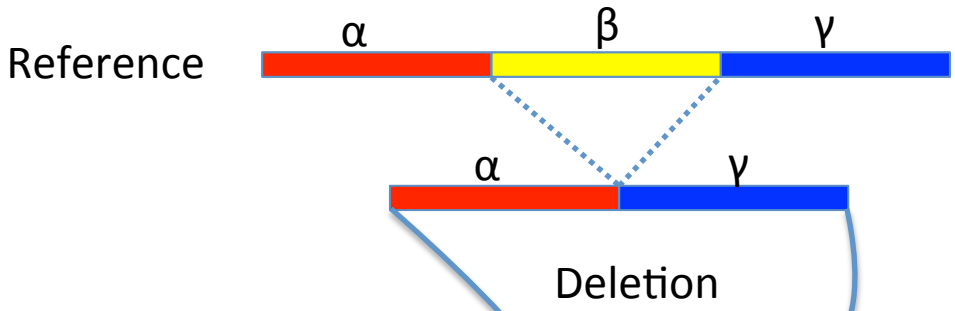
Integrated Genotyping of Structural Variation in NGS Data

Xian Fan

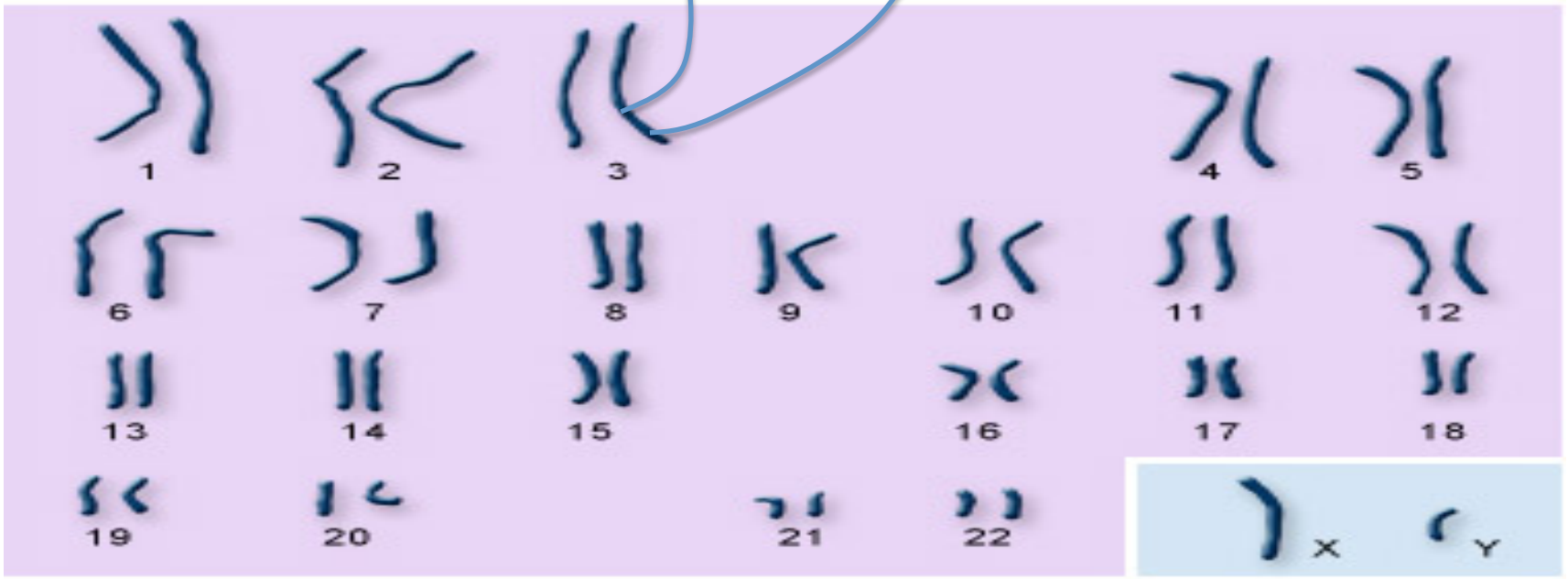
Luay Nakhleh

Ken Chen

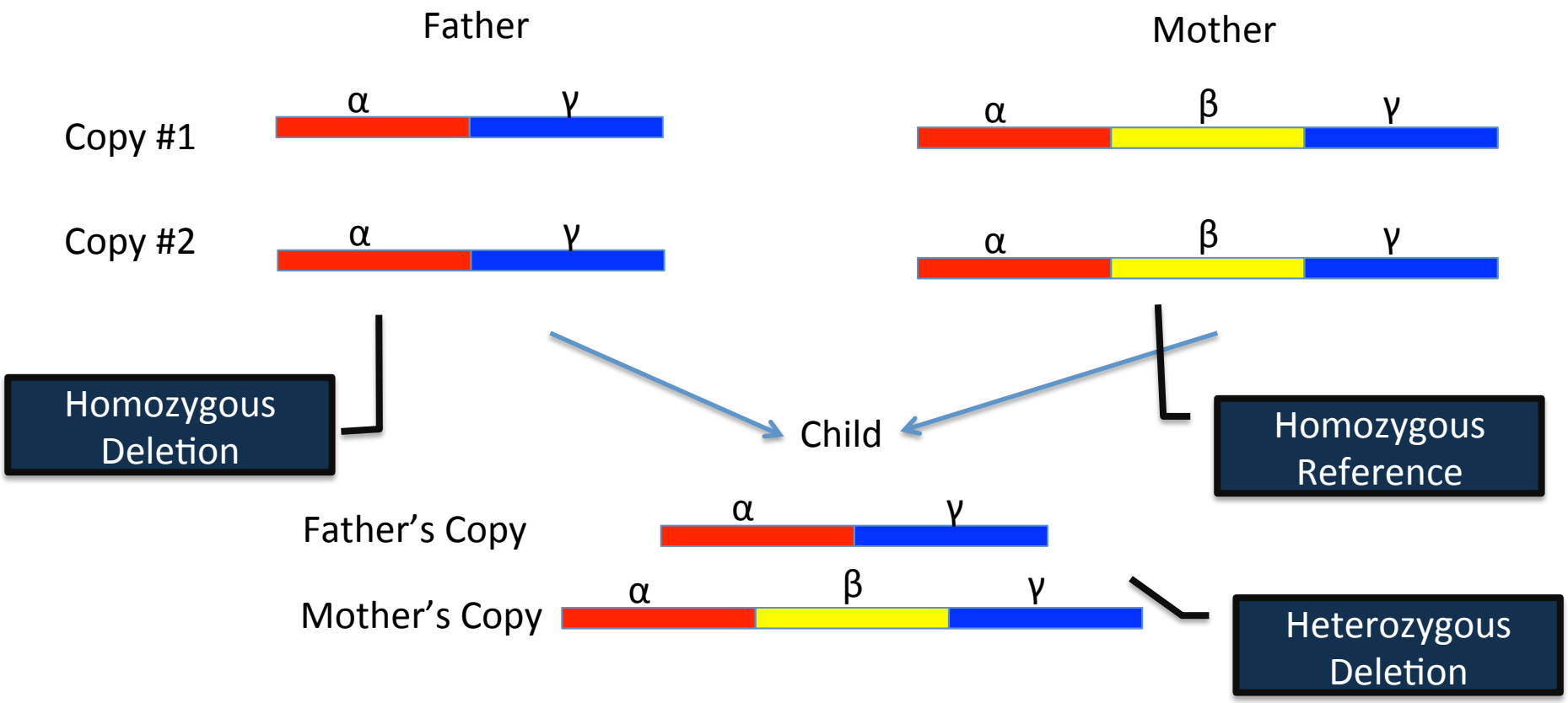
Structural Variation – Variation in Chromosomal Structure



Mr Unknowns' genome:



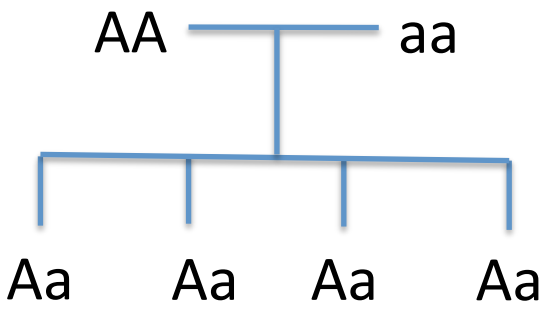
Structural Variation Genotyping



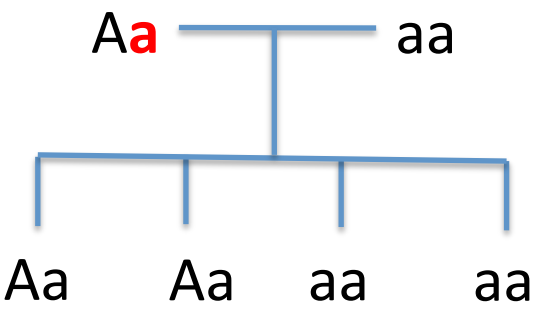
- Genotyping: find variation states in a genomic region.

Genotyping is Important to Genetic Disease Inheritance

A: dominant



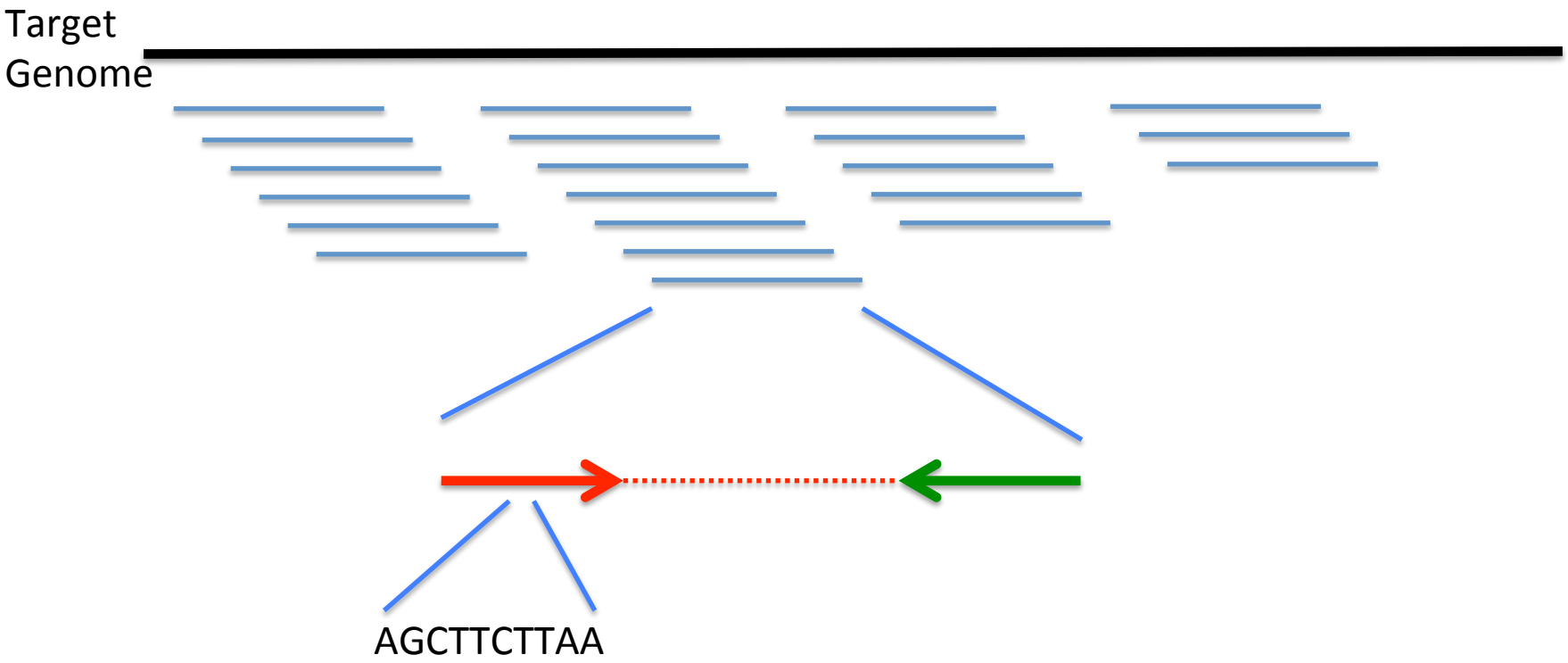
100%



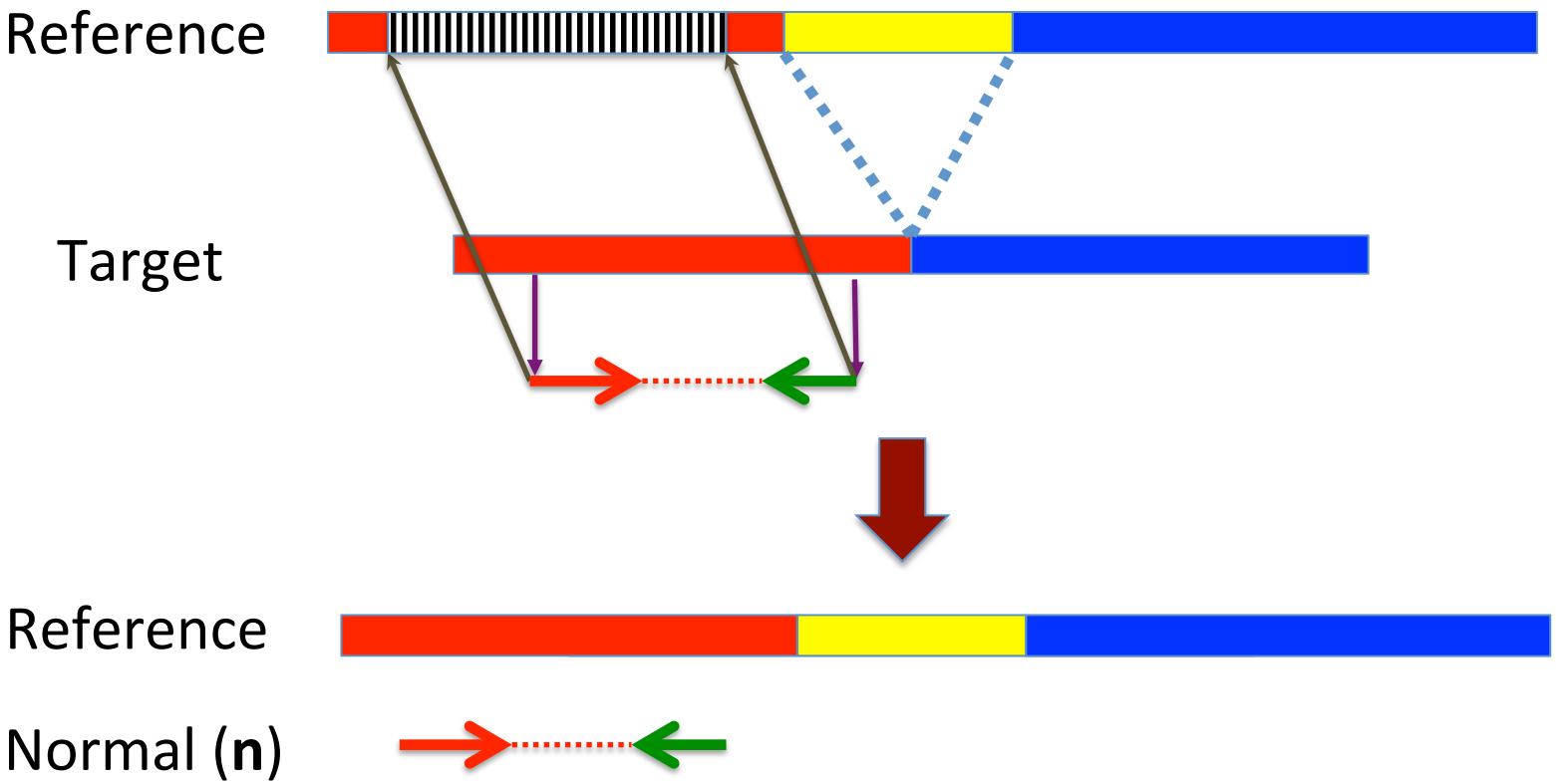
50%

Next Generation Sequencing (NGS)

- Many paired-end reads are randomly sequenced from the target genome.

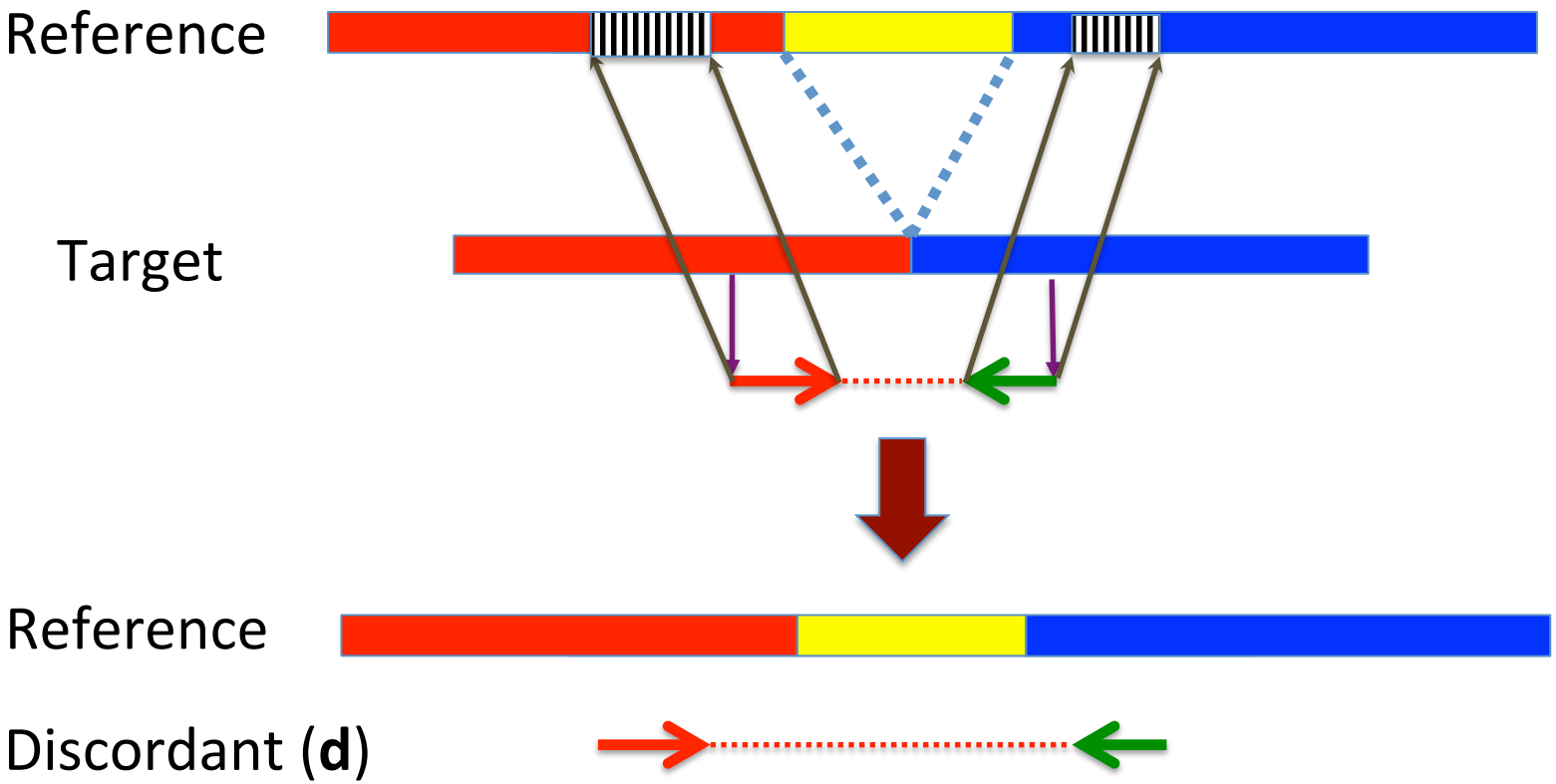


Preprocessing: Alignment to the Reference – Normal Reads



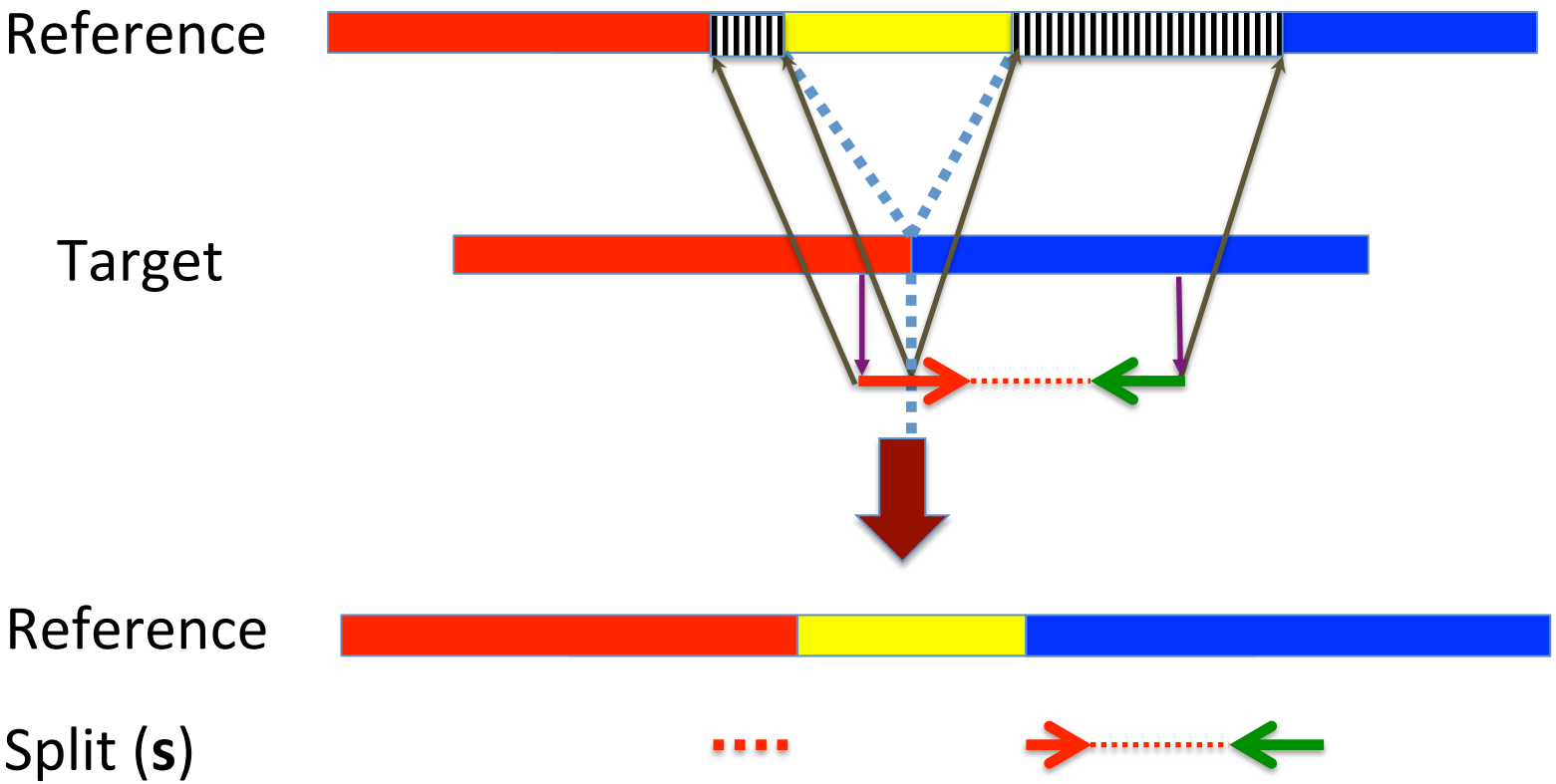
Normal segment: normal distance between left and right reads.

Preprocessing: Alignment to the Reference – Discordant Reads



Discordant segment: abnormal distance between left and right reads.

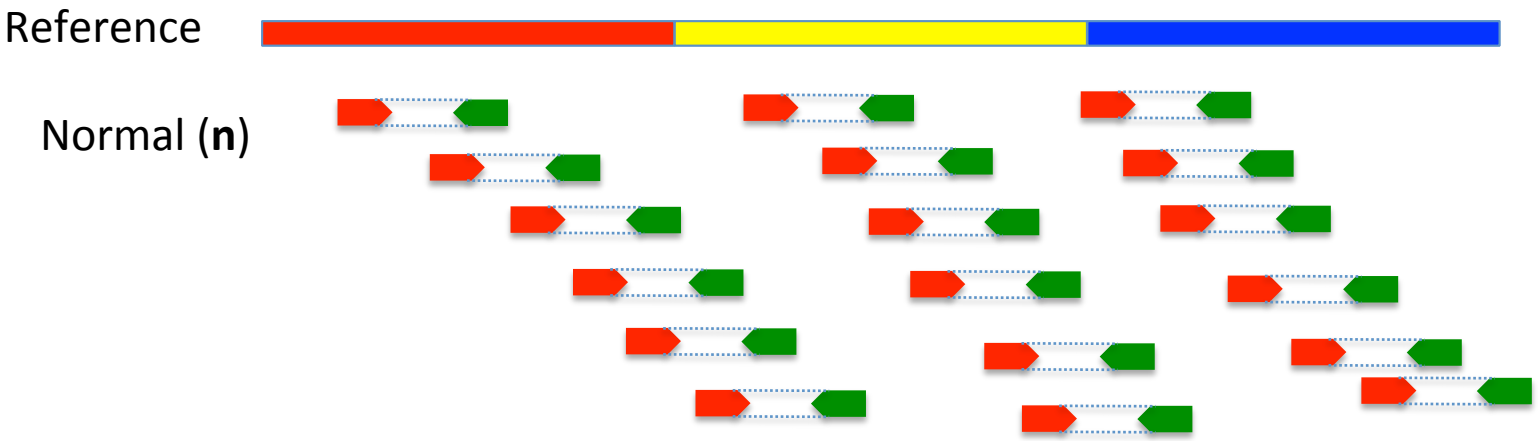
Preprocessing: Alignment to the Reference – Split Reads



Split segment: split into two in left or right read.

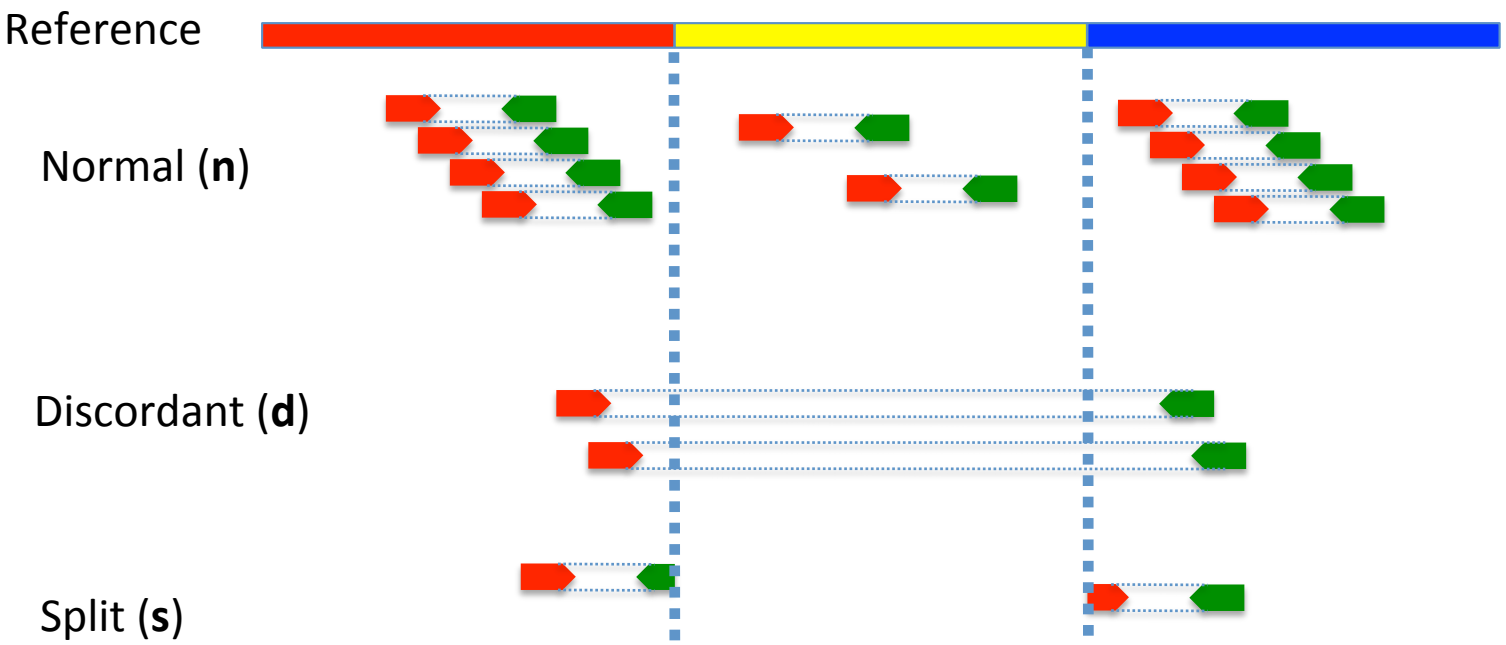
A Standard Homozygous Reference – No Deletion

- Two copies of normal reads inside deleted region.
- No discordant and split reads



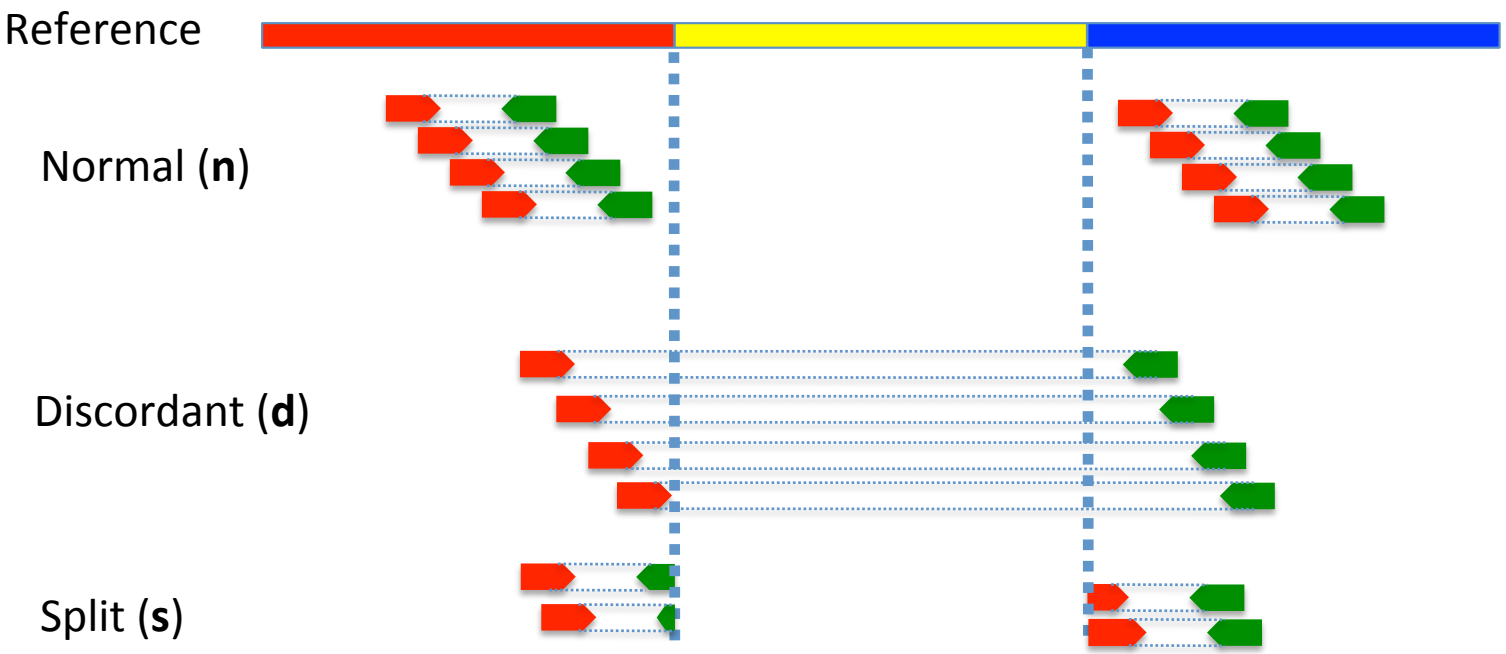
A Standard Heterozygous Deletion

- One copy of normal reads inside deleted region.
- A few discordant and split reads



A Standard Homozygous Deletion

- Zero copy of normal reads inside deleted region.
- A few discordant and split reads (twice as heterozygous deletion)



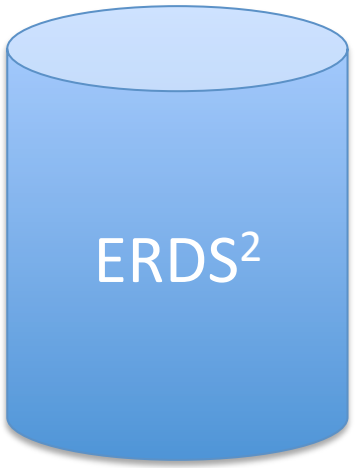
Problem

- Given:
  - **n**, **d** and **s** for a particular loci in a sample genome
- Goal:
  - Estimate the genotype of the deletion
    - homozygous deletion
    - heterozygous deletion
    - homozygous reference

Combining **n**, **d**, and **s**

- Random sampling and noise
  - Coverage is over-dispersed.
- Imperfect alignment and repetitive sequence
  - Reads aligned to this region might come from somewhere else.
  - One or more of **n**, **d** and **s** is not trustable.
- Emerging papers address how to combine these three signals.

Literature of Combination



- Mixture Gaussian on population genomes
- Bayesian Framework for combination of **n**, **d**, **s**

- Decision tree
- Applicable only to high coverage data

1. Handsaker, R. E., Korn, J. M., Nemesh, J. & McCarroll, S. A. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat Genet* **43**, 269-276 (2011).
2. Zhu, M. *et al.* Using ERDS to infer copy-number variants in high-coverage genomes. *American journal of human genetics* **91**, 408-421 (2012).

We Proposed BreakDown: A General Framework for Genotyping

Genotype likelihood of a variant:

$$L = P(D \mid G)$$

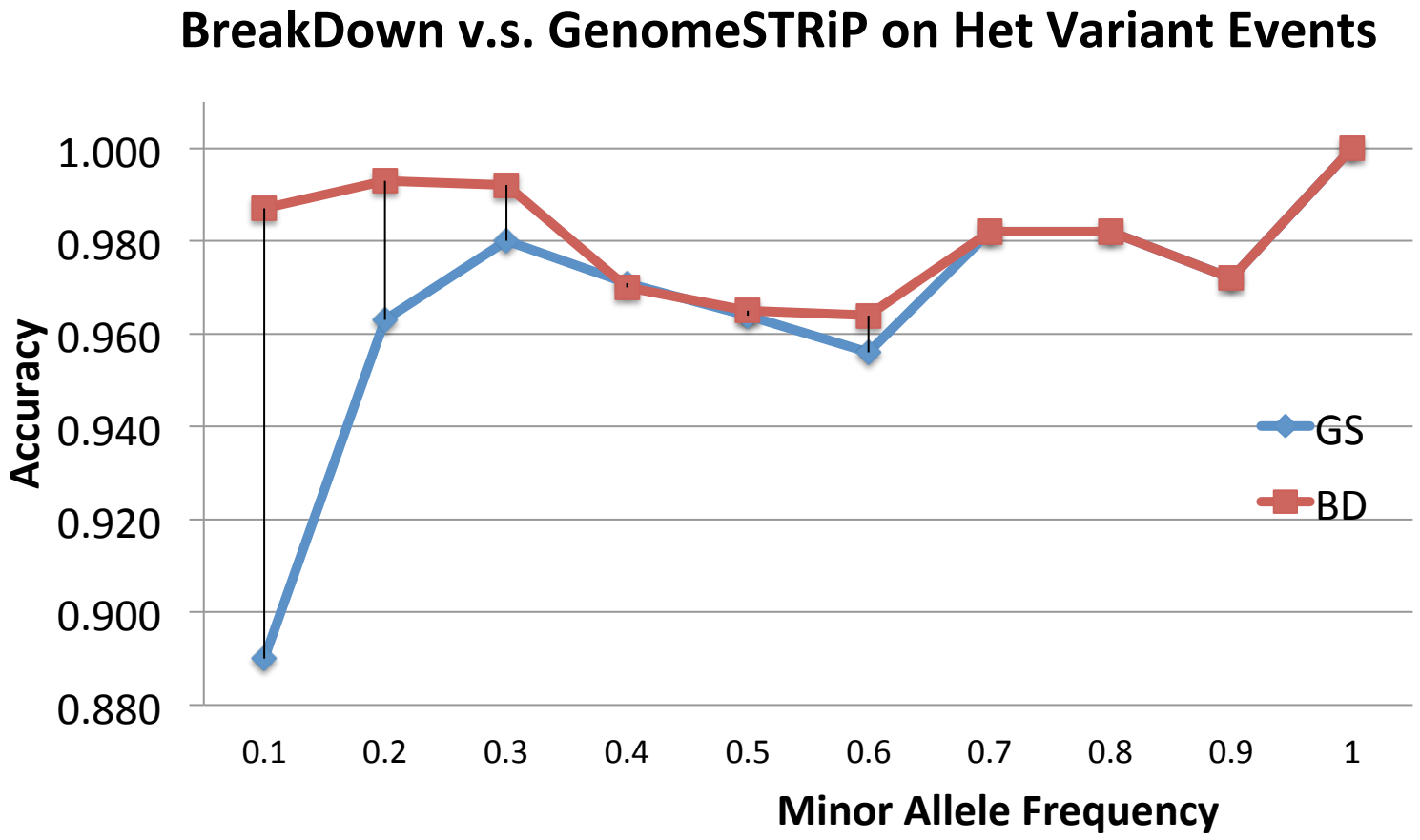
$$= P(n \mid G) \bullet P(d \mid G) \bullet P(s \mid G)$$

D: Aligned reads in a window encompassing the variant (**n**, **d**, **s**)

G: {homozygous deletion, heterozygous deletion, homozygous reference}

Results – 1000 Genomes Project

- 43 normal low coverage samples (2X – 8X)



Results – Cancer Data

- Metastatic melanoma cell line COLO-829 (41X sequence coverage) and matched normal one (32X).
- Discovered 82 novel loss of heterozygosity calls (normal is heterozygous whereas tumor is not).
- Some overlap with important genes.

Conclusion

- BreakDown: statistically tackle SV genotyping by integrating three sources of information.
- Applicable to
  - Both single samples and populations;
  - Both high and low coverage data.
- Applied to both normal and cancer genomes.
  - High accuracy achieved on normal genome;
  - Novel discovery obtained in cancer genome.

Thanks!

- Questions?