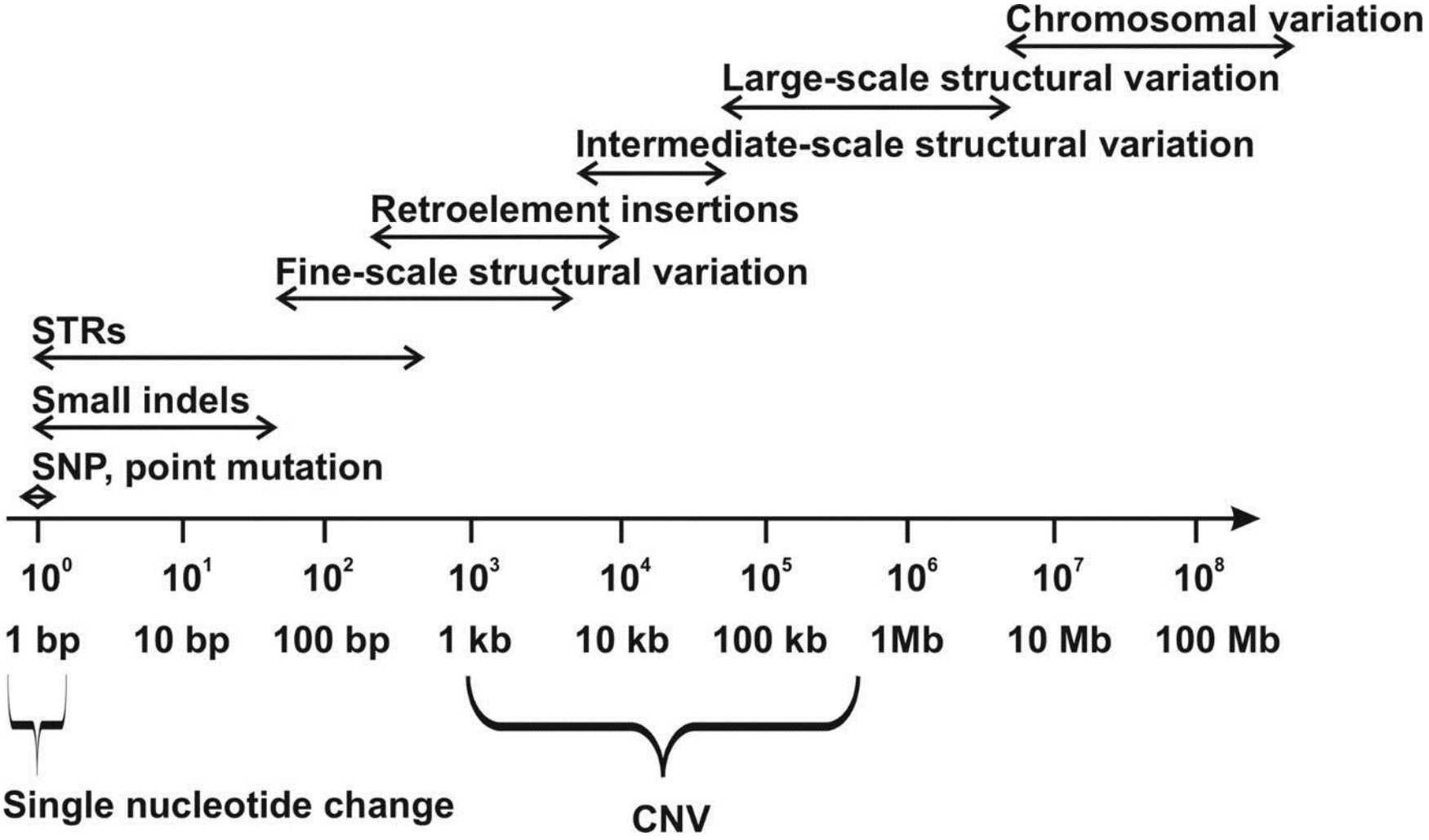
Detection of structural variants in human diseases

2019 Dragon Star Bioinformatics Course (Day 3)

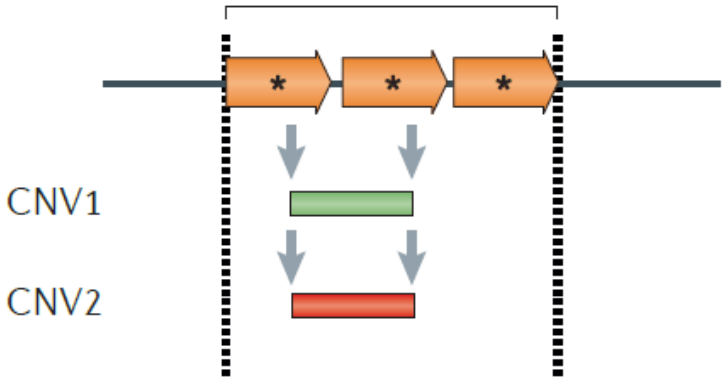
Human genetic variation



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 and Lupski, Nat Rev Genet, 2016

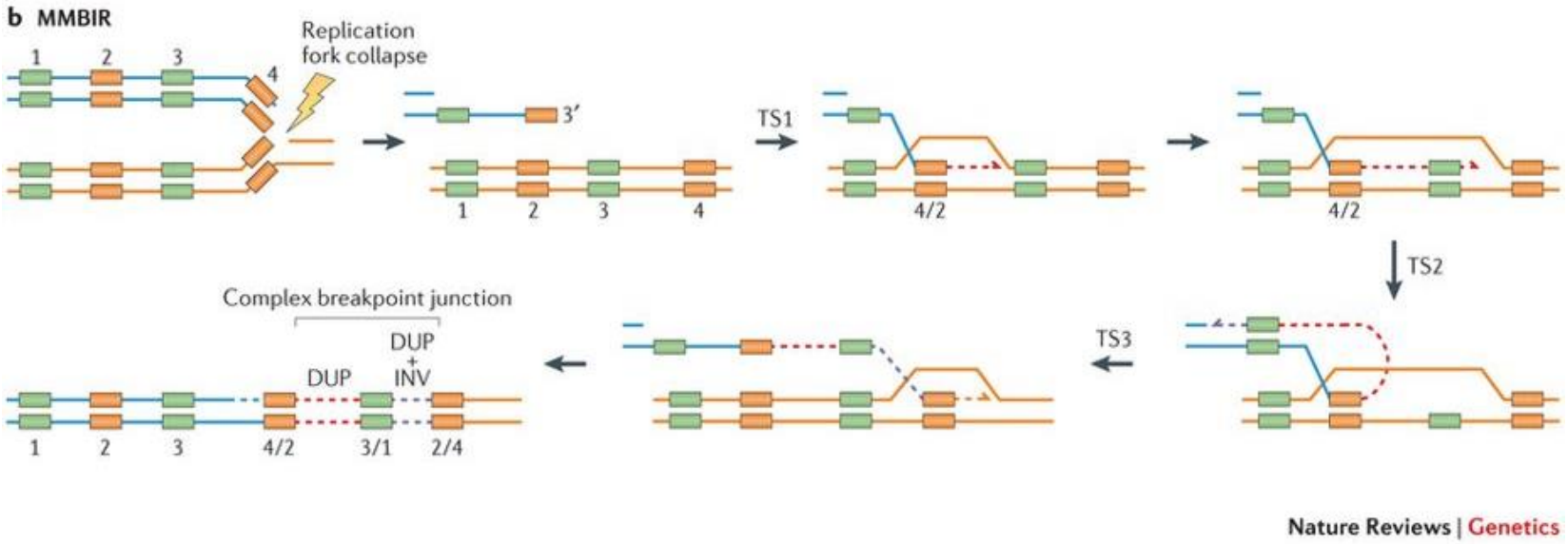
Mechanisms underlying structural variant formation

- **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)

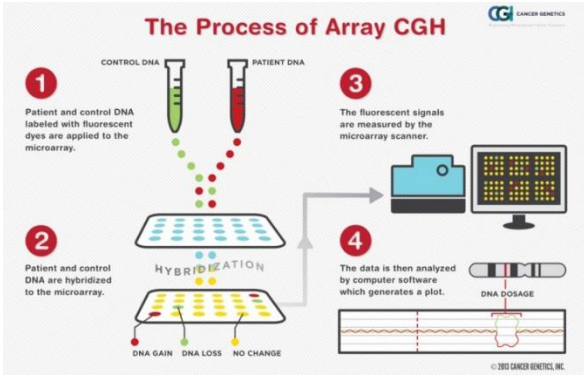


Carvalho, C.M. and J.R. Lupski. *Nat Rev Genet,* 2016.

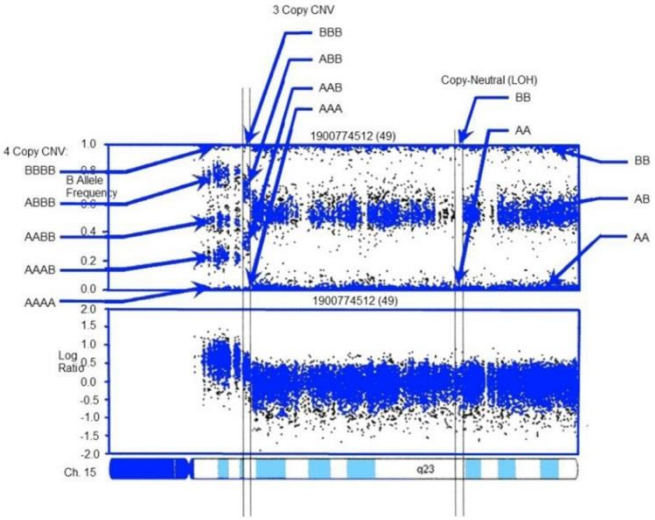
Technologies for CNV Detection

- 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

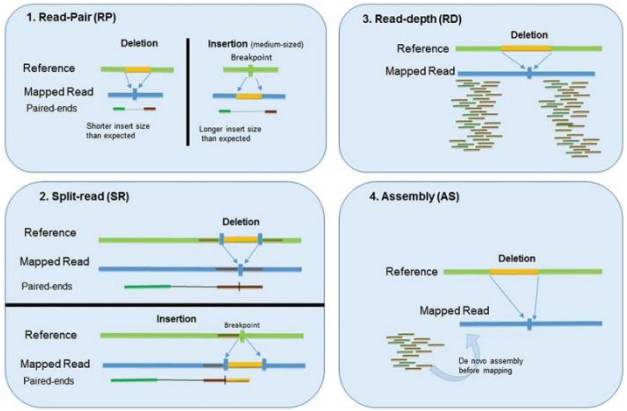
Array CGH



SNP array



Next-generation sequencing

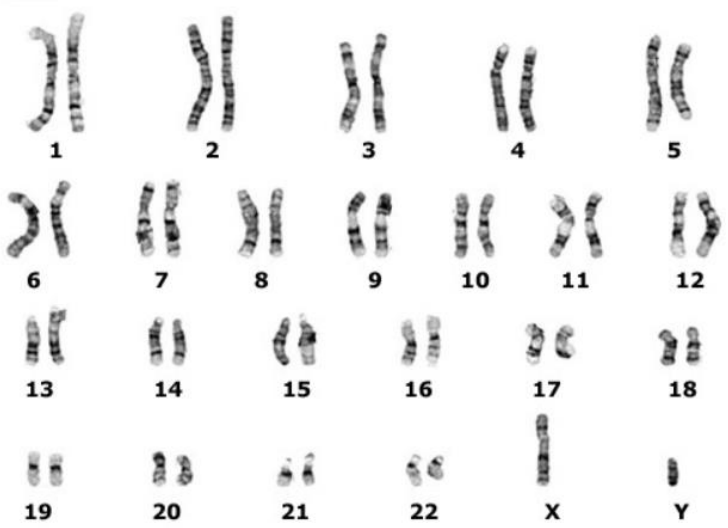


Commonly used cytogenetic techniques

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

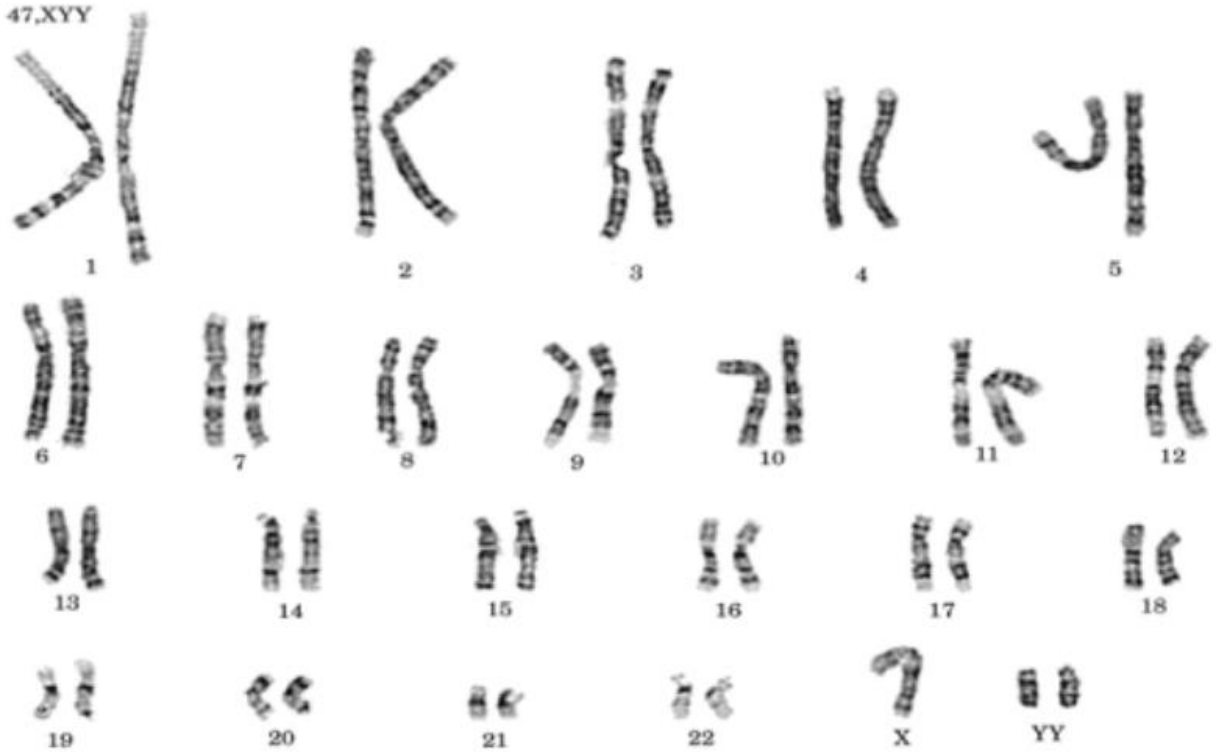
Cytogenetic techniques

- 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

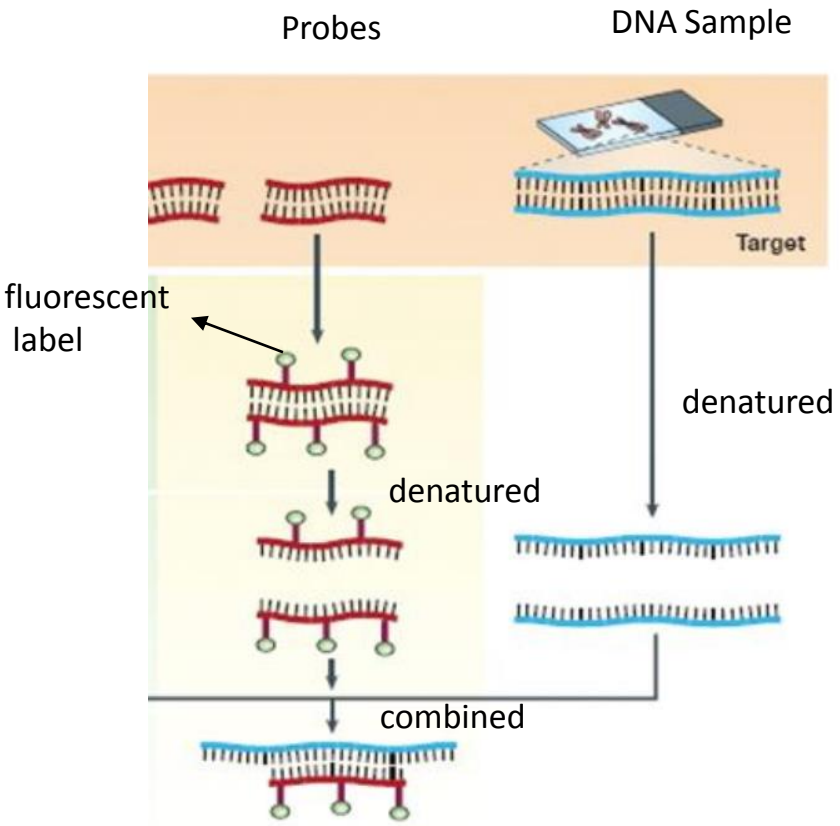
Detection of chromosomal abnormalities by cytogenetic techniques



Polipalli, J Clin Diagn Res. 2016

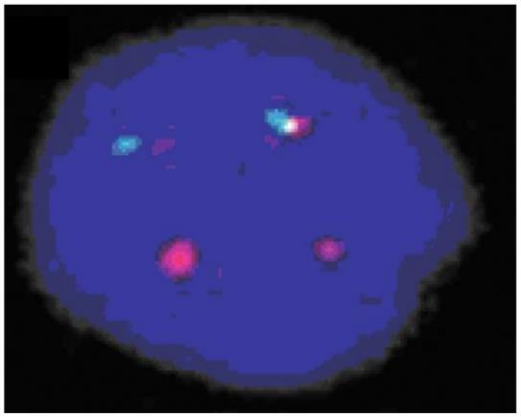
Cytogenetic techniques

- 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.



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)



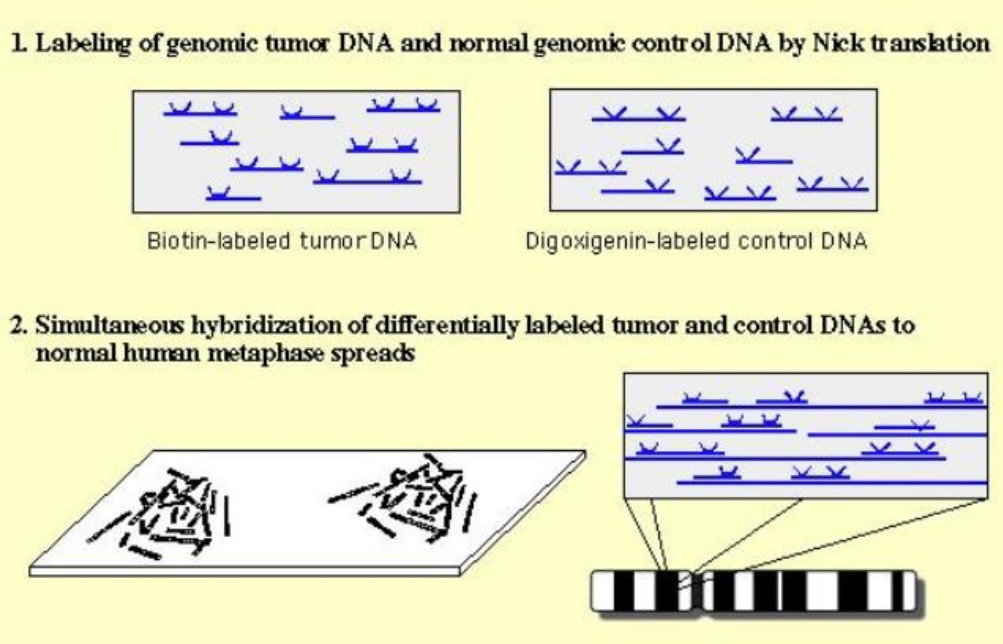
Using interphase FISH to detect the *BCR/ABL* translocation.

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.

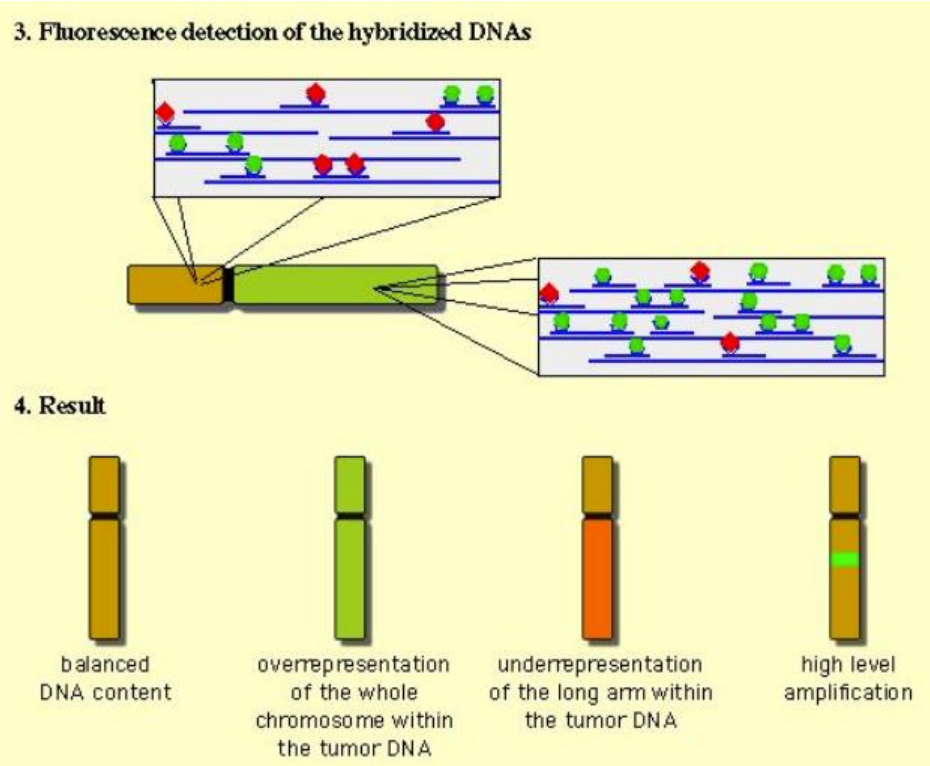
Bioscience Horizons: The International Journal of Student Research,
Volume 3, Issue 1, 1 March 2010, Pages 85–95

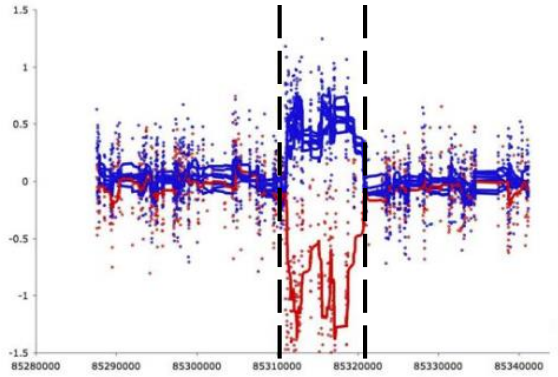
Comparative genomic hybridization (CGH)

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



Comparative genomic hybridization (CGH)



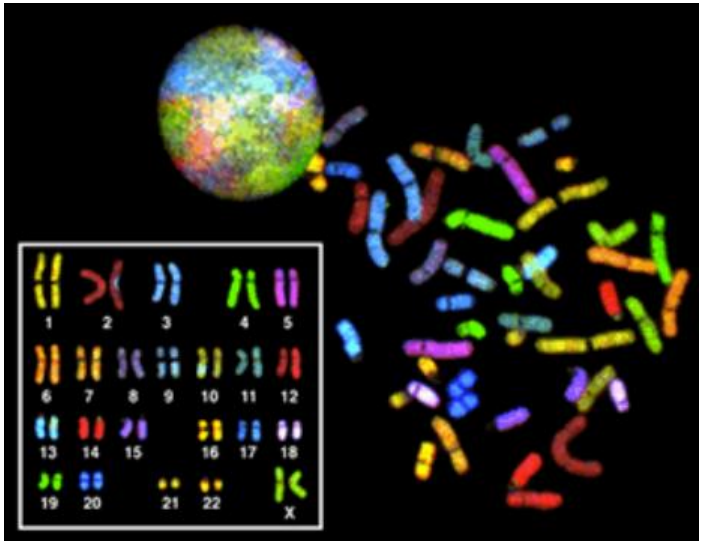


Blue line: individuals with two copies
Red line: individual with zero copy

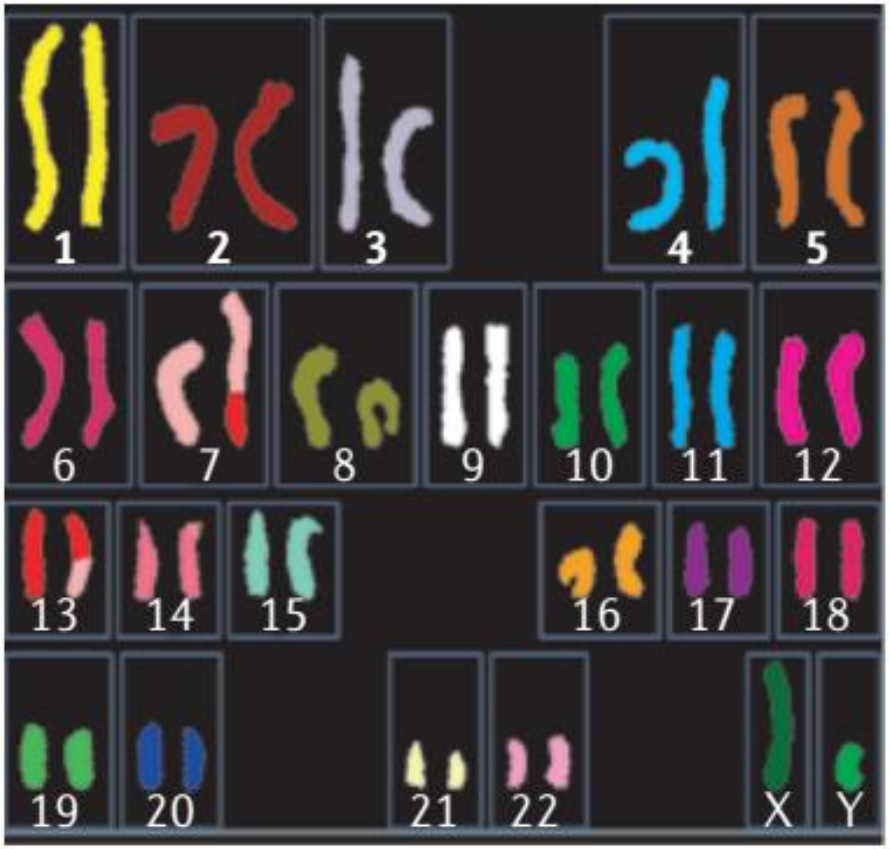
Spectral karyotyping (SKY)

- 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.

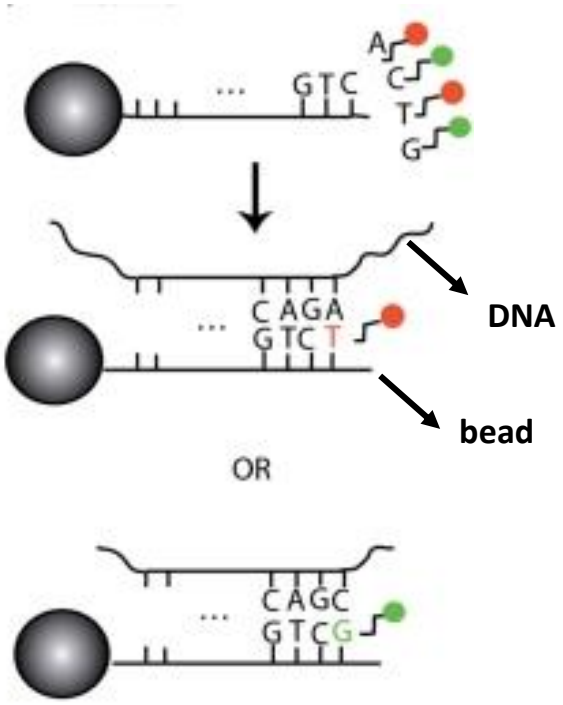
Feuk et al. Nature Reviews Genetics, 2006

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

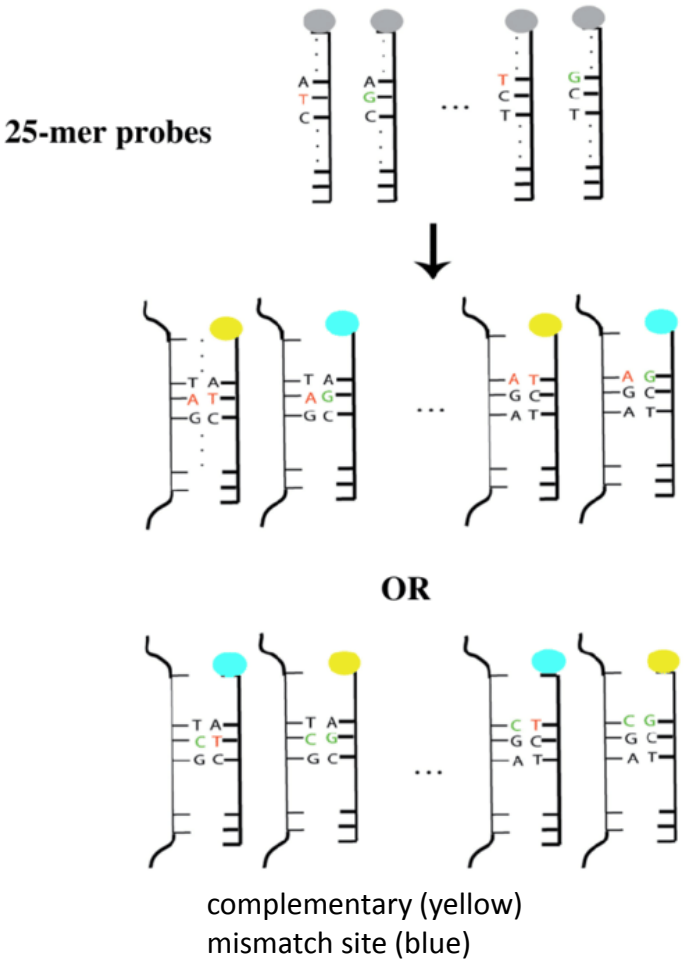


Nucleic Acids Res. 2009 Jul; 37(13): 4181–4193.

- 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

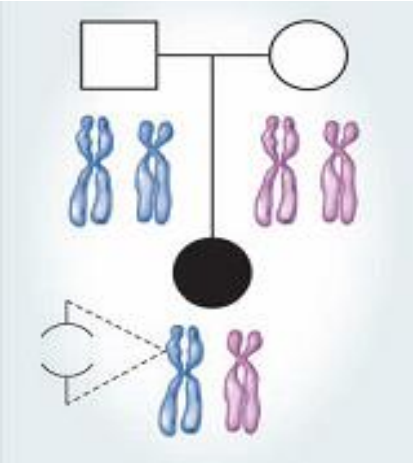
- In the Affymetrix assay, there are 25-mer probes for both alleles.
  - Assuming there are two alleles (e.g. A-Allele and B-allele) at a particular site.
  - The DNA can bind to both probes
  - But will have much higher affinity for the perfectly matched probe.
  - For example,
    - if the DNA is B-allele,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



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

- What we know are: LRR and BAF
- What we want to know is: copy number

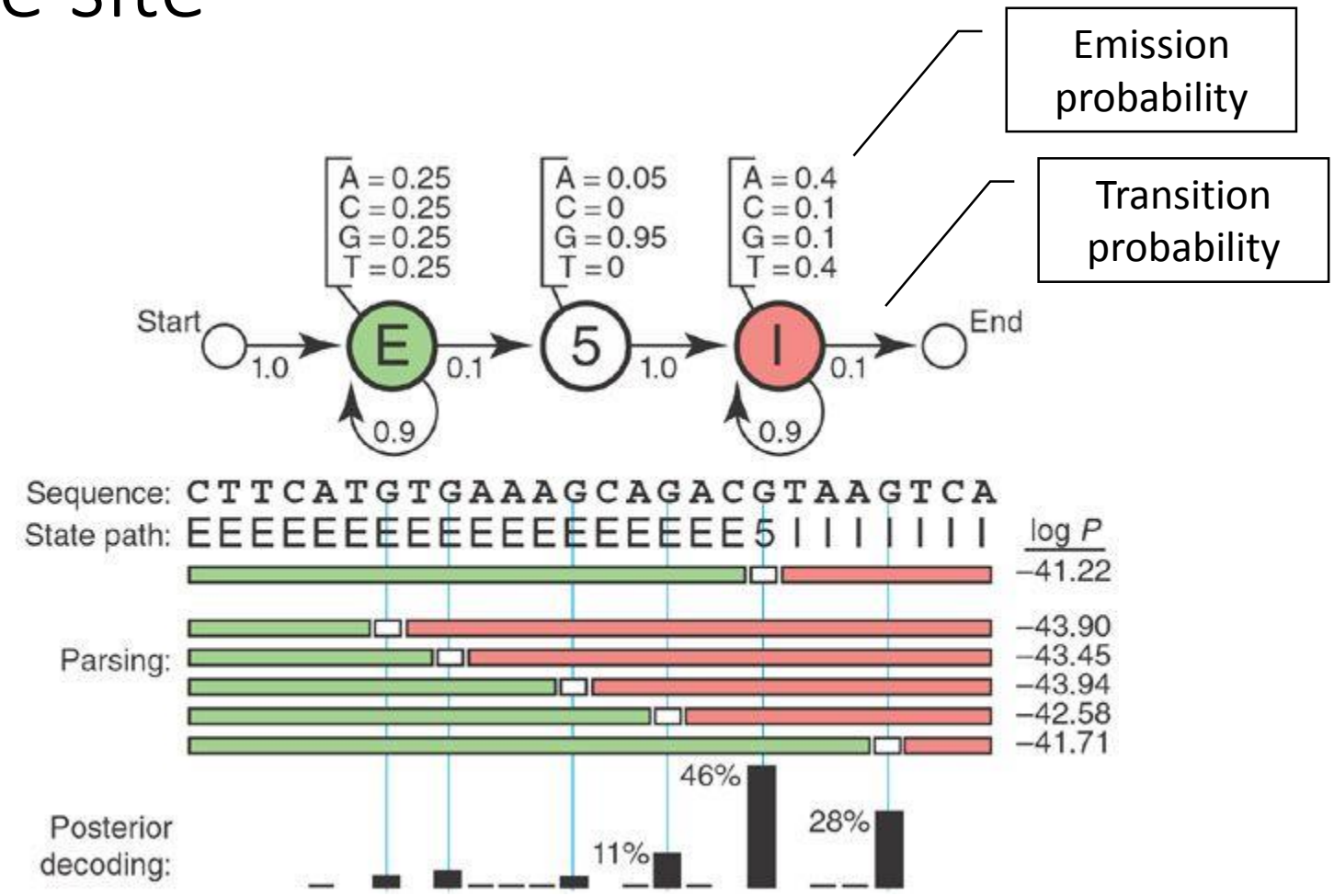
- PennCNV is an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data

Wang et al , Genome Research, 2007

What is HMM?

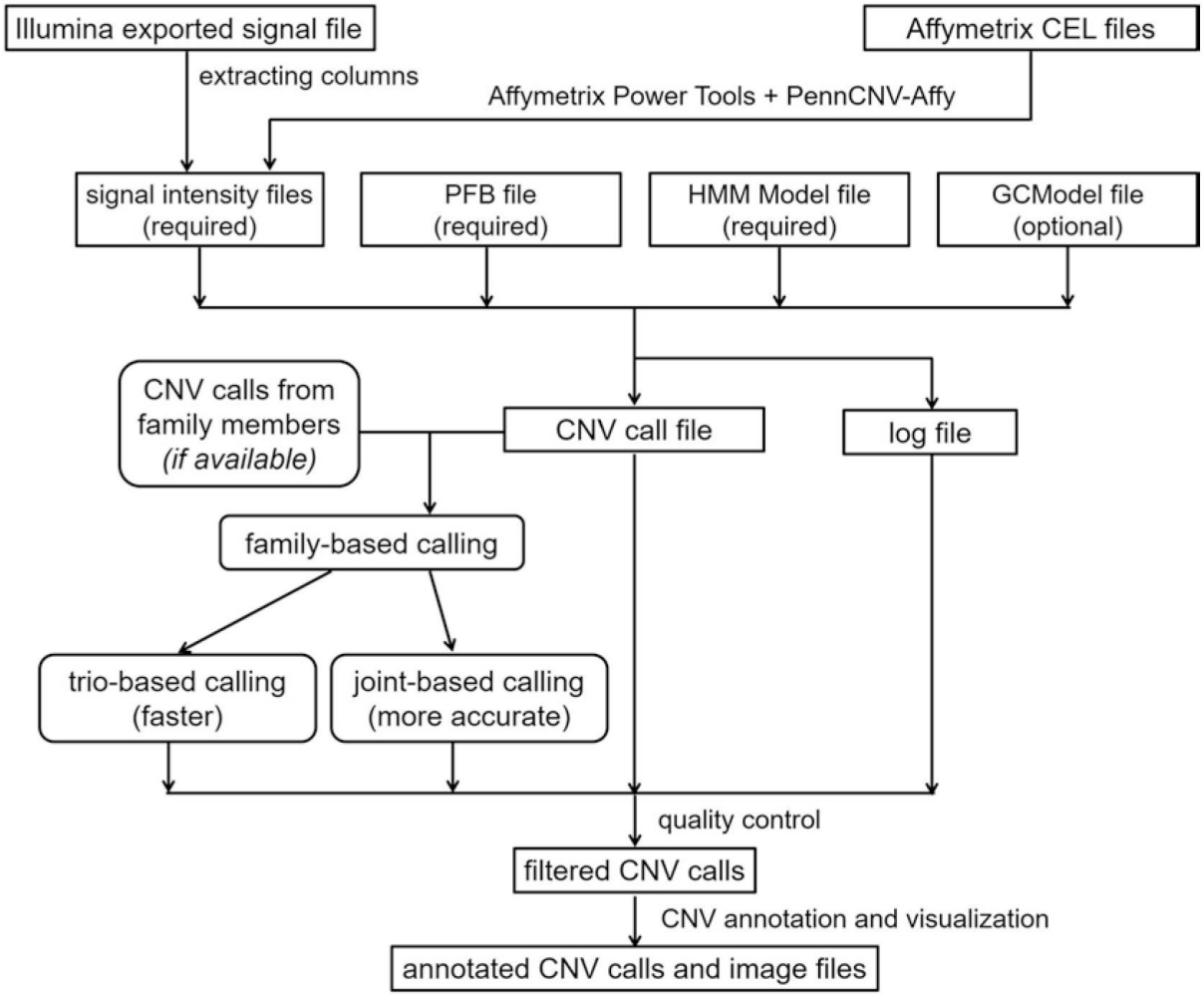
- Markov chain: a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event
- HMM: a statistical model in which the system being modeled is assumed to be a Markov process with unobservable (i.e. hidden) states.

An example of HMM to predict 5' splice site

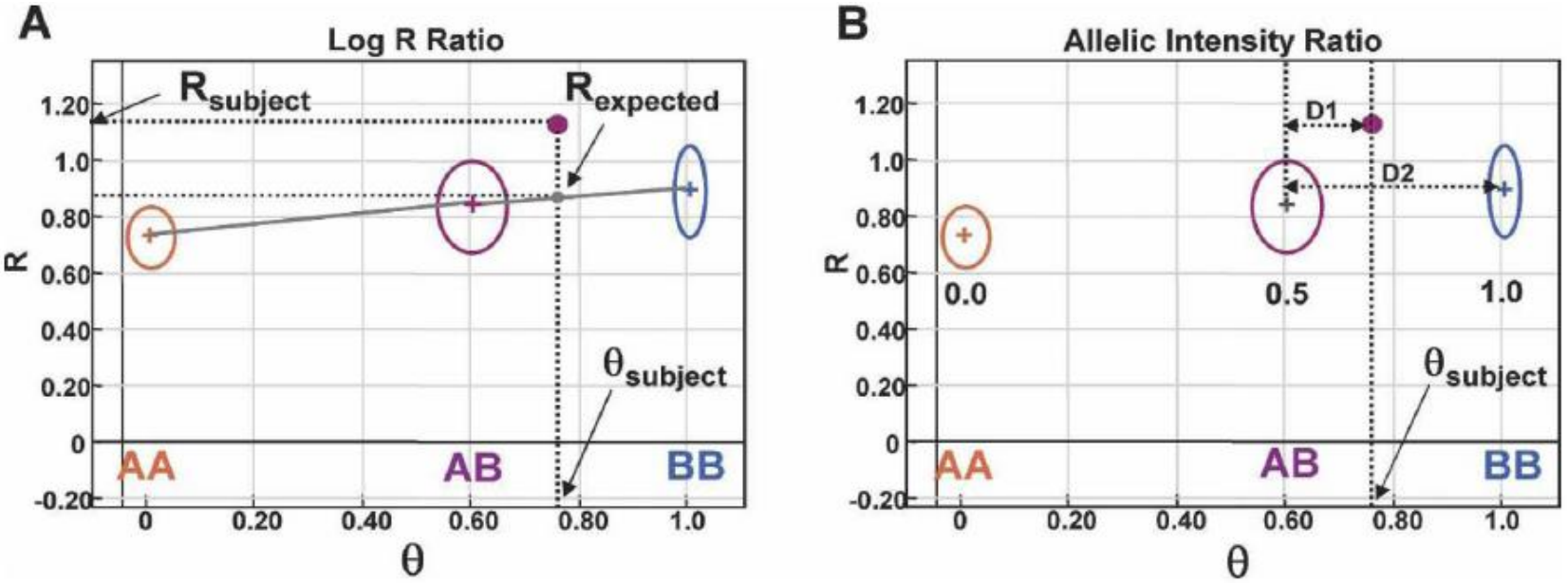


Eddy et al , Nature Biotechnology, 2004

PennCNV Flowchart



SNP Signal Intensities



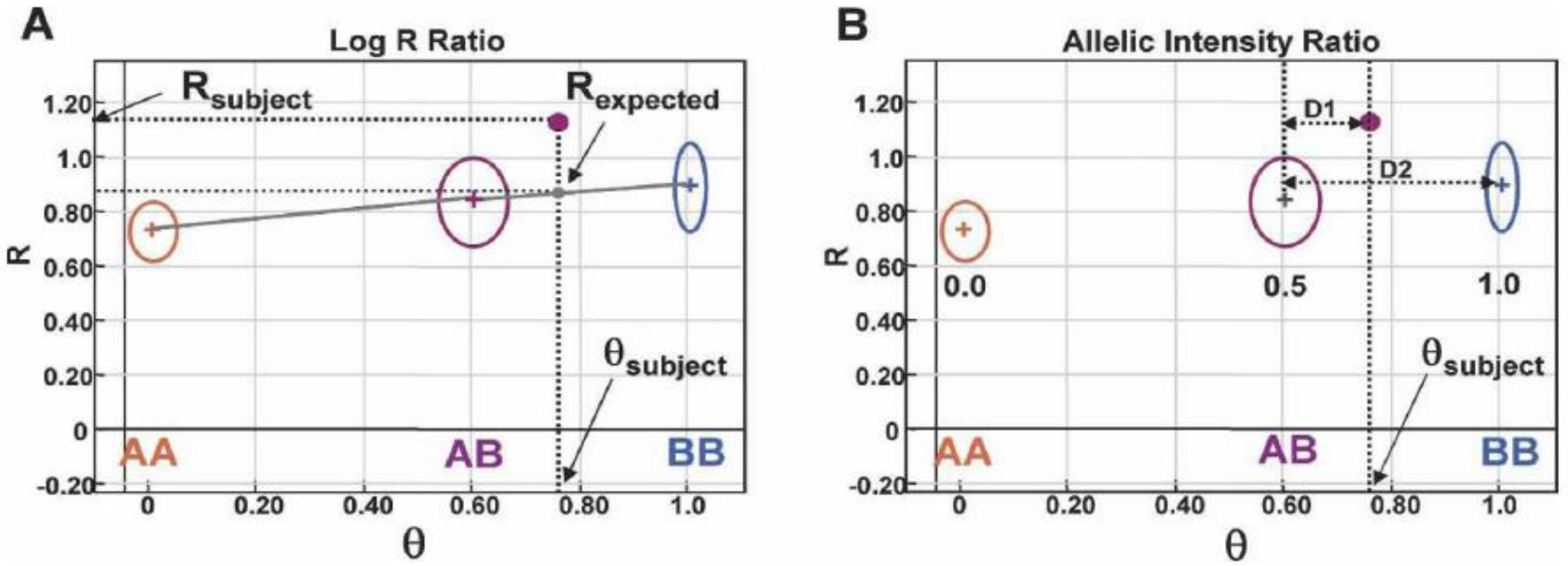
$$R = X_A + X_B, \quad \theta = (2/\pi) \times \arctan(X_B/X_A)$$

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

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

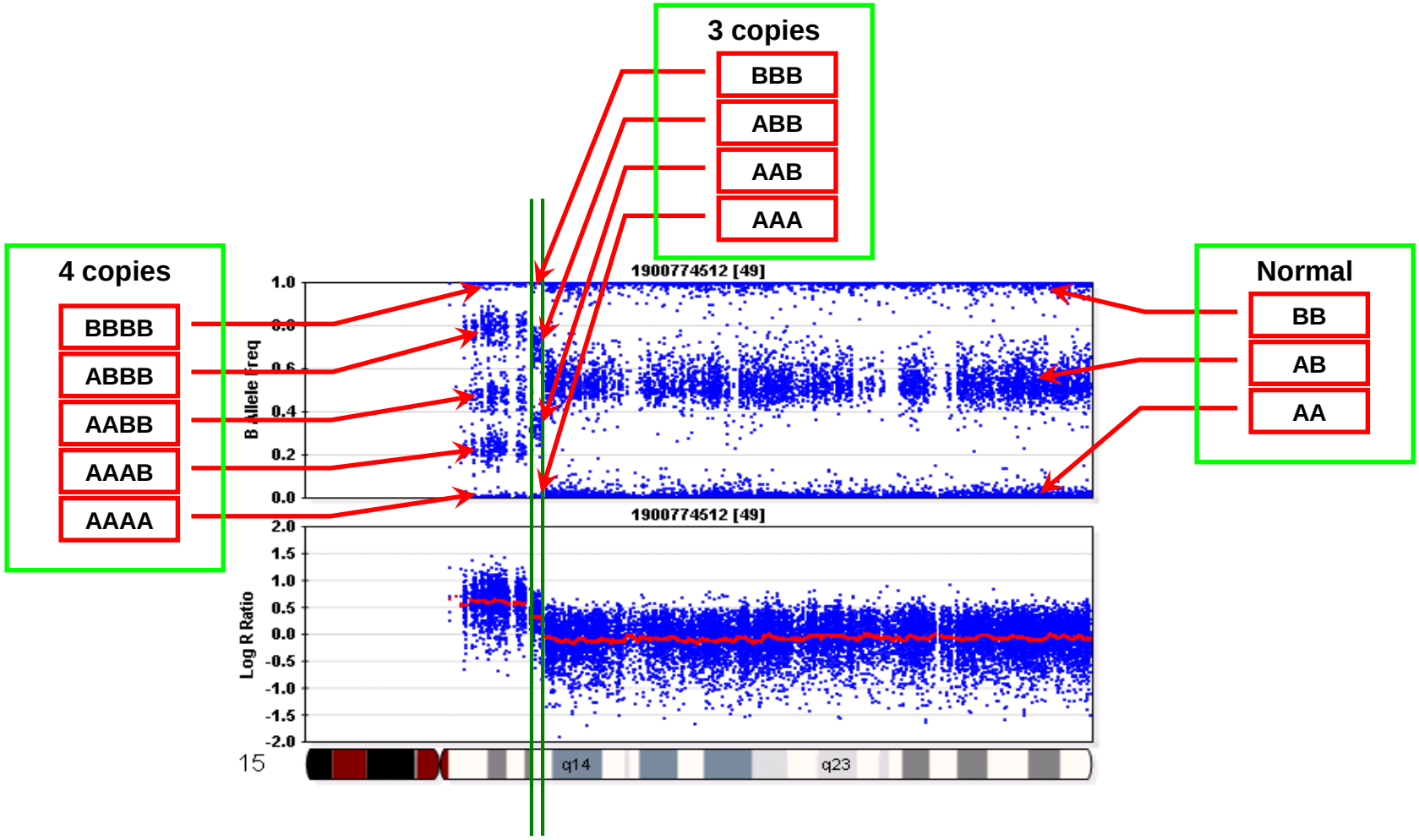
$R_{expected}$: calculated based on a reference dataset assuming copy number = 2

SNP Signal Intensities

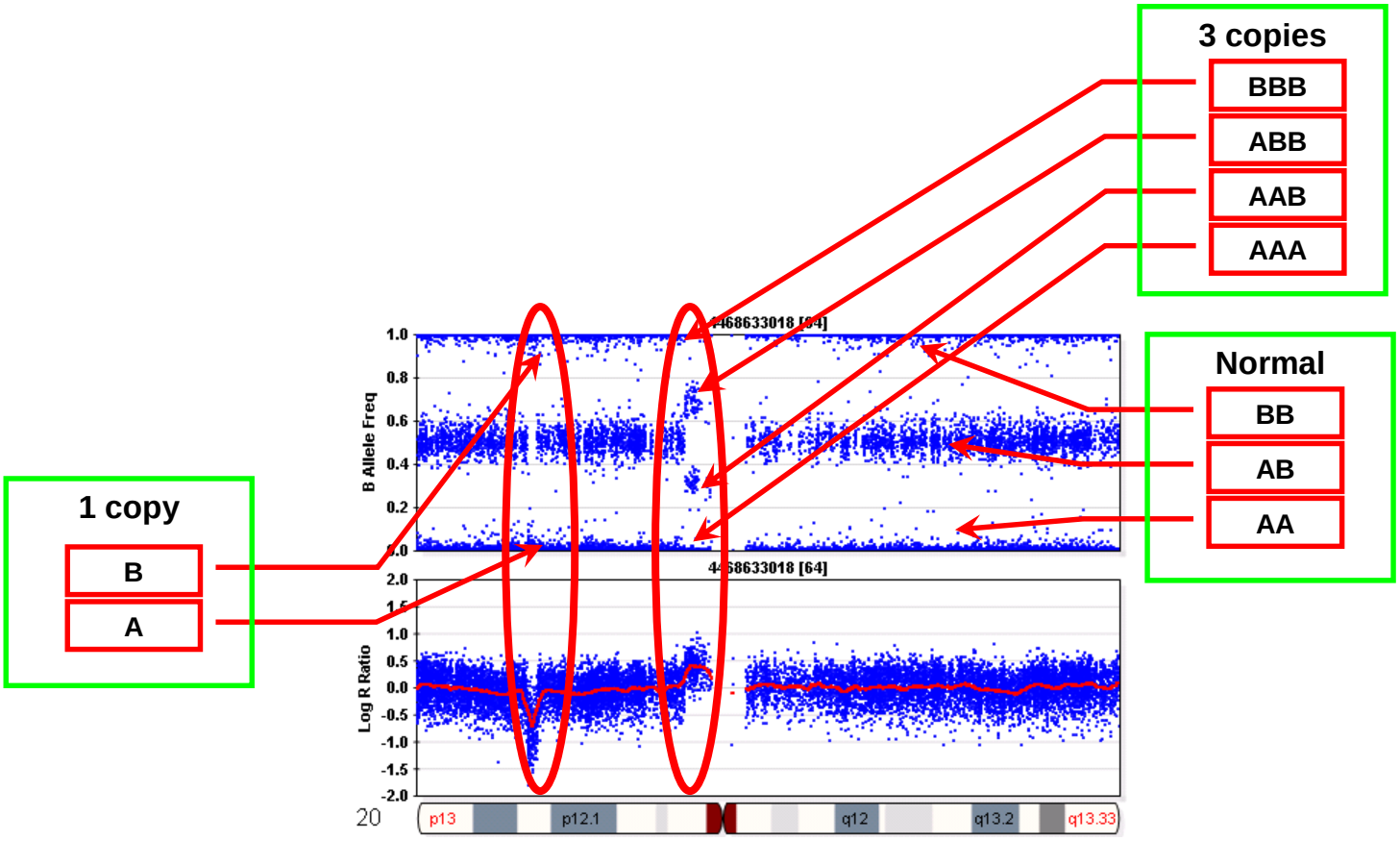


$$BAF = \begin{cases} 0, & \text{if } \theta_{\text{subject}} < \theta_{AA} \\ 0.5(\theta_{\text{subject}} - \theta_{AA})/(\theta_{AB} - \theta_{AA}), & \text{if } \theta_{AA} \le \theta_{\text{subject}} \le \theta_{AB} \\ 0.5 + 0.5(\theta_{\text{subject}} - \theta_{AB})/(\theta_{BB} - \theta_{AB}),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

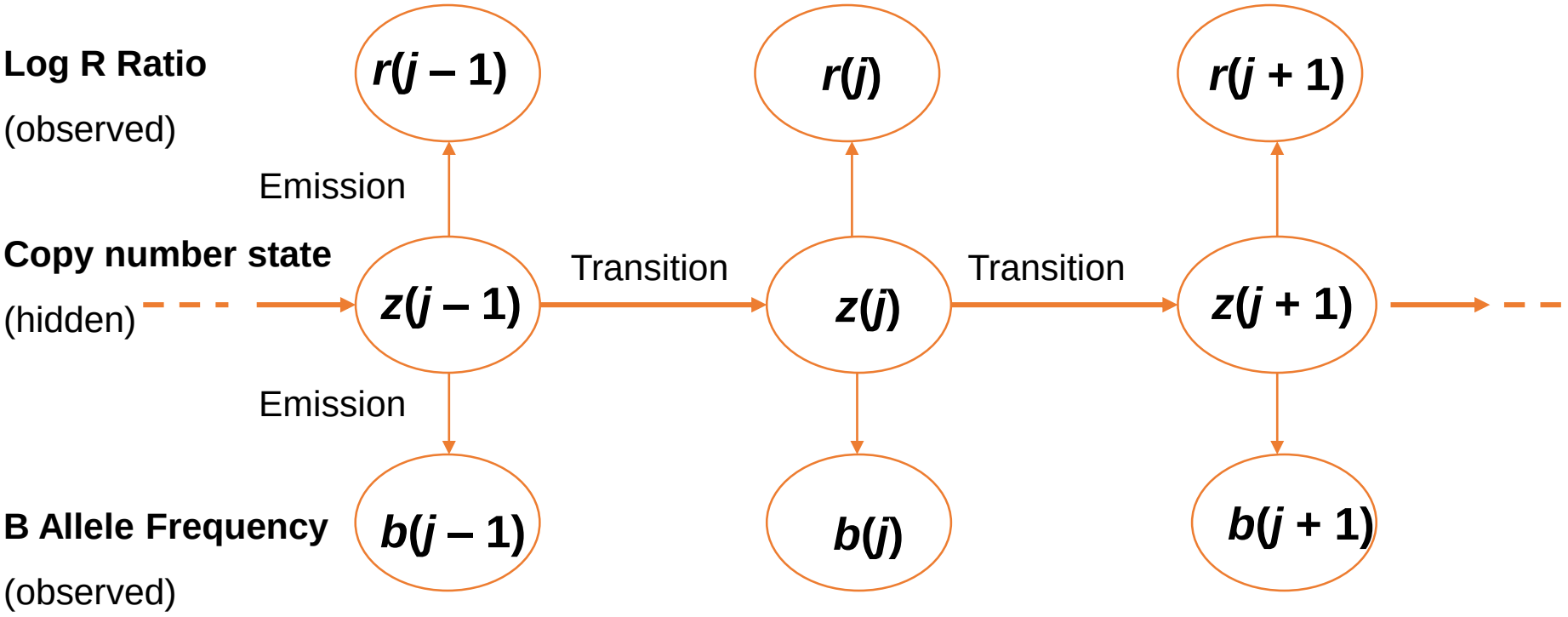
Visualization of CNVs



Visualization of CNVs



Hidden Markov Model in PennCNV



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)

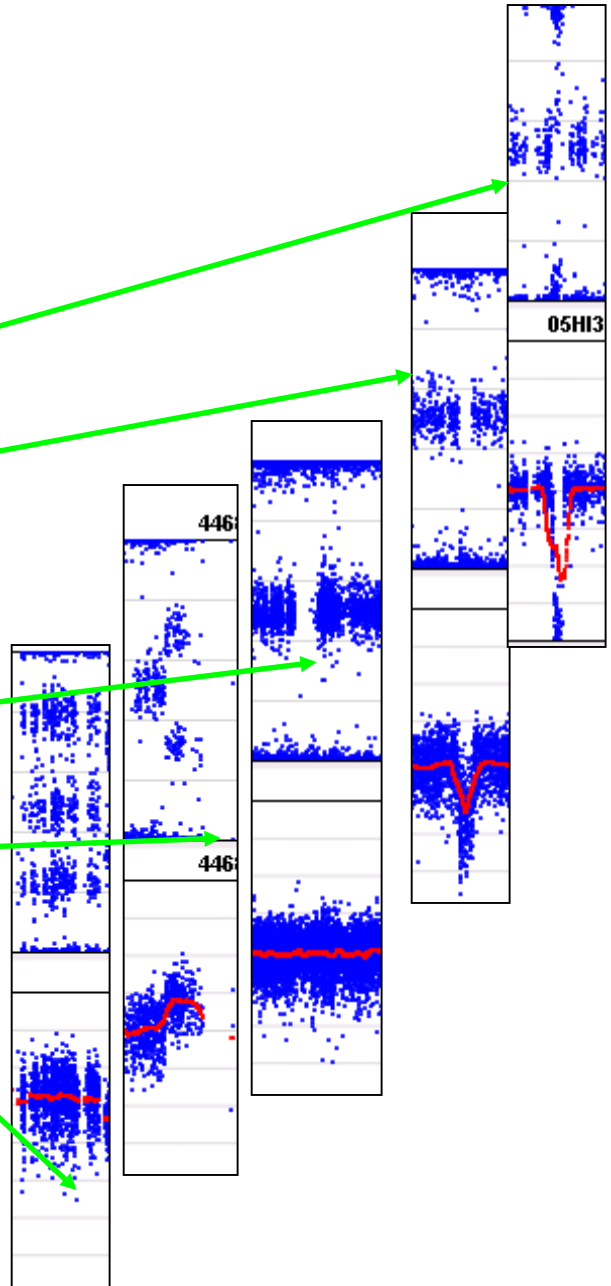


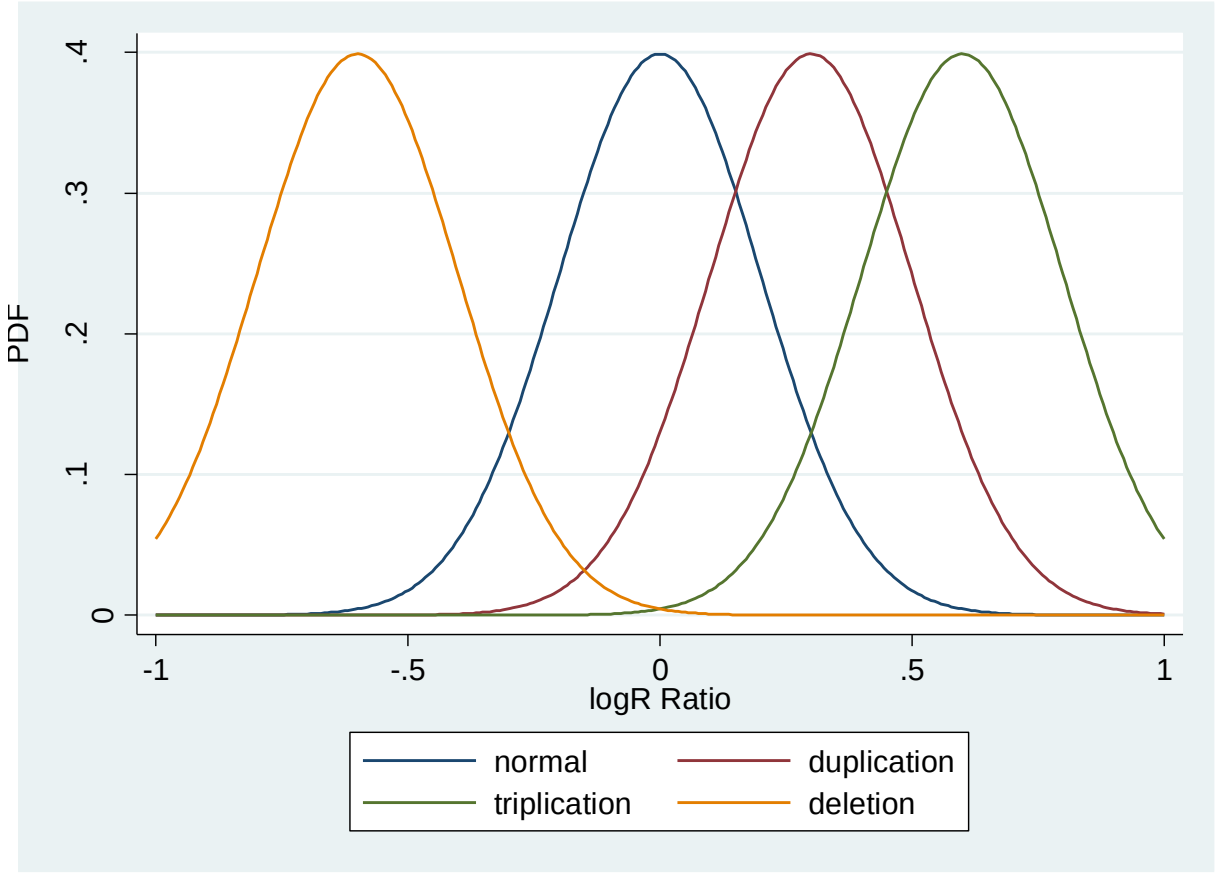
Table 1. Hidden states, copy numbers, and their descriptions

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

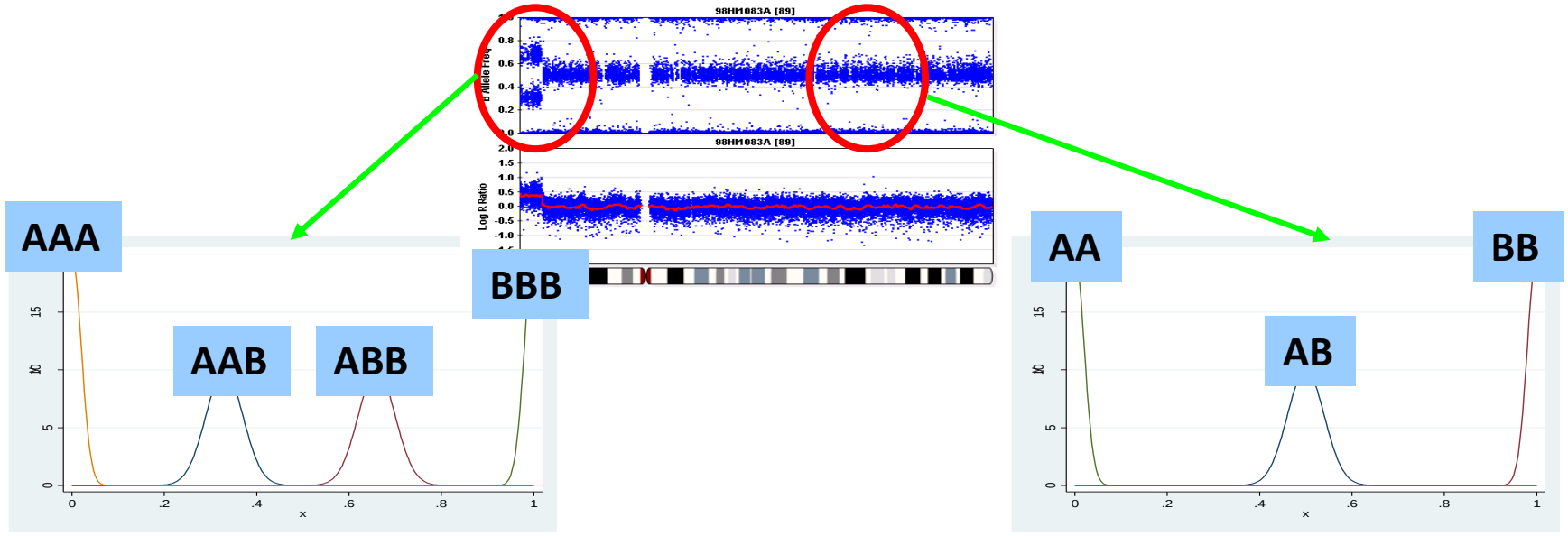
Each state has a different distribution of CNV genotypes.

Emission Probability of LRR

- Given a copy number state, LRR is normally distributed



Emission Probability of BAF



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

CNV Calling

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

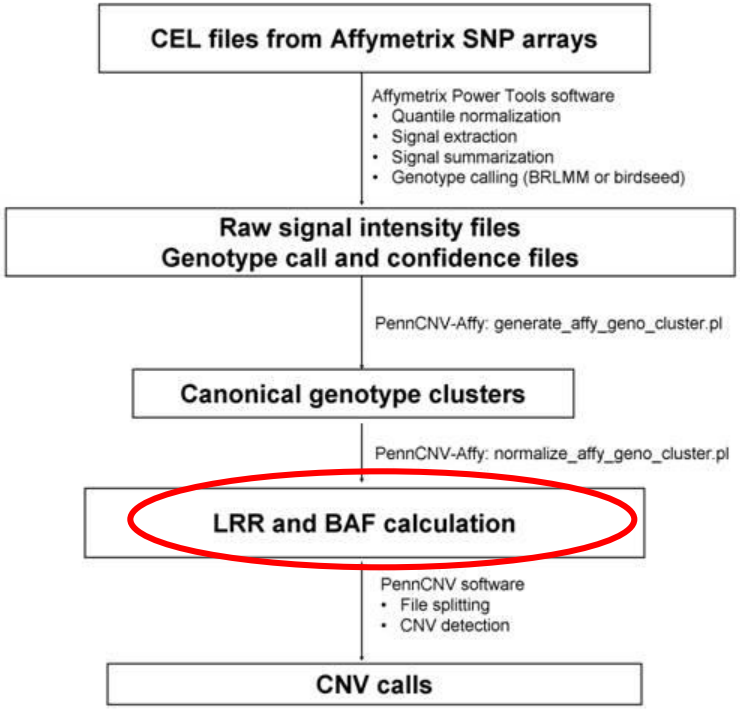
- Viterbi algorithm for calling
  - Calculate the most likely path in HMM (a path of state 1-6 for each SNP marker)
  - Collect any non-normal state path as the CNV calls
  - Example:

```
Most likely CN sequence is:
22222222111111122222222222222222222224444422222222222
           CNV!                                  CNV!
```

Other Types of Signal Data

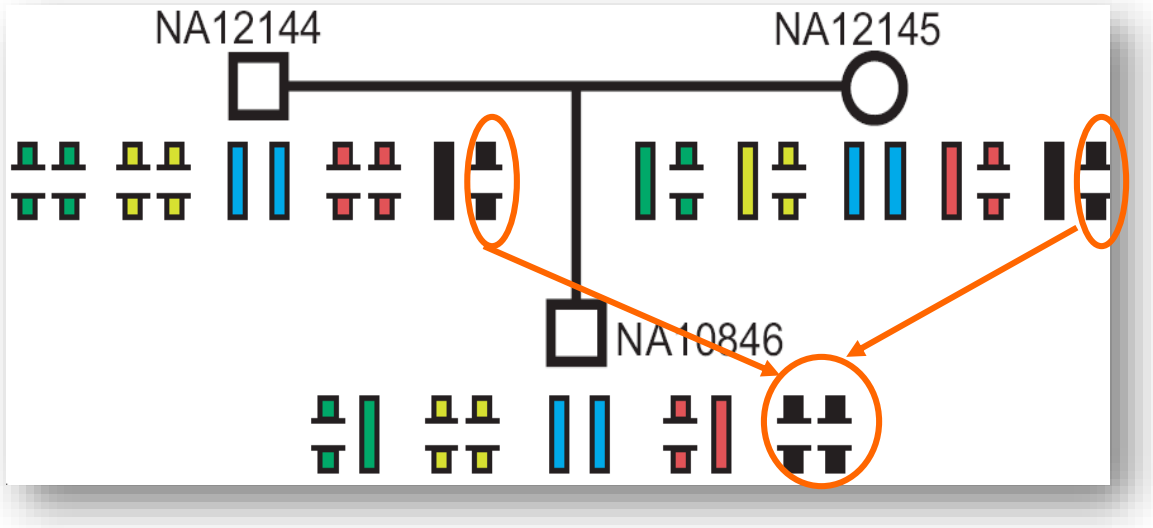
- PennCNV can be applied to data from other technical platforms:
  - Transformation of signal data to LRR/BAF:
    - Affymetrix whole-genome SNP genotyping array
    - Perlegen whole-genome SNP genotyping array
  - Use information from LRR only:
    - BAC clone based array-CGH
    - Oligonucleotide arrays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

PennCNV-Affy Pipeline



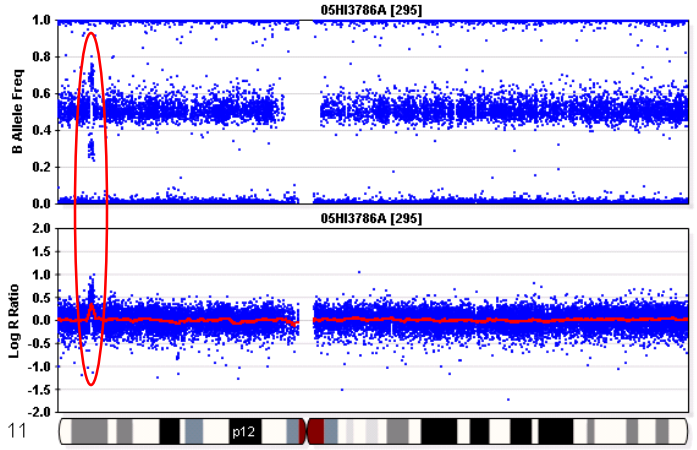
Joint Modeling on Family Data

- Most CNVs demonstrate Mendelian inheritance

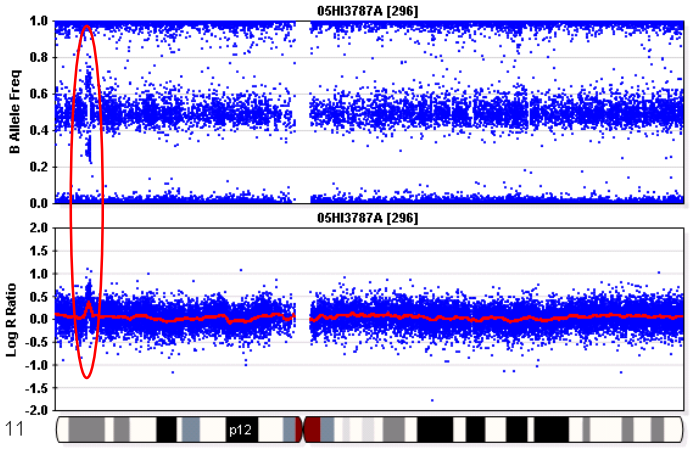


- Incorporate family relationship can potentially improve sensitivity of CNV calling

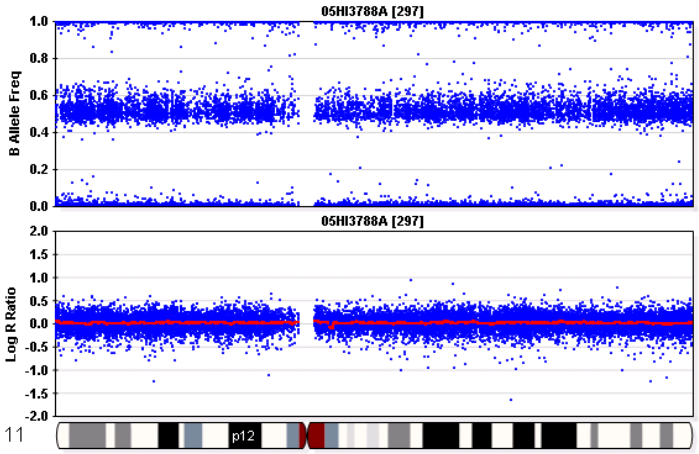
Example of Inherited CNV



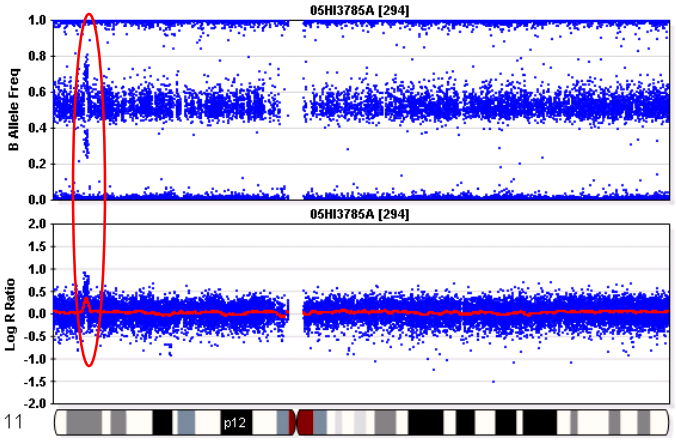
Father



Child 1

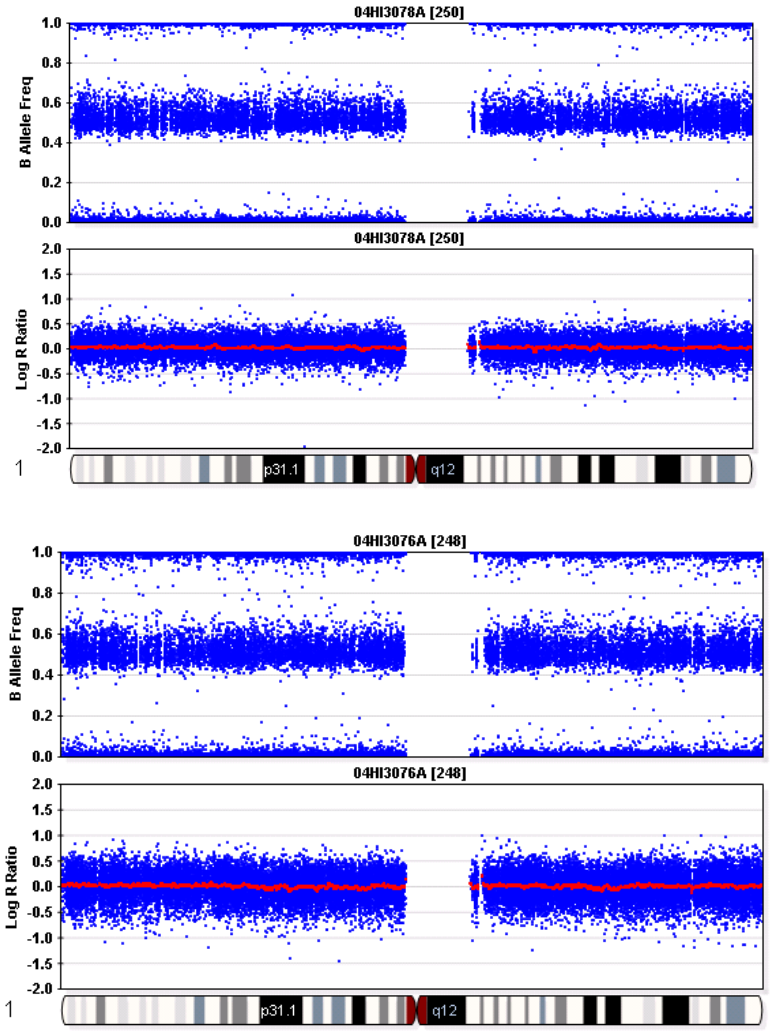


Mother

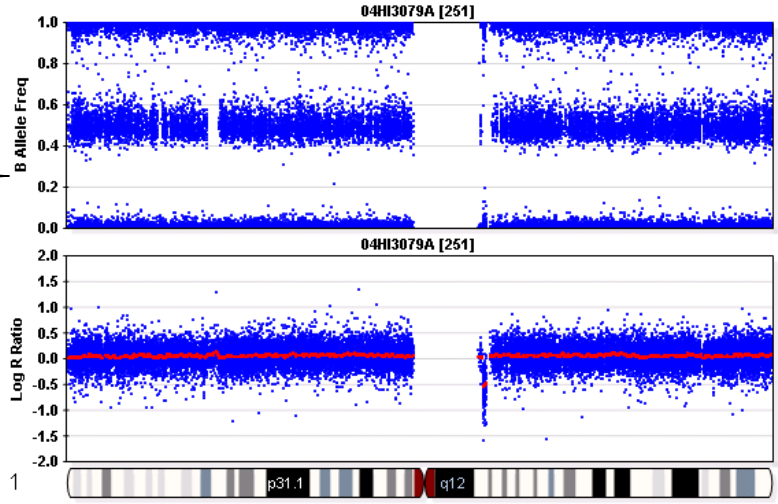


Child 2

Example of *de novo* CNV

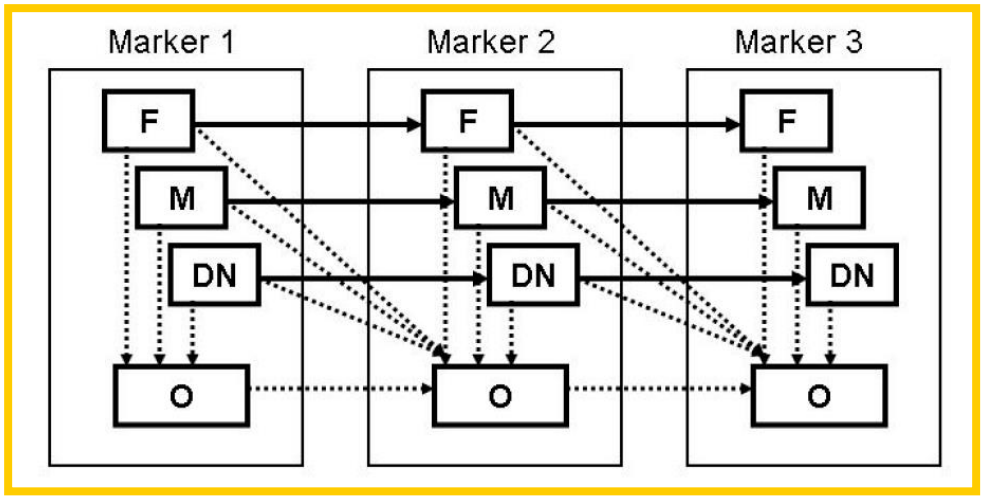


Some CNVs are due to *de novo* events, which occur as germline, somatic or cell line-induced chromosome aberrations in offspring that are not inherited from either parent.



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

$$
\begin{aligned}
&P(r_1,\ldots,r_T,b_1,\ldots,b_T \mid \lambda) \\
&= \sum_{z_1}\cdots\sum_{z_T}\sum_{DN_1}\cdots\sum_{DN_T}\{P(r_1,\ldots,r_T \mid z_1,\ldots,z_T,\lambda)\times \\
&\qquad P(b_1,\ldots,b_T \mid z_1,\ldots,z_T,\lambda)\times \\
&\qquad P(z_1,\ldots,z_T \mid DN_1,\ldots,DN_T,\lambda)\times \\
&\qquad P(DN_1,\ldots,DN_T \mid \lambda)\} \\
&= \sum_{z_1}\cdots\sum_{z_T}\sum_{DN_1}\cdots\sum_{DN_T}\{P(r_1 \mid z_1,\lambda)P(b_1 \mid z_1,\lambda)P(z_1 \mid DN_1,\lambda)P(DN_1 \mid \lambda)\times \\
&\qquad \prod_{j=2}^{T}\left[P(r_i \mid z_i,\lambda)P(b_i \mid z_i,\lambda)P(z_i \mid z_{i-1},DN_i,DN_{i-1},\lambda)P(DN_i \mid DN_{i-1},\lambda)\right]\}.
\end{aligned}
$$

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

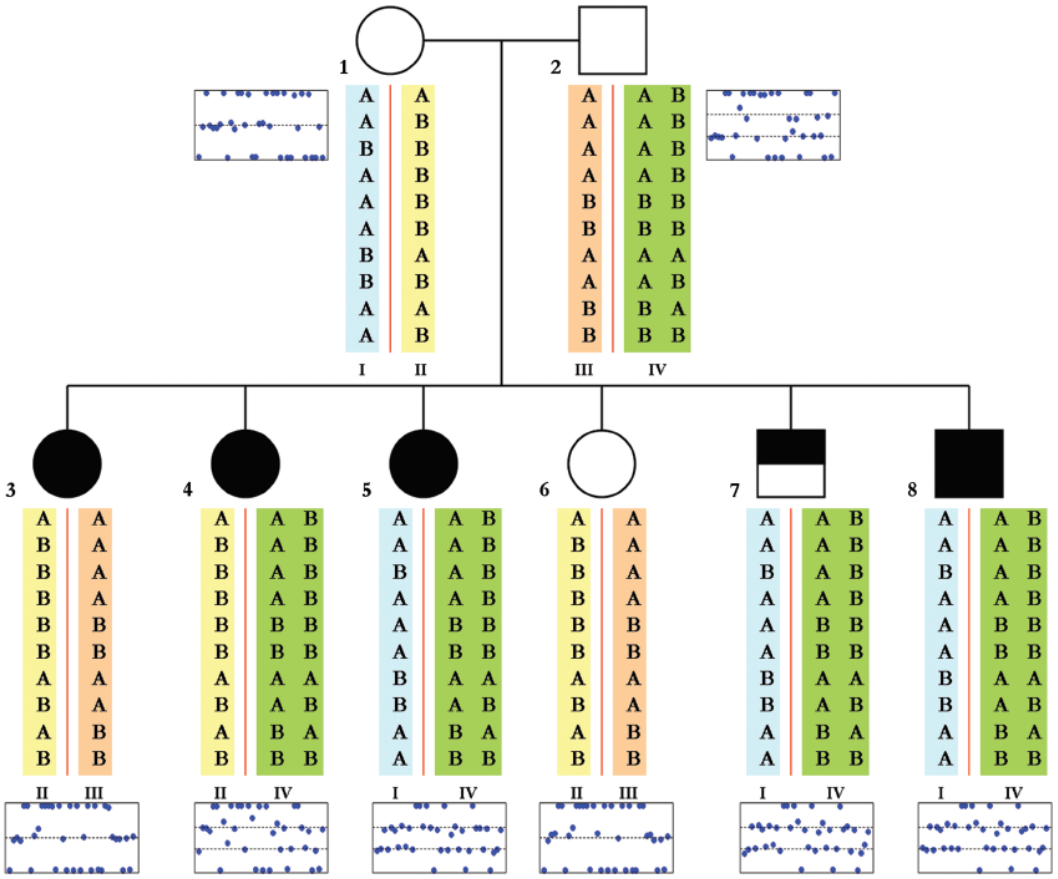
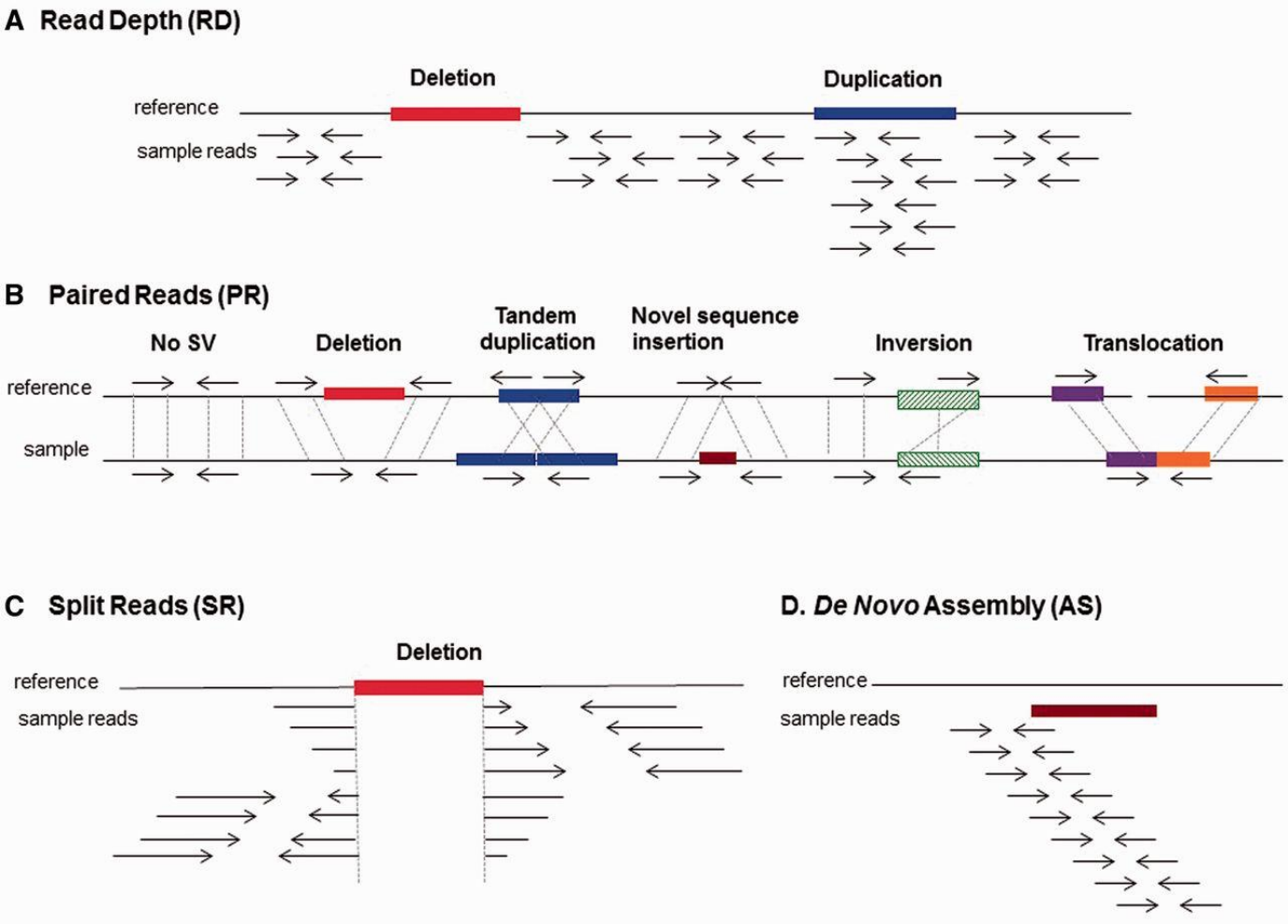


Figure 4. Illustration of a duplication CNV on 10q11.22 that exists in the father and is transmitted to four offspring. The CNV calls are made on six trios separately by the joint-calling algorithm. For each individual, the BAF values for all SNPs within the CNV and the chromosome-specific SNP genotypes (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.

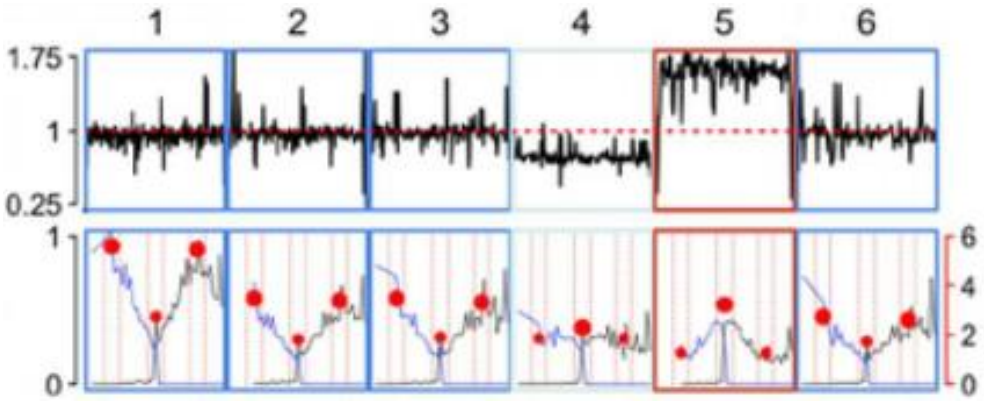
NGS-based SV detection



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

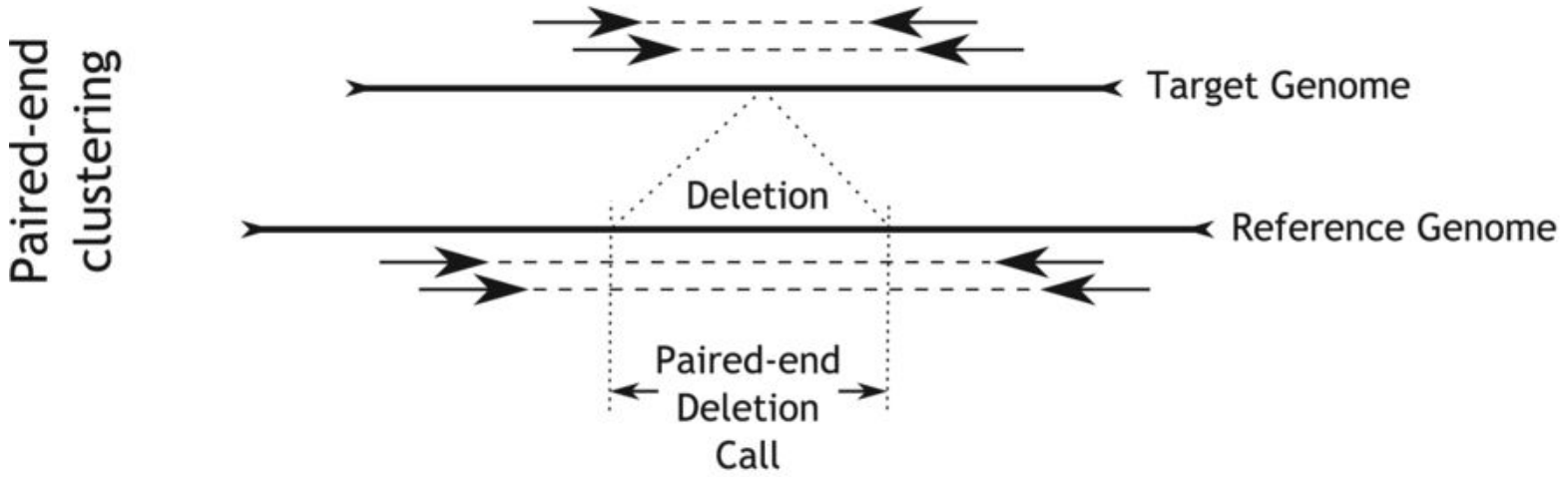
Read count-based methods for SV detection

- 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

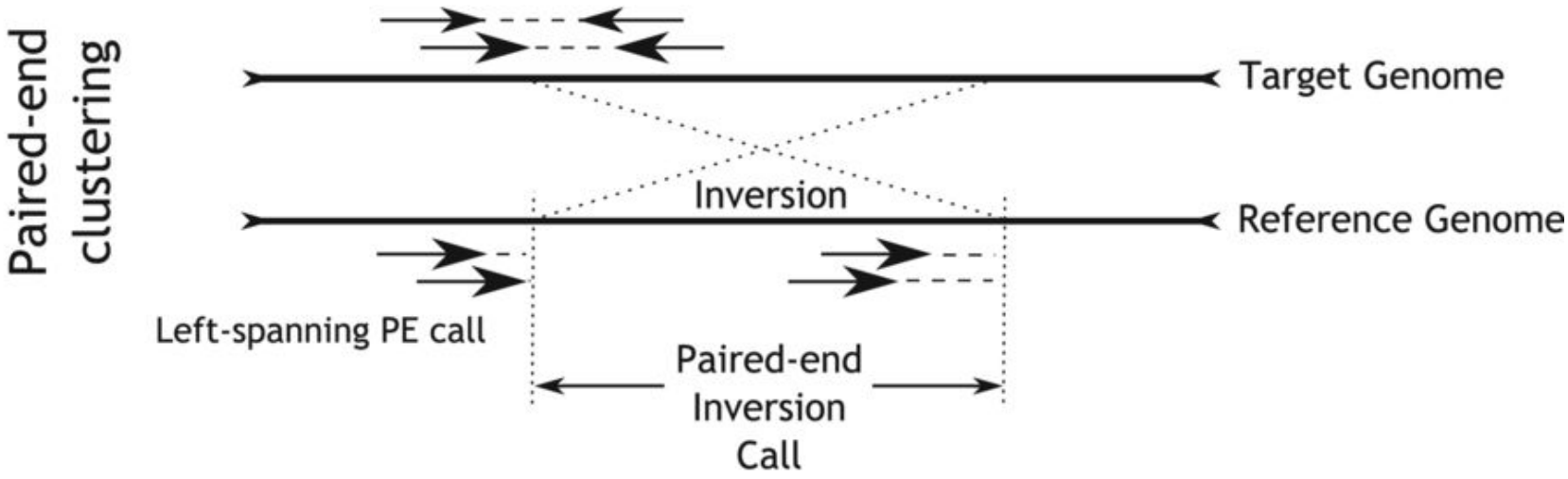
- Widely used software tools: Delly, Lumpy
- Pattern of deletions: large gaps between read pairs:



Rausch T, et al. (2012) Bioinformatics

Detection of SVs from discordant read pairs

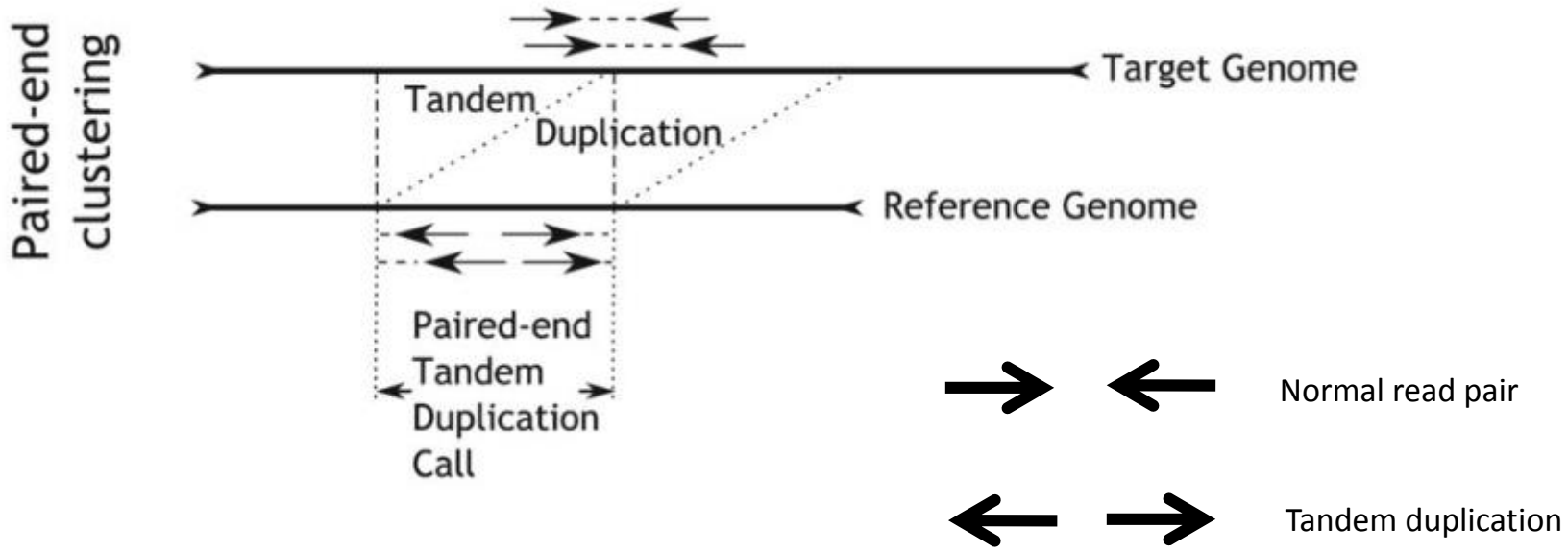
- Pattern of inversions: same orientation between read pairs:



Rausch T, et al. (2012) Bioinformatics

Detection of SVs from discordant read pairs

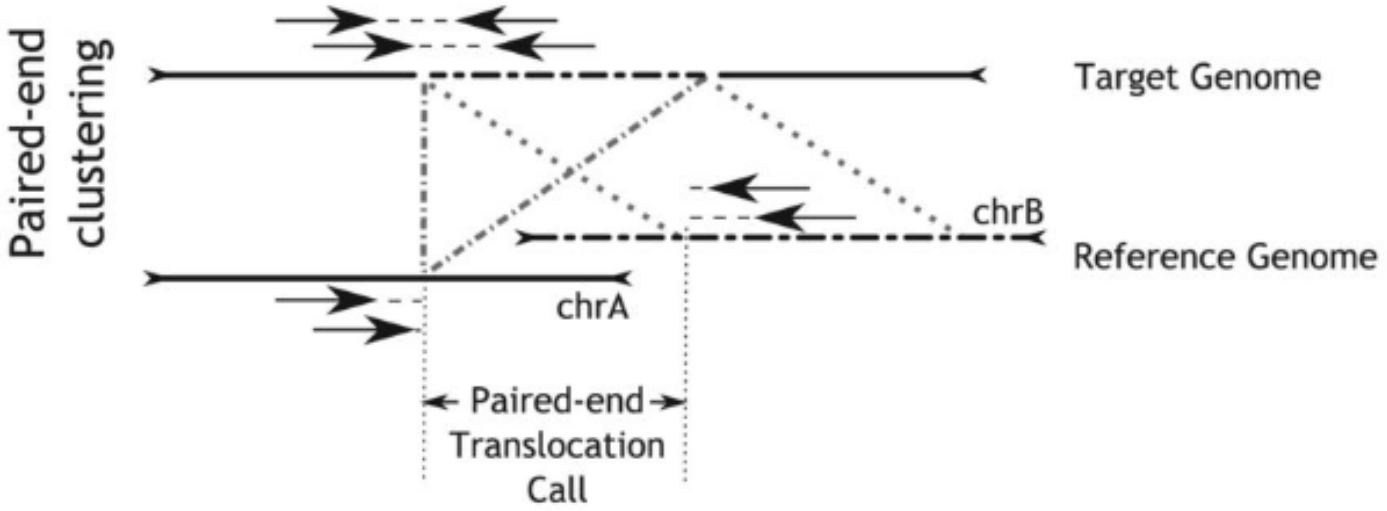
- Pattern of tandem duplication: the first and second read changed their relative order



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)

SV detection from long-read sequencing

- 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

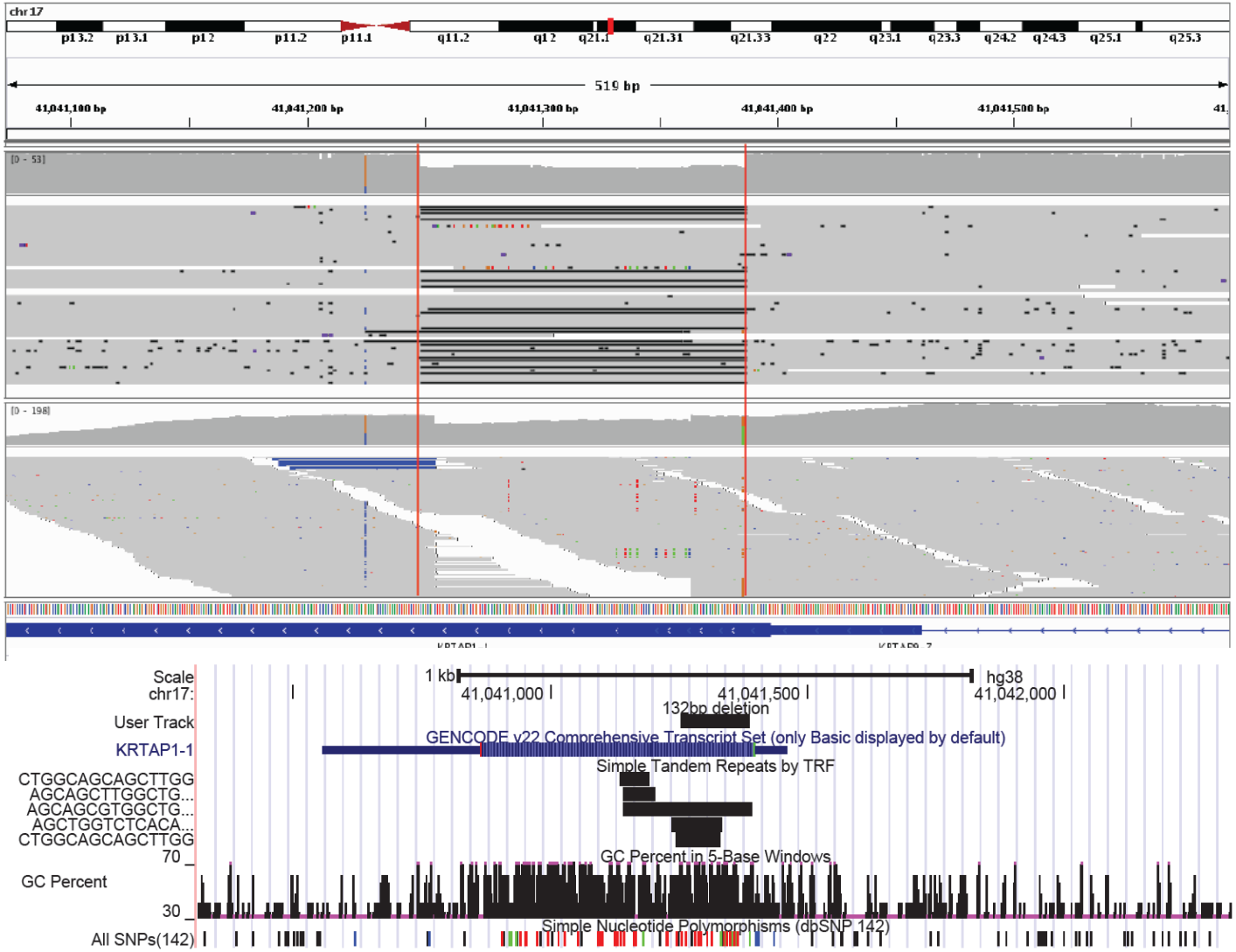
SV detection from long-read sequencing

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

Short/long reads on SV detection

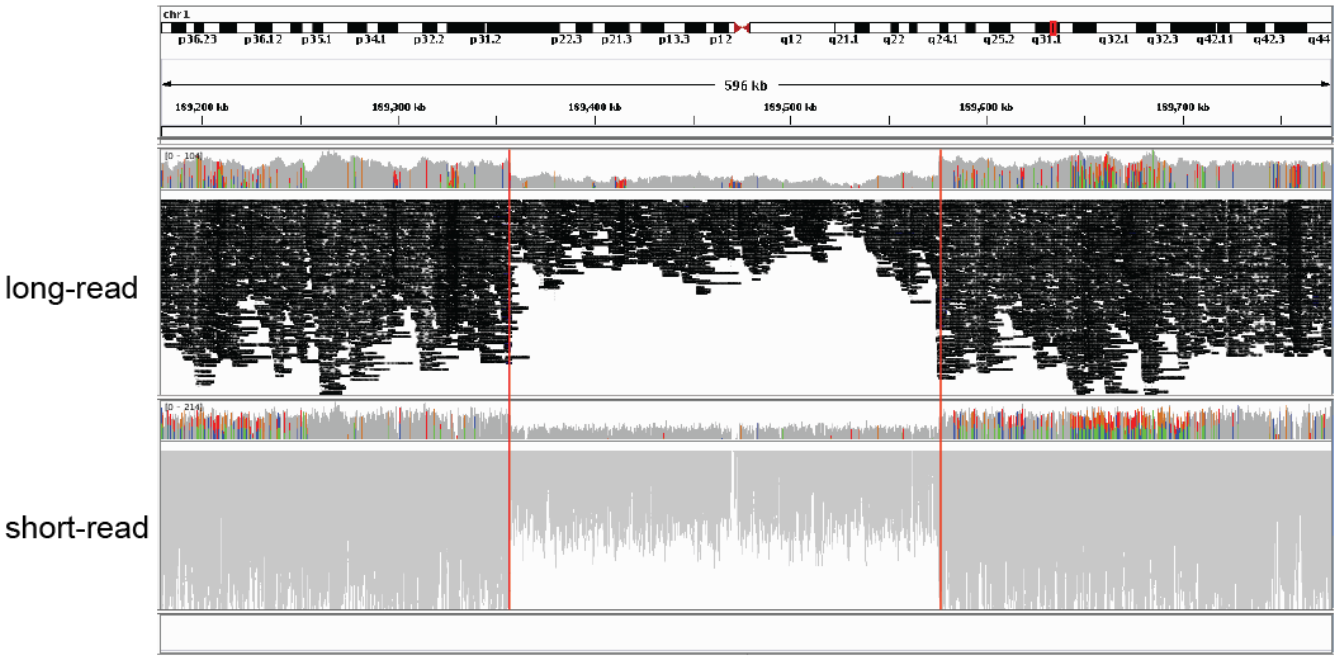
PacBio

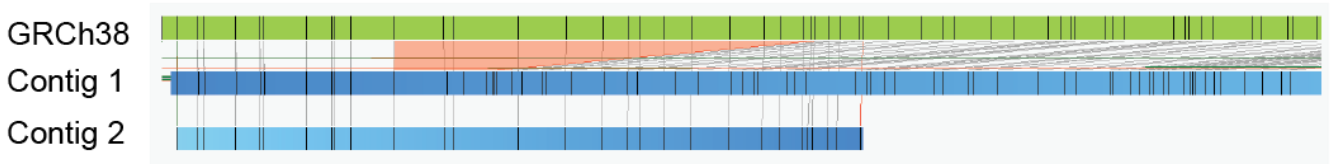
Illumina



Shi L et al, Nature Communications, 2016

Short/long reads on SV detection





Shi L et al, Nature Communications, 2016

Structural Variant Detection with Sniffles

- Step 1: Read filtering
  - To reduce false positive SV calls Sniffles stringently filters for spurious read mappings.
  - A read is discarded if it has a mapping quality lower than 20 (by default) or the ratio of its best and second best alignment score is less than 2.
  - A read is discarded if it shows more than 7 (by default) split read alignments or if every aligned portion of the read does not exceed 1kbp (by default)

Sedlazeck et al, Nat Methods, 2018

Structural Variant Detection with Sniffles

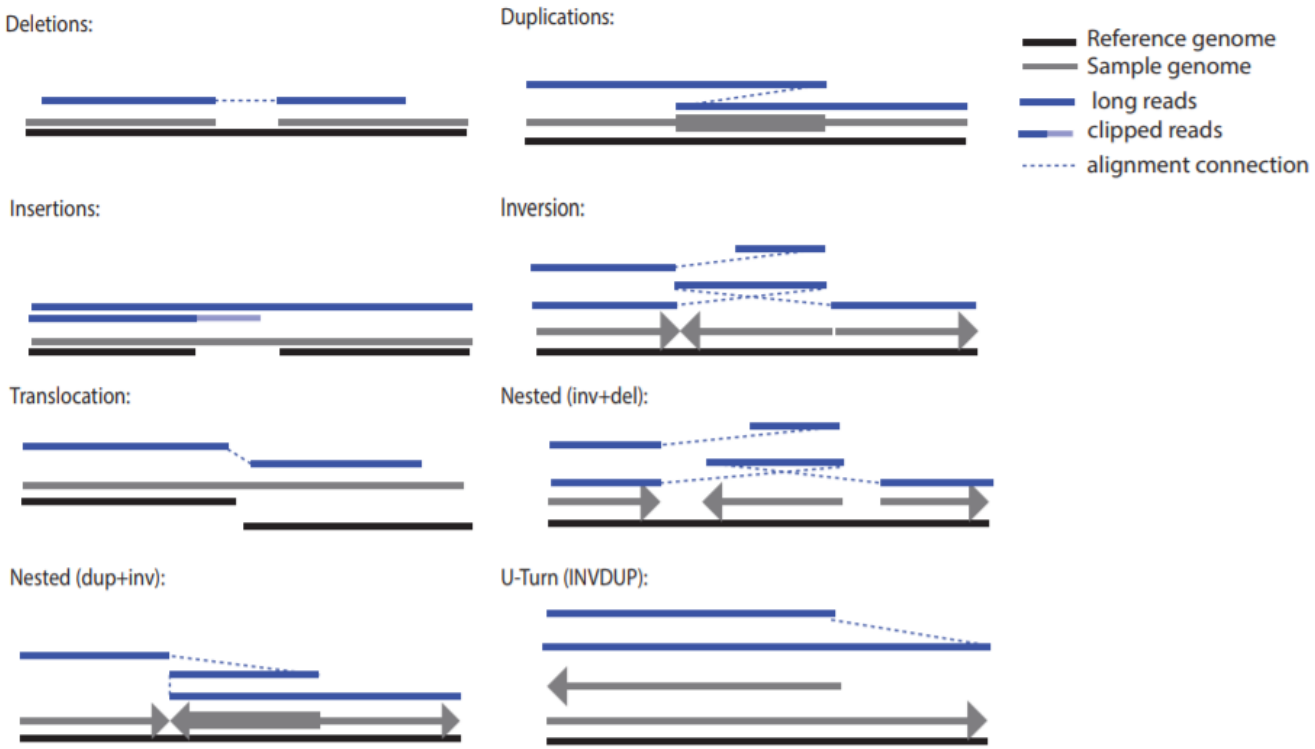
- Step 2: detection of small SV events from read alignments
  - Sniffles scans the read alignments to detect smaller (<1kb) insertions, deletions and regions with an increased number of mismatches and very short (1-5bp) indels.
  - These “noisy regions” often indicate incorrect read mappings caused by SV.

Structural Variant Detection with Sniffles

- Step 3: detection of large SV events from split reads
  - Sniffles processes split read information to identify SV that cannot be represented in a single alignment (large indels, inversions, duplications, translocations).
- Step 4: Merging SV calls that were caused by the same SV.
- Step 5: Identifying and removing spurious SV calls

Sedlazeck et al, Nat Methods, 2018

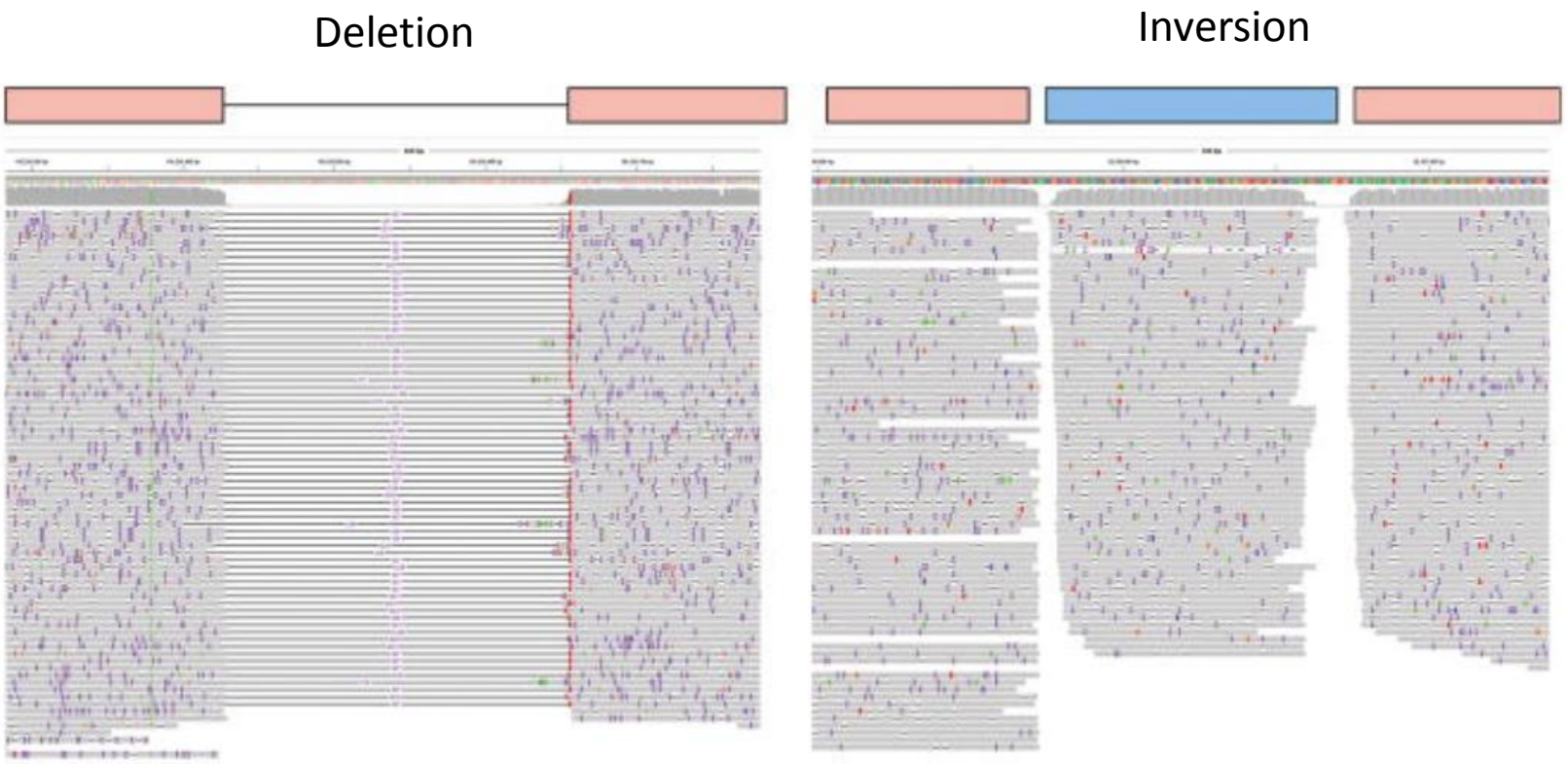
Different SV types that can be detected by Sniffles



Supplementary Figure 2.4. Schematic illustration of the different SV types Sniffles can detect and how the split reads are aligned to be able to detect these events.

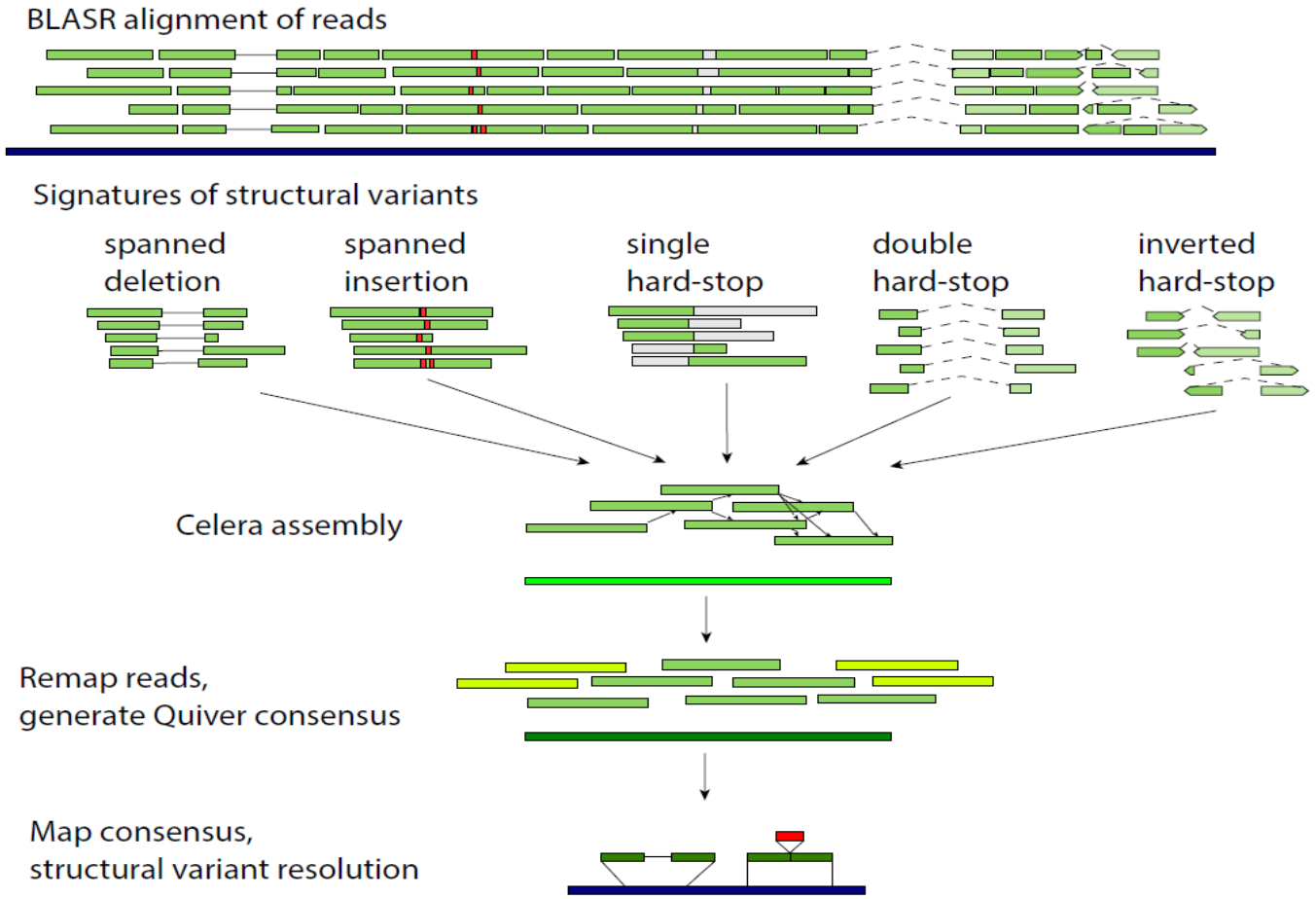
Sedlazeck et al, Nat Methods, 2018

Example of SV detection from long-read sequencing



Sedlazeck et al, Nat Methods, 2018

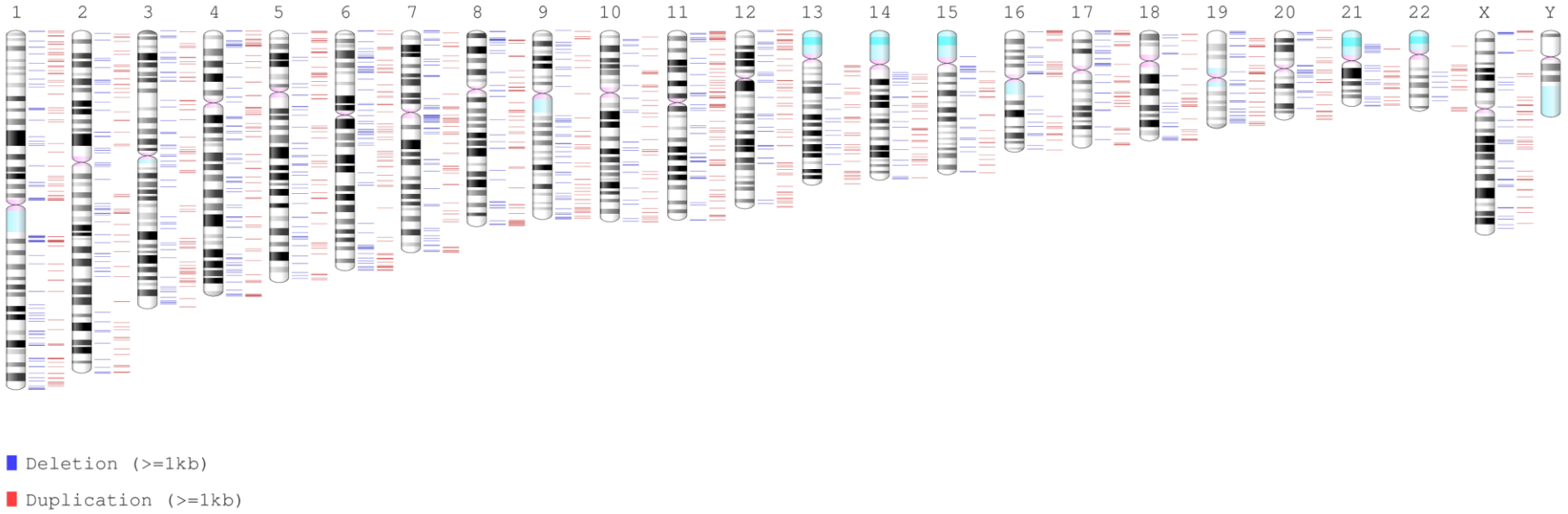
SV detection by Long-read sequencing from local assembly



Chaisson et al, Nature, 2015

SV detection from long-read sequencing via local de novo assembly

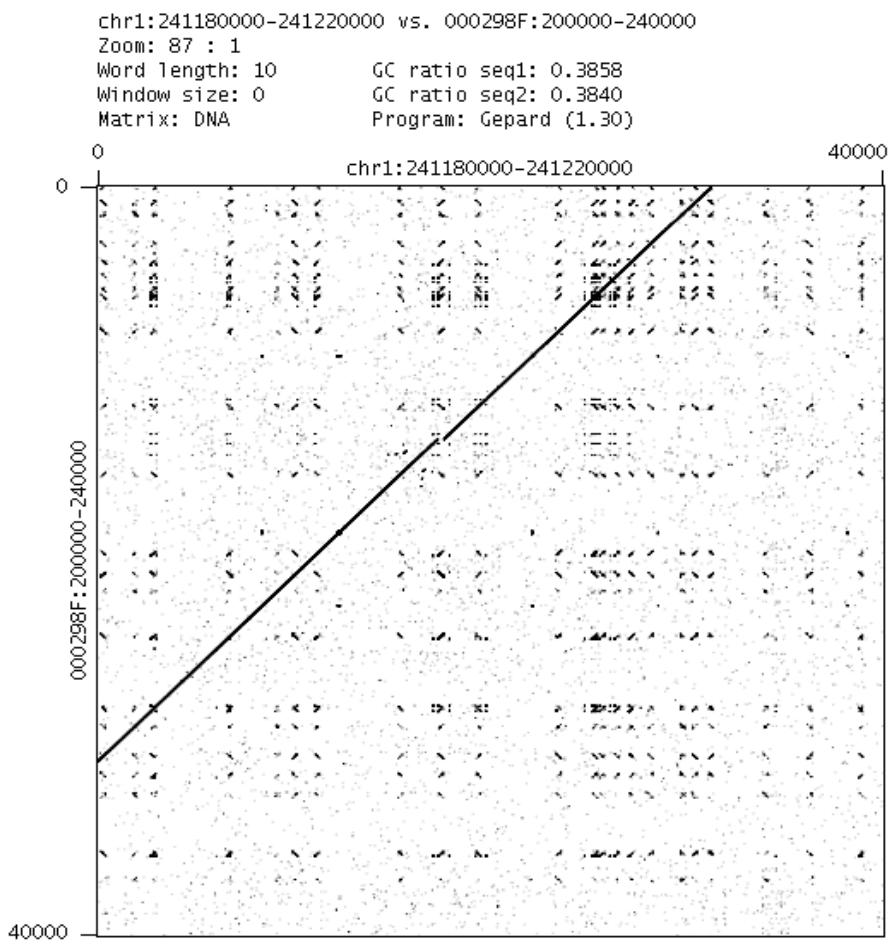
- 9,643 deletions and 10,022 insertions

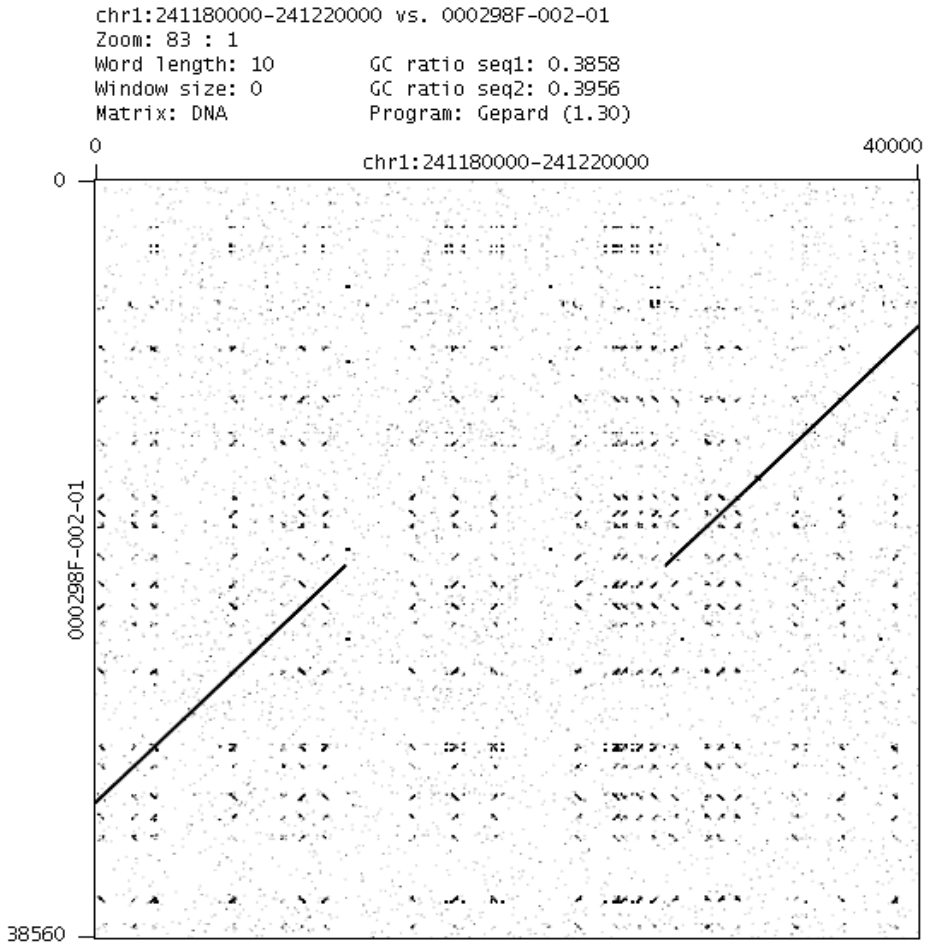


Shi L et al, Nature Communications, 2016

Assembly-based SV detection

- Dot plot of primary and associate contig

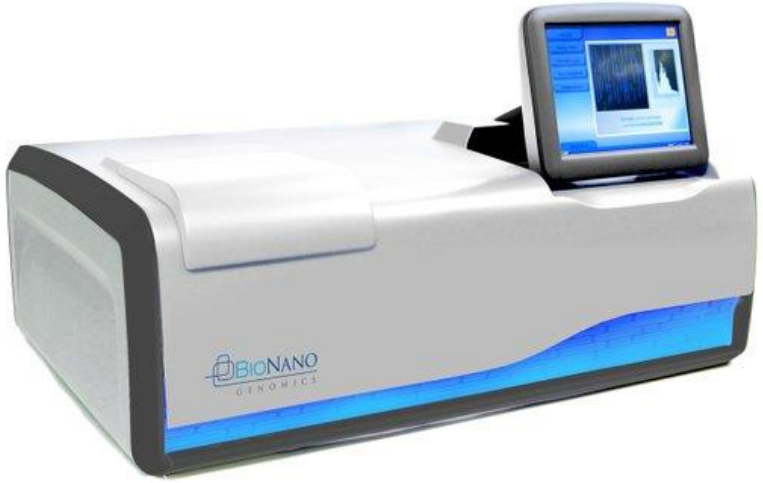


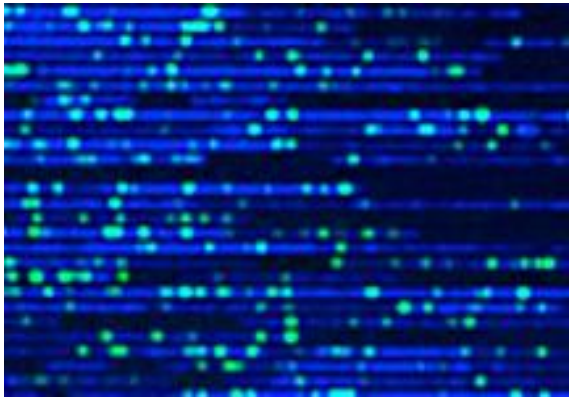


Shi L et al, Nature Communications, 2016

Bionano optical mapping for SV detection

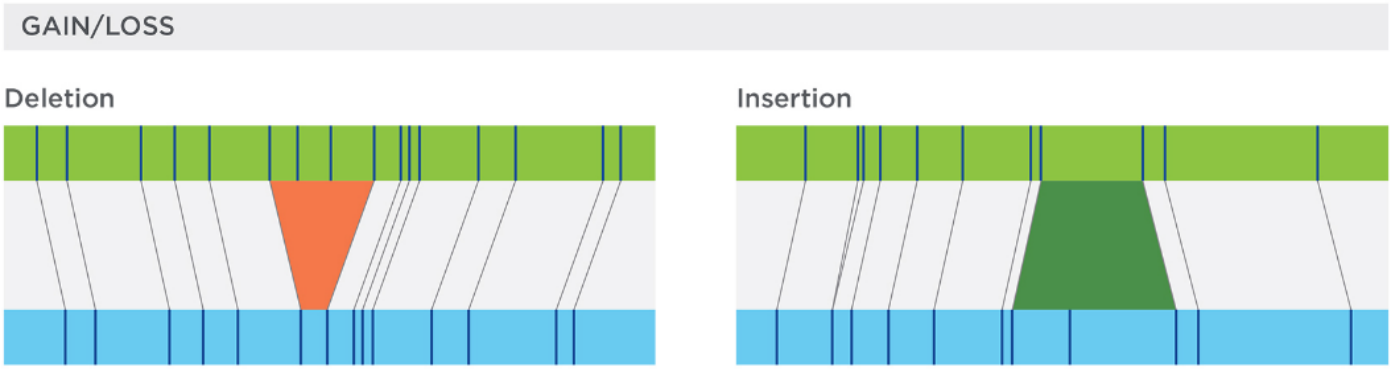
- A nanopore array that detects a characteristic 6 or 7-nucleotide sequence along very long genomic segments





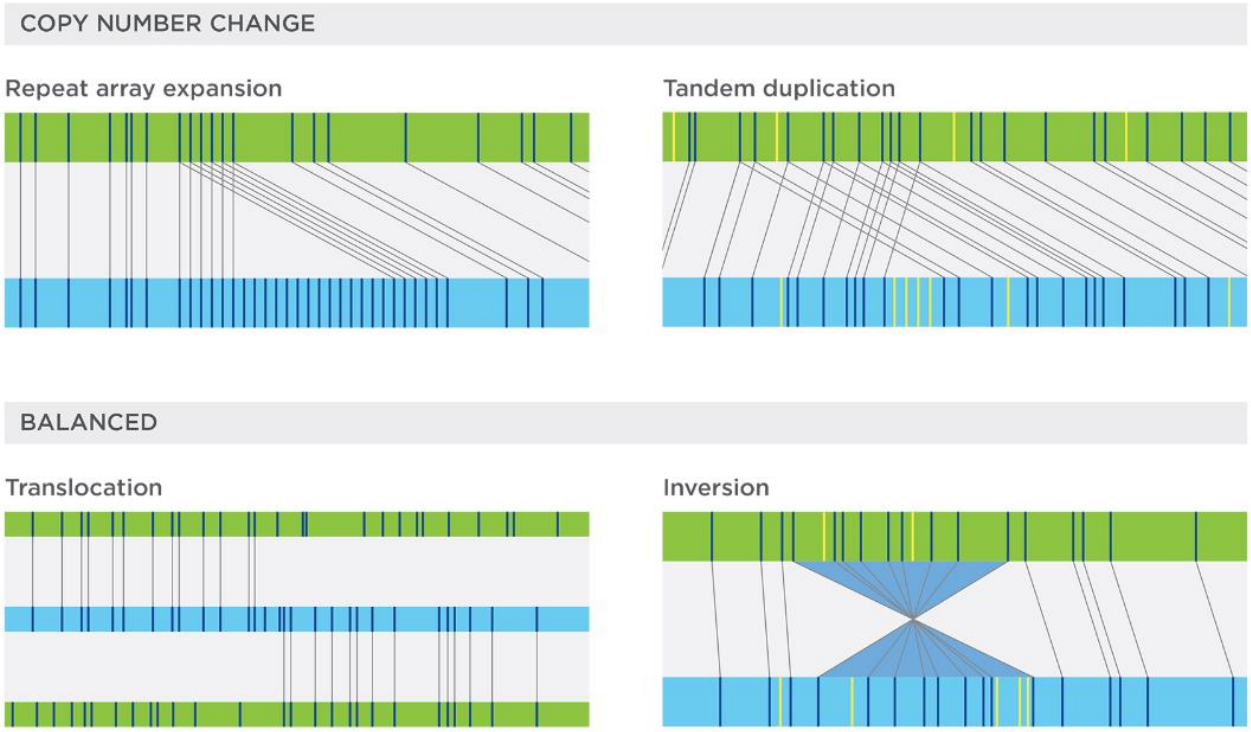
SV detection from Single-molecule optical mapping

- 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