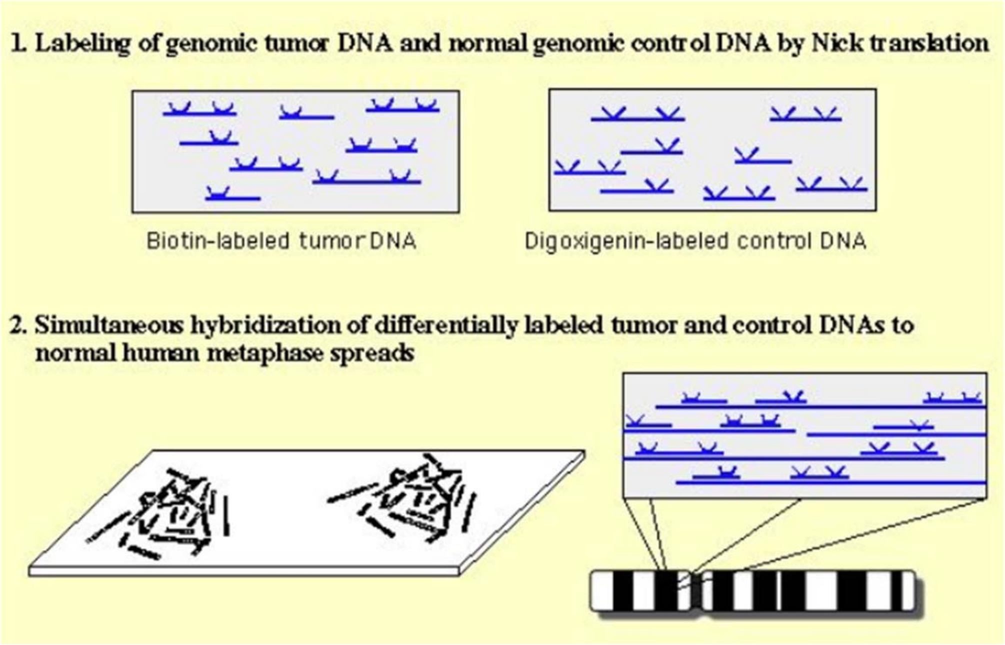
Using interphase FISH to detect the *BCR/ABL* translocation.

Green signal indicates the presence of the *BCR* gene
Red signals indicate the presence of the *ABL* gene
Red-green fusion (yellow) signal confirms *BCR/ABL* translocation.

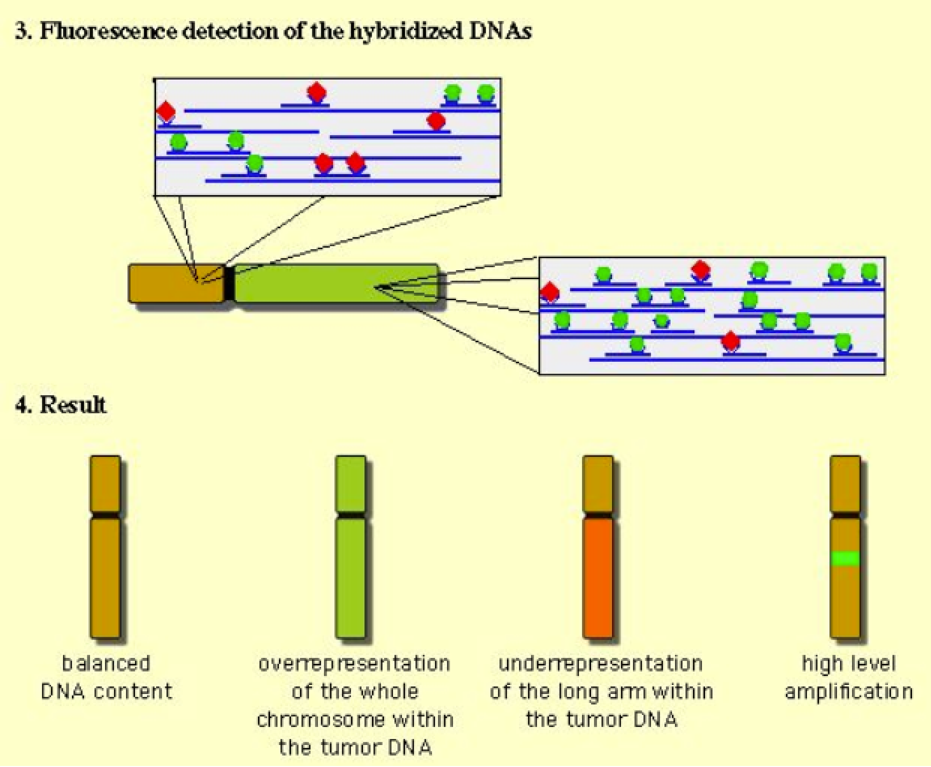
Bioscience Horizons: The International Journal of Student Research,
Volume 3, Issue 1, 1 March 2010, Pages 85–95,

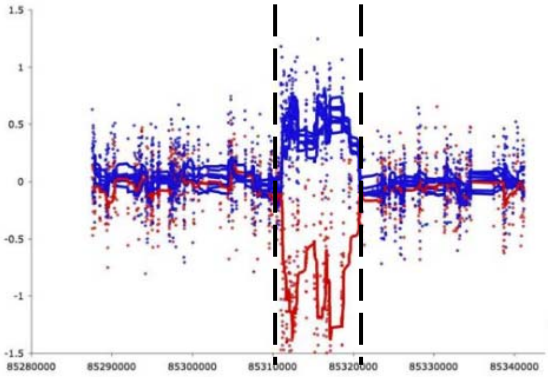
Comparative genomic hybridization (CGH)

- A molecular cytogenetic method for detection of copy number variations (CNVs)
  - A reference sample is used as a control.


1. Labeling of genomic tumor DNA and normal genomic control DNA by Nick translation
Biotin-labeled tumor DNA
Digoxigenin-labeled control DNA
2. Simultaneous hybridization of differentially labeled tumor and control DNAs to normal human metaphase spreads

Comparative genomic hybridization (CGH)


3. Fluorescence detection of the hybridized DNAs
4. Result
balanced DNA content
overrepresentation of the whole chromosome within the tumor DNA
underrepresentation of the long arm within the tumor DNA
high level amplification

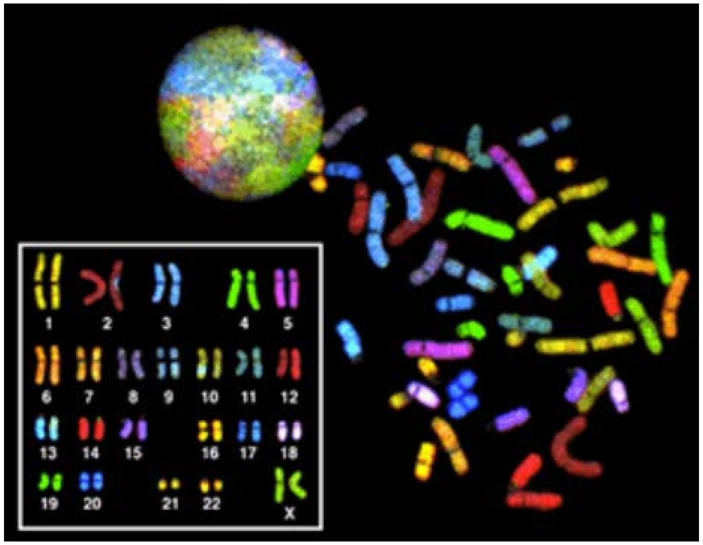


Blue line: individuals with two copies
Red line: individual with zero copy

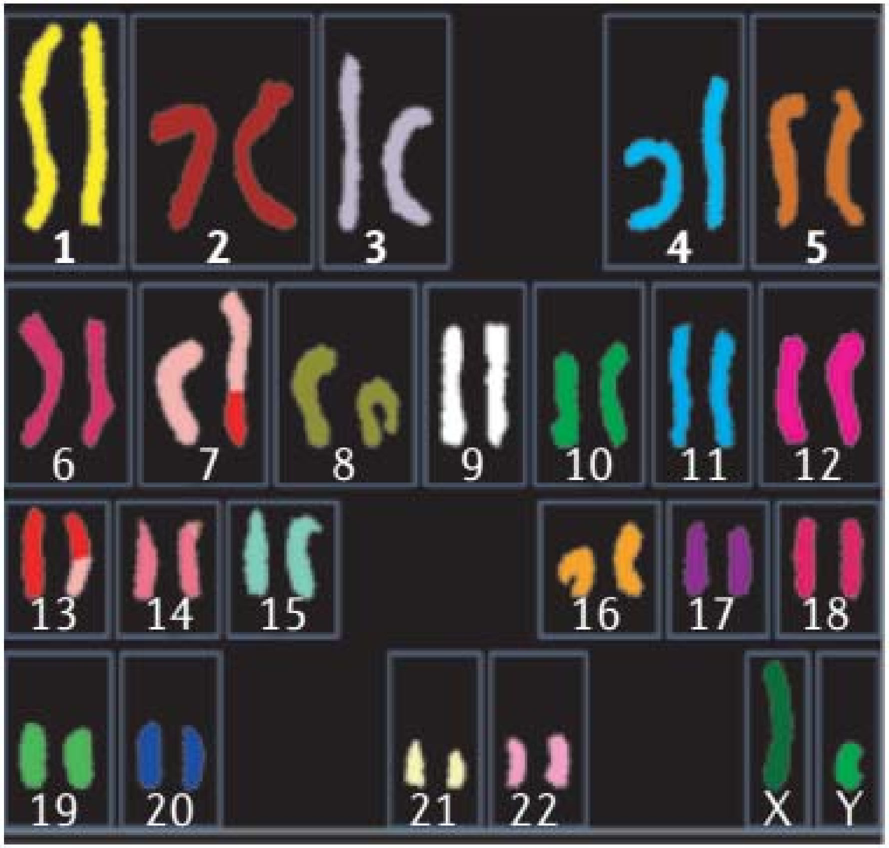
Spectral karyotyping (SKY)

- Spectral karyotyping (SKY) is a laboratory technique
  - Allows the visualization of all the human chromosomes at one time by "painting" each pair of chromosomes in a different fluorescent color.

SKY also uses fluorescent probes. Each probe is complementary to a unique region of one chromosome. The probes that bind to different chromosomes are designed to have different fluorescent color.



Detection of interchromosomal translocations using cytogenetic techniques (SKY)



The example shows the detection of a t(7;13) translocation.

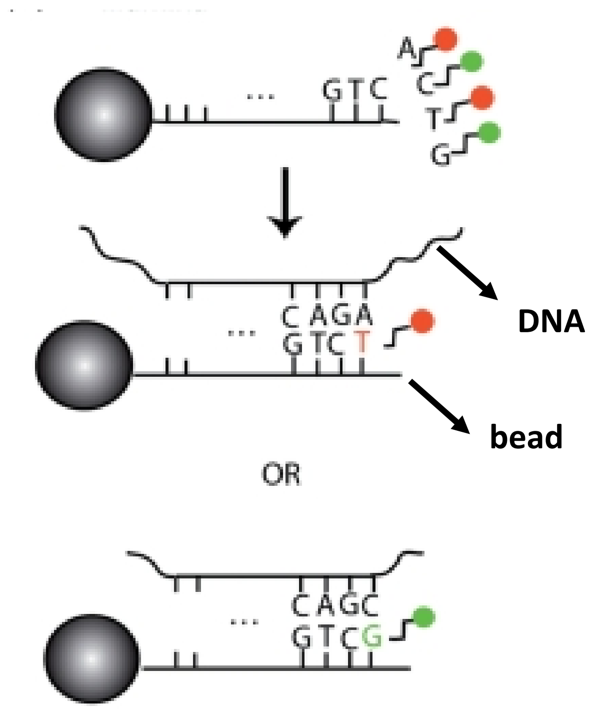
Feuk L. et al. Nature Reviews Genetics. 2006

SNP genotyping arrays

- SNP genotyping array is a type of DNA microarray which is used to detect SNPs.

- Two major SNP array companies:
  - Affymetrix arrays
  - Illumina arrays

Illumina SNP array technology

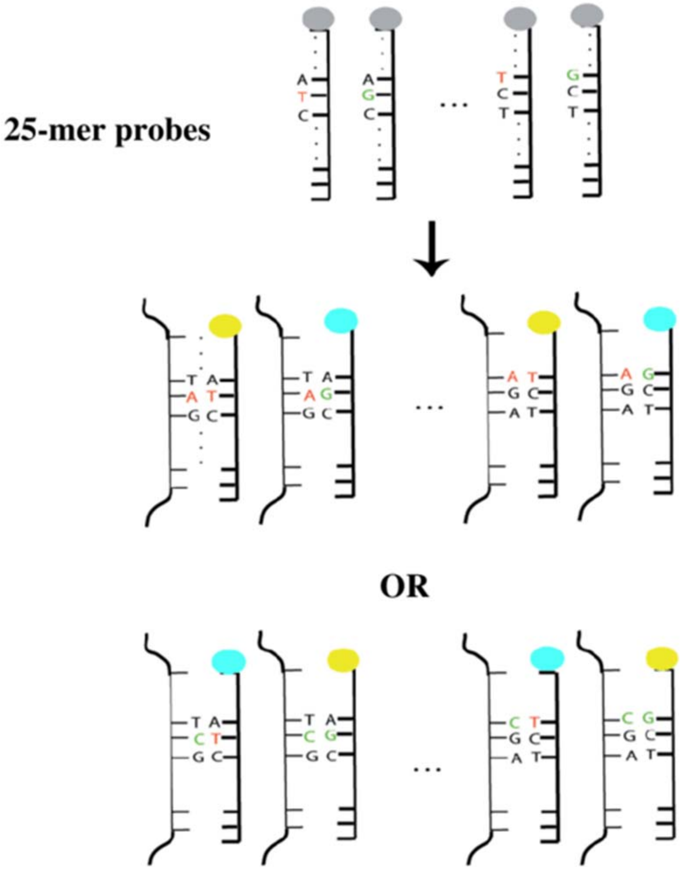


Nucleic Acids Res. 2009 Jul; 37(13): 4181–4193.

- In the Illumina array, attached to each Illumina bead is a 50-mer sequence complementary to the sequence adjacent to the SNP site.
- The single-base extension (T or G) that is complementary to the allele carried by the DNA (A or C, respectively) then binds and results in the appropriately-colored signal (red or green, respectively).

Affymetrix SNP array technology

- In the Affymetrix assay, there are 25-mer probes for both alleles.
    - Assuming there are two alleles (e.g. A-Allele and B-allele) at a particular site.
    - The DNA can bind to both probes
    - But will have much higher affinity for the perfectly matched probe.
    - For example,
        - if the DNA is B-allele, it will binds to both probes
        - But have much higher binding affinity to the probe of B-allele.
        - Therefore, the signal of B-allele probe is much higher than A-allele probe.

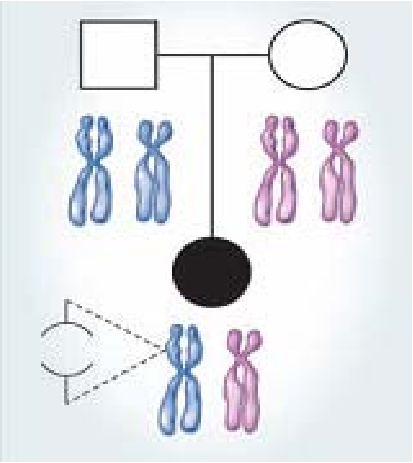


complementary (yellow)
mismatch site (blue)

LaFramboise et al, Nucleic Acids Res, 2009

CNV Detection

- There is a need to develop a high-resolution CNV detection algorithm using high-density SNP genotyping data:
  - Identify location of the CNVs.
  - Estimate the copy numbers.
  - Model family relationships.
  - Incorporate *de novo* events.



Log R Ratio (LRR) and B Allele Frequency (BAF)

- For both platforms, the computational algorithms convert the raw signals into Log R Ratio (LRR) and B Allele Frequency (BAF).
- LRR is a measure of normalized total signal intensity.
- BAF is a measure of normalized allelic intensity ratio.

- The combination of LRR and BAF can be used together to determine different copy numbers and to differentiate copy-neutral LOH regions from normal copy regions.

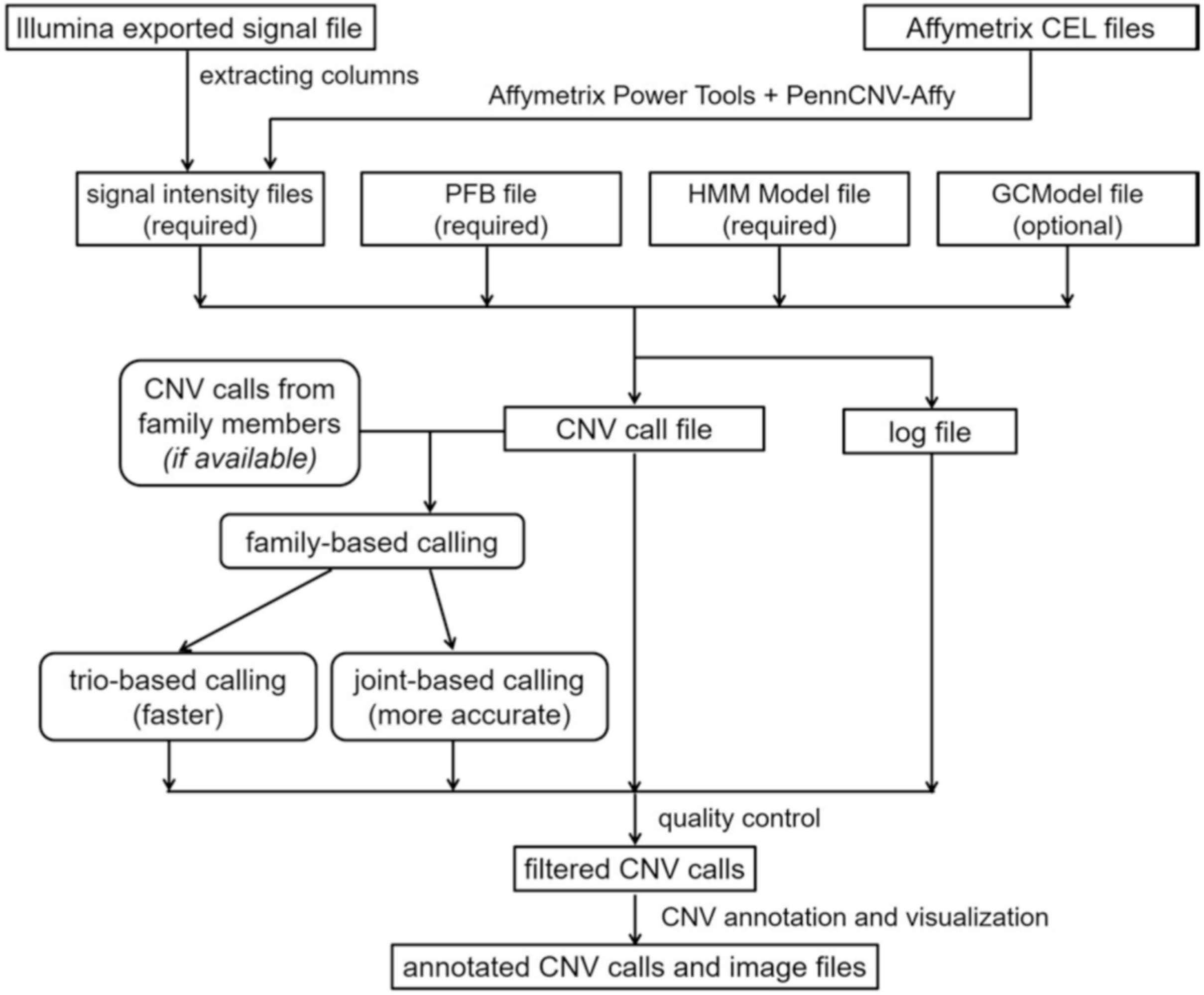
Detection of CNVs from SNP arrays using PennCNV

PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data

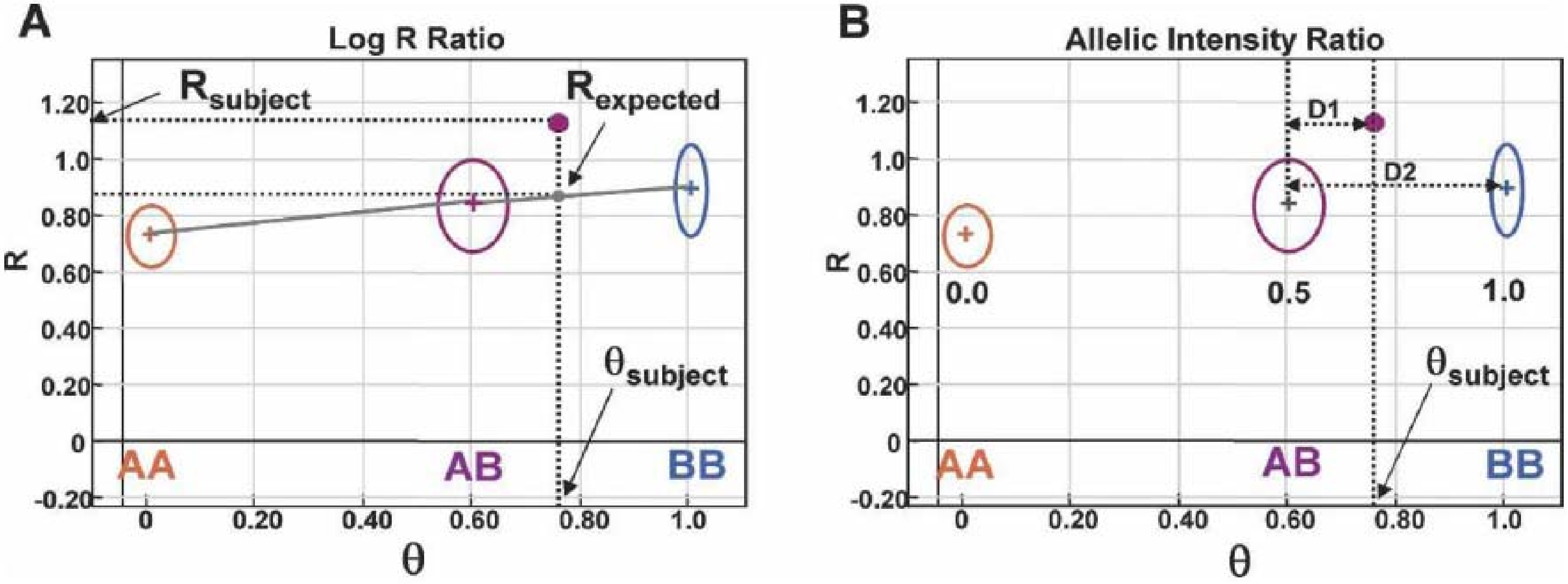
Kai Wang,[1] Mingyao Li,[2] Dexter Hadley,[1,3] Rui Liu,[1] Joseph Glessner,[4] Struan F.A. Grant,[4] Hakon Hakonarson,[4] and Maja Bucan[1,5]

- Hidden Markov Model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with hidden states.
- What we know are: LRR and BAF
- What we want to know is: copy number

PennCNV Flowchart


Illumina exported signal file
extracting columns
Affymetrix CEL files
Affymetrix Power Tools + PennCNV-Affy
signal intensity files (required)
PFB file (required)
HMM Model file (required)
GCModel file (optional)
CNV calls from family members *(if available)*
CNV call file
log file
family-based calling
trio-based calling (faster)
joint-based calling (more accurate)
quality control
filtered CNV calls
CNV annotation and visualization
annotated CNV calls and image files

SNP Signal Intensities



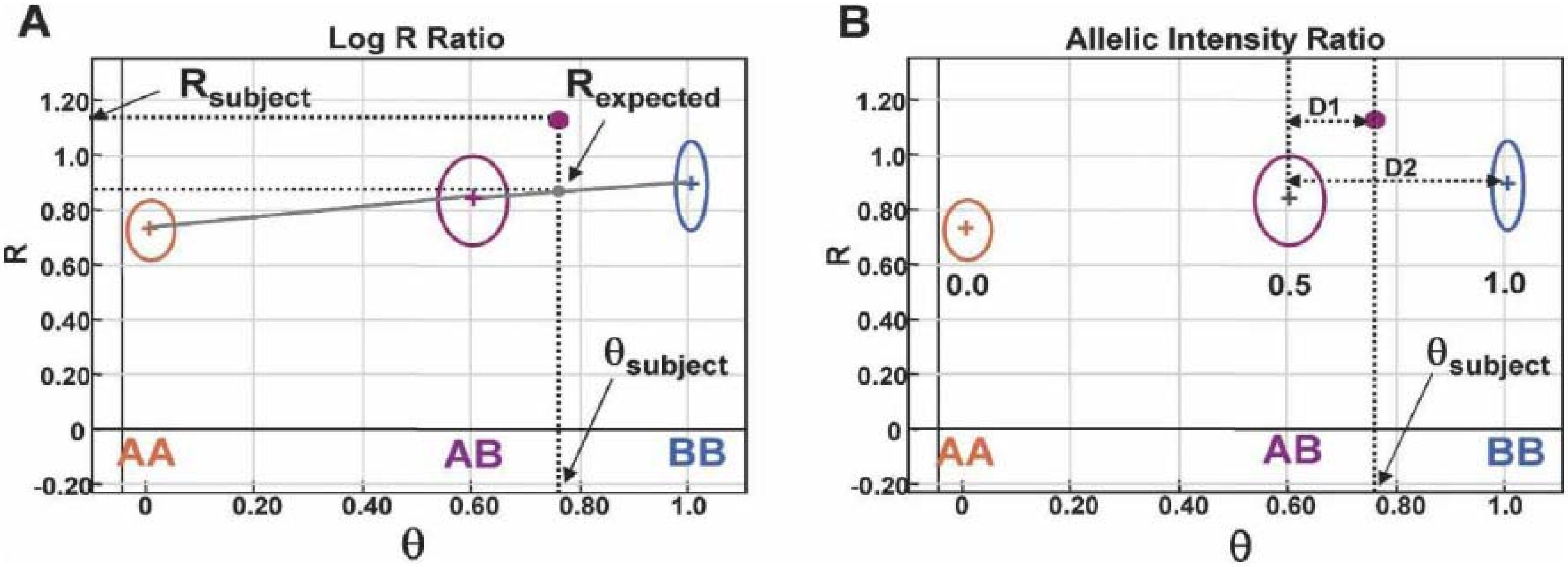
$$R = X_A + X_B, \quad \theta = (2/\pi) \times \arctan(X_B/X_A)$$

$$LRR = \log_2 \; R_{subject}/R_{expected}$$

X_A and X_B: normalized signal intensities for alleles A and B

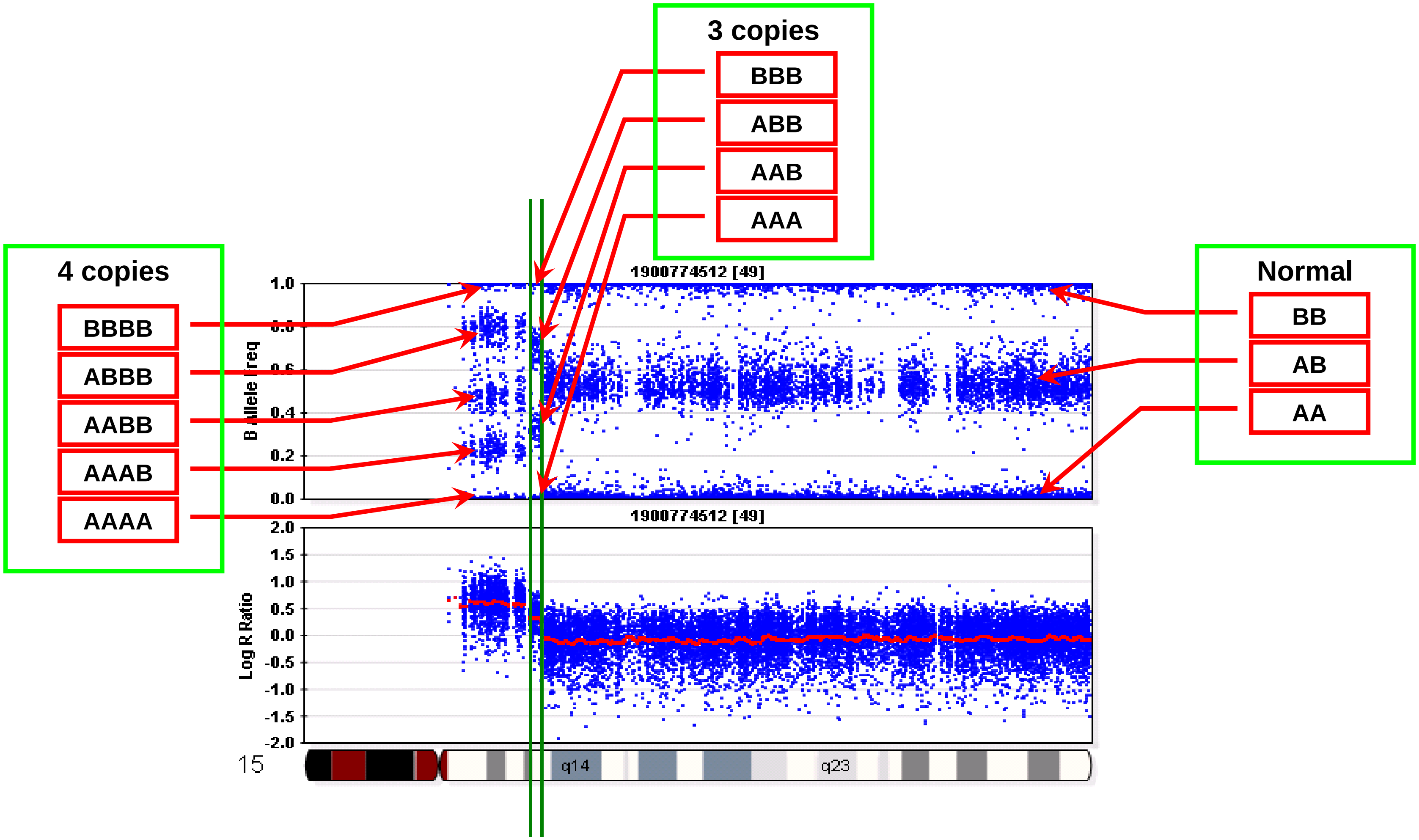
$R_{expected}$: calculated based on a reference dataset assuming copy number = 2

SNP Signal Intensities

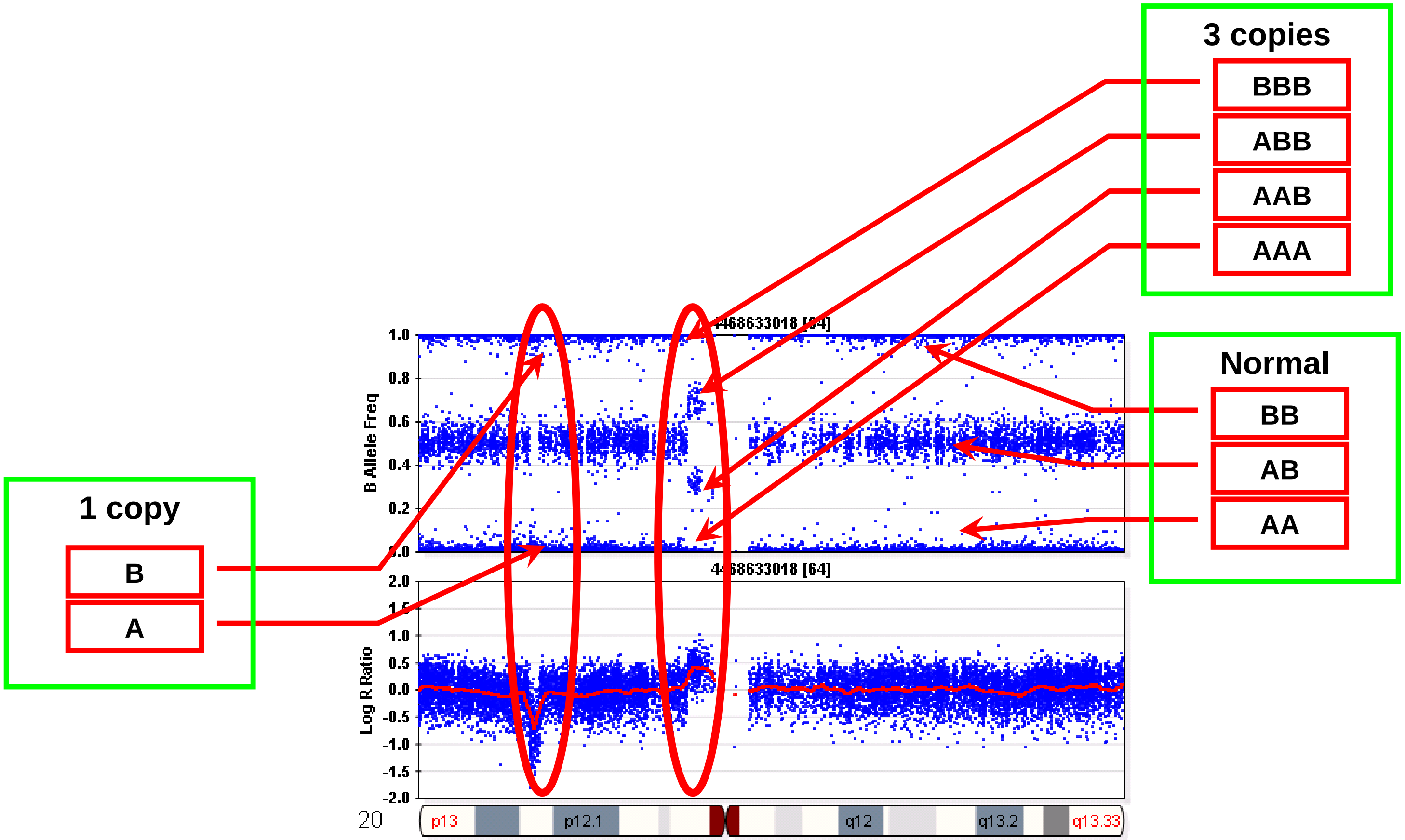


$$BAF = \begin{cases} 0, & \text{if } \theta_{\text{subject}} < \theta_{AA} \\ 0.5(\theta_{\text{subject}} - \theta_{AA})/(\theta_{AB} - \theta_{AA}), & \text{if } \theta_{AA} \le \theta_{\text{subject}} \le \theta_{AB} \\ 0.5 + 0.5(\theta_{\text{subject}} - \theta_{AB})/(\theta_{BB} - \theta_{AB}), & \text{if } \theta_{AB} \le \theta_{\text{subject}} \le \theta_{BB} \\ 1, & \text{if } \theta_{\text{subject}} > \theta_{BB} \end{cases}$$

Visualization of CNVs



Visualization of CNVs



1 copy

B

A

3 copies

BBB

ABB

AAB

AAA

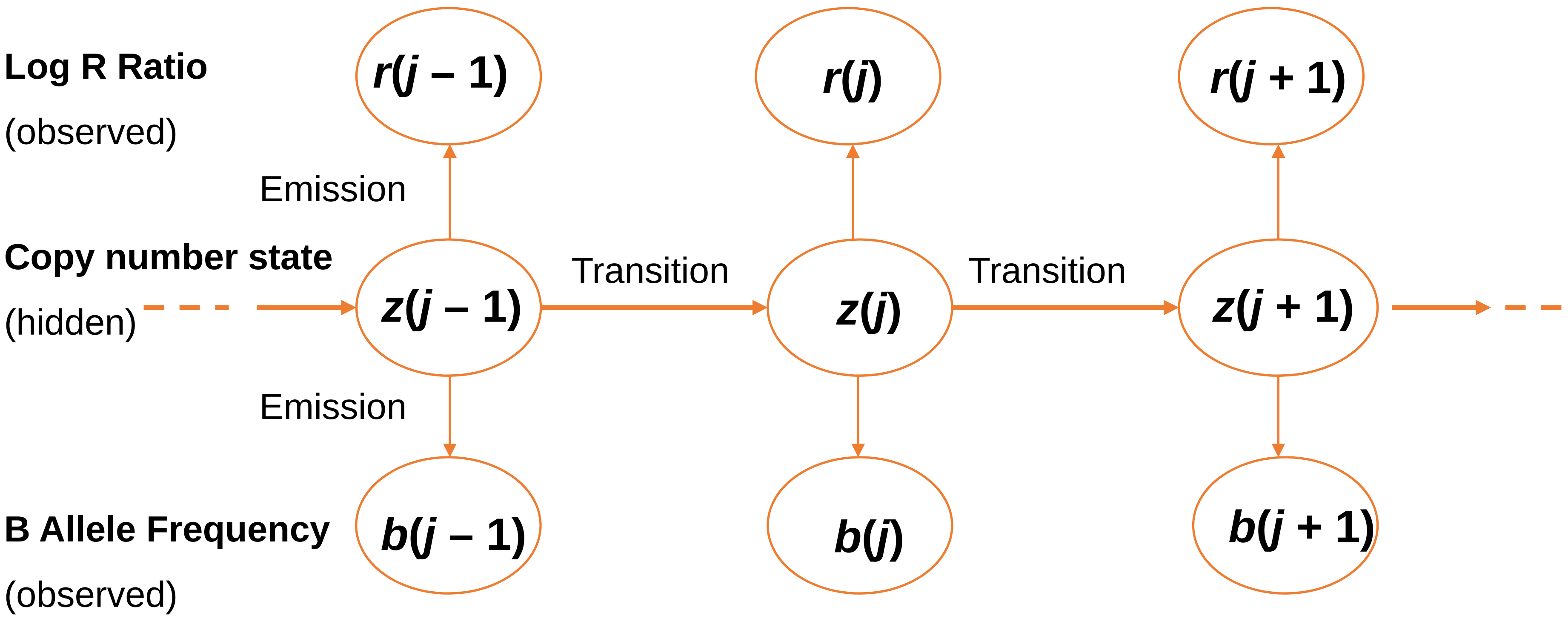
Normal

BB

AB

AA

Hidden Markov Model in PennCNV



Log R Ratio

(observed)

Copy number state

(hidden)

B Allele Frequency

(observed)

Copy Number States

6 States:

- State1: CNV=0 (double deletions)
- State2: CNV=1 (single deletion)
- State3: CNV=2 (normal)
- State4: CNV=2 (normal with LOH)
- State5: CNV=3 (single duplication)
- State6: CNV=4 (double duplications)

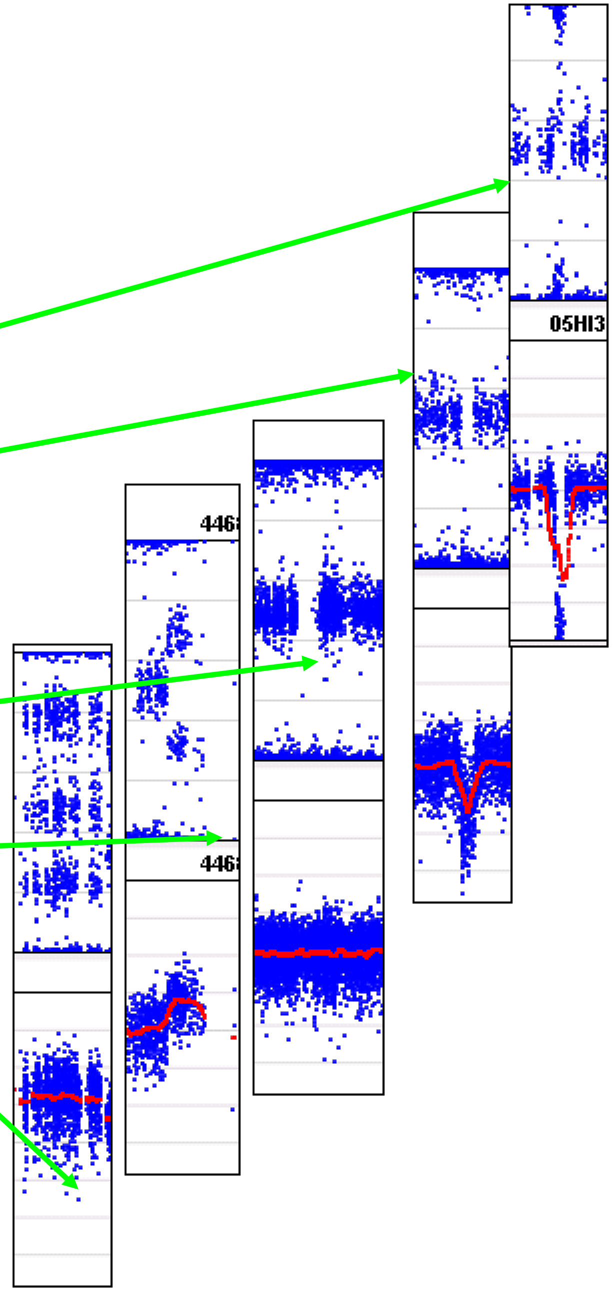


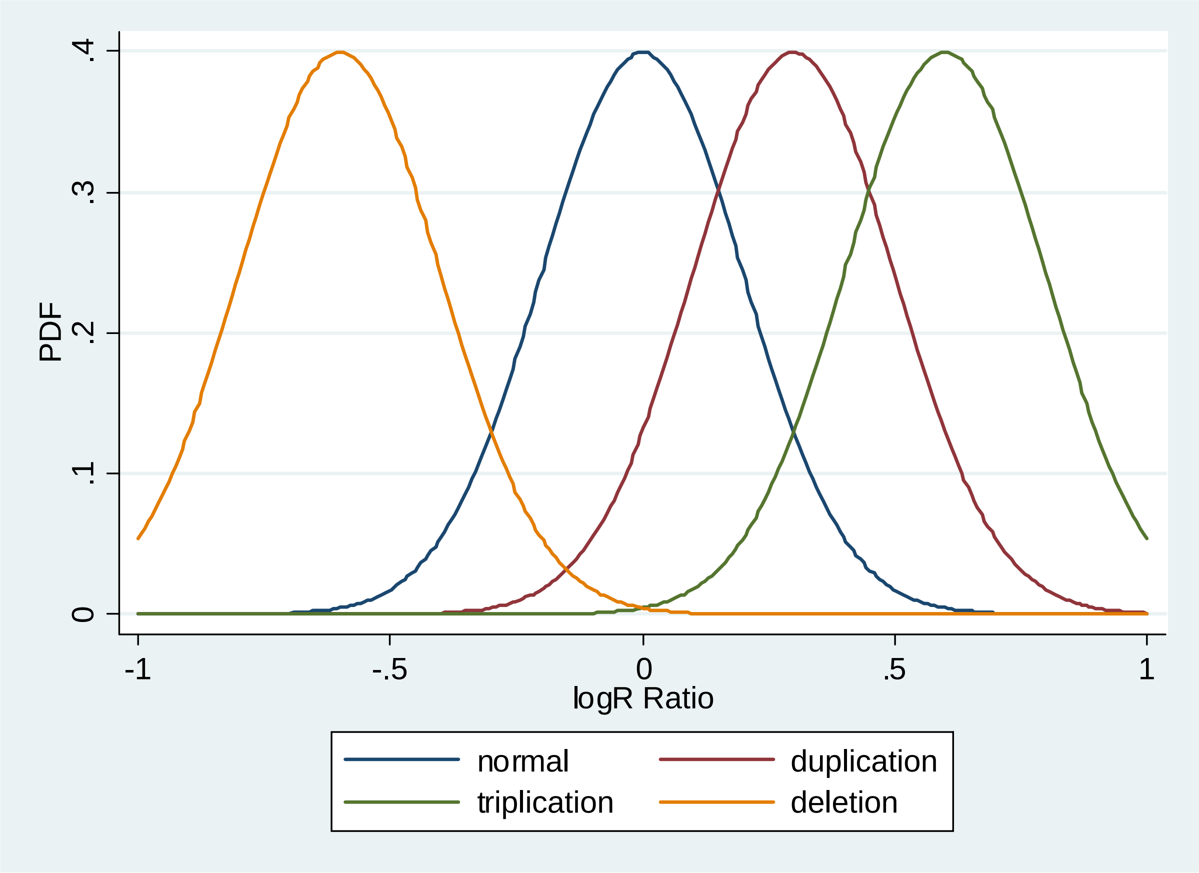
Table 1. Hidden states, copy numbers, and their descriptions

Copy no. state	Total copy no.	Description (for autosome)	CNV genotypes
1	0	Deletion of two copies	Null
2	1	Deletion of one copy	A, B
3	2	Normal state	AA, AB, BB
4	2	Copy-neutral with LOH	AA, BB
5	3	Single copy duplication	AAA, AAB, ABB, BBB
6	4	Double copy duplication	AAAA, AAAB, AABB, ABBB, BBBB

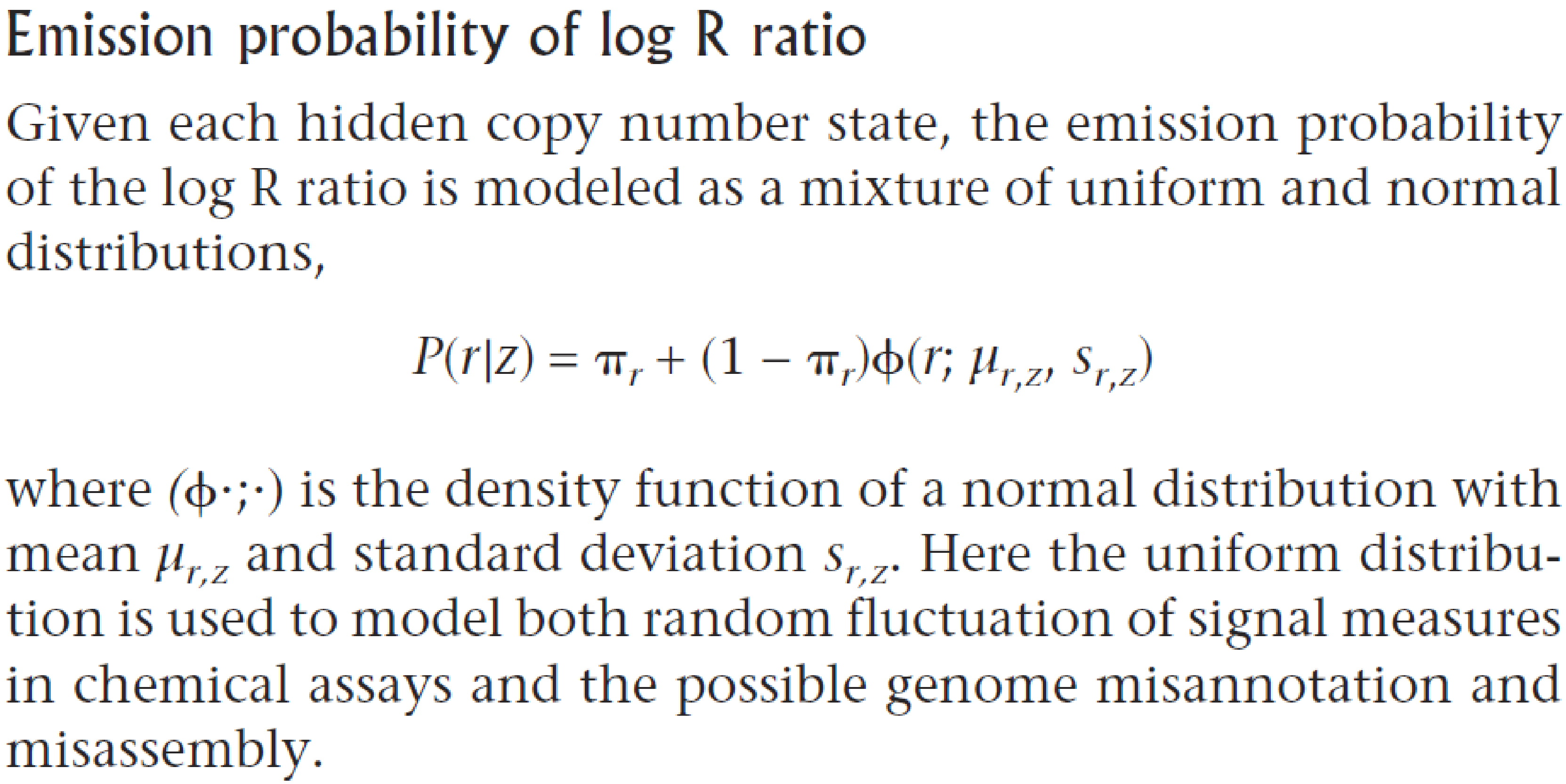
Each state has a different distribution of CNV genotypes.

Emission Probability of LRR

- Given a copy number state, LRR is normally distributed



Emission Probability of LRR



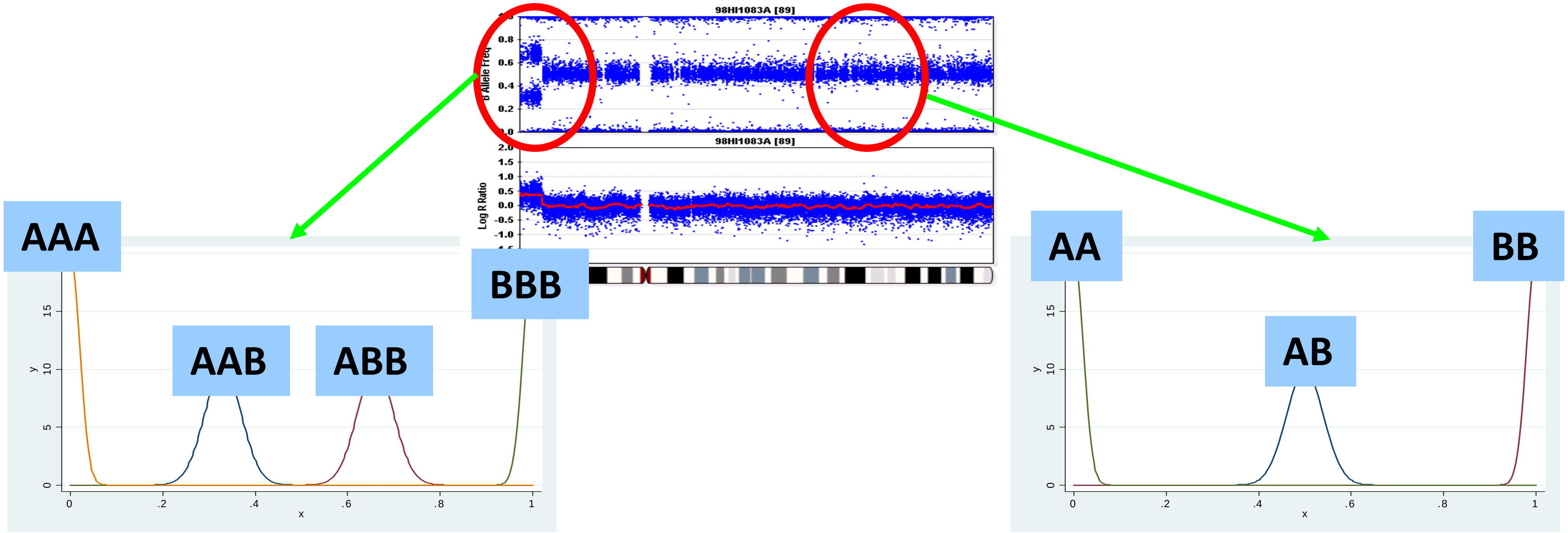
Emission probability of log R ratio

Given each hidden copy number state, the emission probability of the log R ratio is modeled as a mixture of uniform and normal distributions,

$$P(r|z) = \pi_r + (1 - \pi_r)\phi(r; \mu_{r,z}, s_{r,z})$$

where $(\phi\cdot;\cdot)$ is the density function of a normal distribution with mean $\mu_{r,z}$ and standard deviation $s_{r,z}$. Here the uniform distribution is used to model both random fluctuation of signal measures in chemical assays and the possible genome misannotation and misassembly.

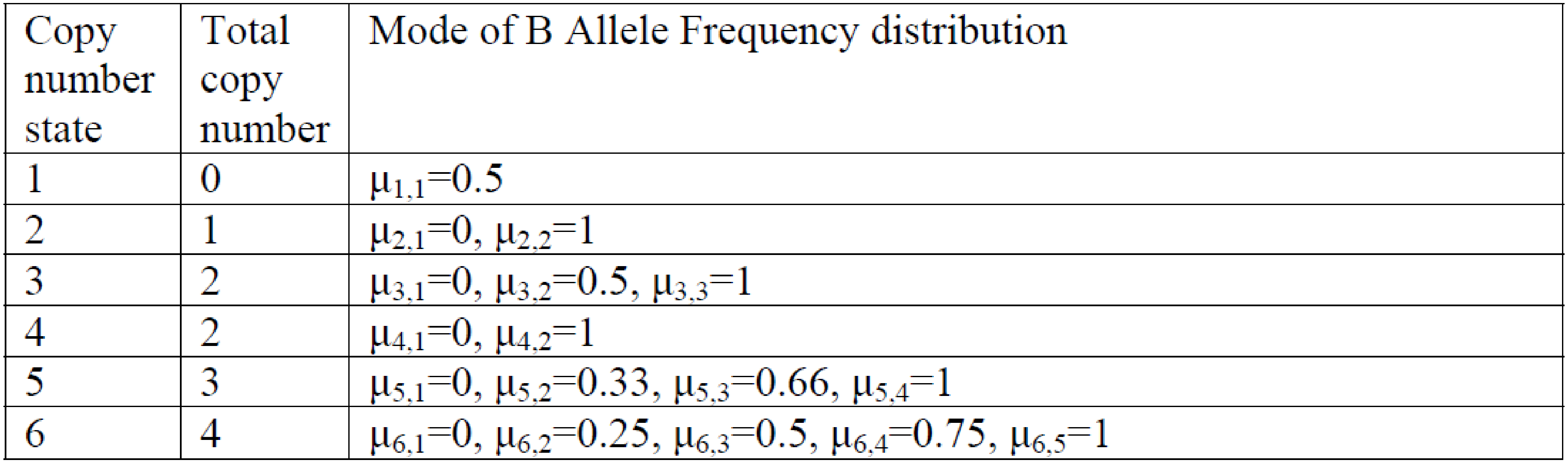
Emission Probability of BAF



$$P(b_j \mid z_j, \lambda) = \sum_g P(b_j \mid g, z_j, \lambda) P(g \mid z_j, \lambda)$$

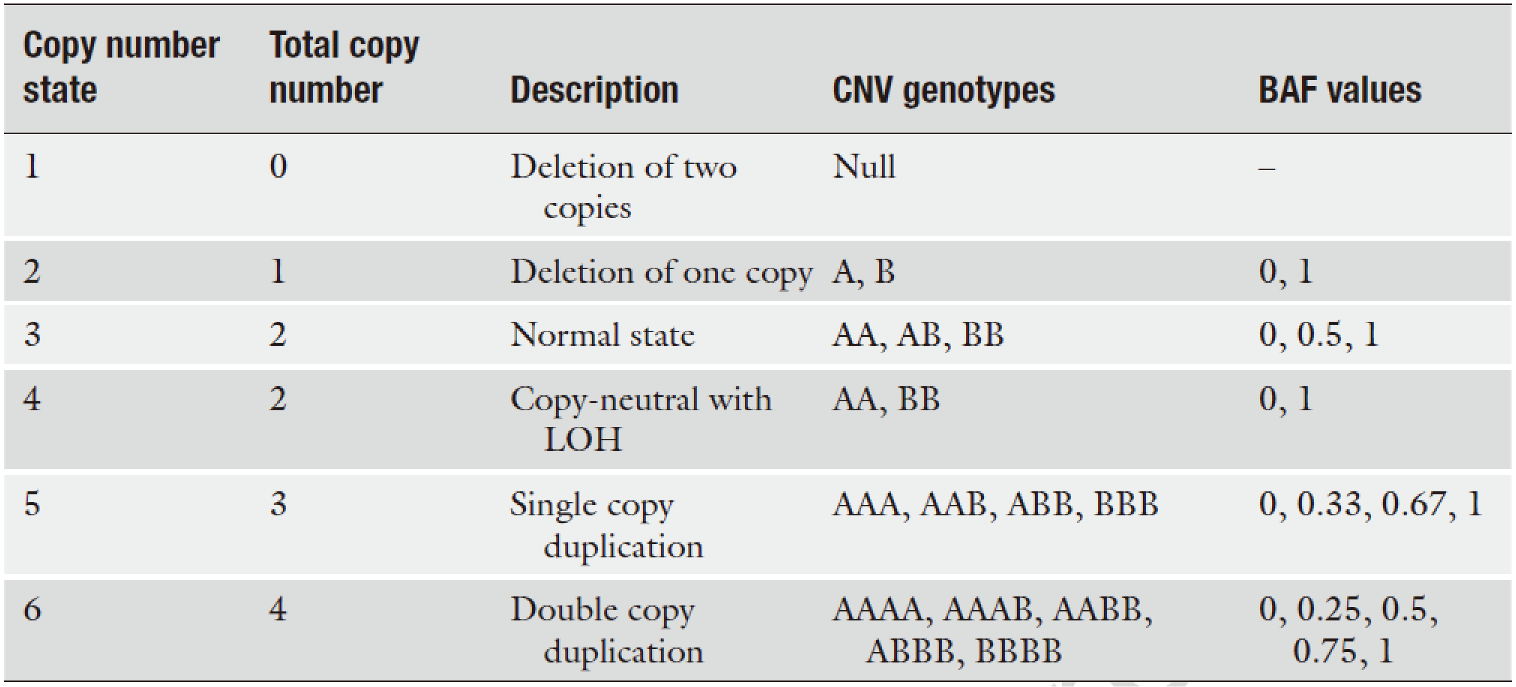
$$P(b_j \mid g, z_j, \lambda) = \begin{cases} \phi(b_j; \mu_{BAF,z_j,g}, \sigma_{BAF,z_j,g}), & \text{if } 0 < g < C(z_j) \\ I_{\{b_j=0\}} M_0 + I_{\{0<b_j<1\}} \phi(b_j; \mu_{BAF,z_j,g}, \sigma_{BAF,z_j,g}), & \text{if } g = 0 \\ I_{\{b_j=1\}} M_1 + I_{\{0<b_j<1\}} \phi(b_j; \mu_{BAF,z_j,g}, \sigma_{BAF,z_j,g}), & \text{if } g = C(z_j) \end{cases}$$

Emission Probability of BAF



Copy number state	Total copy number	Mode of B Allele Frequency distribution
1	0	$\mu_{1,1}$=0.5
2	1	$\mu_{2,1}$=0, $\mu_{2,2}$=1
3	2	$\mu_{3,1}$=0, $\mu_{3,2}$=0.5, $\mu_{3,3}$=1
4	2	$\mu_{4,1}$=0, $\mu_{4,2}$=1
5	3	$\mu_{5,1}$=0, $\mu_{5,2}$=0.33, $\mu_{5,3}$=0.66, $\mu_{5,4}$=1
6	4	$\mu_{6,1}$=0, $\mu_{6,2}$=0.25, $\mu_{6,3}$=0.5, $\mu_{6,4}$=0.75, $\mu_{6,5}$=1

Hidden states, copy numbers, CNV genotypes, and their descriptions



Copy number state	Total copy number	Description	CNV genotypes	BAF values
1	0	Deletion of two copies	Null	–
2	1	Deletion of one copy	A, B	0, 1
3	2	Normal state	AA, AB, BB	0, 0.5, 1
4	2	Copy-neutral with LOH	AA, BB	0, 1
5	3	Single copy duplication	AAA, AAB, ABB, BBB	0, 0.33, 0.67, 1
6	4	Double copy duplication	AAAA, AAAB, AABB, ABBB, BBBB	0, 0.25, 0.5, 0.75, 1

Transition Probability

Transition of copy number states from SNP *j* to SNP *j* + 1

$$P(z_{j+1}=l \mid z_j=h,\lambda)=\begin{cases}1-\sum_{s\neq h} p_{h,s}(1-e^{-d_j/D}) & \text{if } l=h \\ p_{h,l}(1-e^{-d_j/D}) & \text{if } l\neq h\end{cases}$$

d_j: physical distance between SNP *j* and SNP *j* + 1

D: standardizing constant

Likelihood

Assume *M* SNPs are genotyped, then the likelihood of LRR and BAF is

$$P(r_1,\ldots,r_M,b_1,\ldots,b_M)=\sum_{z_1}\cdots\sum_{z_M}P(r_1,\ldots,r_M,b_1,\ldots,b_M\mid z_1,\ldots,z_M)P(z_1,\ldots,z_M)$$

Assume conditional independence between LRR and BAF given copy number state, then

$$P(r_1,\ldots,r_M,b_1,\ldots,b_M)=\sum_{z_1}\cdots\sum_{z_M}\left\{\left(\prod_{i=1}^{M}P(r_i\mid z_i)P(b_i\mid z_i)\right)\left(P(z_1)\prod_{i=2}^{M}P(z_i\mid z_{i-1})\right)\right\}$$

CNV Calling

- Use Viterbi algorithm to infer the most likely state path z = (z_1, …, z_M), by maximizing P(z|r, b, λ).

- Calculation is speed up using Baum's forward-backward algorithm.

- A CNV is called whenever a stretch of states different from the normal state is observed.

- Algorithm is implemented in software PennCNV.

http://penncnv.openbioinformatics.org/en/latest/

CNV Calling

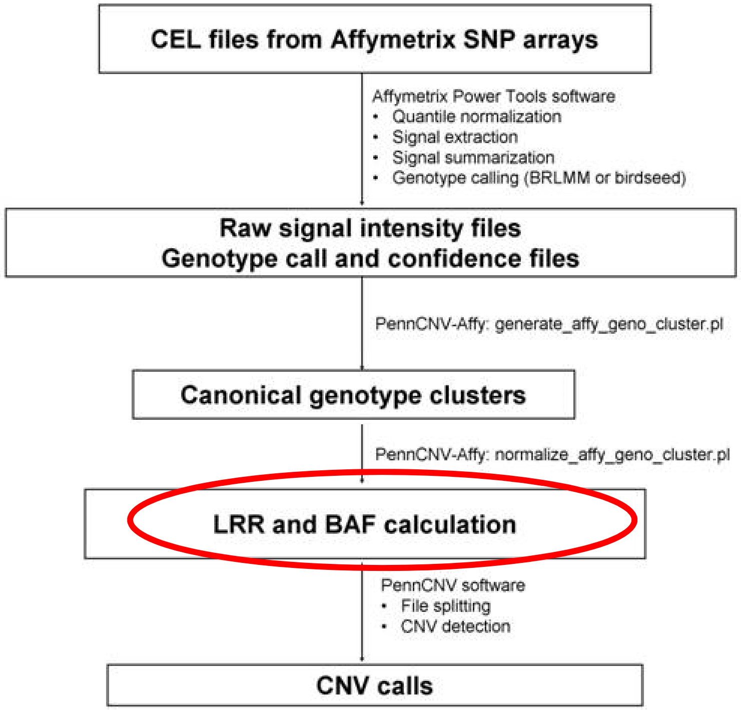
- Viterbi algorithm for calling
  - Calculate the most likely path in HMM (a path of state 1-6 for each SNP marker)
  - Collect any non-normal state path as the CNV calls
  - Example:

```
Most likely CN sequence is:
22222222111111112222222222222222222222224444442222222222222
           CNV!                           CNV!
```

Other Types of Signal Data

- PennCNV can be applied to data from other technical platforms:
  - Transformation of signal data to LRR/BAF:
    - Affymetrix whole-genome SNP genotyping array
    - Perlegen whole-genome SNP genotyping array
  - Use information from LRR only:
    - BAC clone based array-CGH
    - Oligonucleotide arrays
    - Non-polymorphic markers in recent SNP genotyping arrays

PennCNV-Affy Pipeline



CEL files from Affymetrix SNP arrays

Affymetrix Power Tools software
- Quantile normalization
- Signal extraction
- Signal summarization
- Genotype calling (BRLMM or birdseed)

Raw signal intensity files
Genotype call and confidence files

PennCNV-Affy: generate_affy_geno_cluster.pl

Canonical genotype clusters

PennCNV-Affy: normalize_affy_geno_cluster.pl

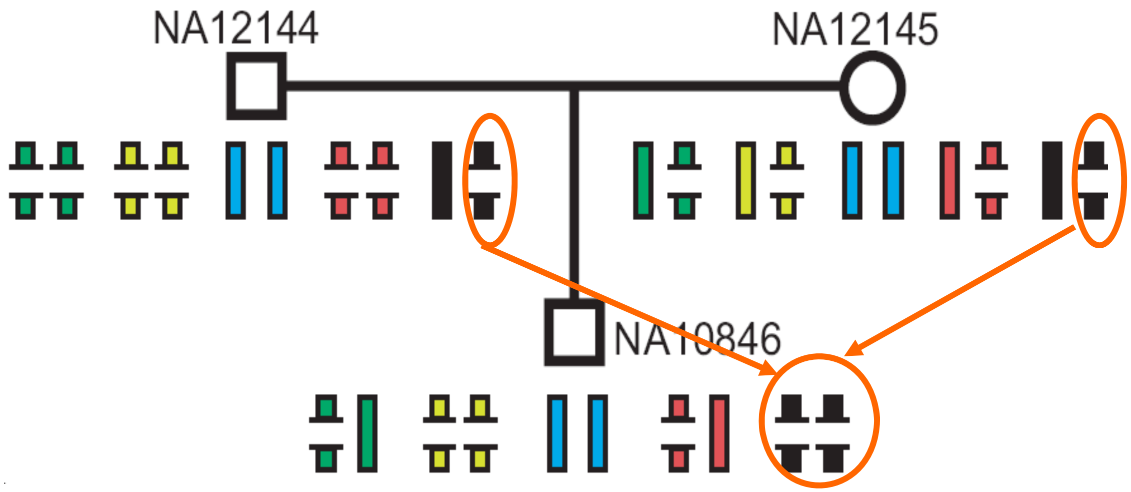
LRR and BAF calculation

PennCNV software
- File splitting
- CNV detection

CNV calls

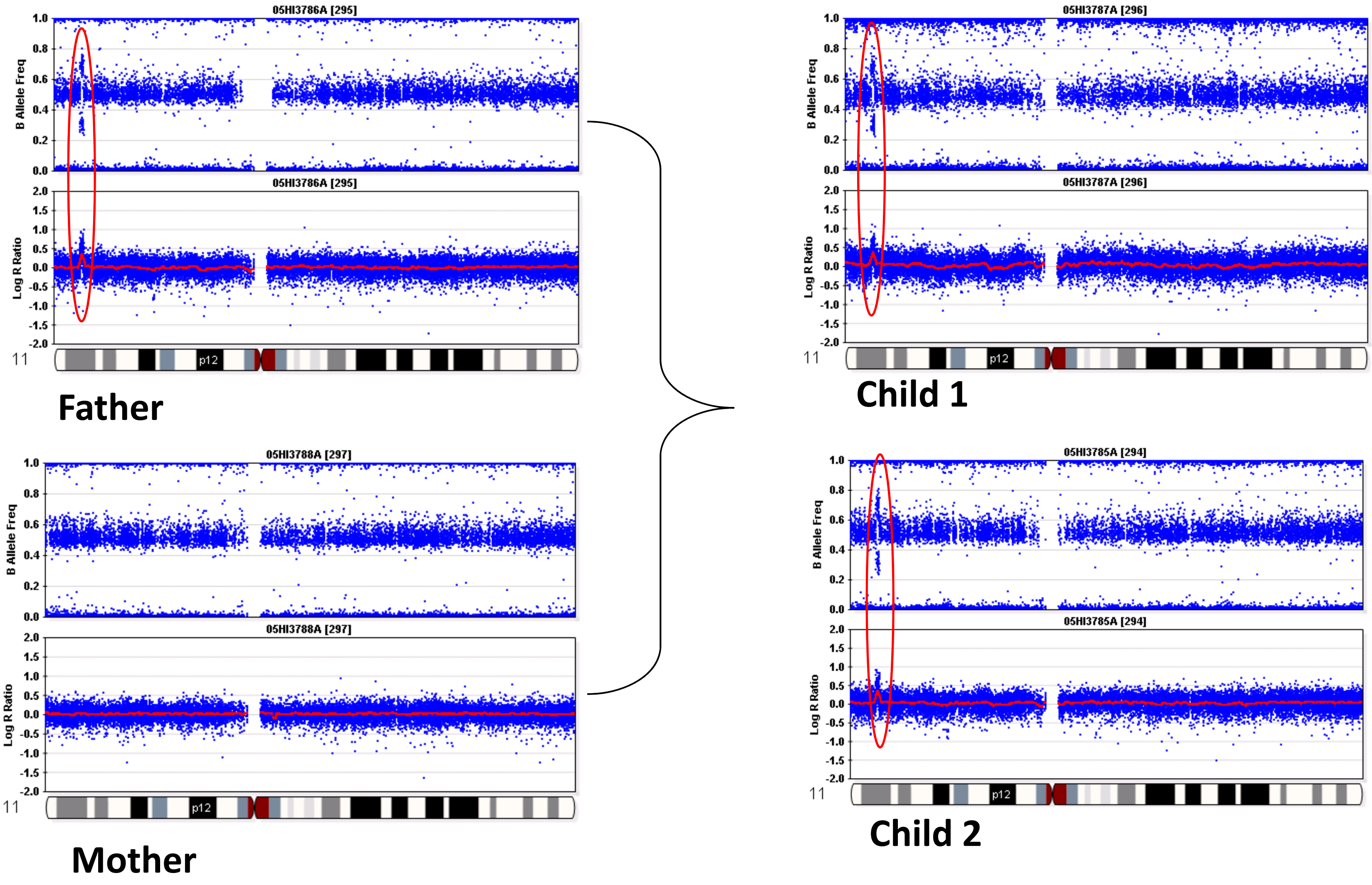
Joint Modeling on Family Data

- Most CNVs demonstrate Mendelian inheritance



- Incorporate family relationship can potentially improve sensitivity of CNV calling

Example of Inherited CNV



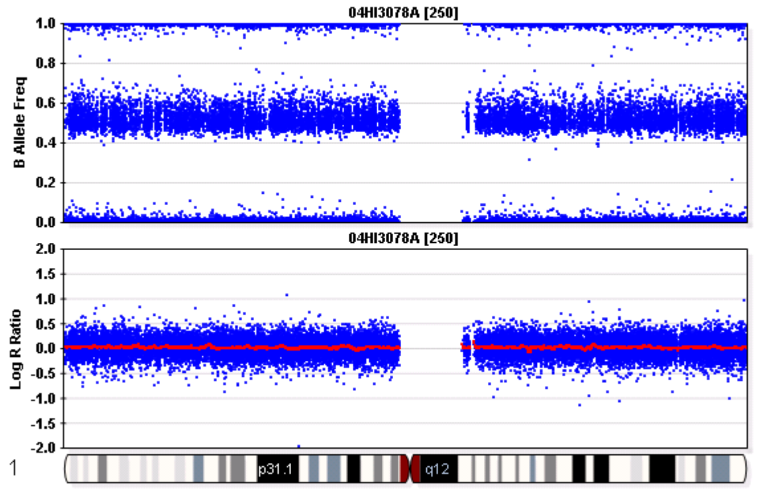
Father

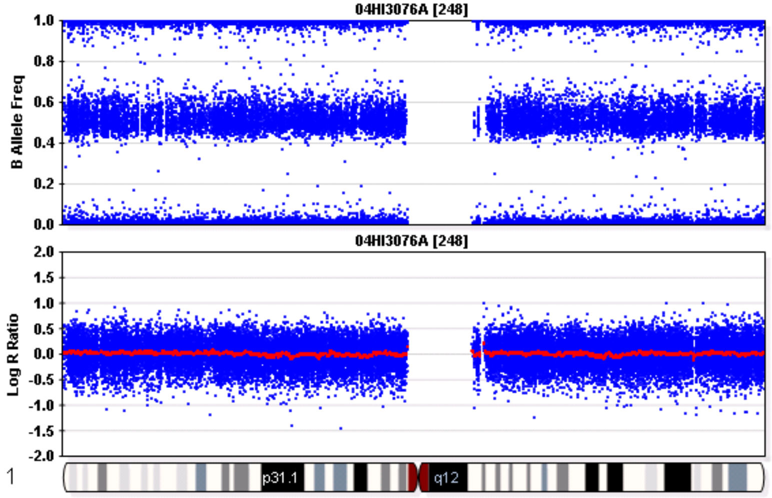
Mother

Child 1

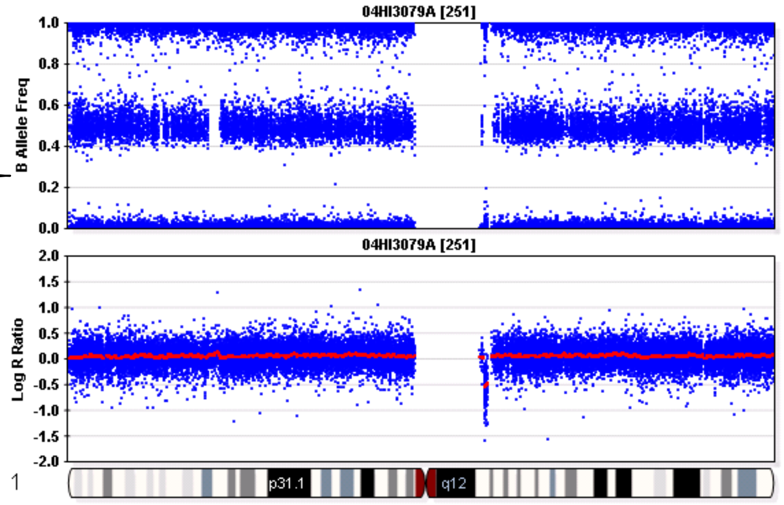
Child 2

Example of *de novo* CNV



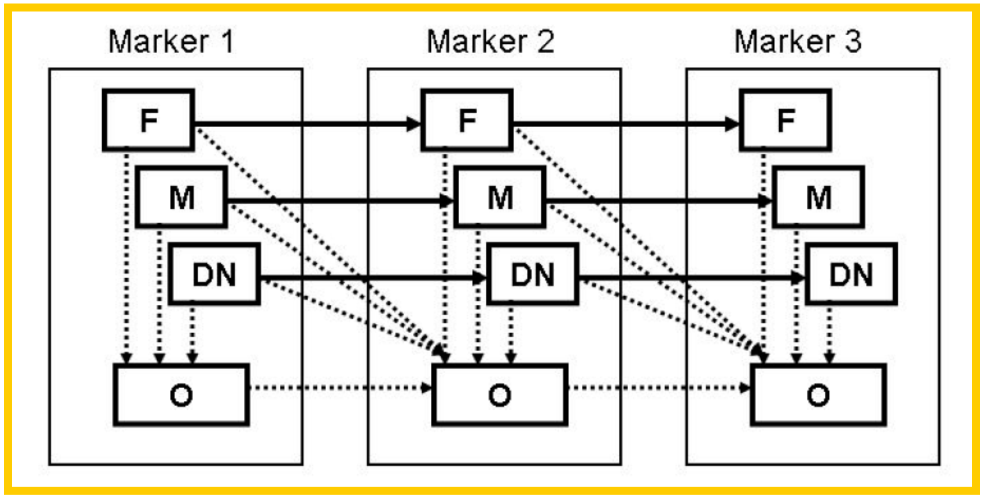


Some CNVs are due to *de novo* events, which occur as germline, somatic or cell line-induced chromosome aberrations in offspring that are not inherited from either parent.



Joint modeling of the CNVs in a trio

- A HMM that jointly models a trio simultaneously
- Do not assume that CNV region is already known



F: father; **M**: mother; **O**: offspring; **DN**: *de novo* event status.

Likelihood of Signal Intensities

$$
\begin{aligned}
&P(r_1,\ldots,r_T,b_1,\ldots,b_T \mid \lambda)\\
&=\sum_{z_1}\cdots\sum_{z_T}\sum_{DN_1}\cdots\sum_{DN_T}\{P(r_1,\ldots,r_T \mid z_1,\ldots,z_T,\lambda)\times\\
&\qquad P(b_1,\ldots,b_T \mid z_1,\ldots,z_T,\lambda)\times\\
&\qquad P(z_1,\ldots,z_T \mid DN_1,\ldots,DN_T,\lambda)\times\\
&\qquad P(DN_1,\ldots,DN_T \mid \lambda)\}\\
&=\sum_{z_1}\cdots\sum_{z_T}\sum_{DN_1}\cdots\sum_{DN_T}\Big\{\boxed{P(r_1 \mid z_1,\lambda)}\boxed{P(b_1 \mid z_1,\lambda)}\boxed{P(z_1 \mid DN_1,\lambda)}\boxed{P(DN_1 \mid \lambda)}\times\\
&\qquad \prod_{j=2}^{T}\boxed{P(r_j \mid z_j,\lambda)}\boxed{P(b_j \mid z_j,\lambda)}\boxed{P(z_j \mid z_{j-1},DN_j,DN_{j-1},\lambda)}\boxed{P(DN_j \mid DN_{j-1},\lambda)}\Big\}.
\end{aligned}
$$

Initial prob of CN states

Initial prob of de novo event status

Emission prob of LRR

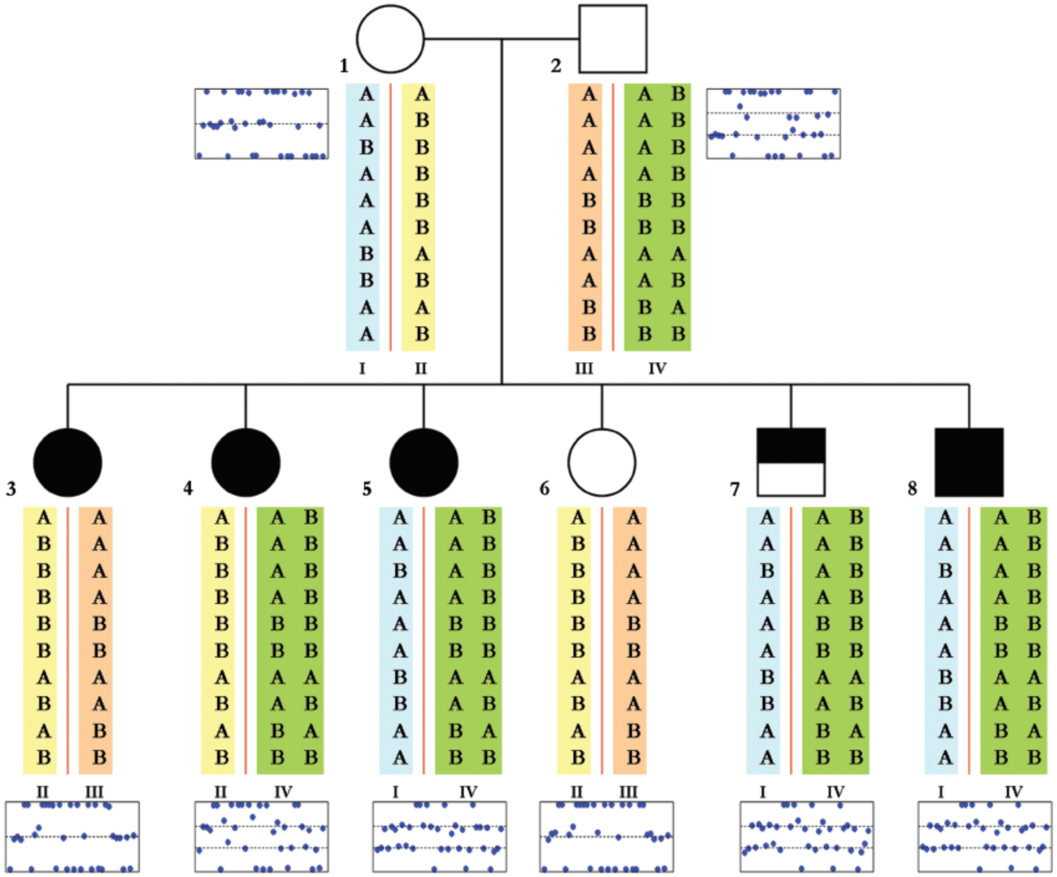
Emission prob of BAF

Transition prob of CN states

Transition prob of de novo event status

By treating the trio as a unit, this calling algorithm can avoid generating calls that are Mendelian inconsistent but preserve the ability to allow *de novo* events.

Inferring chromosome-specific copy numbers

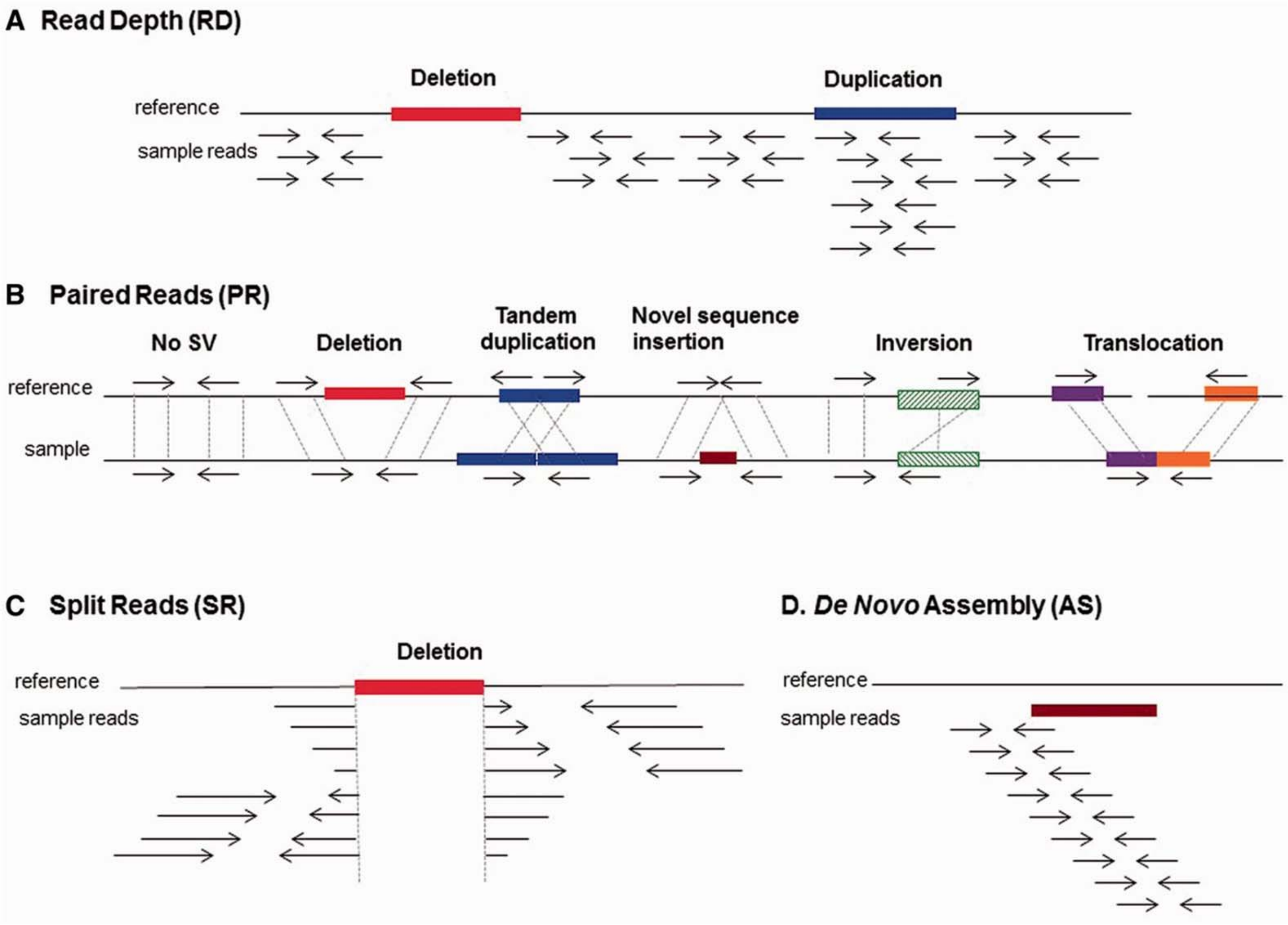


1		2	
I	II	III	IV
A	A	A	A B
A	B	A	A B
B	B	A	A B
A	B	A	A B
A	B	B	B B
A	B	B	B B
B	A	A	A A
B	B	A	A B
A	A	B	B A
A	B	B	B B

3		4		5		6		7		8	
II	III	II	IV	I	IV	II	III	I	IV	I	IV
A	A	A	A B	A	A B	A	A	A	A B	A	A B
B	A	B	A B	A	A B	B	A	A	A B	A	A B
B	A	B	A B	B	A B	B	A	B	A B	B	A B
B	A	B	A B	A	A B	B	A	A	A B	A	A B
B	B	B	B B	A	B B	B	B	A	B B	A	B B
B	B	B	B B	A	B B	B	B	A	B B	A	B B
A	A	A	A A	B	A A	A	A	B	A A	B	A A
B	A	B	A B	B	A B	B	A	B	A B	B	A B
A	B	A	B A	A	B A	A	B	A	B A	A	B A
B	B	B	B B	A	B B	B	B	A	B B	A	B B

Figure 4. Illustration of a duplication CNV on 10q11.22 that exists in the father and is transmitted to four offspring. The CNV calls are made on six trios separately by the joint-calling algorithm. For each individual, the BAF values for all SNPs within the CNV and the chromosome-specific SNP genotypes (for the first 10 SNPs) are displayed, and the SNP genotypes for the entire region are listed at Supplementary Table 4. The four different parental CNV haplotypes are marked by different colors and denoted by I through IV beneath the genotypes. Combining information from total copy number and the SNP genotypes, we can infer the SNP allele compositions within each homologous chromosome confidently for each offspring.

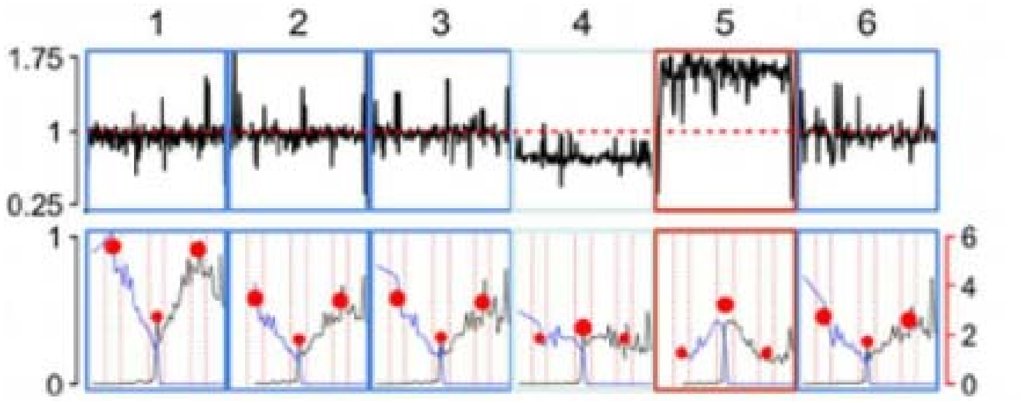
NGS-based SV detection



Escaramís G, et al. Briefings in Functional Genomics, 2015

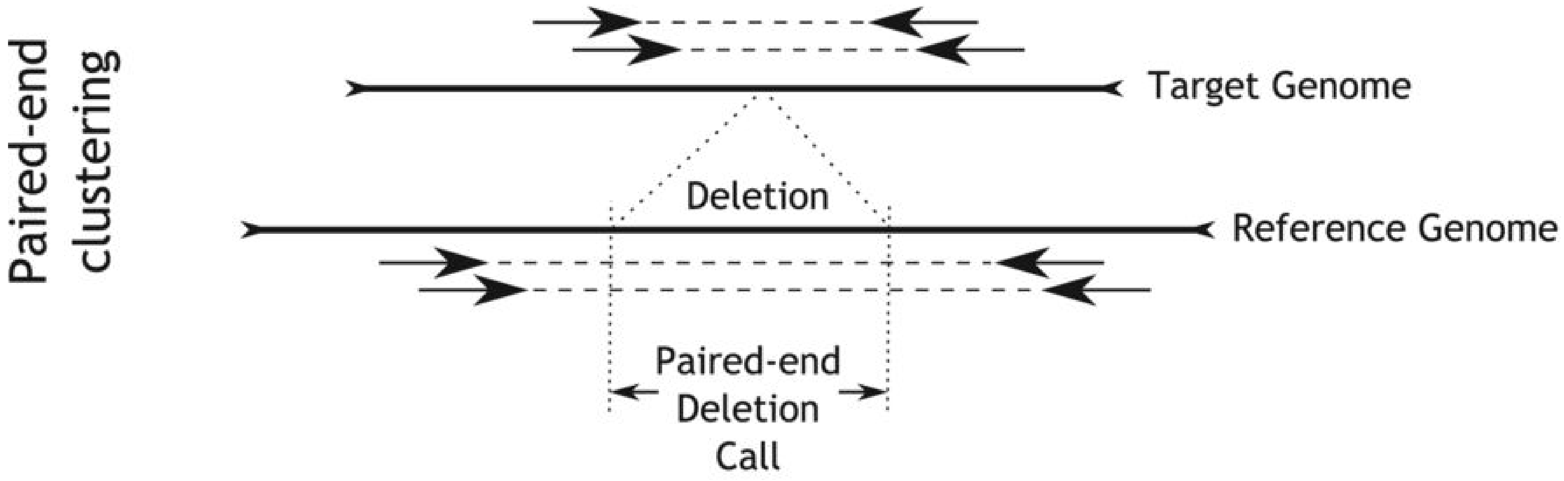
Read count-based methods for SV detection

- Detect the change of read count/sequencing coverage in a certain region.
- Examples of software tools: CNVnator, BIC-SEQ2, PennCNV-Seq
- Limitation:
    - 1) Only detects unbalanced events (copy number variation).
    - 2) Cannot resolve breakpoints at base pair resolution.



Detection of SVs from discordant read pairs

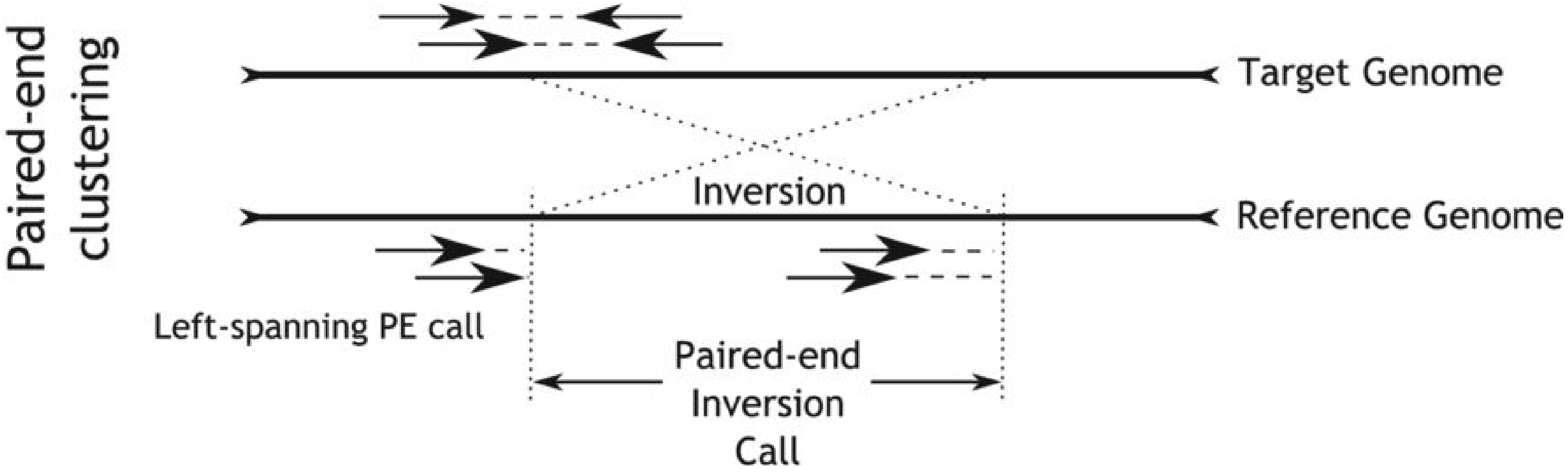
- Widely used software tools: Delly, Lumpy
- Pattern of deletions: large gaps between read pairs:



Rausch T, et al. (2012) Bioinformatics

Detection of SVs from discordant read pairs

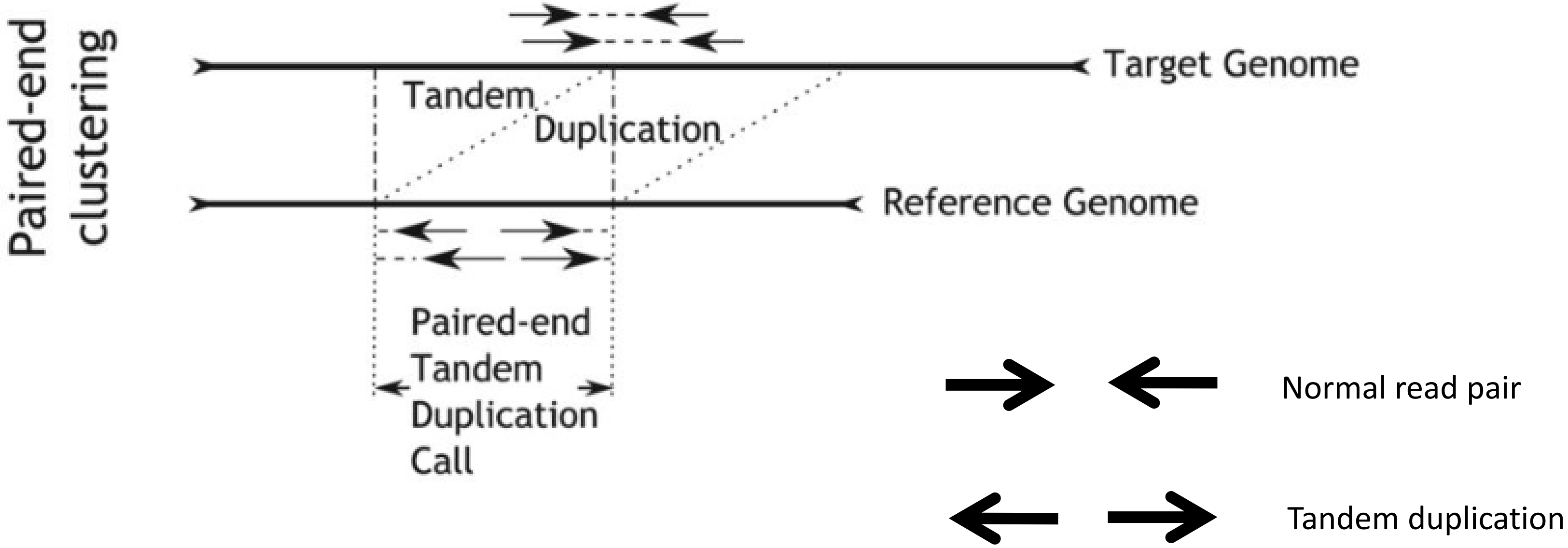
- Pattern of inversions: same orientation between read pairs:



Rausch T, et al. (2012) Bioinformatics

Detection of SVs from discordant read pairs

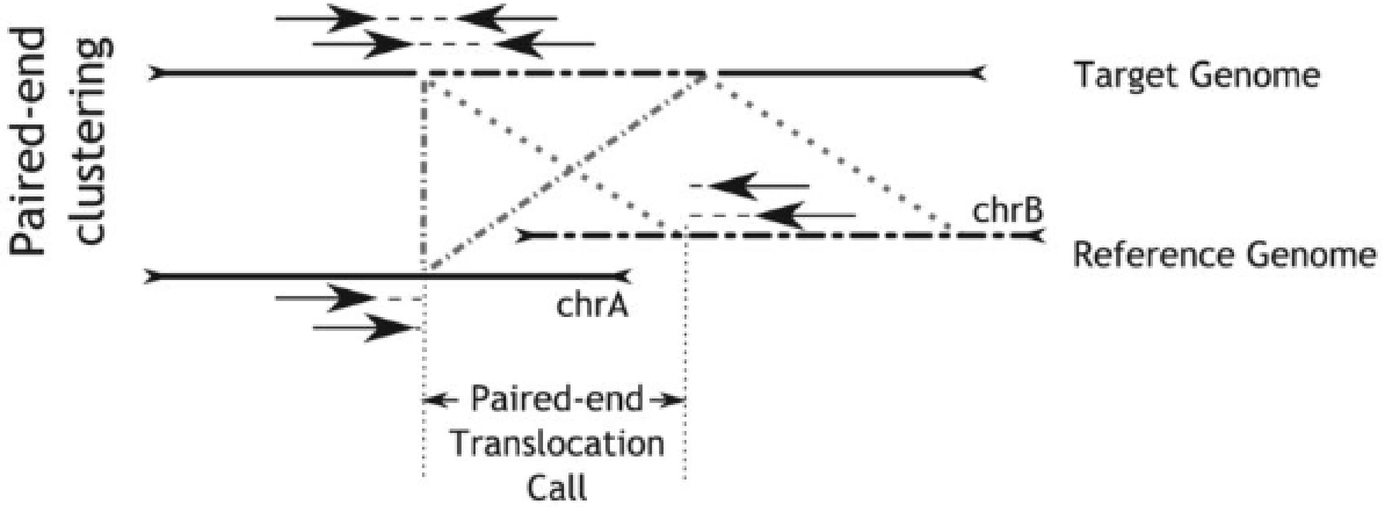
- Pattern of tandem duplication: the first and second read changed their relative order



Rausch T, et al. (2012) Bioinformatics

Detection of SVs from discordant read pairs

- Pattern of translocations: paired-ends mapping to different chromosomes

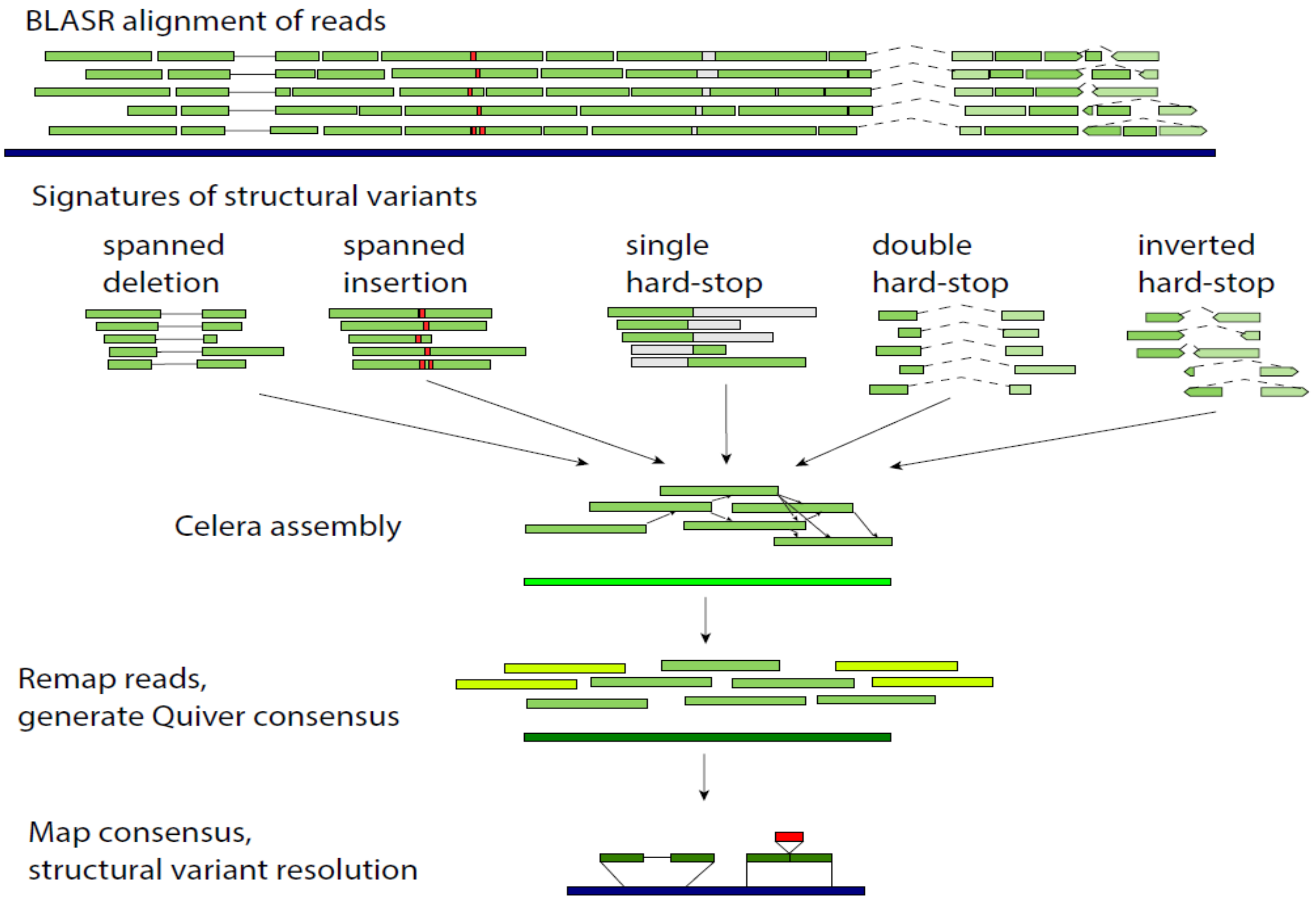


Rausch T, et al. (2012) Bioinformatics

Detection of SVs using assembly-based methods

- De novo sequence assembly (AS) enables the fine-scale discovery of SVs, including novel (non-reference) sequence insertions
- Either global or local assembly may be used to discover SVs
- Example tools:
    - SvABA (genome-wide detection of structural variants and indels by local assembly)
    - novoBreak (local assembly for breakpoint detection in cancer genomes)
    - TIGRA (a targeted iterative graph routing assembler for breakpoint assembly)

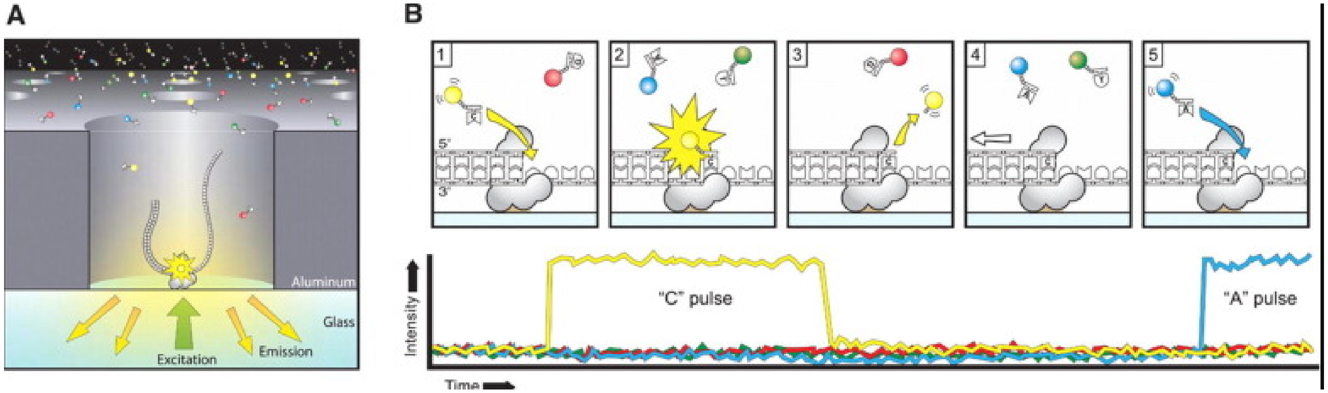
Conceptual overview on local assembly



Chaisson et al, Nature, 2015

Long-read sequencing technologies

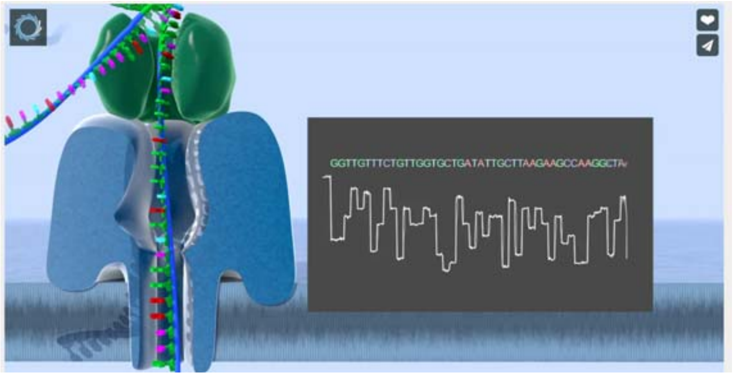
- **PacBio (Pacific Biosciences) sequencing:**
- Typically generates reads with N50 read length of >10kb.
- Error rate: about 15% (CLS) or 1% (CCS)



A double stranded DNA diffuses into a unit called ZMW, and the adaptor binds to a polymerase immobilized at the bottom.
Each of the four nucleotides is labeled with a different fluorescent dye (indicated in red, yellow, green, and blue, respectively for G, C, T, and A) so that they have distinct emission spectrums

Long-read sequencing technologies

- **Oxford Nanopore Technologies**:
- Regular libraries generate reads with N50 length ~20-30 kb
- Ultra-long libraries generate reads with N50 length to >100 kb but with lower throughput
- Error rate: on average 15%. The error rate in some regions (especially homopolymer regions) could be higher.



The DNA molecule passes through a protein pore bounded in a membrane and the current changes are used to infer the DNA sequence.

Mapping long reads to the reference genome

- Multiple alignment tools have been developed to map long reads to the reference genome.
  - Minimap2: a ultra-fast long read alignment tool.
  - NGMLR: an aligner that is specifically developed for SV discovery.
  - BLASR: a aligner developed for PacBio reads
  - BWA-MEM: an early aligner for long reads, could be replaced by Minimap2

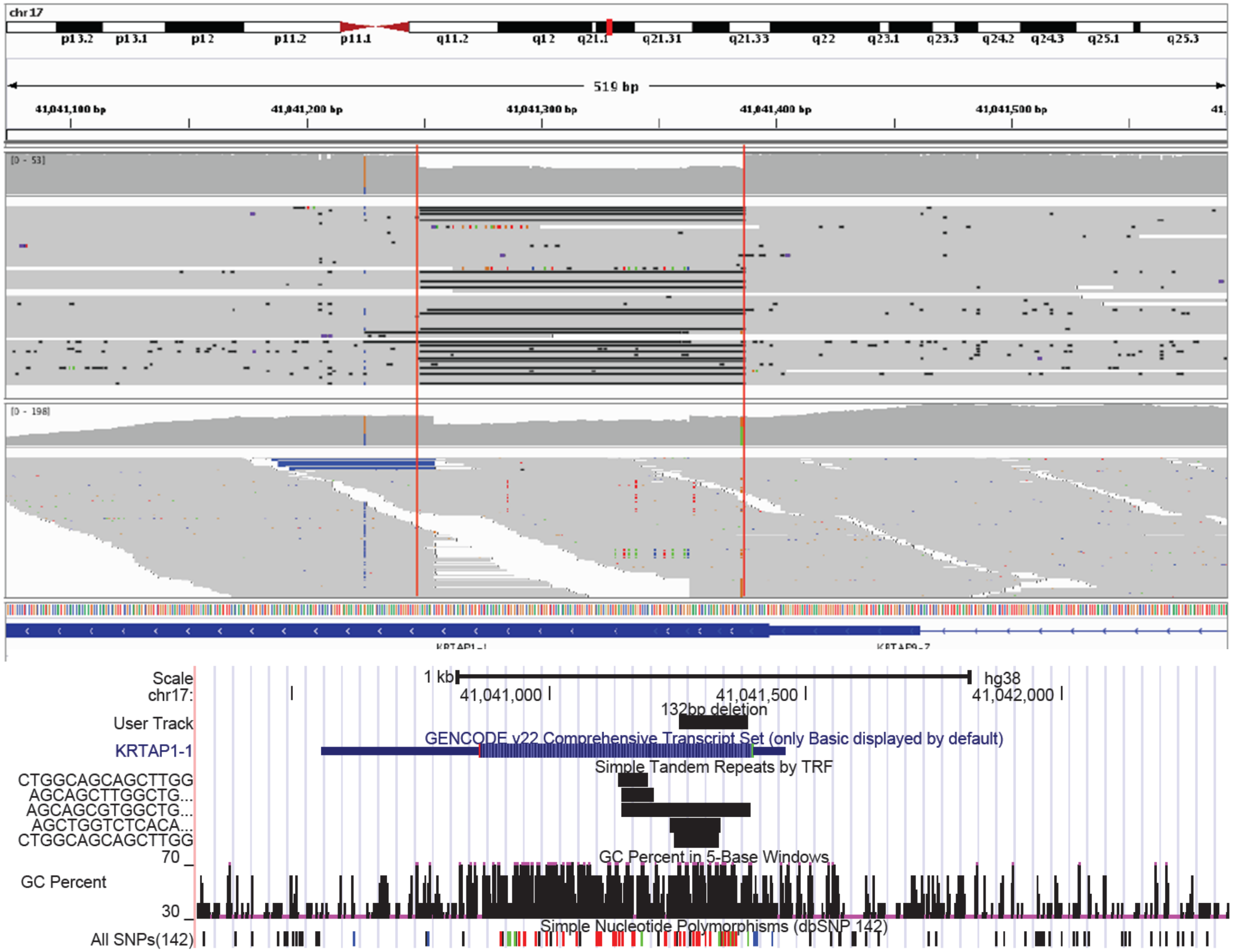
SV detection from long-read sequencing

- Several tools have been developed to detect SVs from long read sequencing. This is an area under active development, and novel software tools are constantly being developed and published
- SV callers for PacBio reads:
  - PBSV
  - SMRT-SV
  - PBHoney
- SV callers for Nanopore reads:
  - NanoSV
  - Sniffles

Short/long reads on SV detection

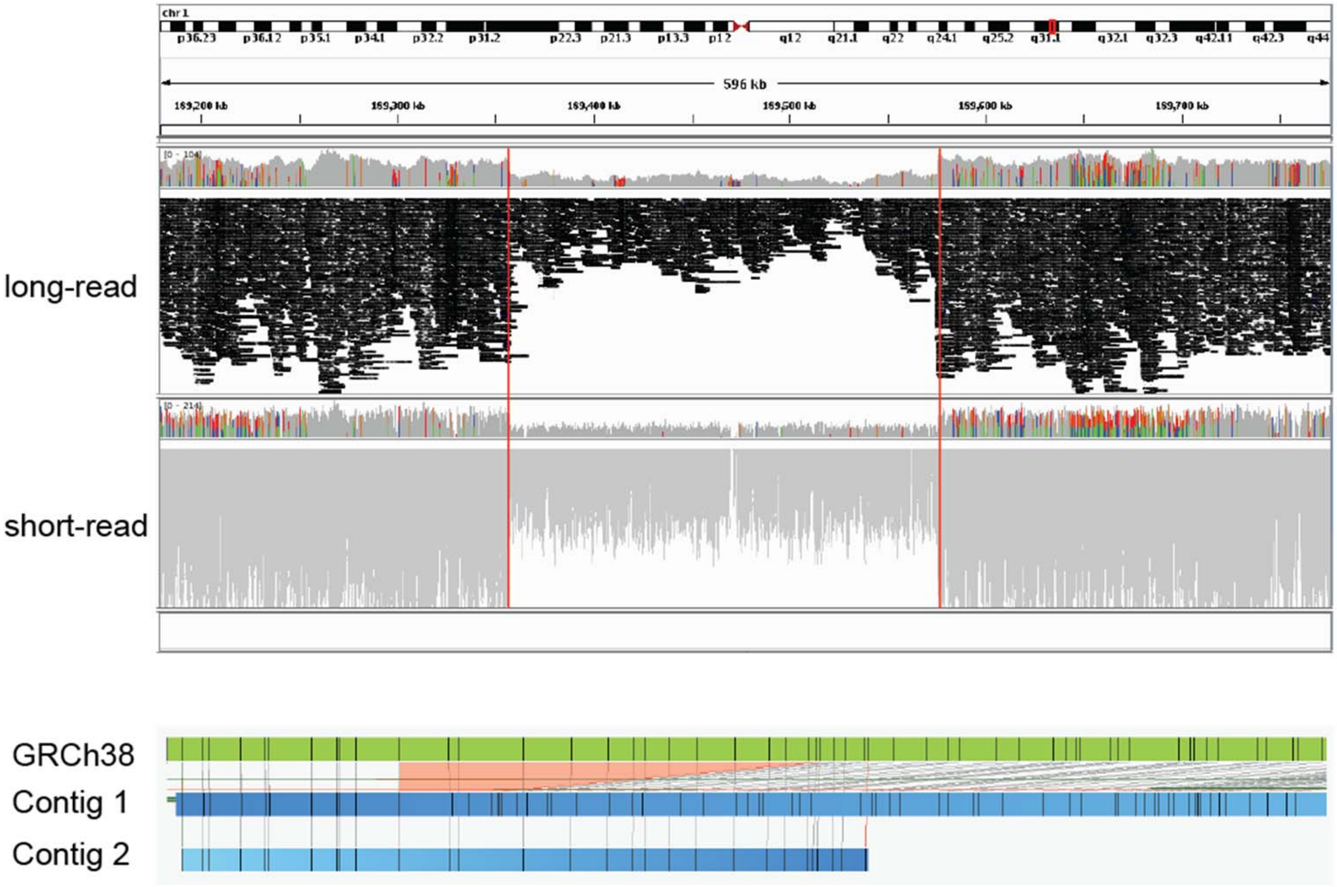
PacBio

Illumina



Shi L et al. (2016) *Nature Communications*

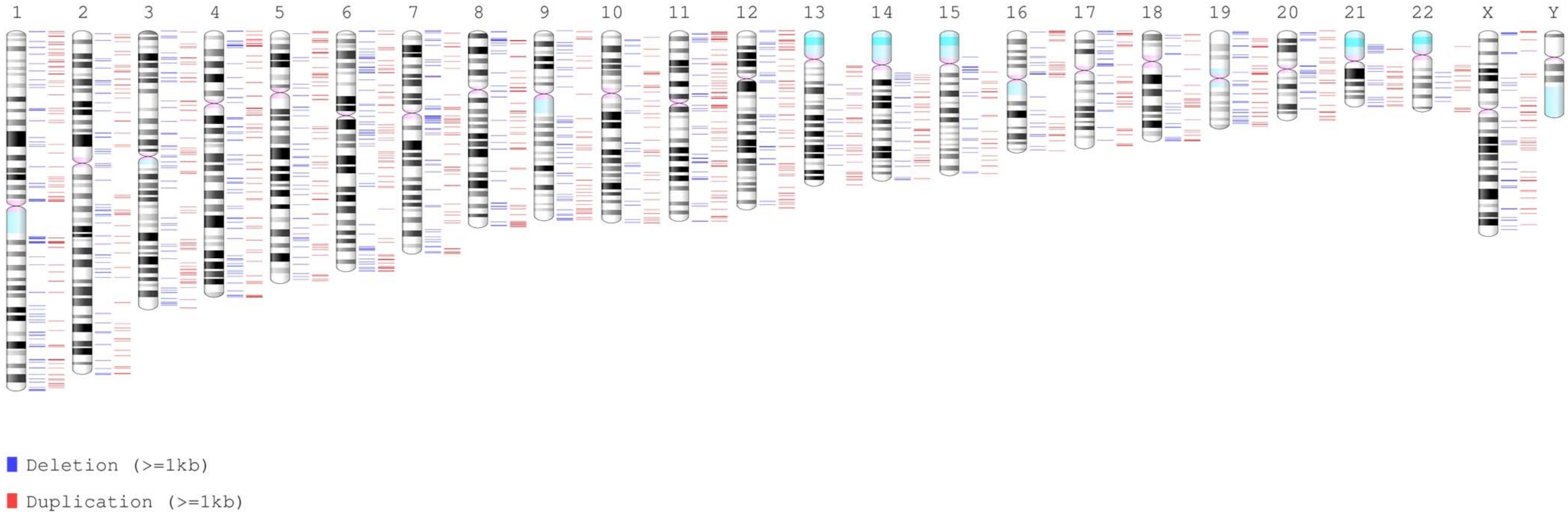
Short/long reads on SV detection



Shi L et al. (2016) *Nature Communications*

Example of SV detection from long-read sequencing

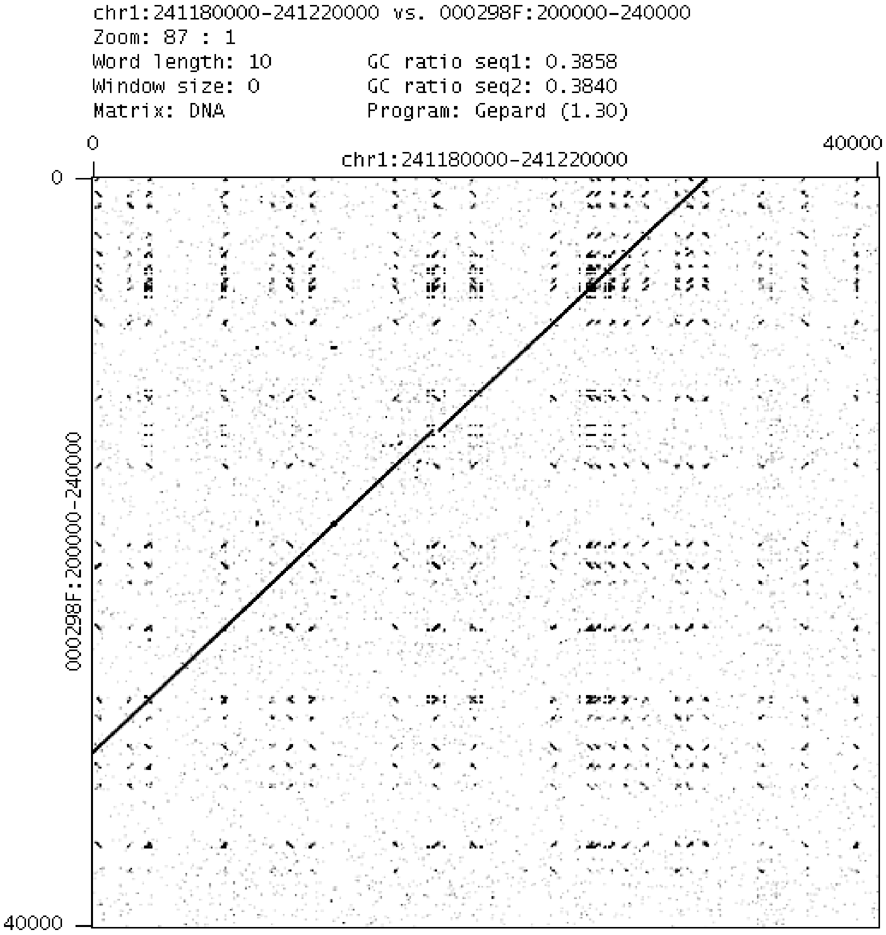
- 9,643 deletions and 10,022 insertions



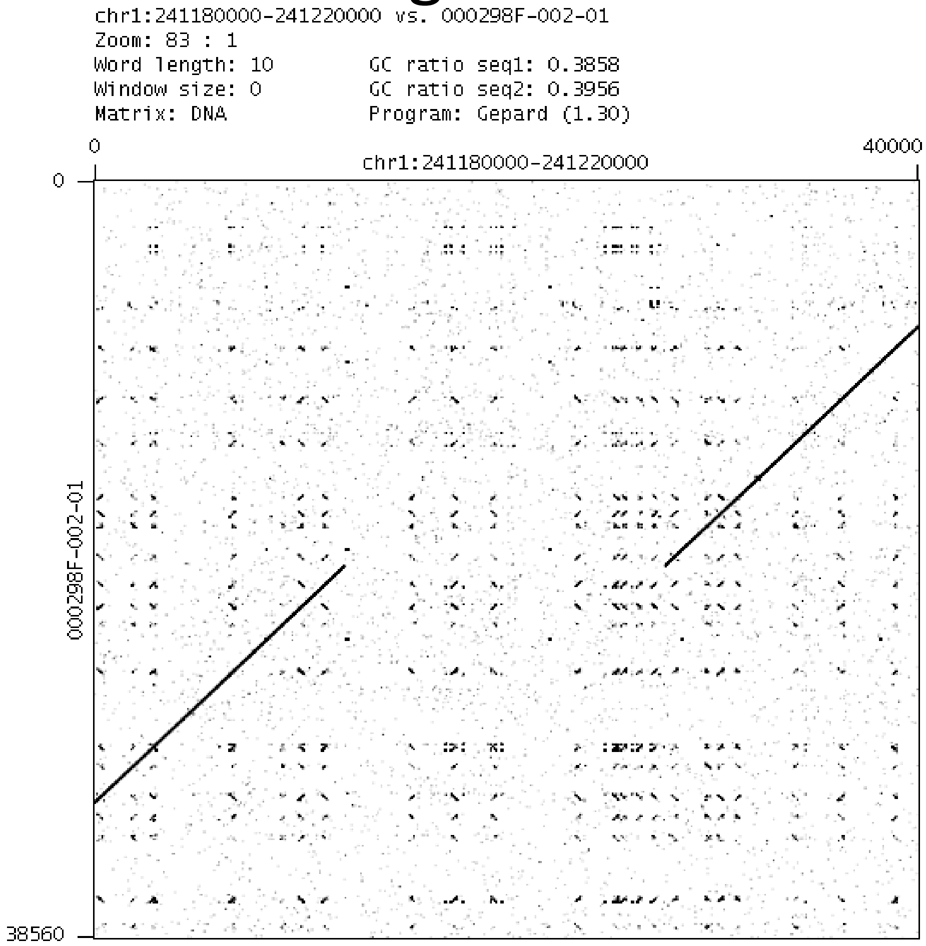
Shi L et al. (2016) *Nature Communications*

Assembly-based SV detection

- Dot plot of primary and associate contig



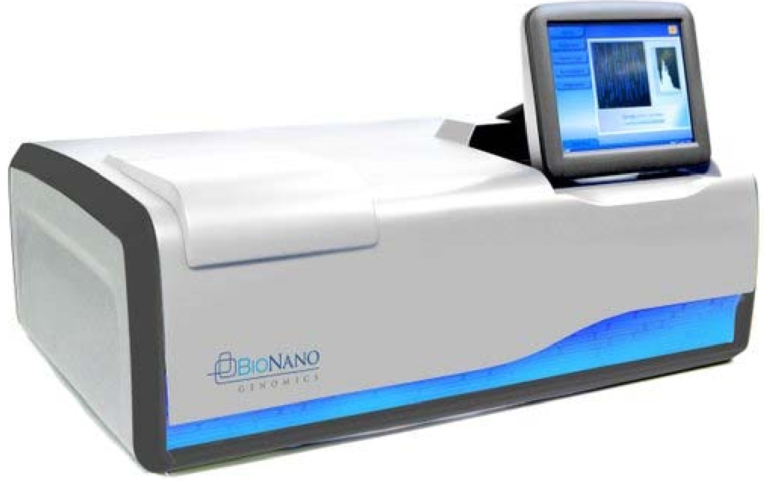
```
chr1:241180000-241220000 vs. 000298F:200000-240000
Zoom: 87 : 1
Word length: 10        GC ratio seq1: 0.3858
Window size: 0         GC ratio seq2: 0.3840
Matrix: DNA            Program: Gepard (1.30)
```

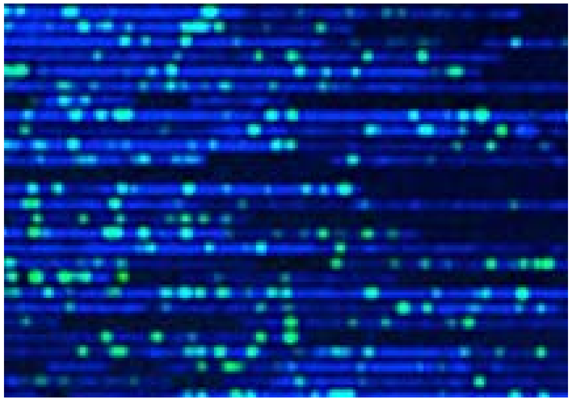


```
chr1:241180000-241220000 vs. 000298F-002-01
Zoom: 83 : 1
Word length: 10        GC ratio seq1: 0.3858
Window size: 0         GC ratio seq2: 0.3956
Matrix: DNA            Program: Gepard (1.30)
```

Bionano optical mapping

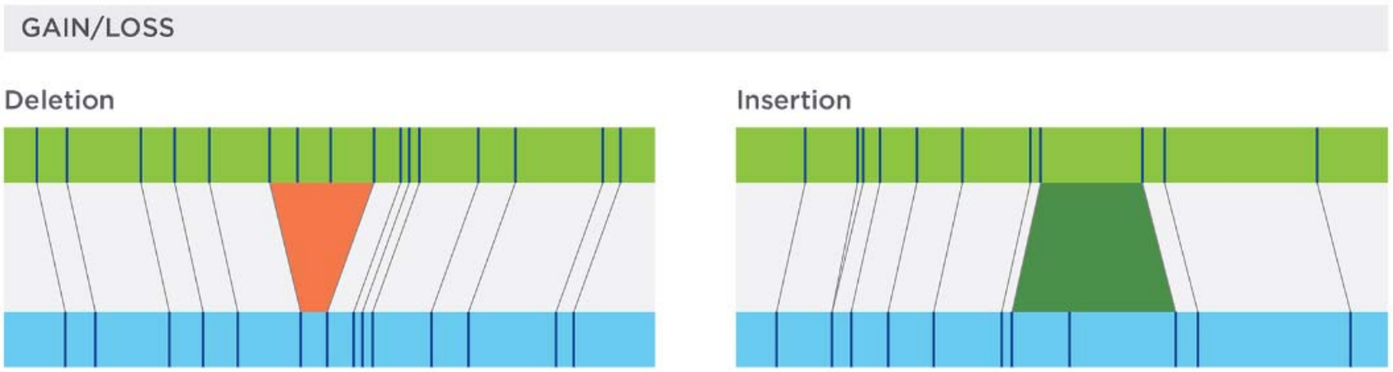
- A nanopore array that detects a characteristic 6 or 7-nucleotide sequence along very long genomic segments





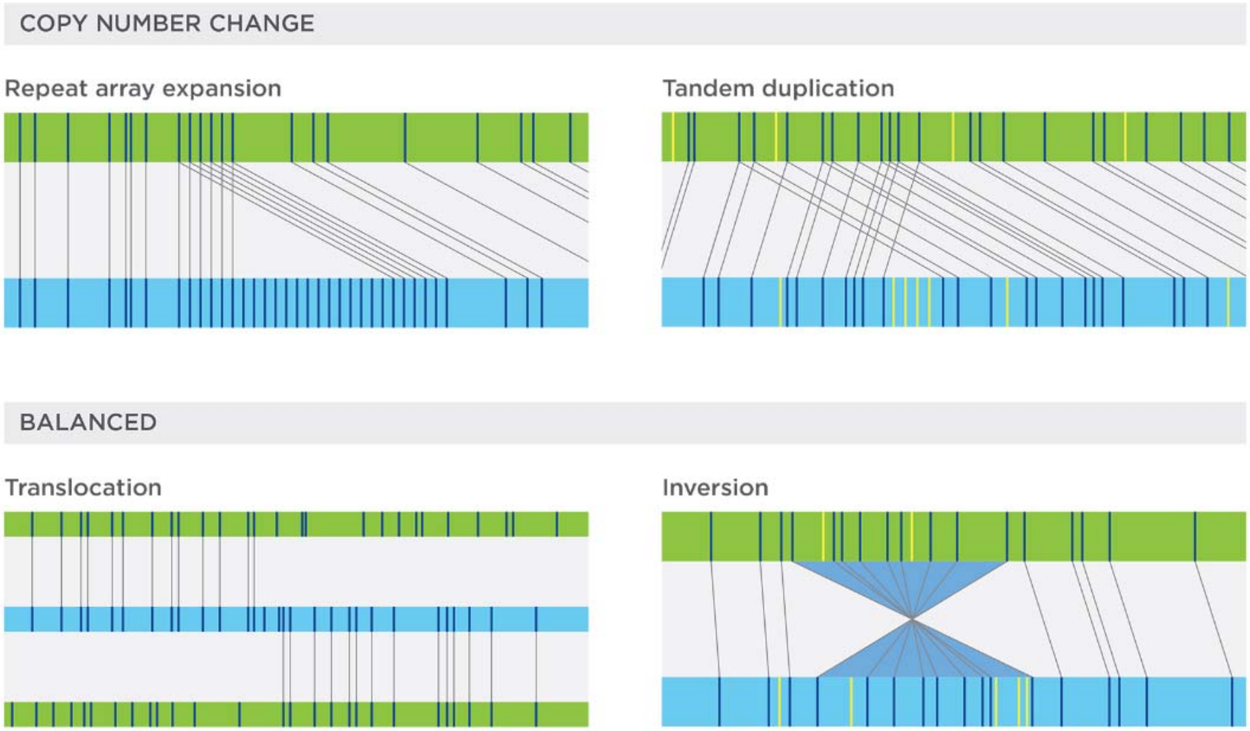
SV detection from Single-molecule optical mapping

- To identify a structural variation, a *de novo* genome map assembly can be aligned to a reference genome.
- By observing changes in label spacing and comparisons of order, position, and orientation of label patterns, SVs can be detected.



https://bionanogenomics.com

SV detection from Single-molecule optical mapping



https://bionanogenomics.com