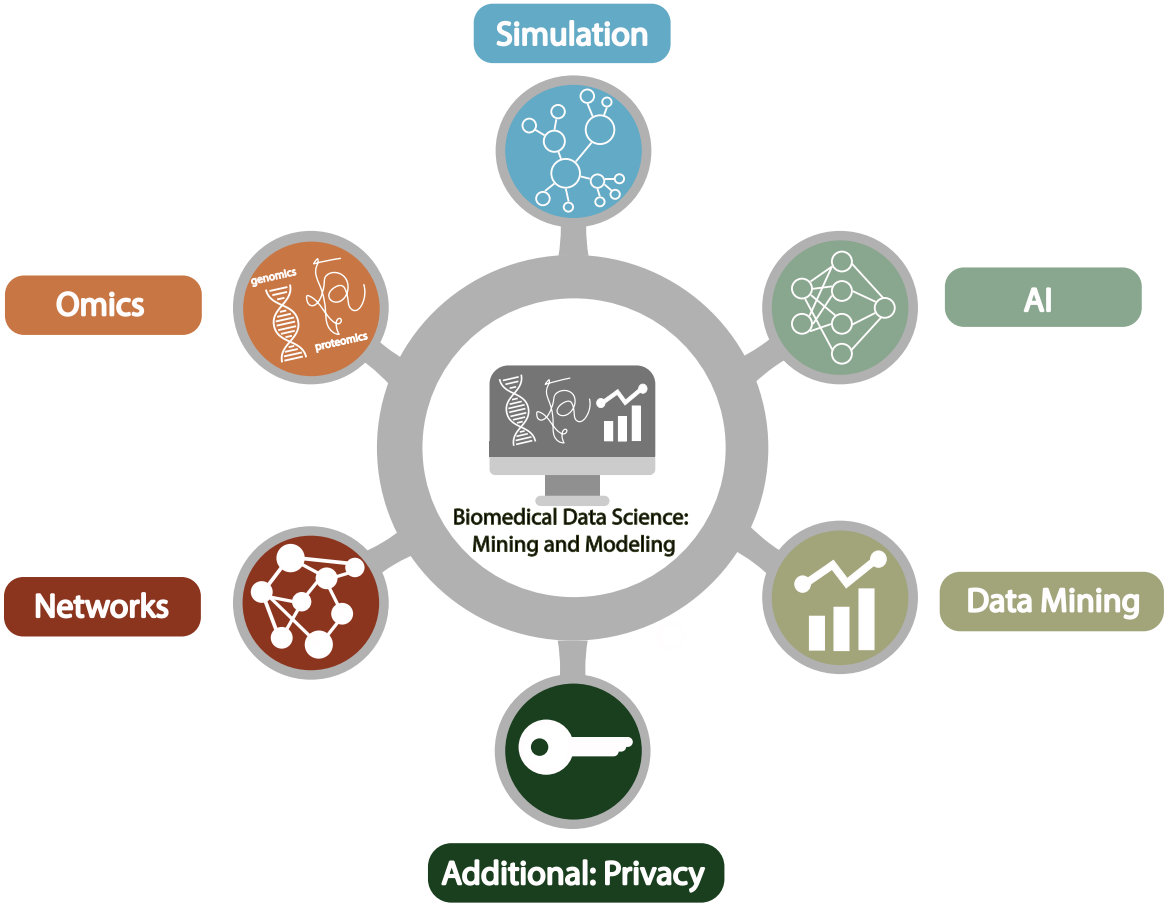
# Biomedical Data Science (GersteinLab.org/courses/452)

## Variant Identification, Focusing on SVs (25m6a)

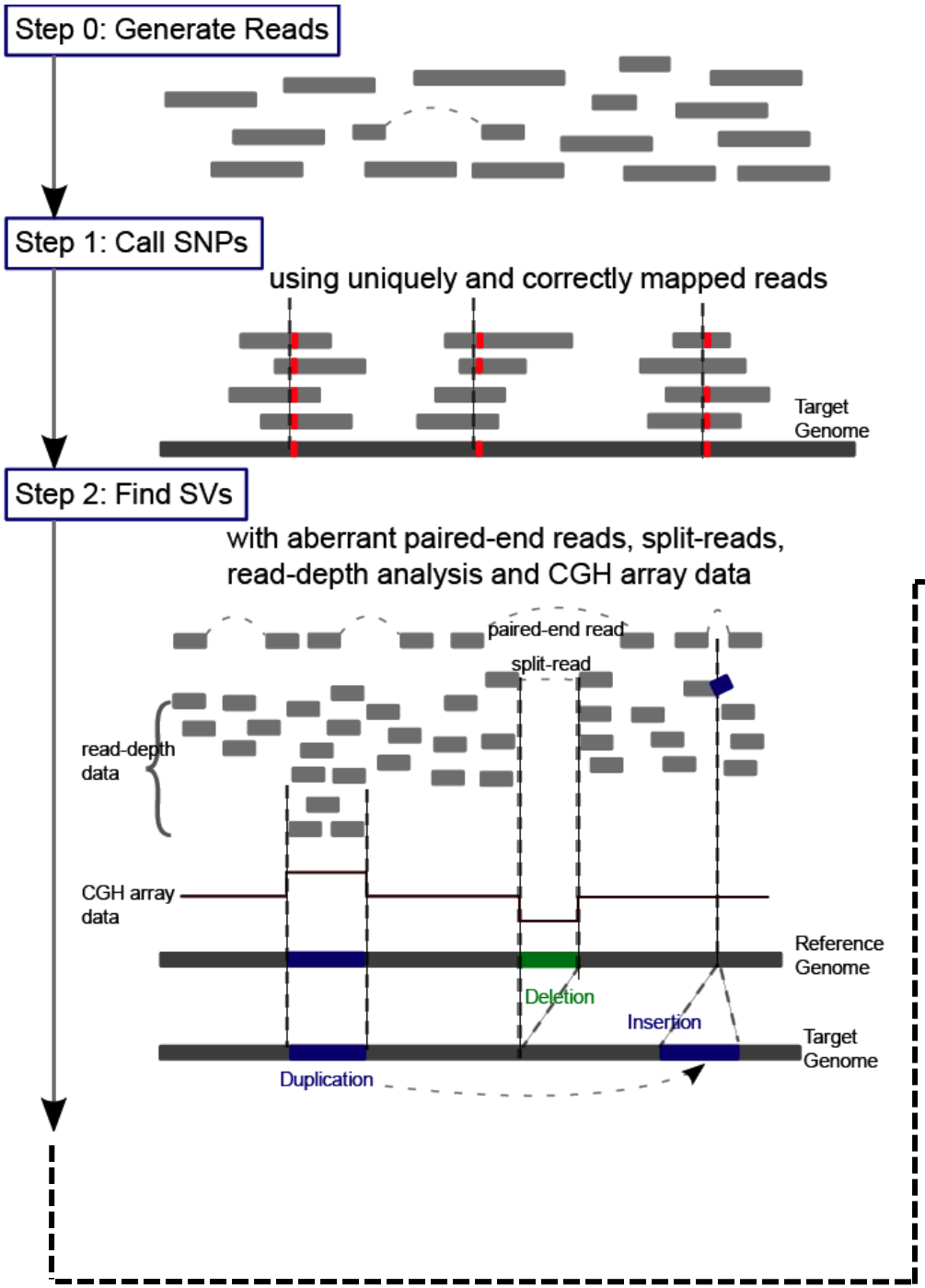



Mark Gerstein
Yale U.

Last edit in spring '25. Unchanged (but with slight edits) from 22m6a & 2021's M6a [which has a video].

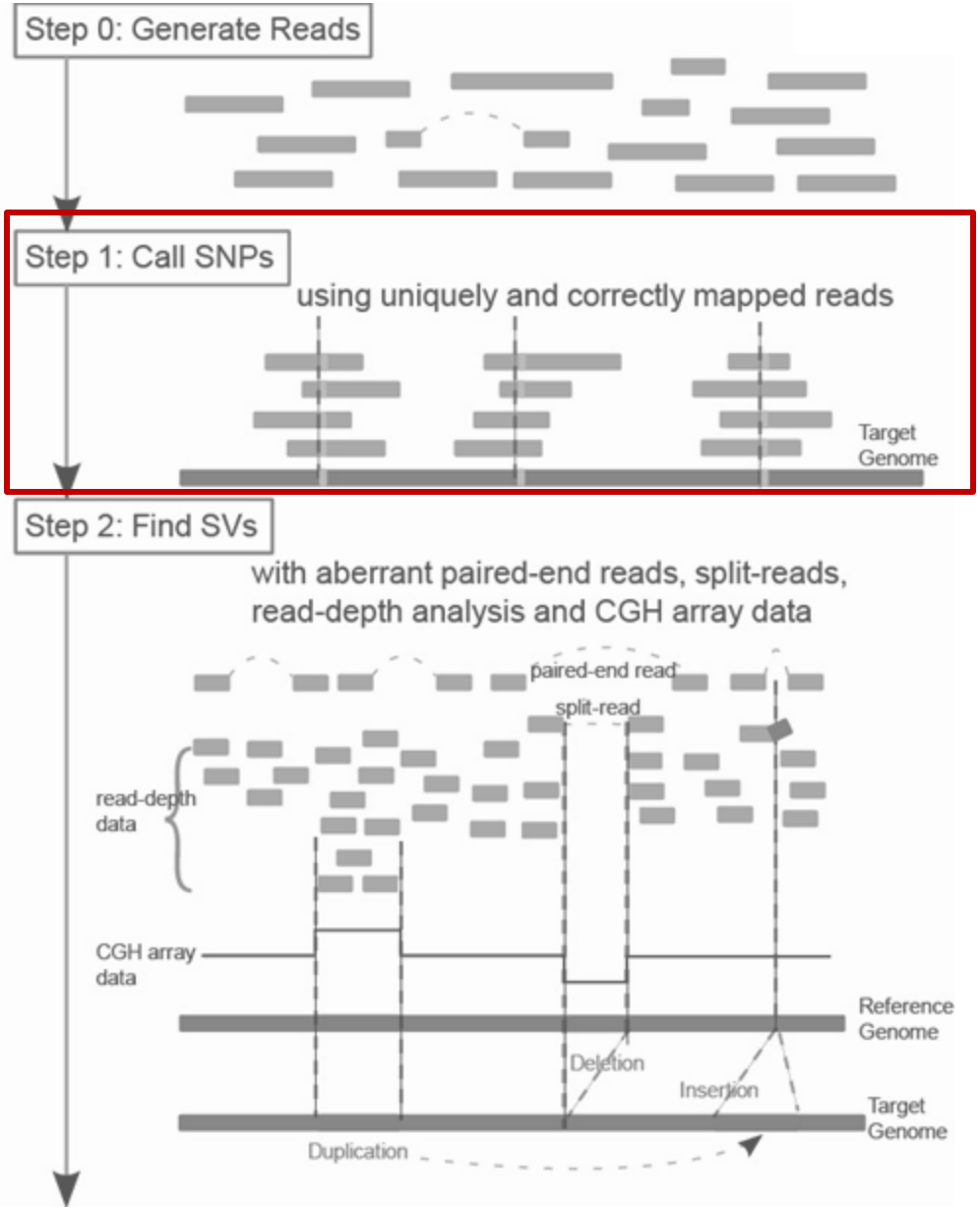
# Main Steps in Genome Resequencing

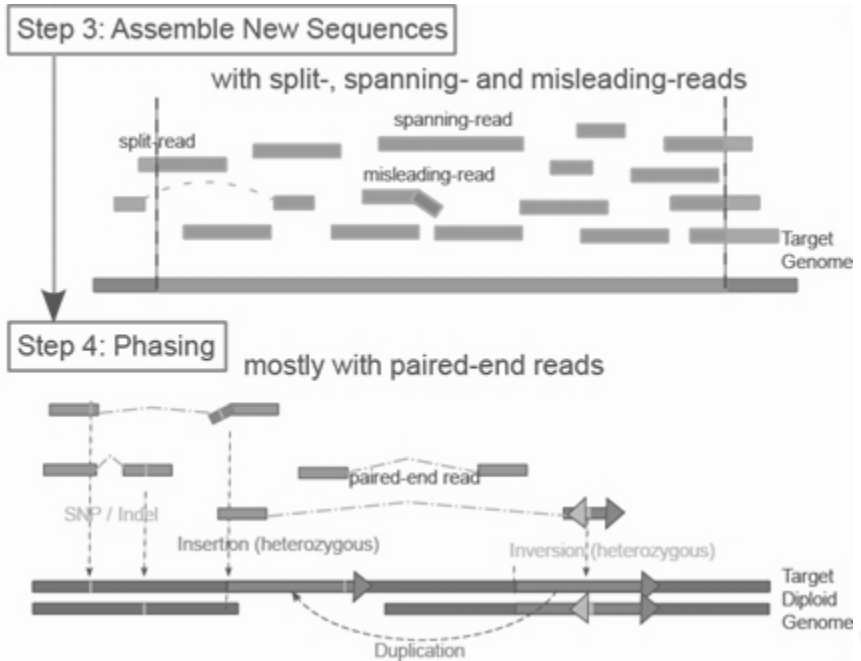
[Snyder et al. Genes & Dev. ('10)]



# Main Steps in Genome Resequencing

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



(More Later)

# Bayes' Theorem to detect genomic variant


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

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

In the above equation:

- $D$ refers to the observed data (the reads)
- $G$ is the genotype whose probability is being calculated
- $G_i$ refers to the $i$th 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.

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

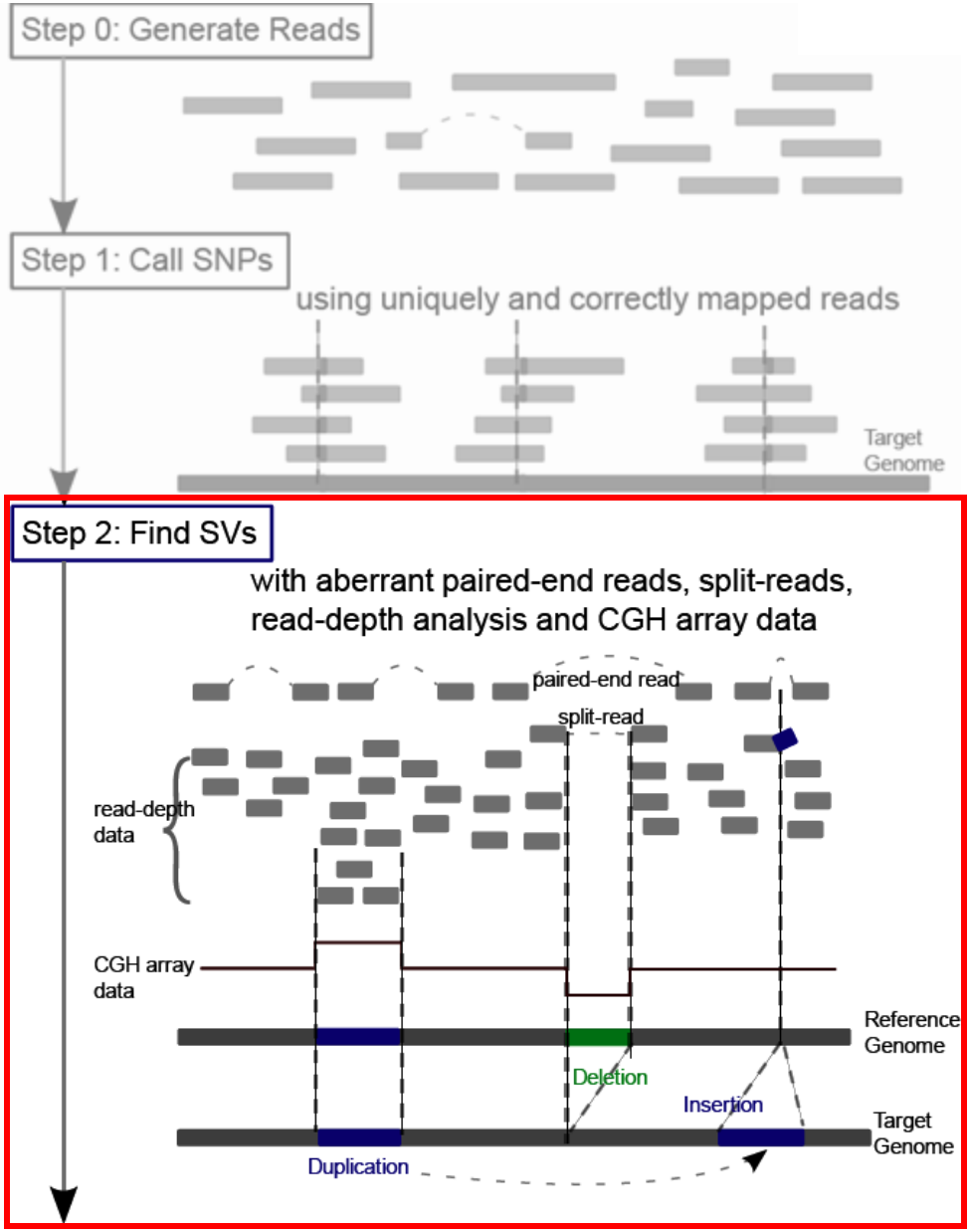
With errors, the calculation is more complicated. (However, the Bayesian formulation becomes more useful.) In general:

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

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

# Main Steps in Genome Resequencing

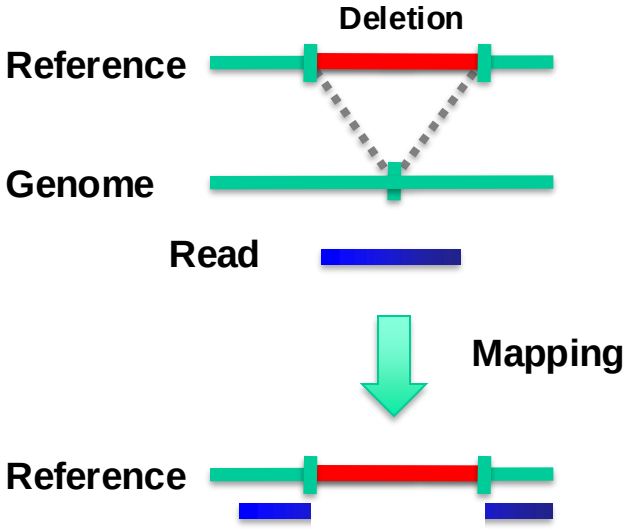
**[Snyder et al. Genes & Dev. ('10)]**



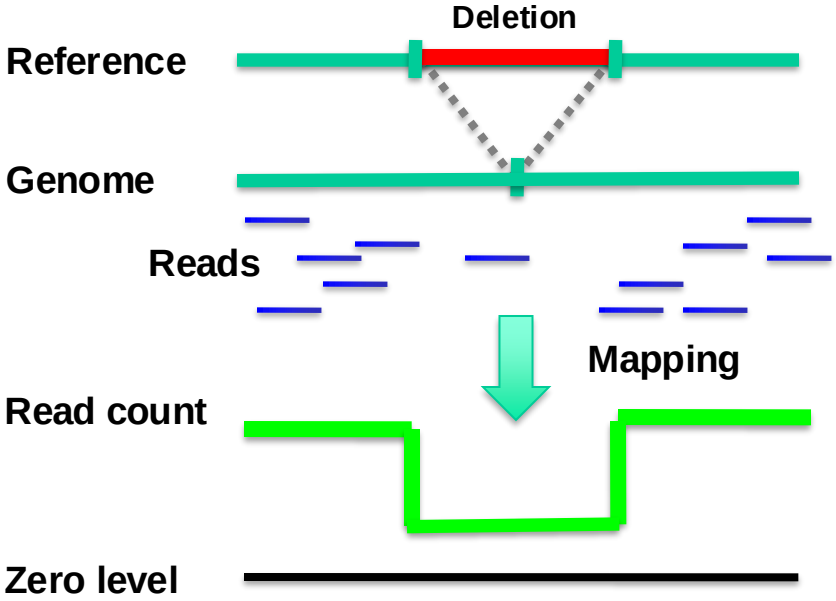
# Methods to Find SVs

## 1. Paired ends


## 2. Split read



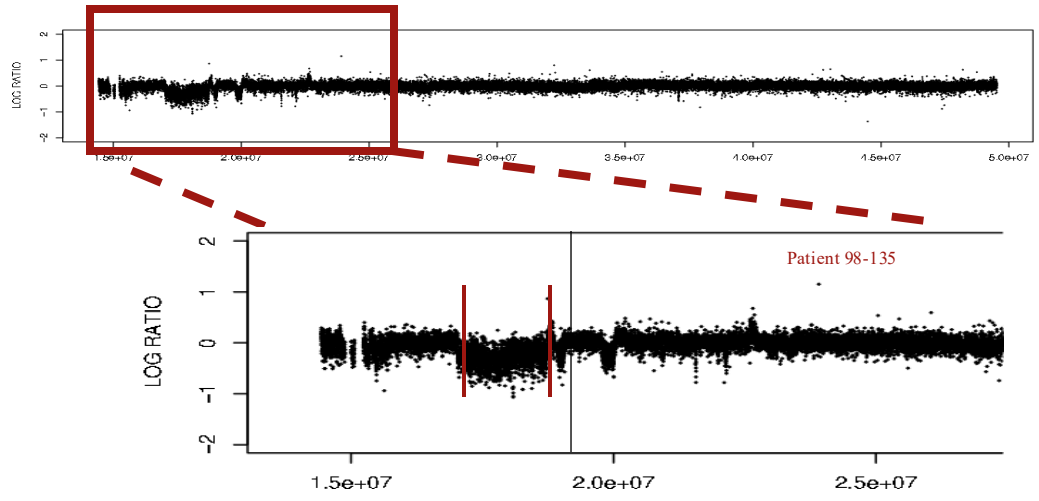

## 3. Read depth (or aCGH)




## 4. Local Reassembly

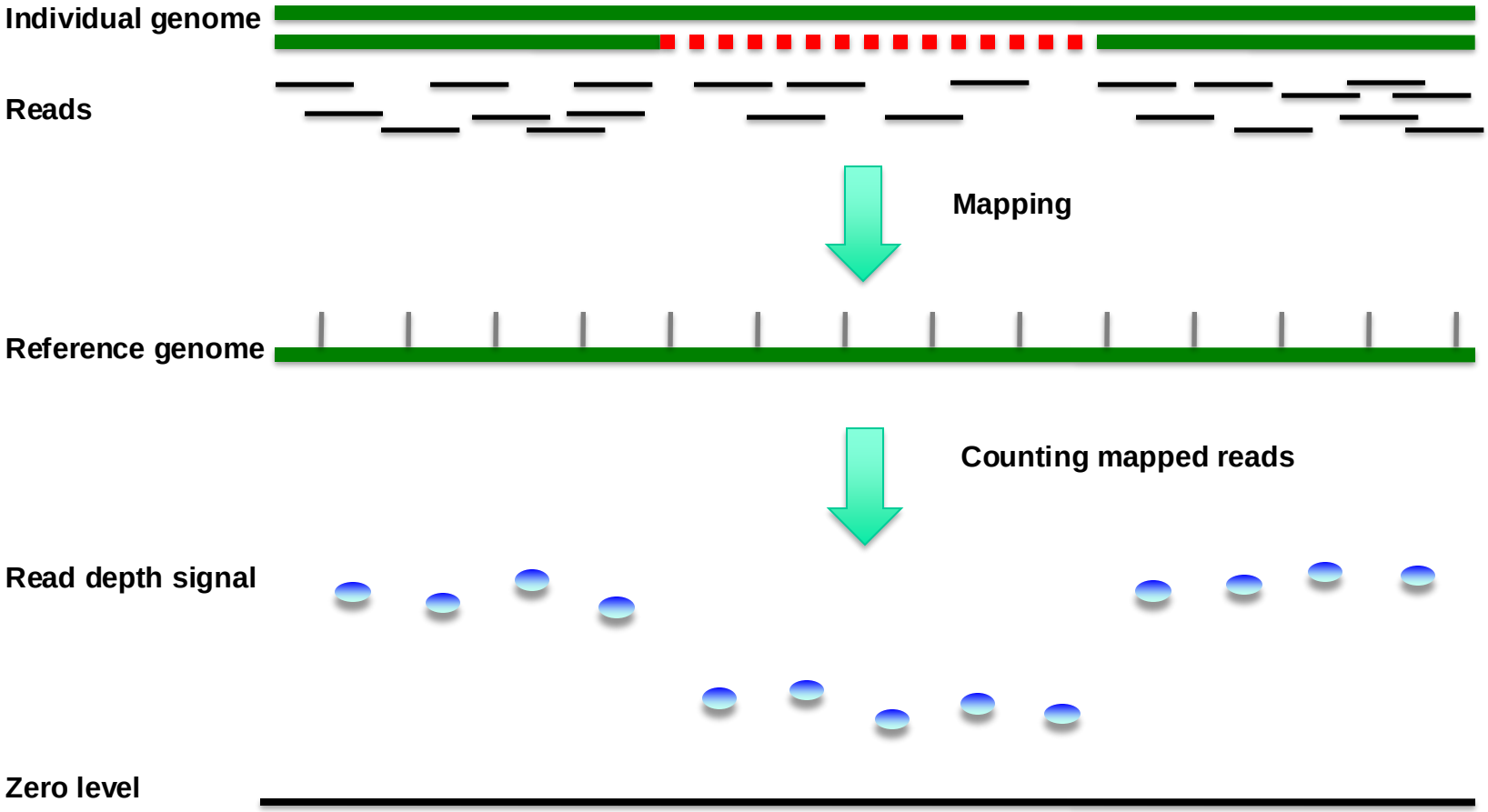
[Snyder et al. Genes & Dev. ('10)]

# Read Depth



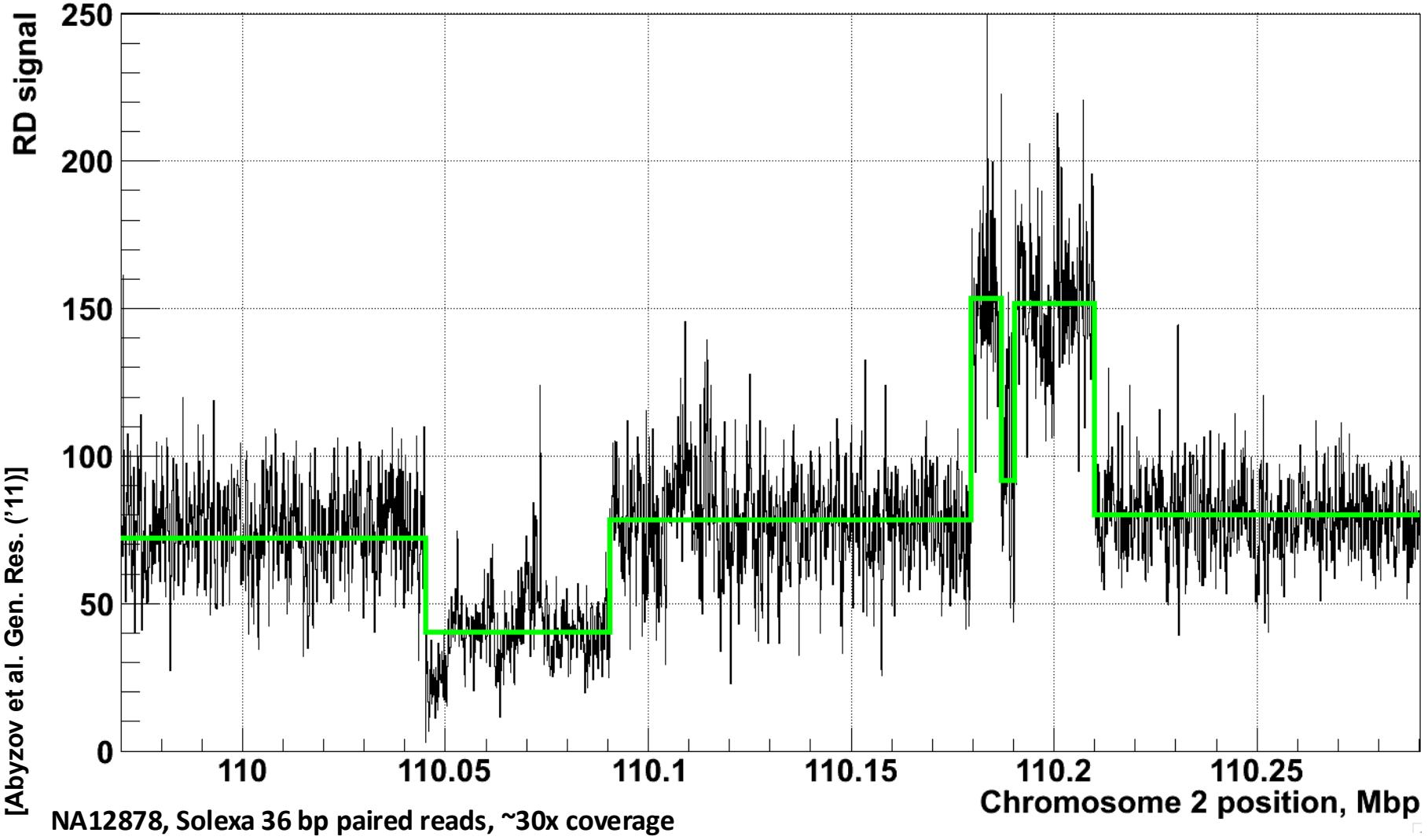

# Array Signal

## Read depth



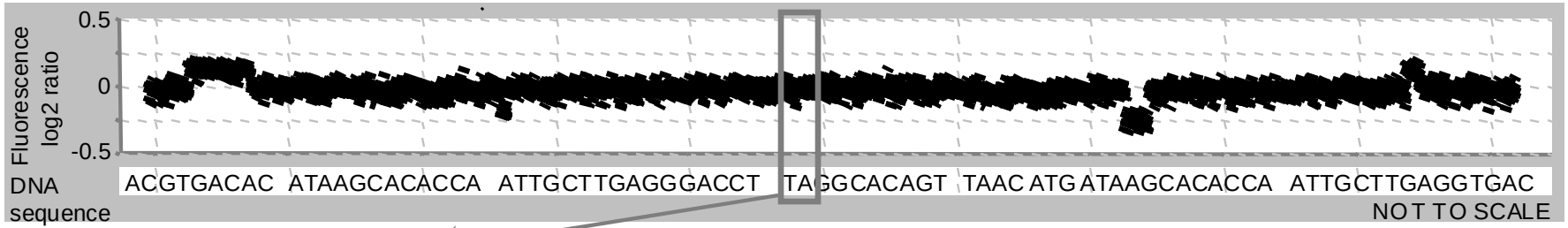

[Urban et al. ('06) PNAS; Wang et al. Gen. Res. ('09); Abyzov et al. Gen. Res. ('11)]

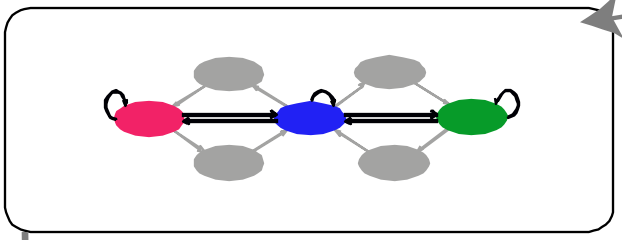
# Example of Application to RD data

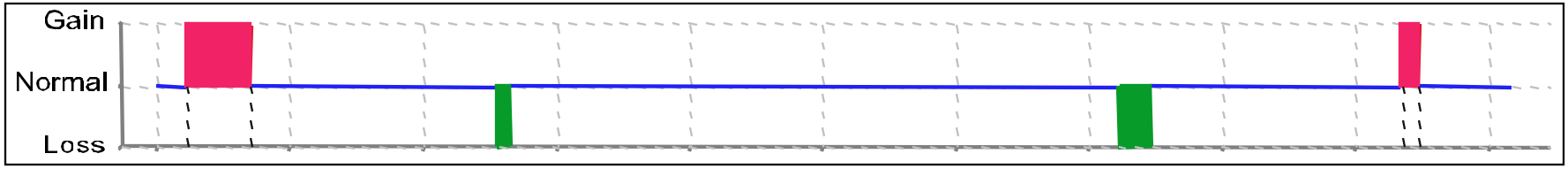


# HMM

- To get highest resolution on breakpoints need to smooth & segment the signal
- BreakPtr: prediction of breakpoints, dosage and cross-hybridization using HMMs

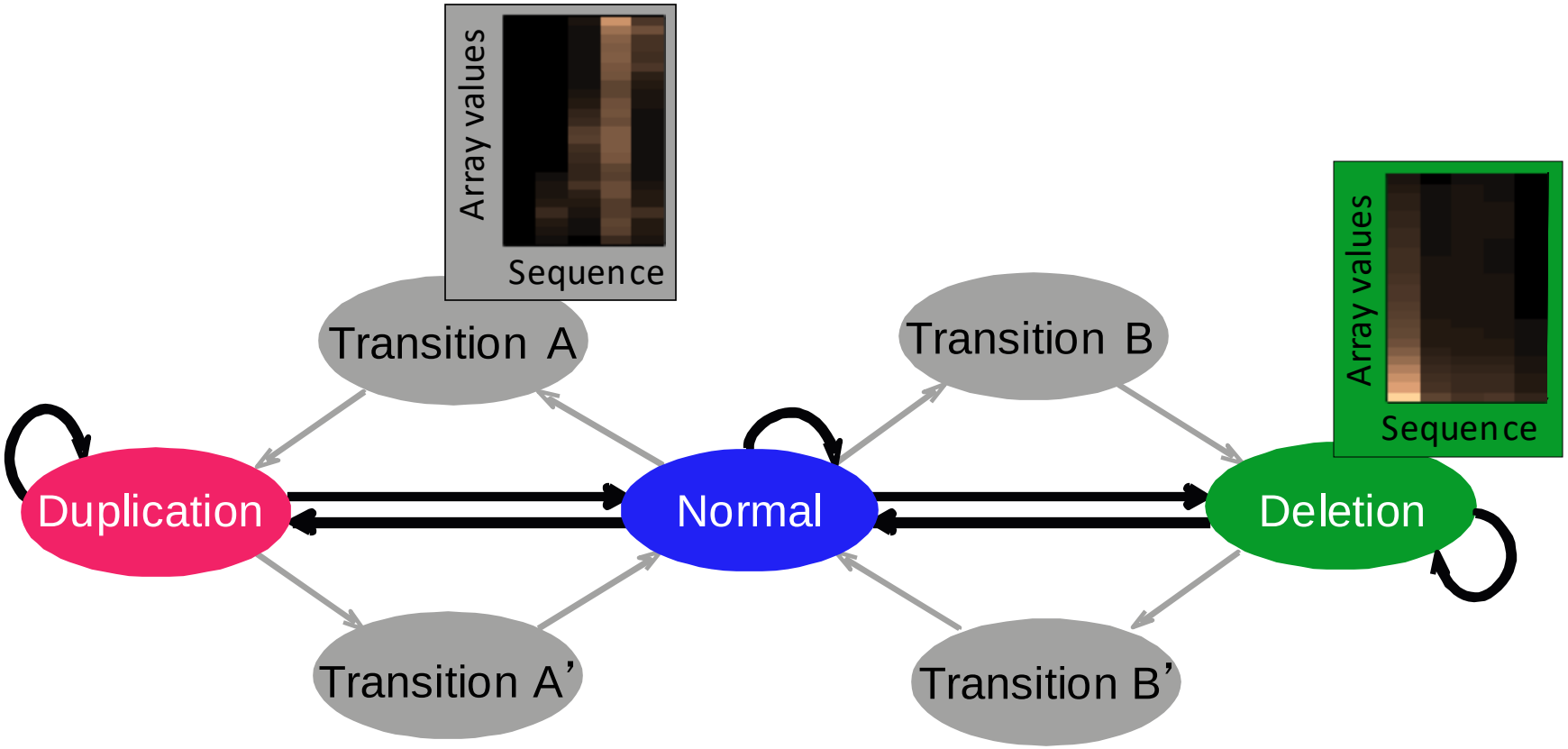





Korbel*, Urban* *et al.*, PNAS (2007)

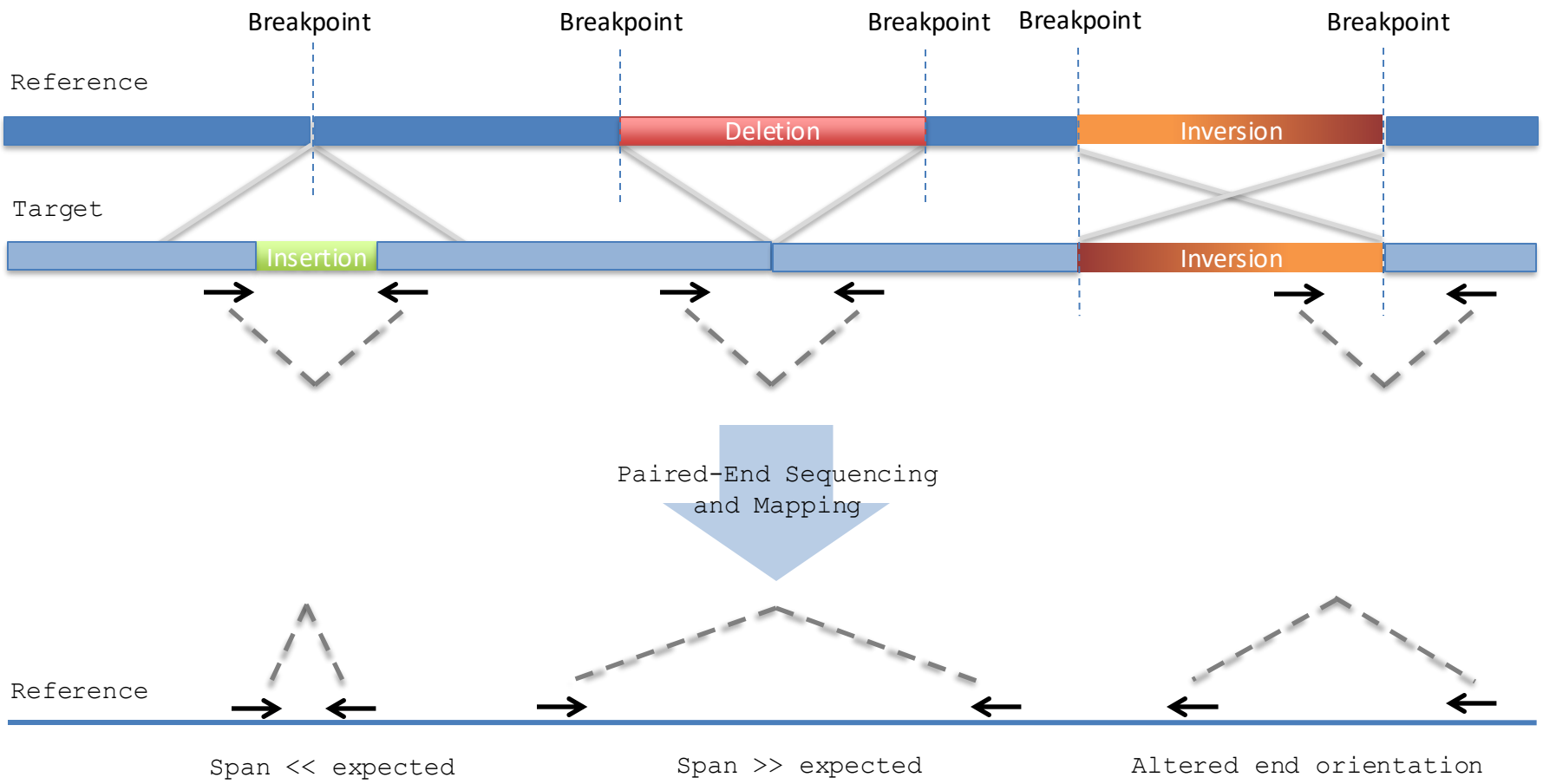
# Statistically integrates array signal and DNA sequence signatures (using a discrete-valued bivariate HMM)




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

# Paired-End

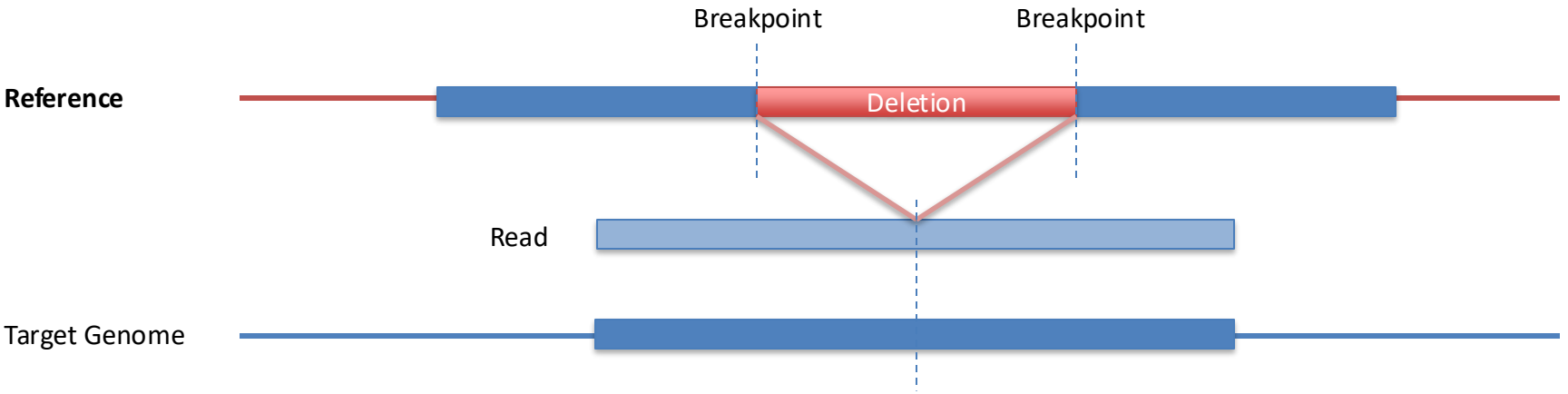
# Paired-End Mapping

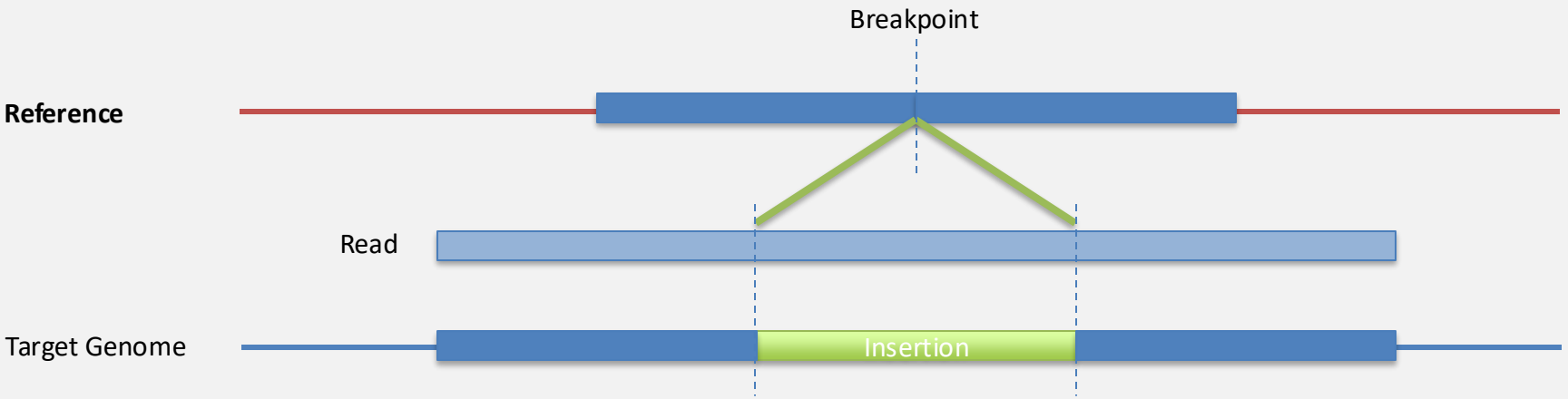



- Both paired-ends map within repeats.
- Limited the distance between pairs; therefore, neither large nor very small rearrangements can be detected

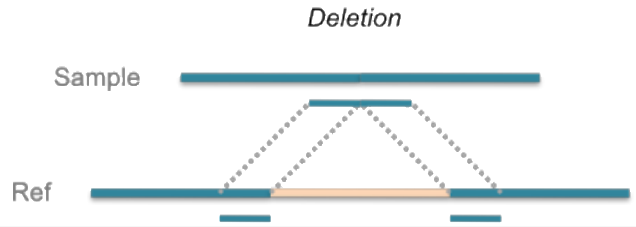
# Split Read

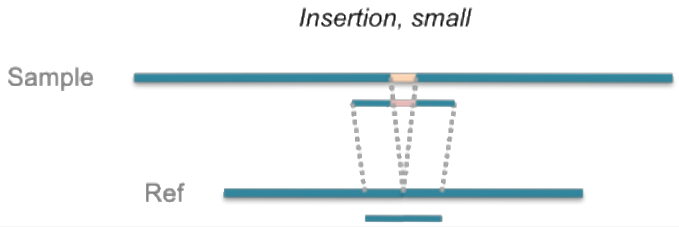
# Split-read Analysis

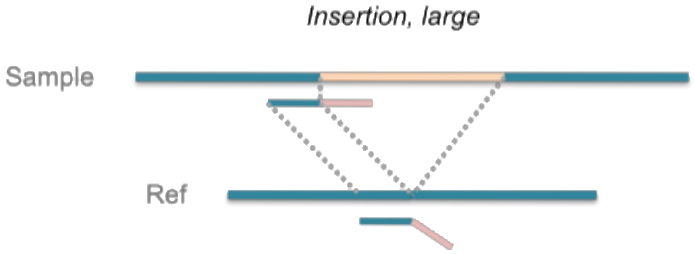


Simple SVs

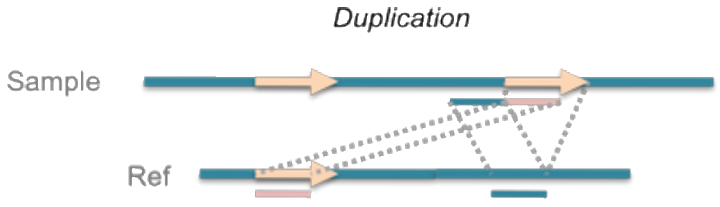


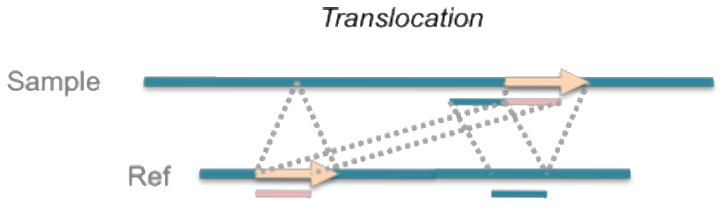



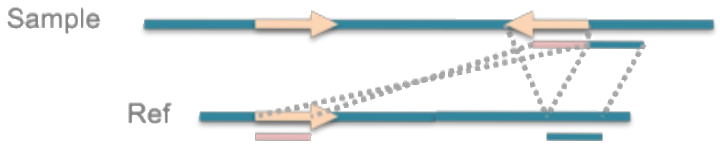



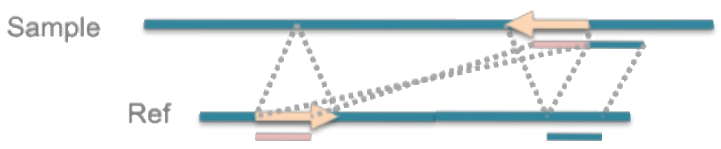

Deletions are the Easiest to Identify

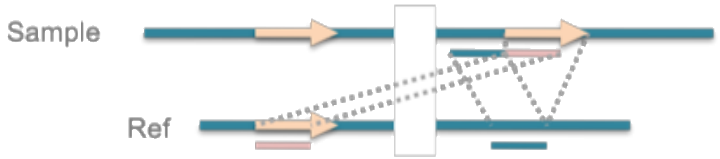
Complex SVs

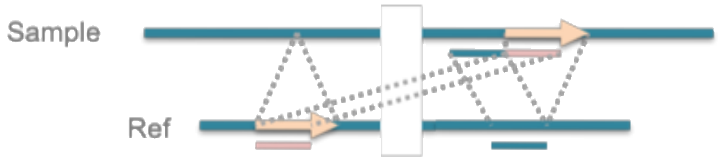



[Zhang et al. ('11) *BMC Genomics*]

# References


- Alkan, C., Coe, B. P., & Eichler, E. E. (2011). Nature Reviews Genetics, 12(5), 363–376. Genome structural variation discovery and genotyping. https://doi.org/10.1038/nrg2958 (Focus on Figs. 1 & 2)