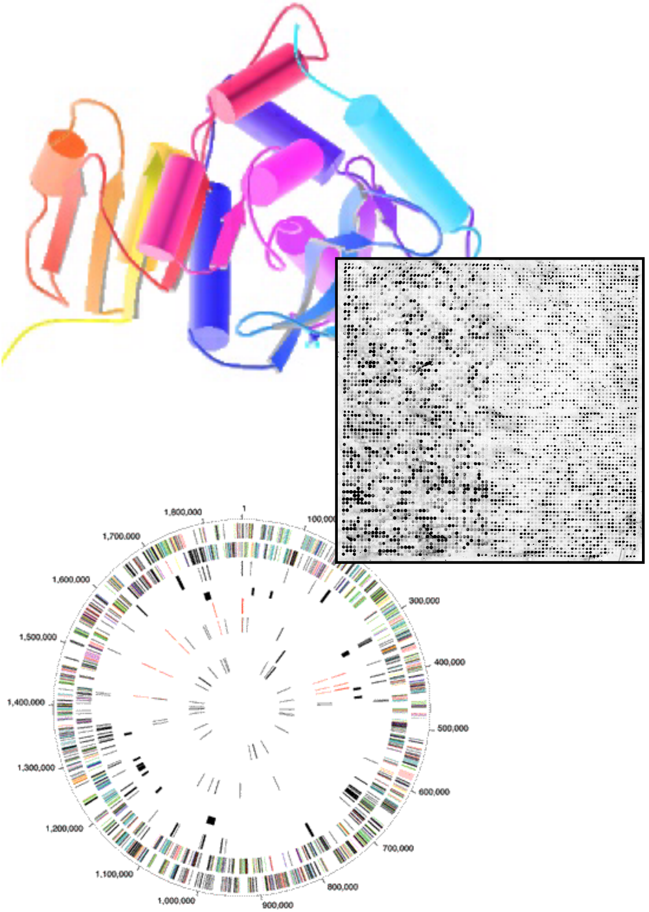
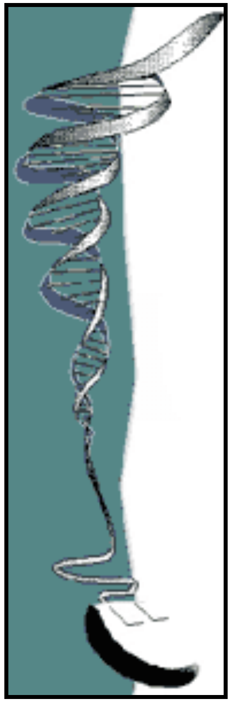
Biomed. Data Sci:

Variant Identification, Focusing on SVs





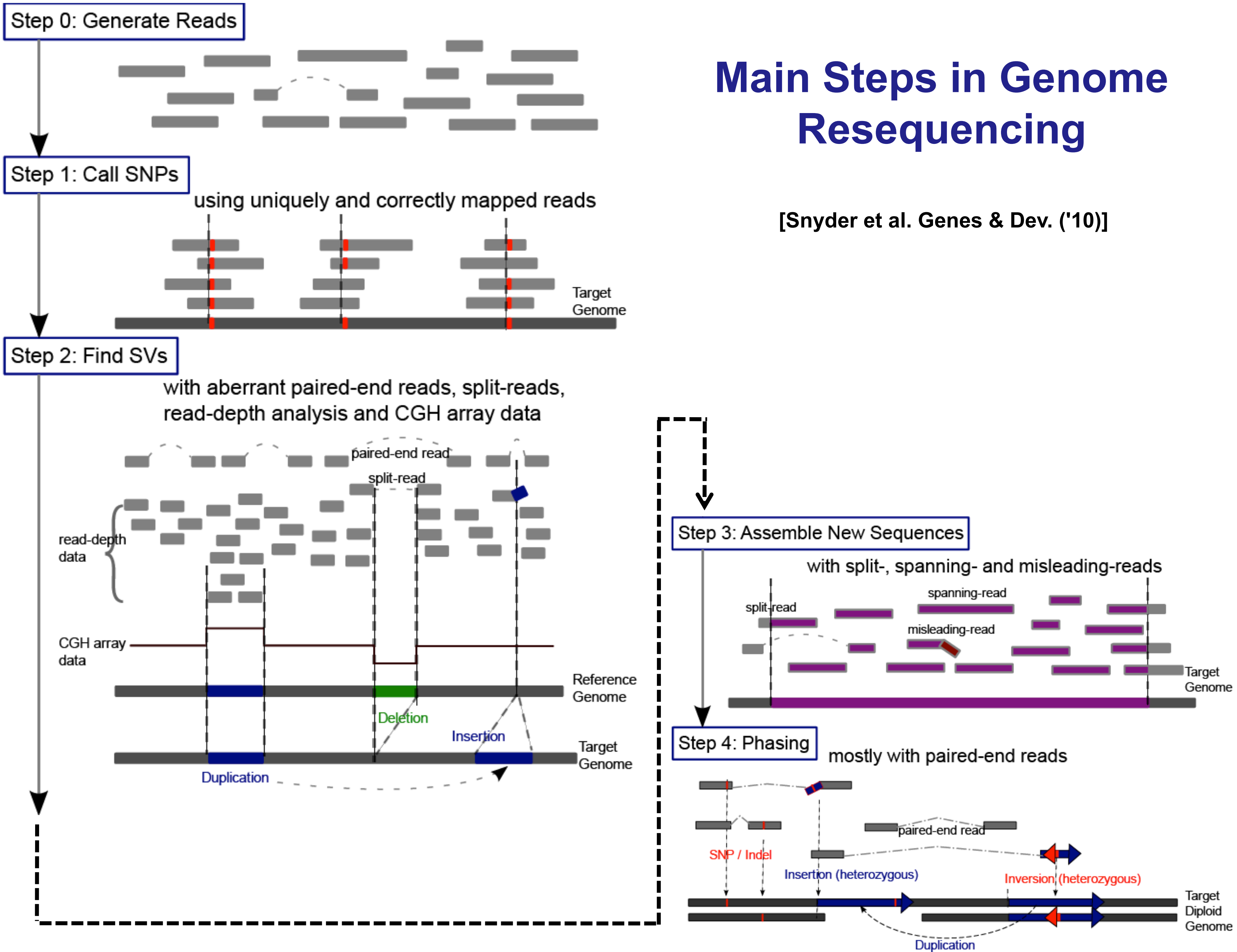
Mark Gerstein, Yale University

gersteinlab.org/courses/452

(last edit in spring ’20, pack #6)

Main Steps in Genome Resequencing

[Snyder et al. Genes & Dev. ('10)]



Step 0: Generate Reads

Step 1: Call SNPs

using uniquely and correctly mapped reads

Step 2: Find SVs

with aberrant paired-end reads, split-reads, read-depth analysis and CGH array data

Step 3: Assemble New Sequences

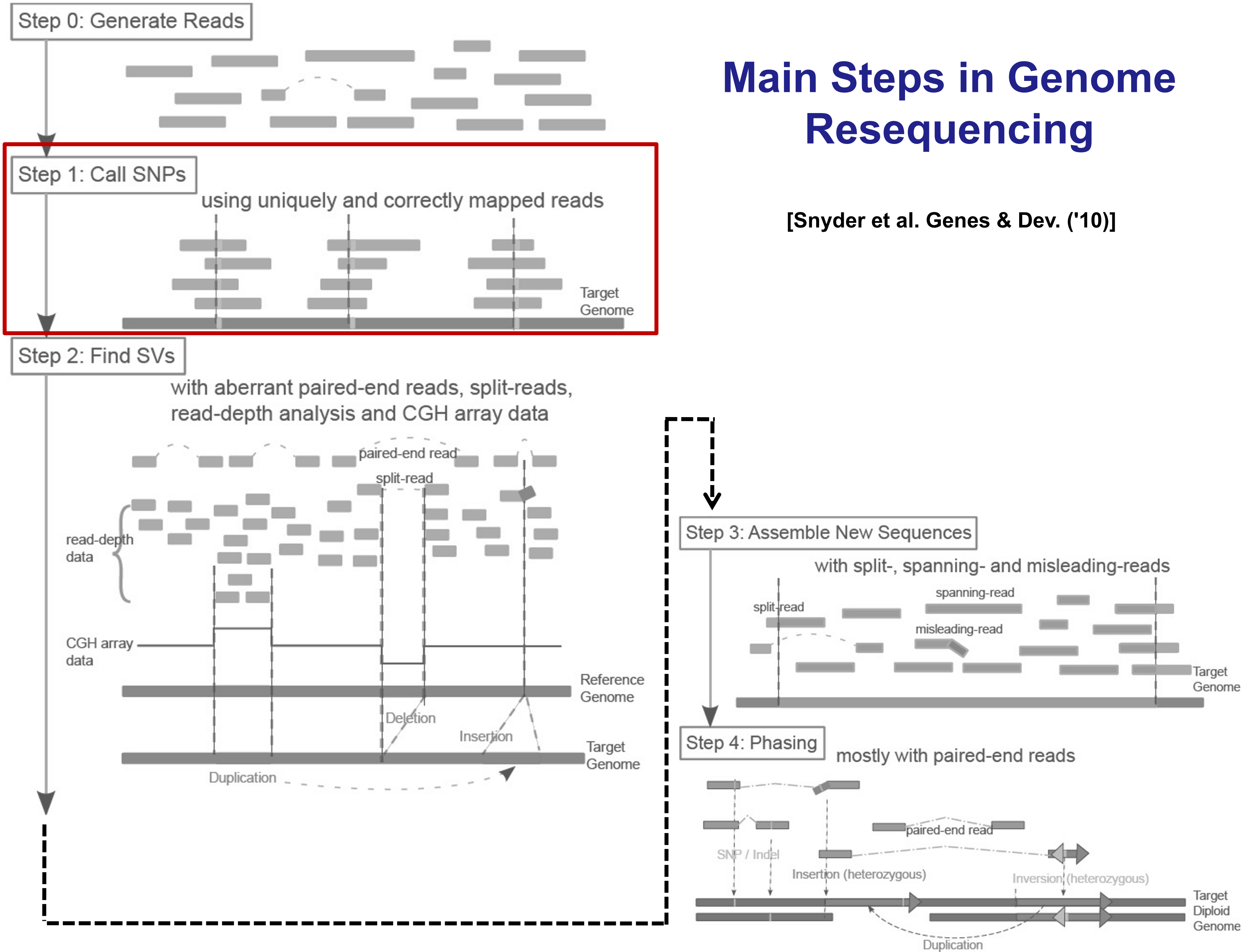
with split-, spanning- and misleading-reads

Step 4: Phasing

mostly with paired-end reads

Main Steps in Genome Resequencing

[Snyder et al. Genes & Dev. ('10)]



Step 0: Generate Reads

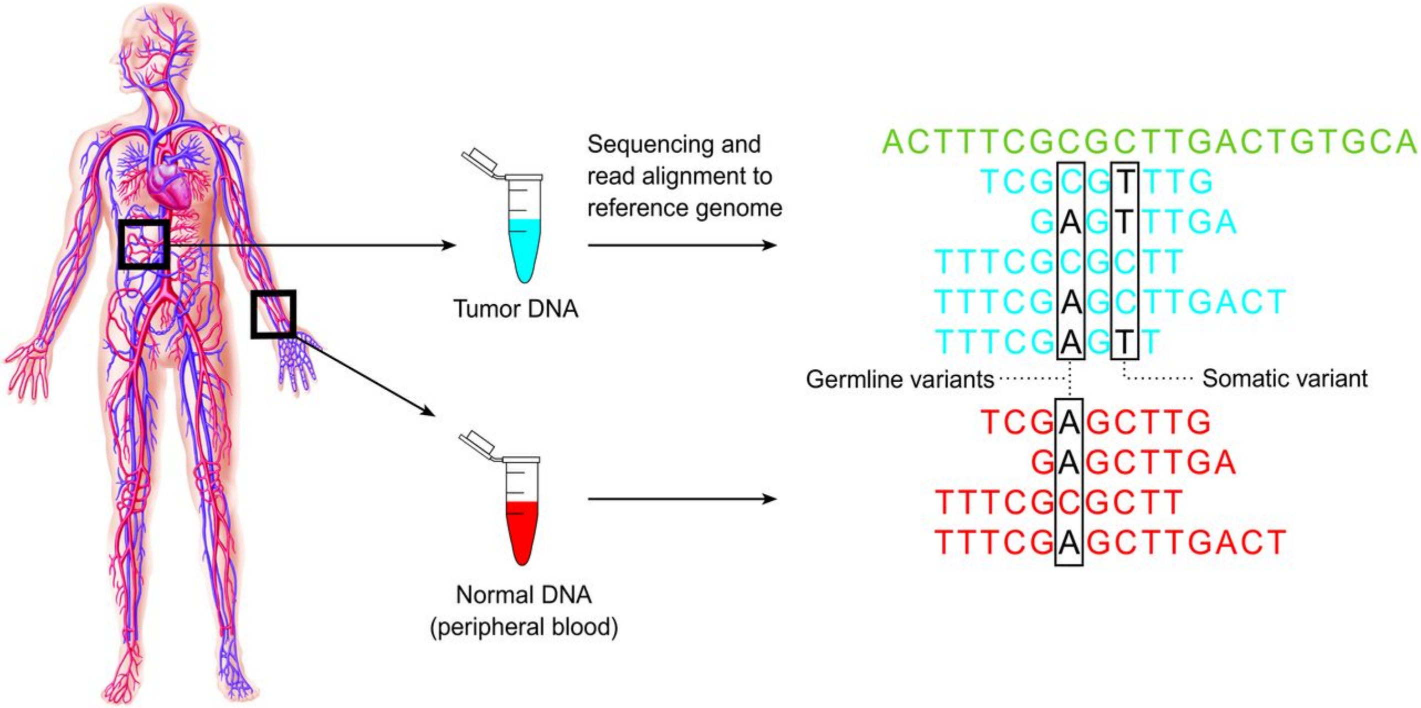
Step 1: Call SNPs using uniquely and correctly mapped reads

Step 2: Find SVs with aberrant paired-end reads, split-reads, read-depth analysis and CGH array data

Step 3: Assemble New Sequences with split-, spanning- and misleading-reads

Step 4: Phasing mostly with paired-end reads

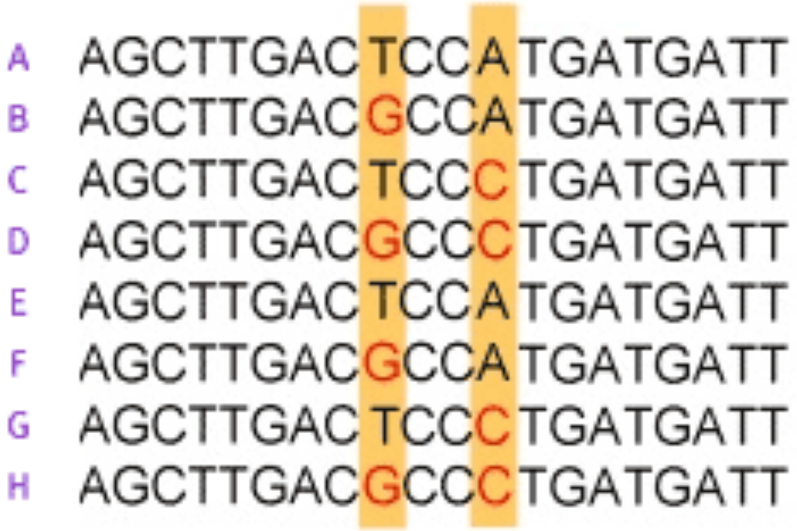
Characterization of genomic variations: somatic vs germline



Sequencing tumor and normal samples from cancer patients provide insight into somatic and germline variation profile.

Samuel et al, Clinical chemistry, 2013

Bayes' Theorem to detect genomic variant



```
A AGCTTGACTCCATGATGATT
B AGCTTGACGCCATGATGATT
C AGCTTGACTCCCTGATGATT
D AGCTTGACGCCCTGATGATT
E AGCTTGACTCCATGATGATT
F AGCTTGACGCCATGATGATT
G AGCTTGACTCCCTGATGATT
H AGCTTGACGCCCTGATGATT
```

$$P(G|D) = \frac{P(D|G)P(G)}{P(D)}$$

$$= \frac{P(D|G)\,P(G)}{\sum_{i=1}^{n} P(D|G_i)\,P(G_i)}$$

In the above equation:

- D refers to the observed data
- G is the genotype whose probability is being calculated
- G_i refers to the ith possible genotype, out of n possibilities

[You et al. ('12) Bioinformatics 28: 643 doi: 10.1093/bioinformatics/bts001]

Calculating the conditional distribution $P(D|G)$:

Assuming an error free model, for each heterozygous SNP site of the diploid genome, covered by K reads, the number of reads i representing one of the two alleles follows binomial distribution.

$$\boldsymbol{P_{err\_free}(D|G) = f(i|k, 0.5) = \binom{k}{i} 0.5^{k}}$$

With errors, the calculation is more complicated.

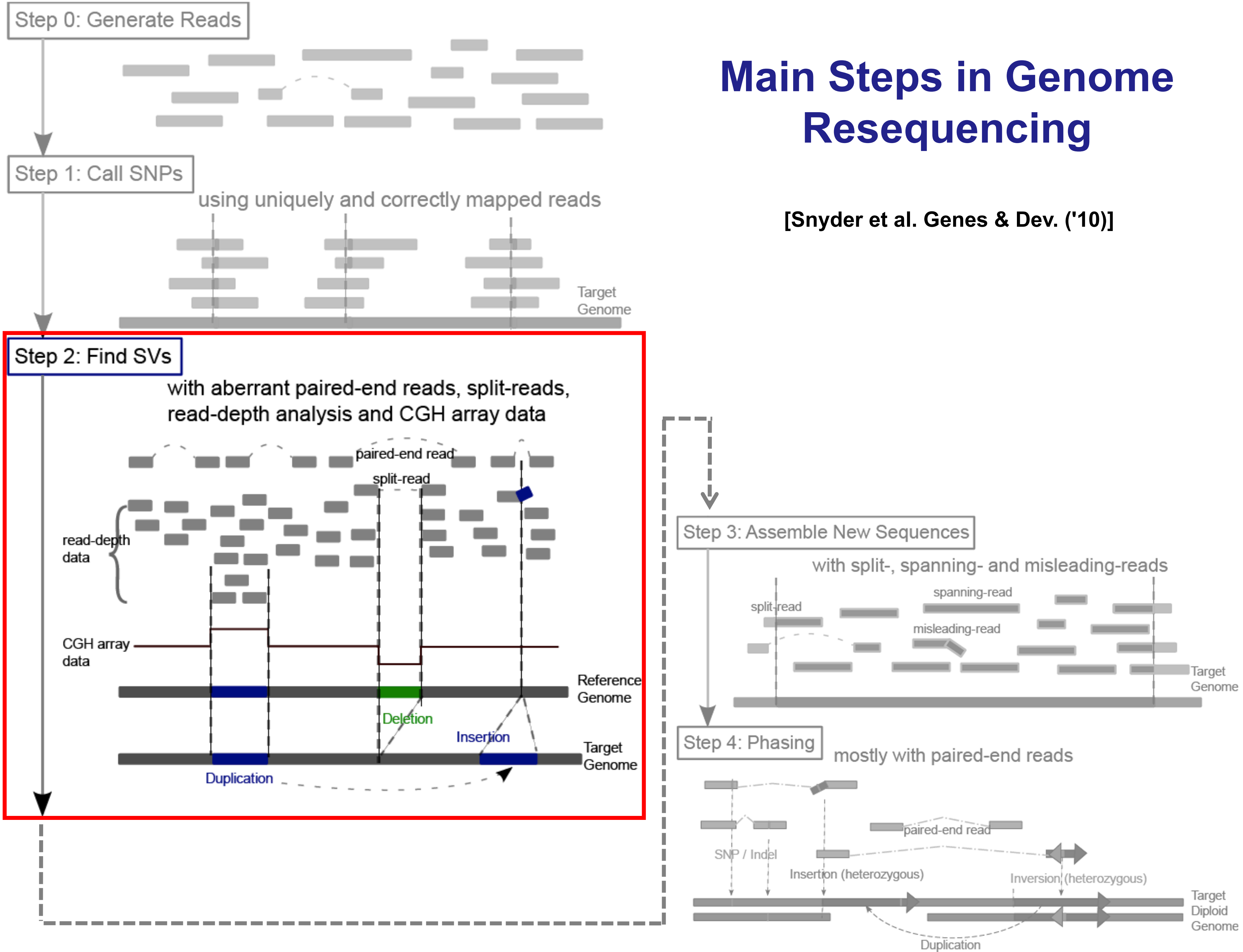
In general:

$$\boldsymbol{P(D|G) = P_{err\_free}(D|G) + P_{err}(D|G)}$$

[You et al. ('12) Bioinformatics 28: 643 doi: 10.1093/bioinformatics/bts001]

Main Steps in Genome Resequencing

[Snyder et al. Genes & Dev. ('10)]



Step 0: Generate Reads

Step 1: Call SNPs

using uniquely and correctly mapped reads

Step 2: Find SVs

with aberrant paired-end reads, split-reads, read-depth analysis and CGH array data

Step 3: Assemble New Sequences

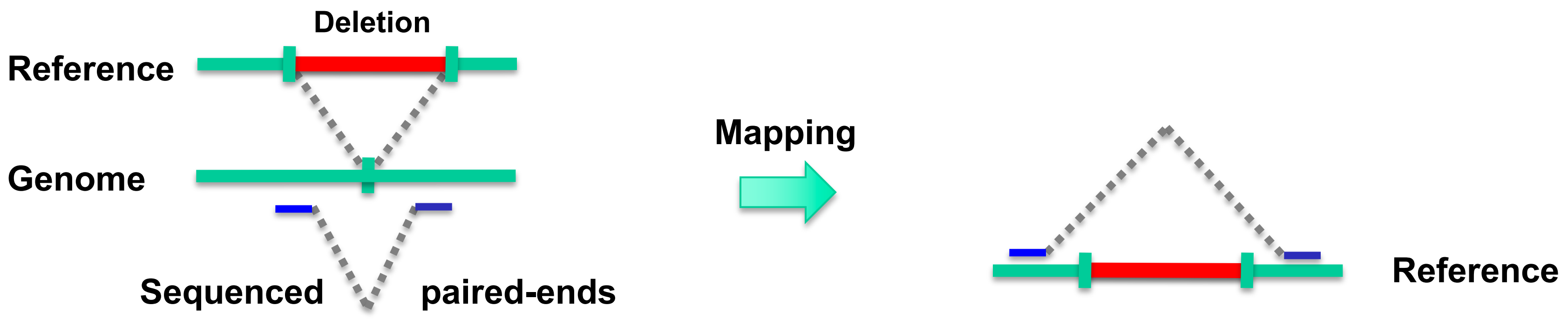
with split-, spanning- and misleading-reads

Step 4: Phasing

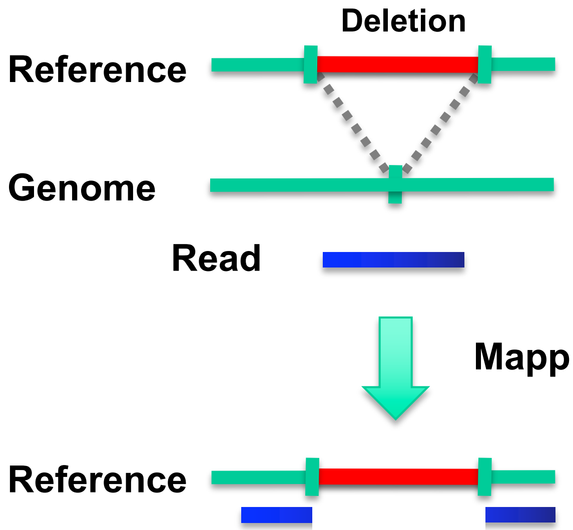
mostly with paired-end reads

Methods to **Find SVs**

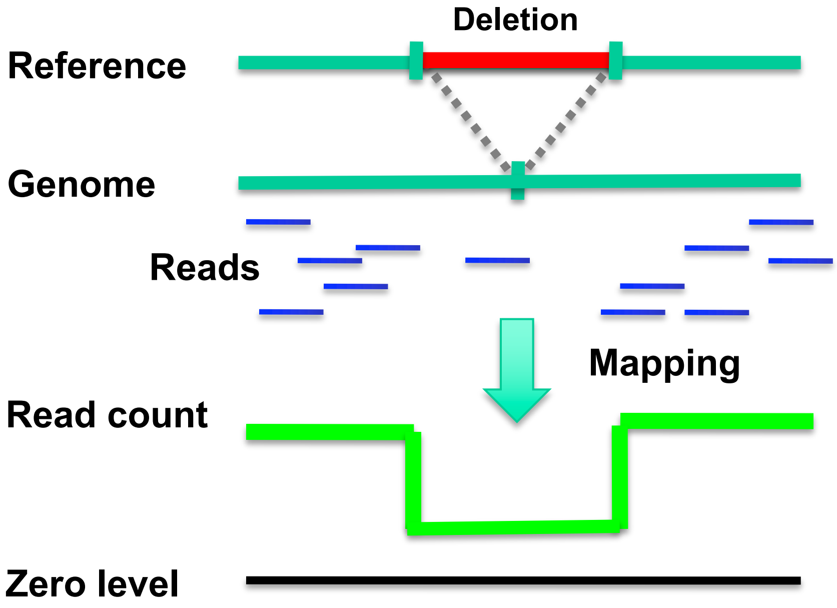
1. Paired ends



2. Split read



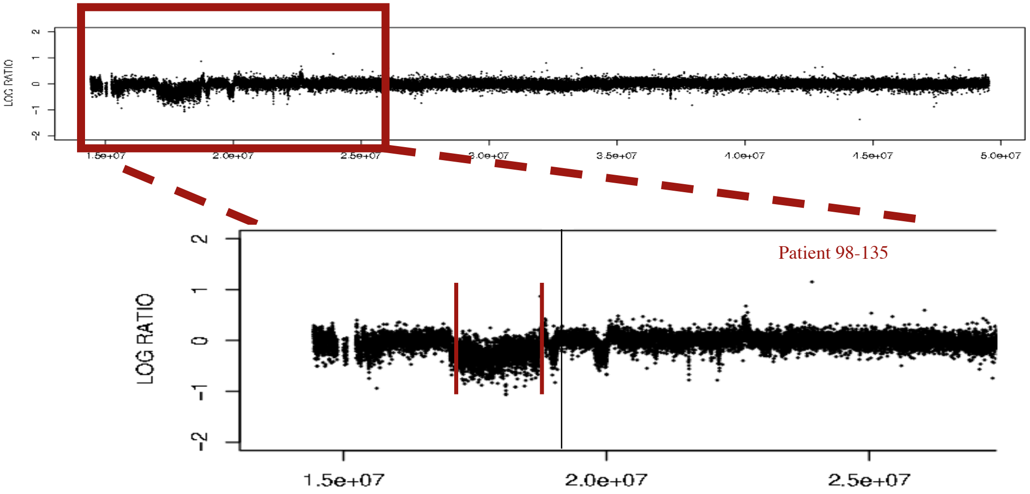
3. Read depth (or aCGH)



4. Local Reassembly

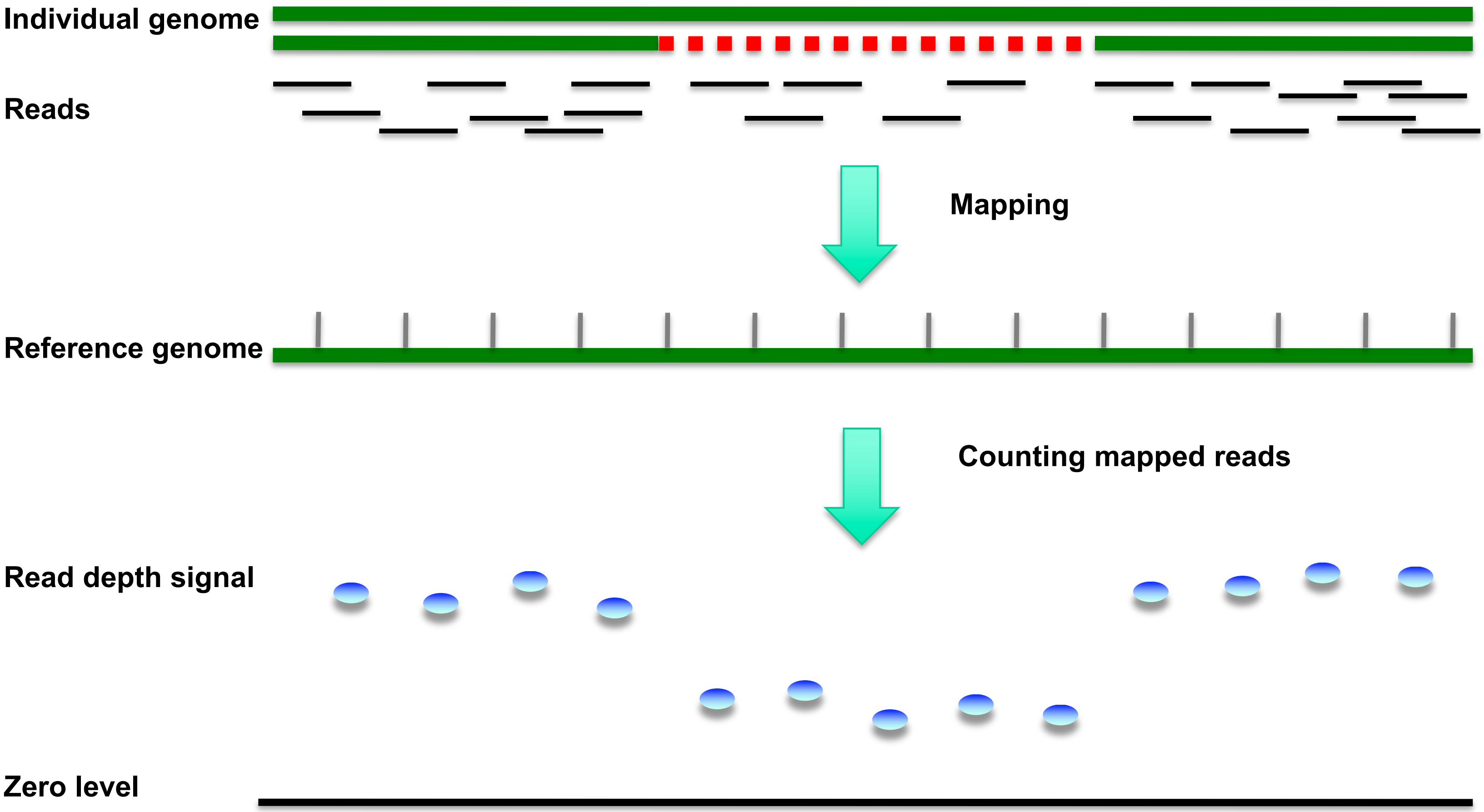
[Snyder et al. Genes & Dev. ('10)]

Read Depth



Array Signal

Read depth



Individual genome

Reads

Mapping

Reference genome

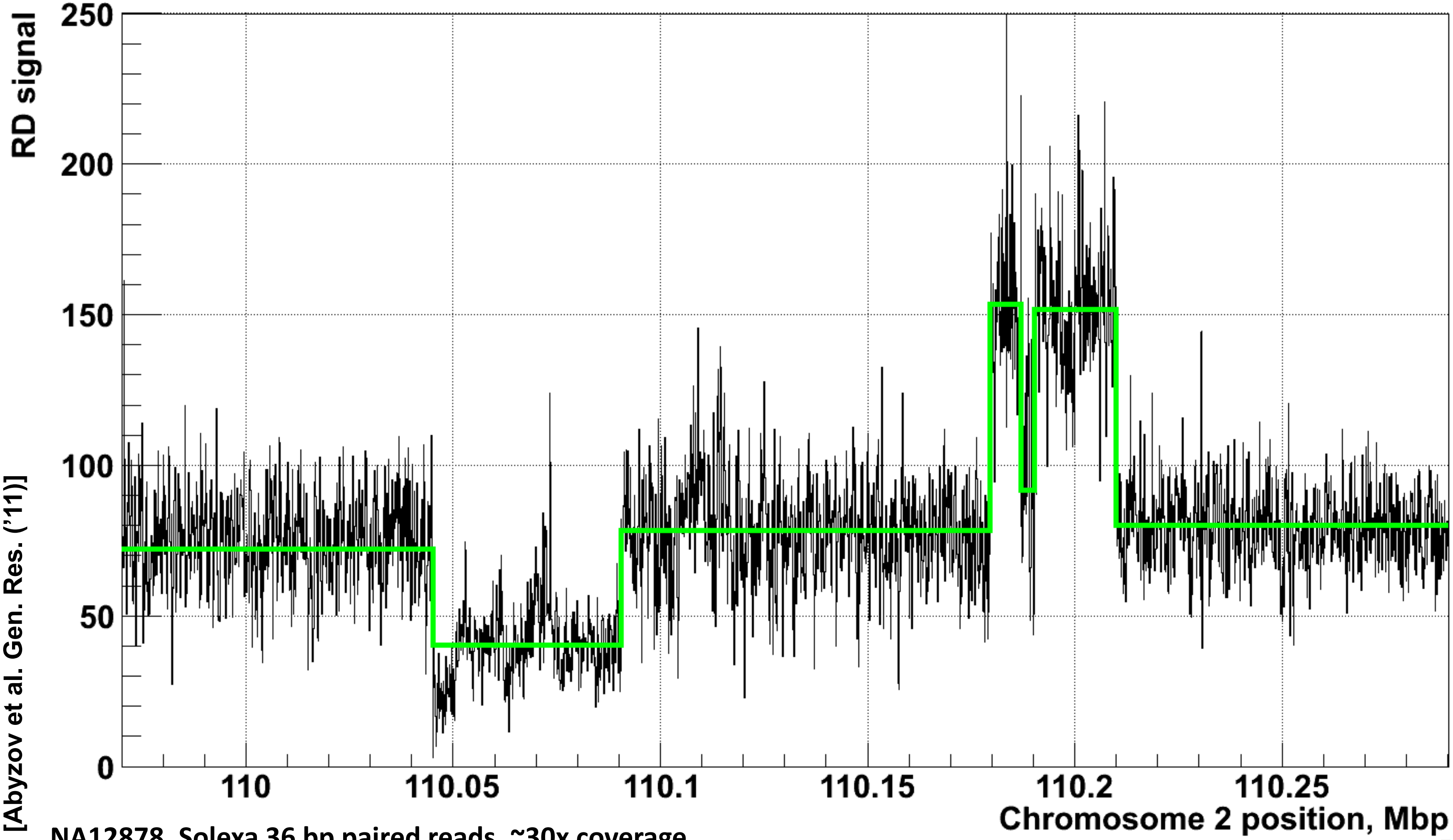
Counting mapped reads

Read depth signal

Zero level

[Urban et al. ('06) PNAS; Wang et al. Gen. Res. ('09); Abyzov et al. Gen. Res. ('11)]

Example of Application to RD data

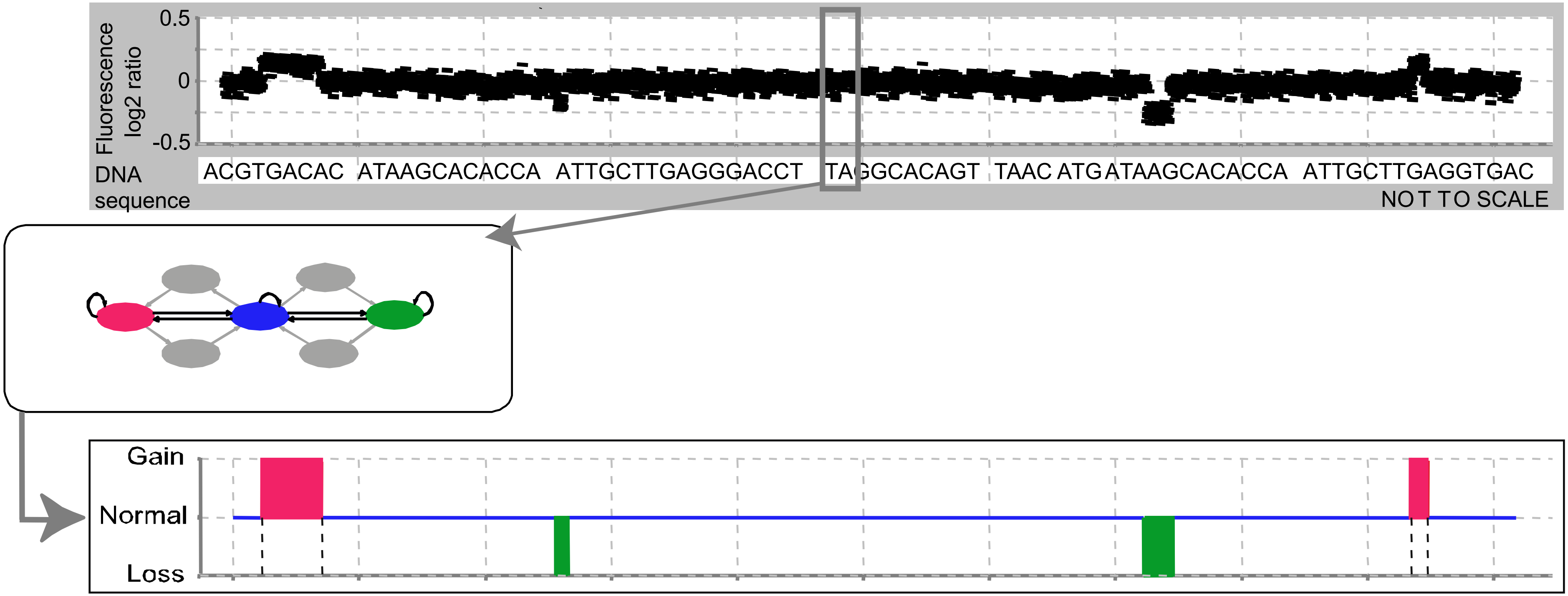


[Abyzov et al. Gen. Res. ('11)]

NA12878, Solexa 36 bp paired reads, ~30x coverage

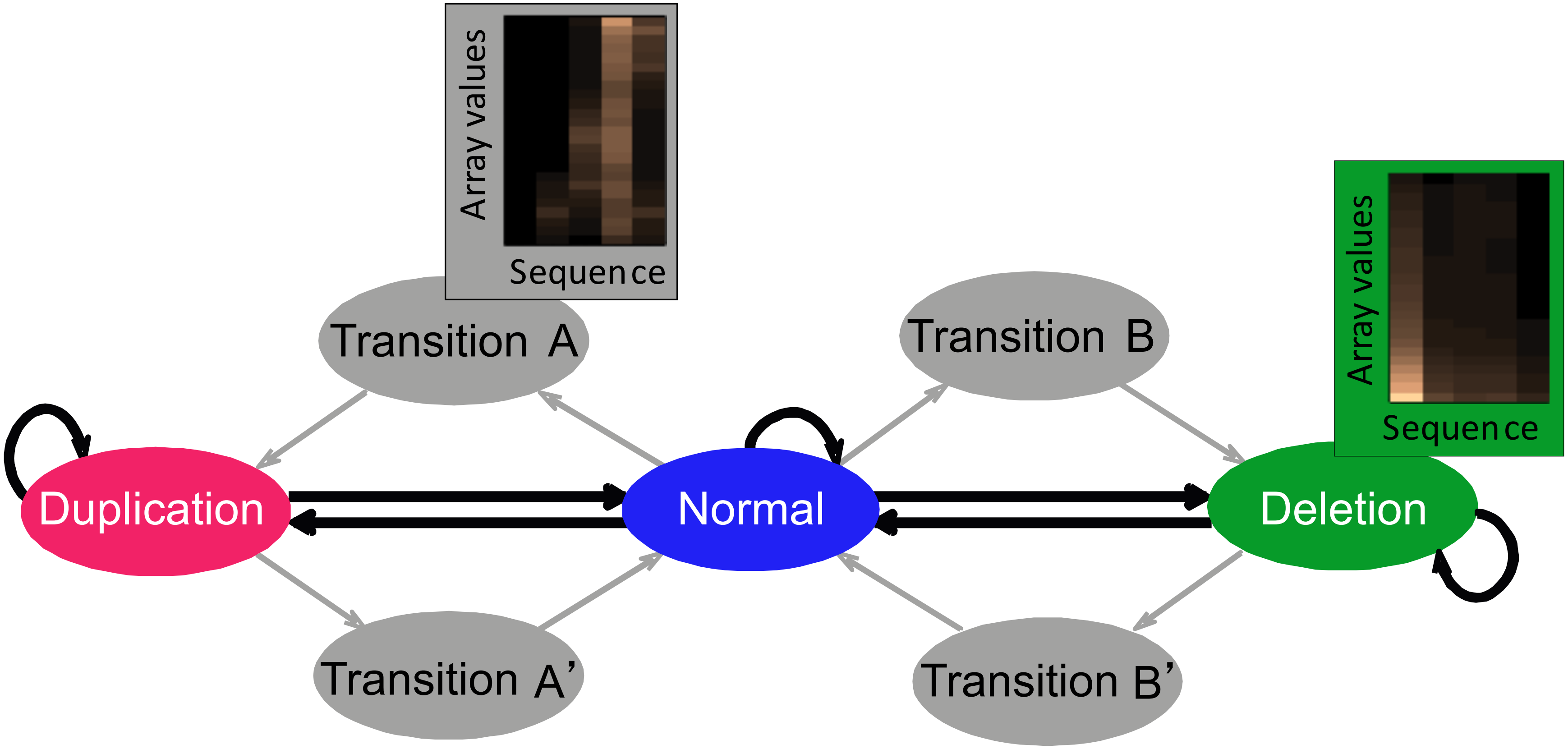
HMM

- To get highest resolution on breakpoints need to smooth & segment the signal
- BreakPtr: prediction of breakpoints, dosage and cross-hybridization using a system based on Hidden Markov Models



Korbel*, Urban* *et al.*, PNAS (2007)

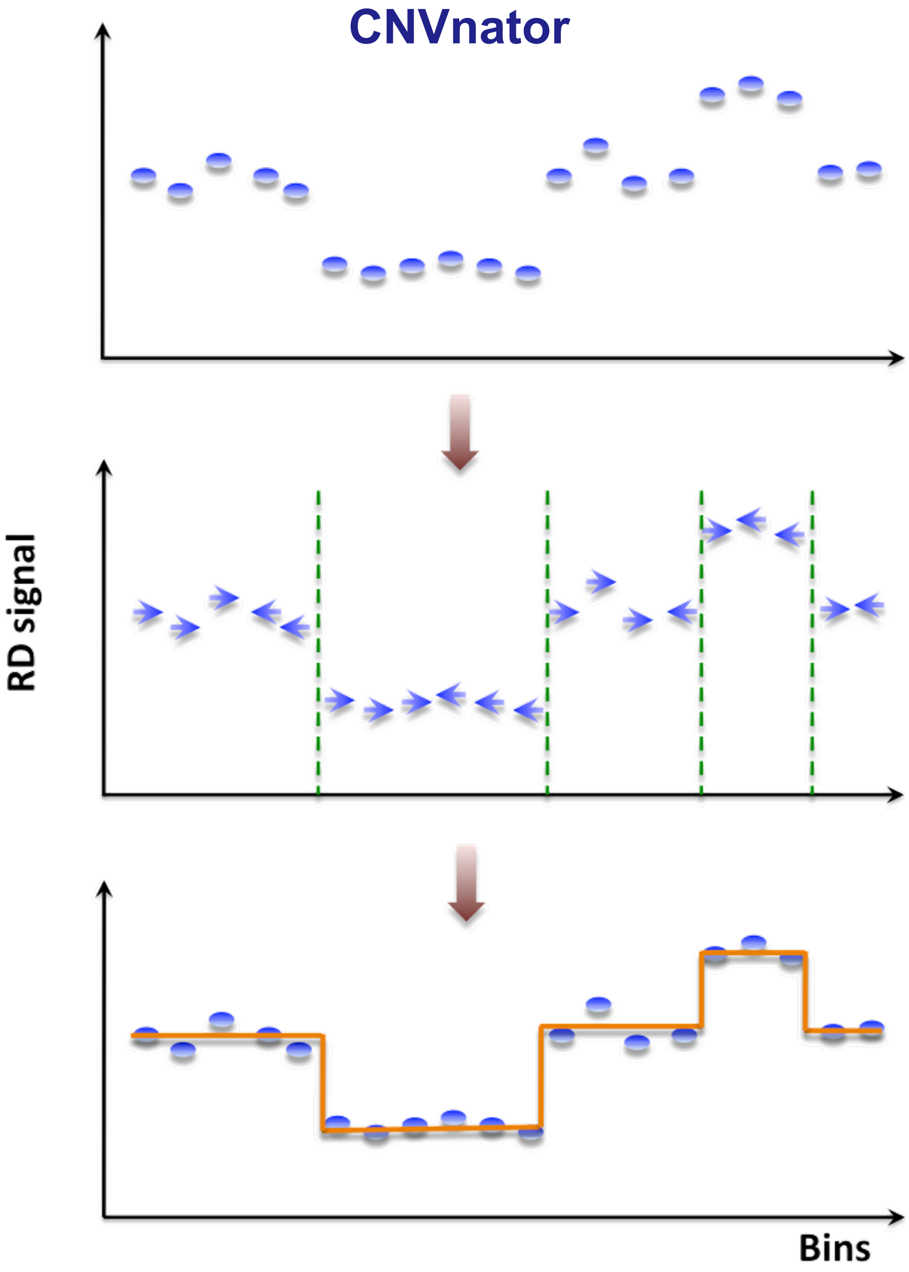
Statistically integrates array signal and DNA sequence signatures (using a discrete-valued bivariate HMM)



Korbel*, Urban* *et al.*, PNAS (2007)

Mean-shift-based (MSB) segmentation: no explicit model

- For each bin attraction (mean-shift) vector points in the direction of bins with most similar RD signal
- No prior assumptions about number, sizes, haplotype, frequency and density of CNV regions
- Not Model-based (e.g. like HMM) with global optimization, distr. assumption & parms. (e.g. num. of segments).
- Achieves discontinuity-preserving smoothing
- Derived from image-processing applications



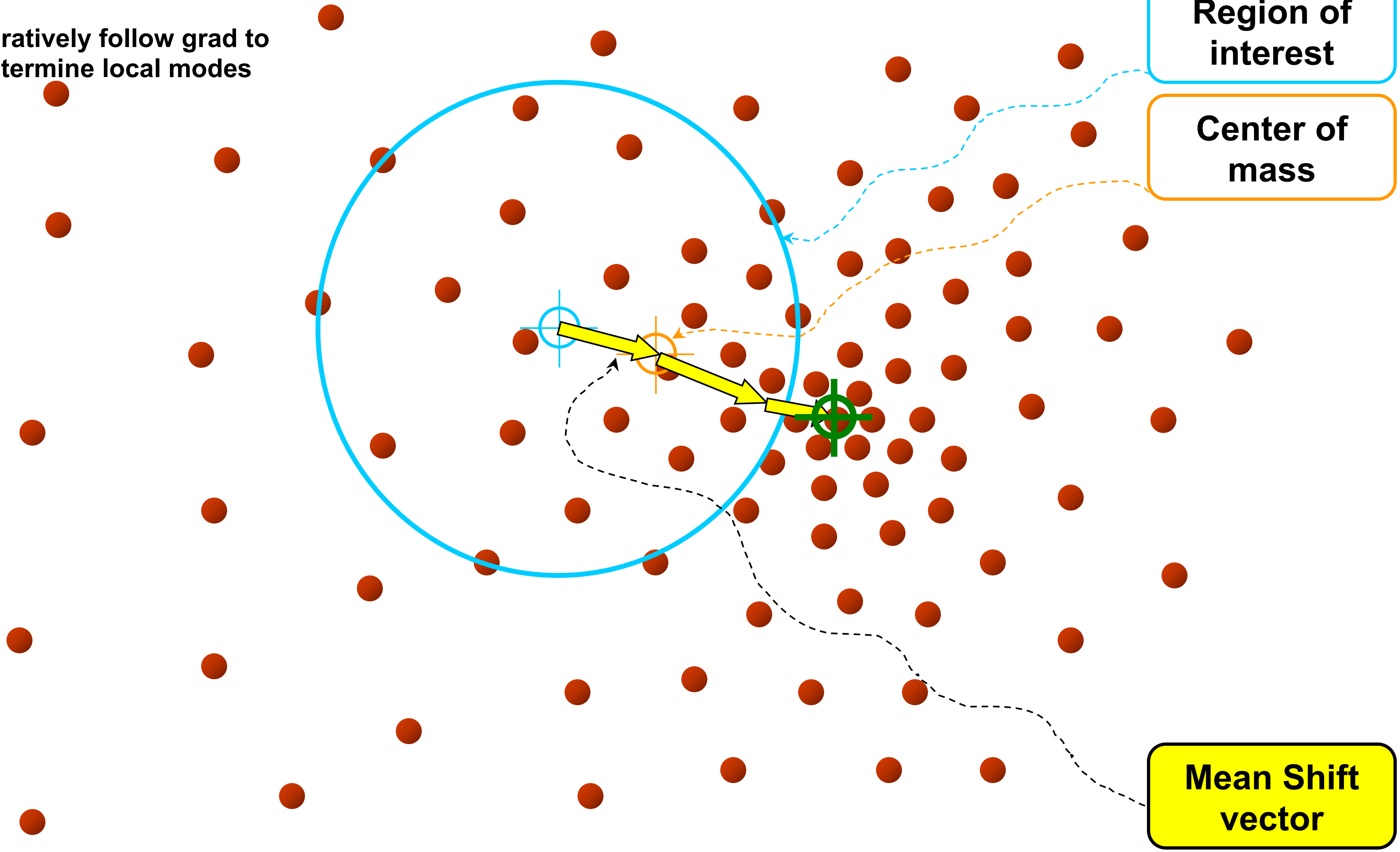
[Abyzov et al. Gen. Res. ('11)]

Intuitive Description of MSB

● Observed depth of coverage counts as samples from PDF

➡ Kernel-based approach to estimate local gradient of PDF

⌖ Iteratively follow grad to determine local modes



Region of interest

Center of mass

Mean Shift vector

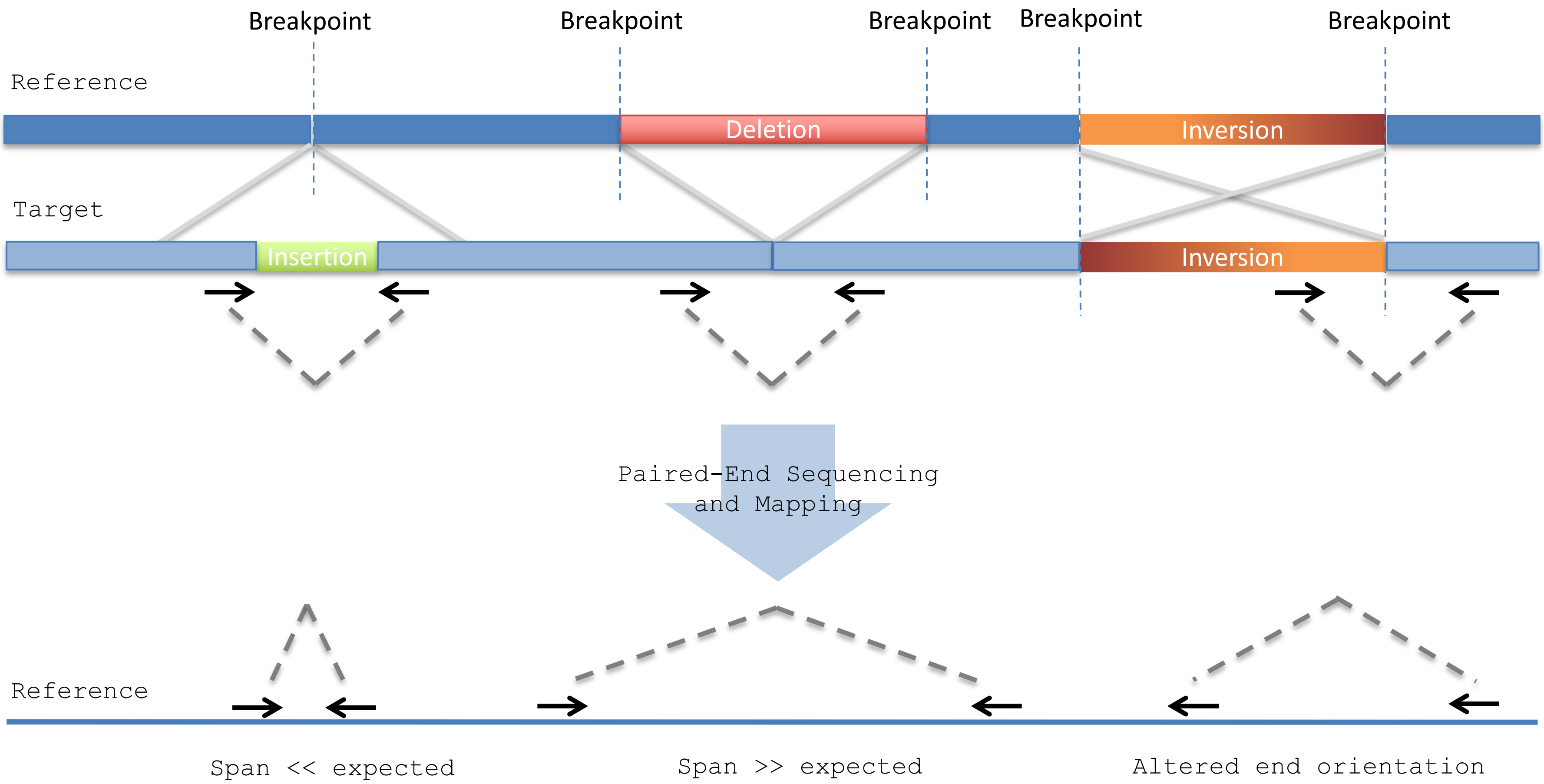
Objective : Find the densest region

Distribution of identical billiard balls

[Adapted from S Ullman et al. **"Advanced Topics in Computer Vision,"** www.wisdom.weizmann.ac.il/~vision/courses/2004_2]

Paired-End

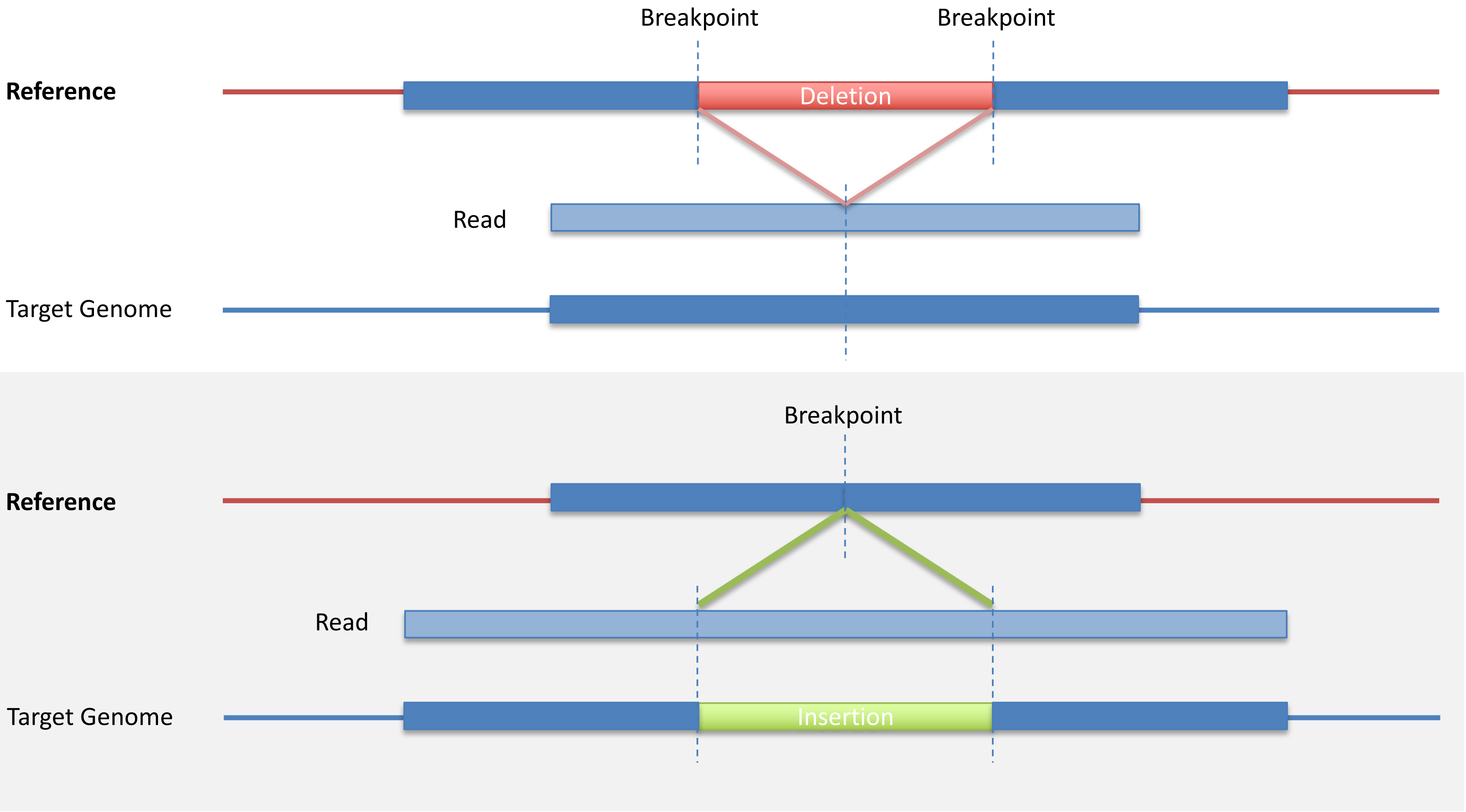
Paired-End Mapping



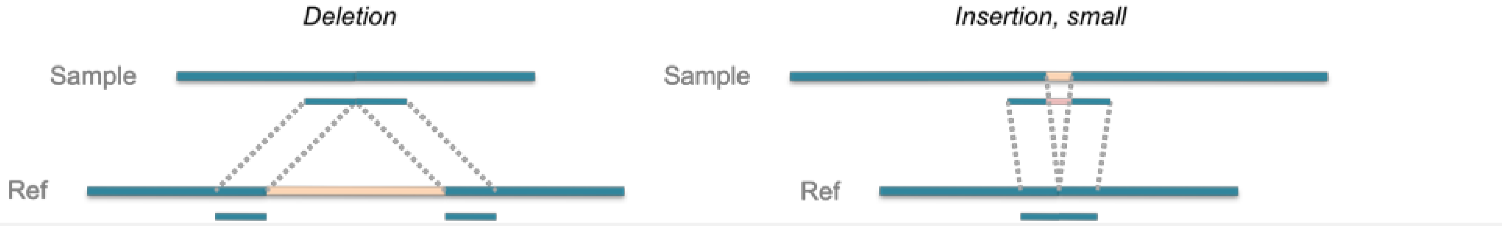
- Both paired-ends map within repeats.
- Limited the distance between pairs; therefore, neither large nor very small rearrangements can be detected

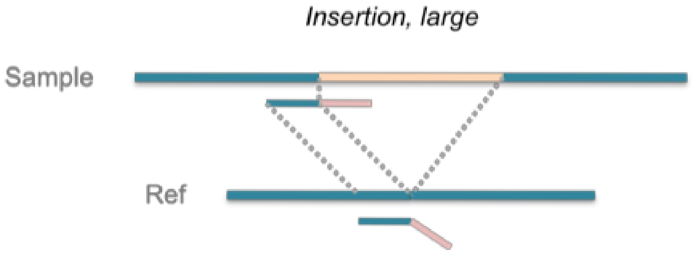
Split Read

Split-read Analysis



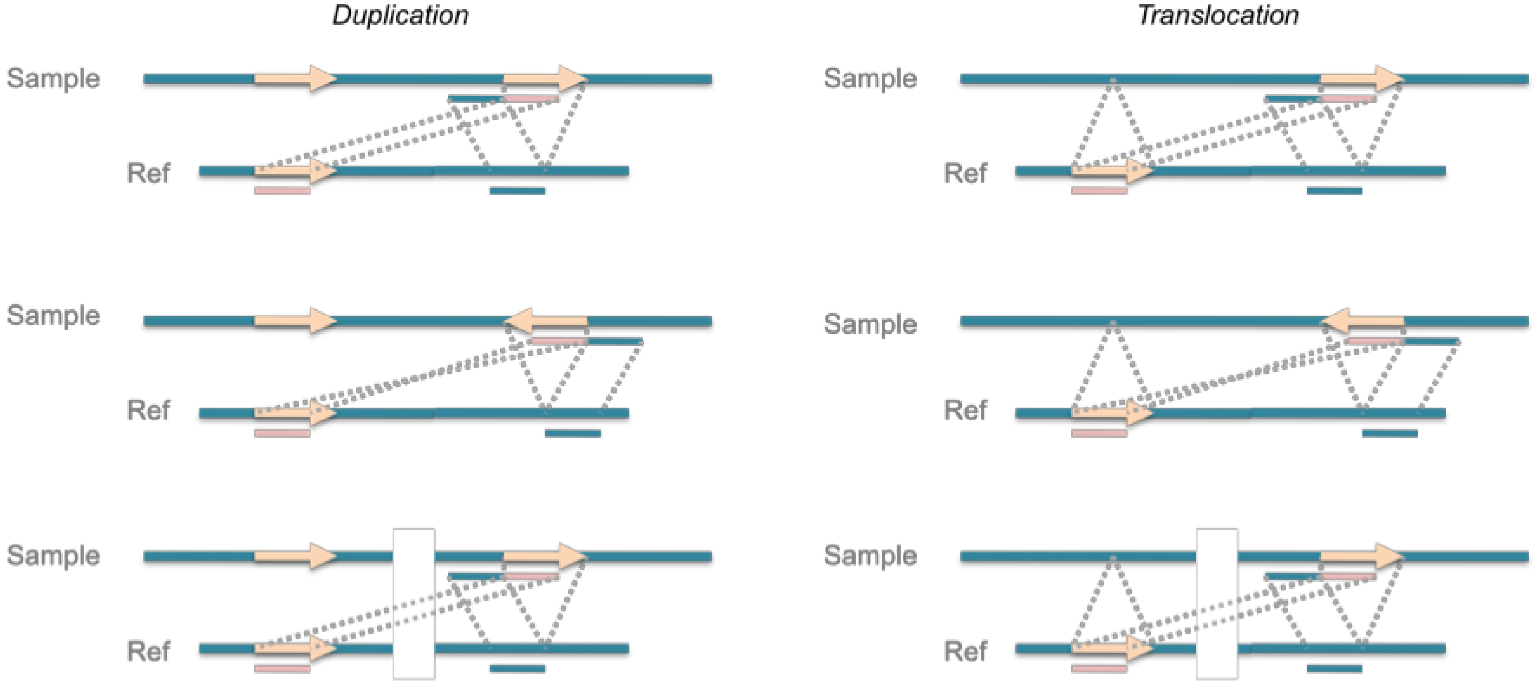
Simple SVs





Deletions are the Easiest to Identify

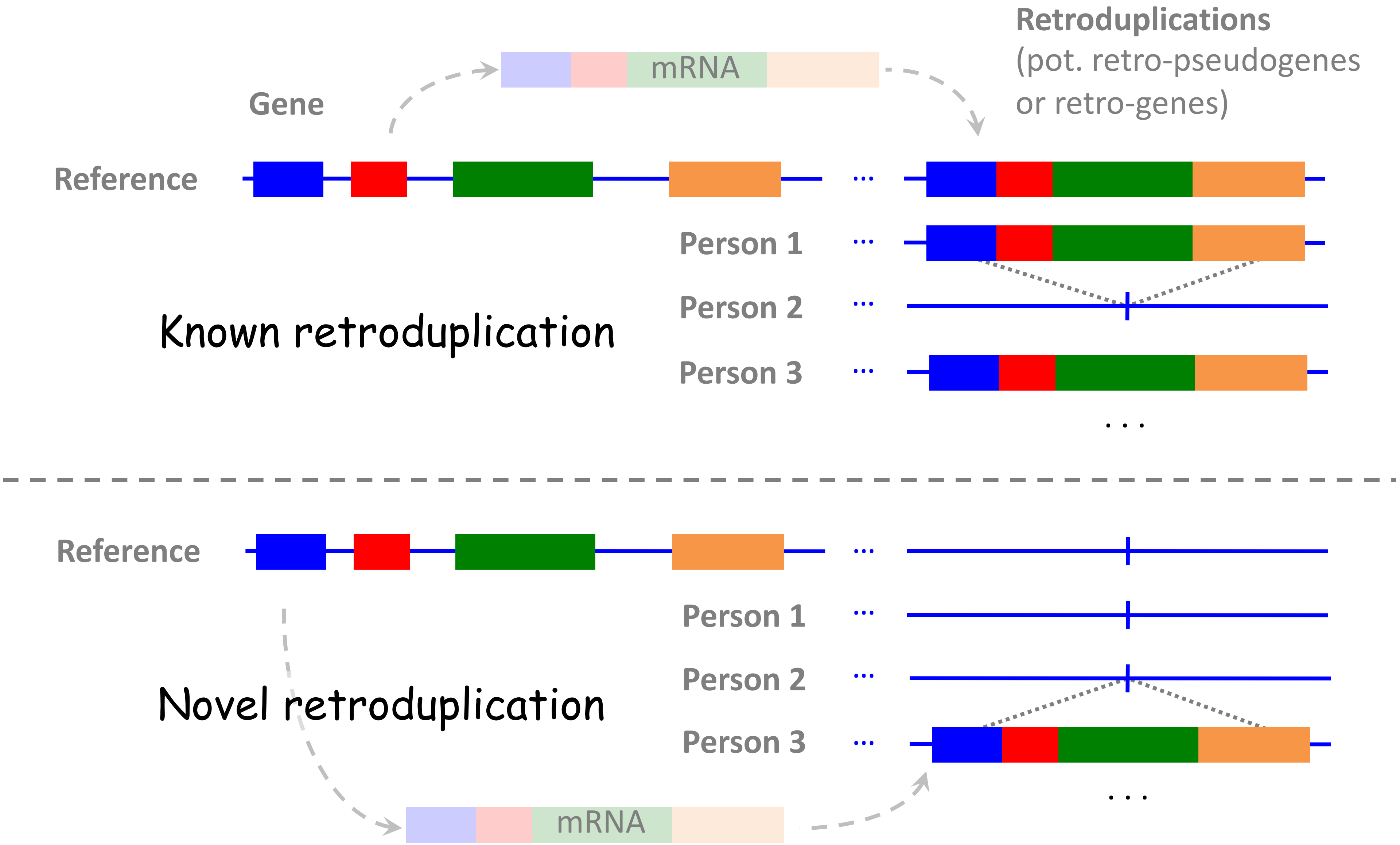
Complex SVs



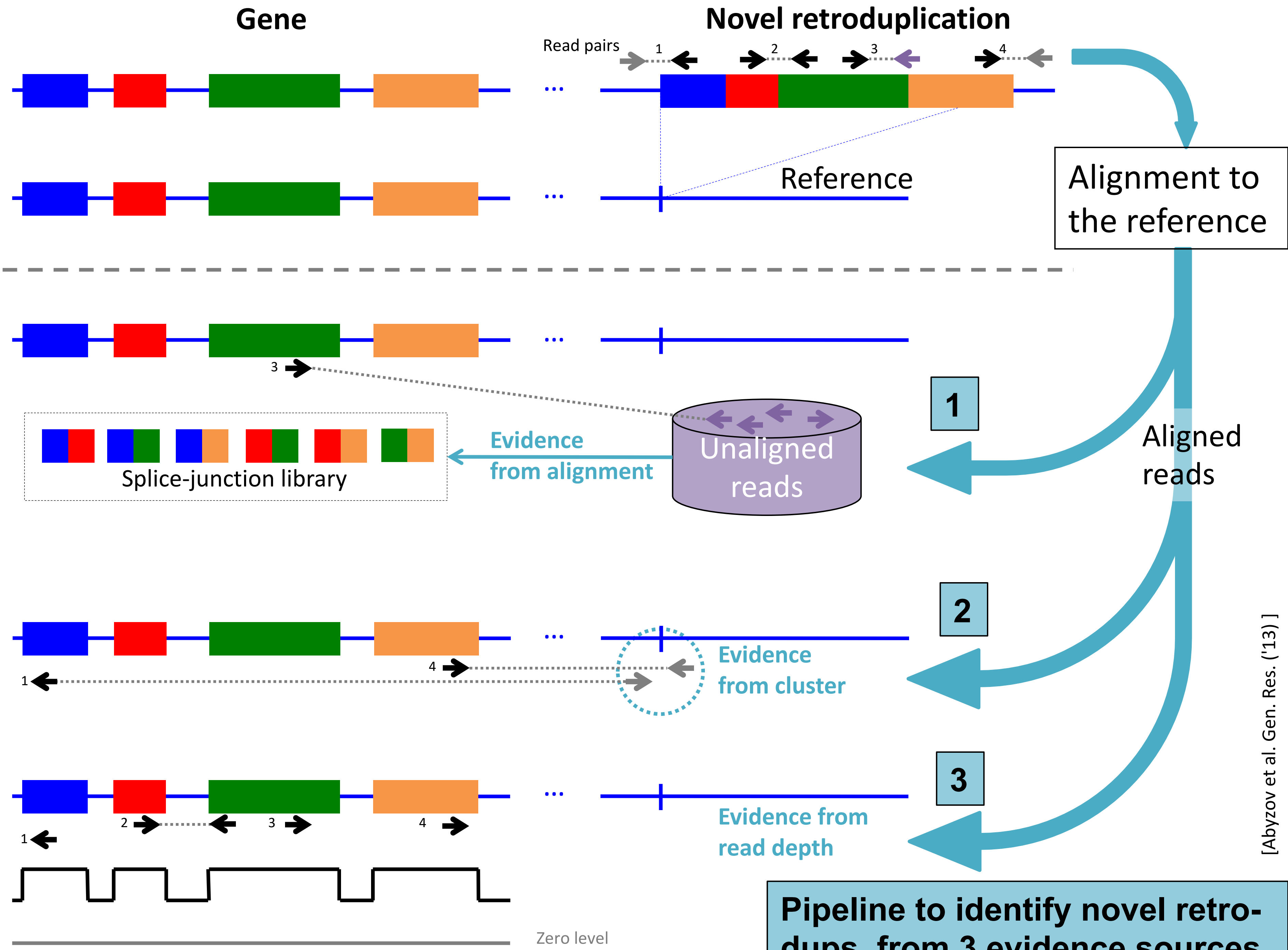
[Zhang et al. ('11) *BMC Genomics*]

RDV & Mobile Elements

Retroduplication variation (RDV)



[Abyzov et al. Gen. Res. ('13)]



[Abyzov et al. Gen. Res. ('13)]

Pipeline to identify novel retro-dups. from 3 evidence sources