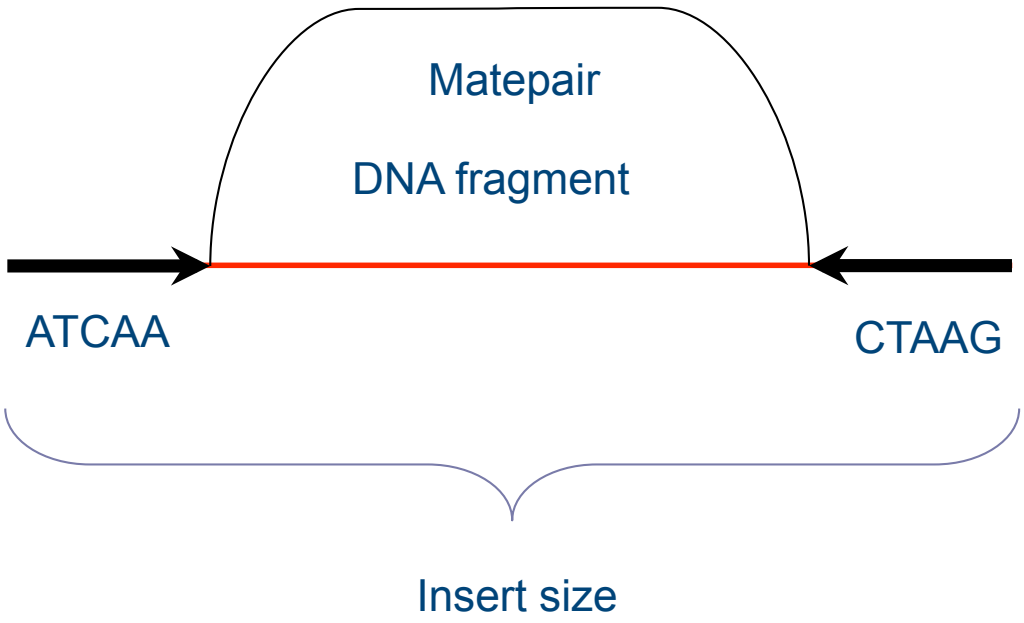
Detecting INDELs and CNVs with High Throughput Sequencing

Michael Brudno

Department of Computer Science
Banting& Best Dept of Medical Research
University of Toronto

NGS 2009, October 1, 2009

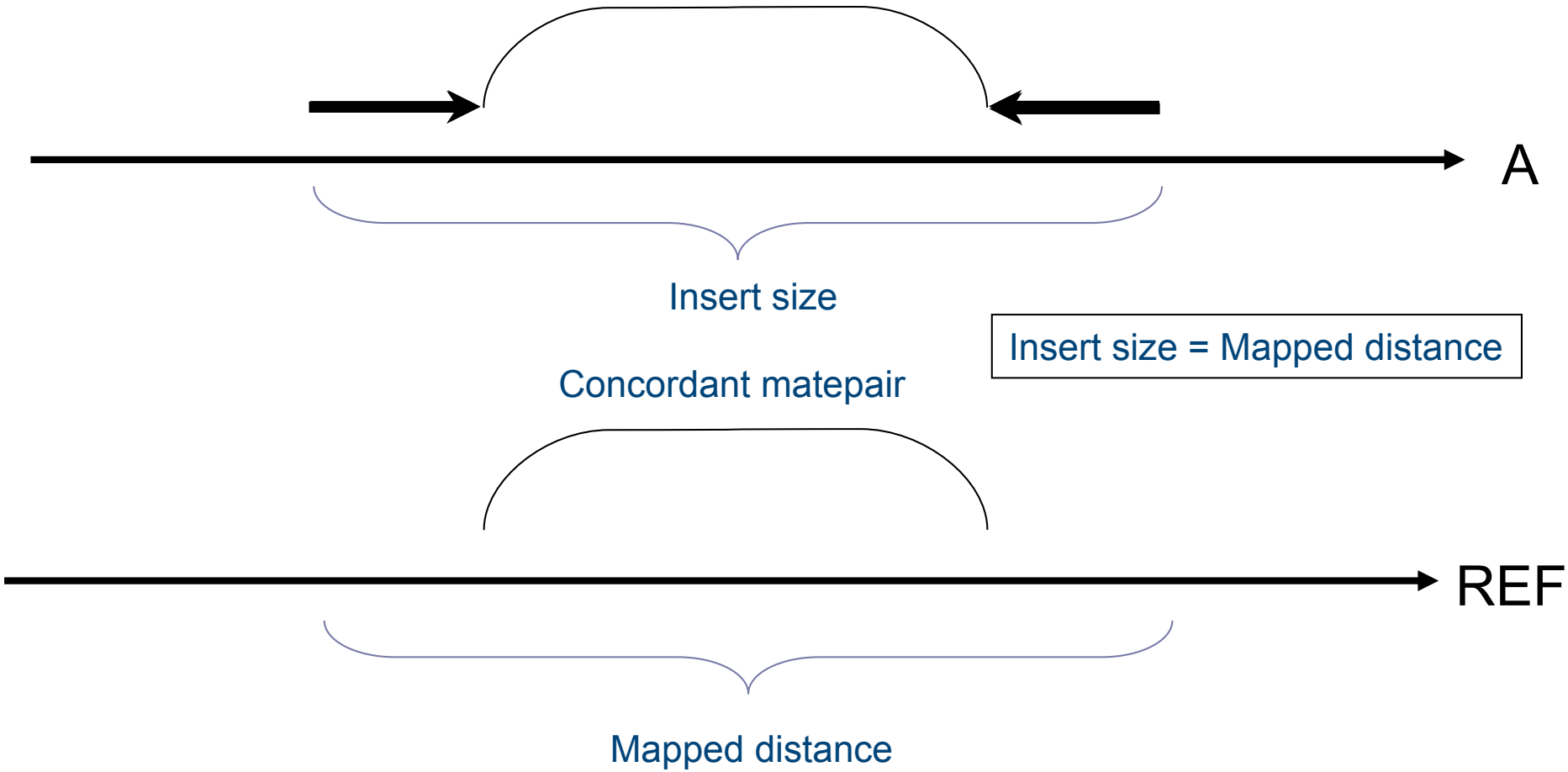
What are Matepairs?



For now, assume insert size is perfect

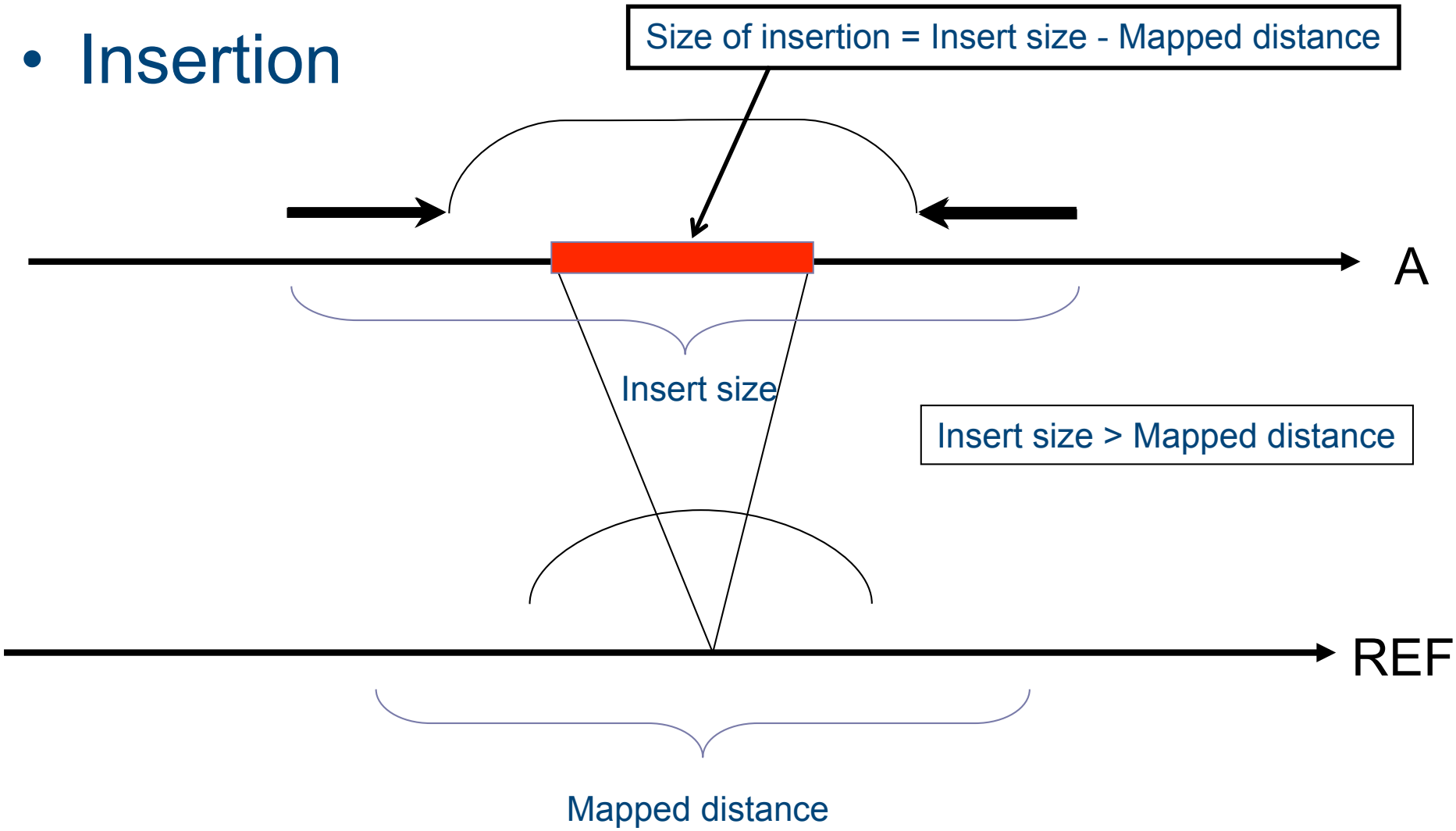
Detecting Structural Variants

- No structural variants



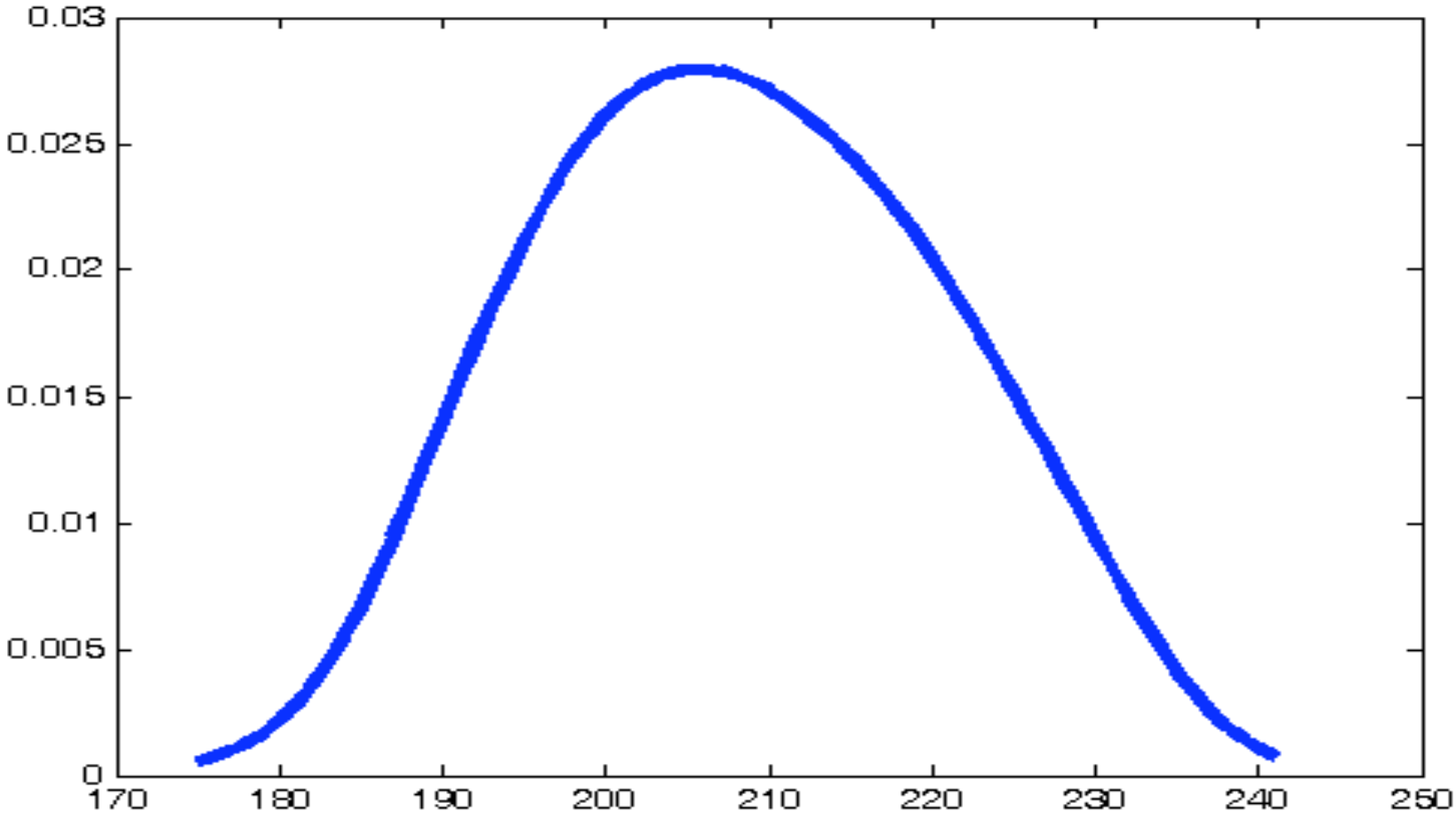
Detecting Structural Variants

- Insertion

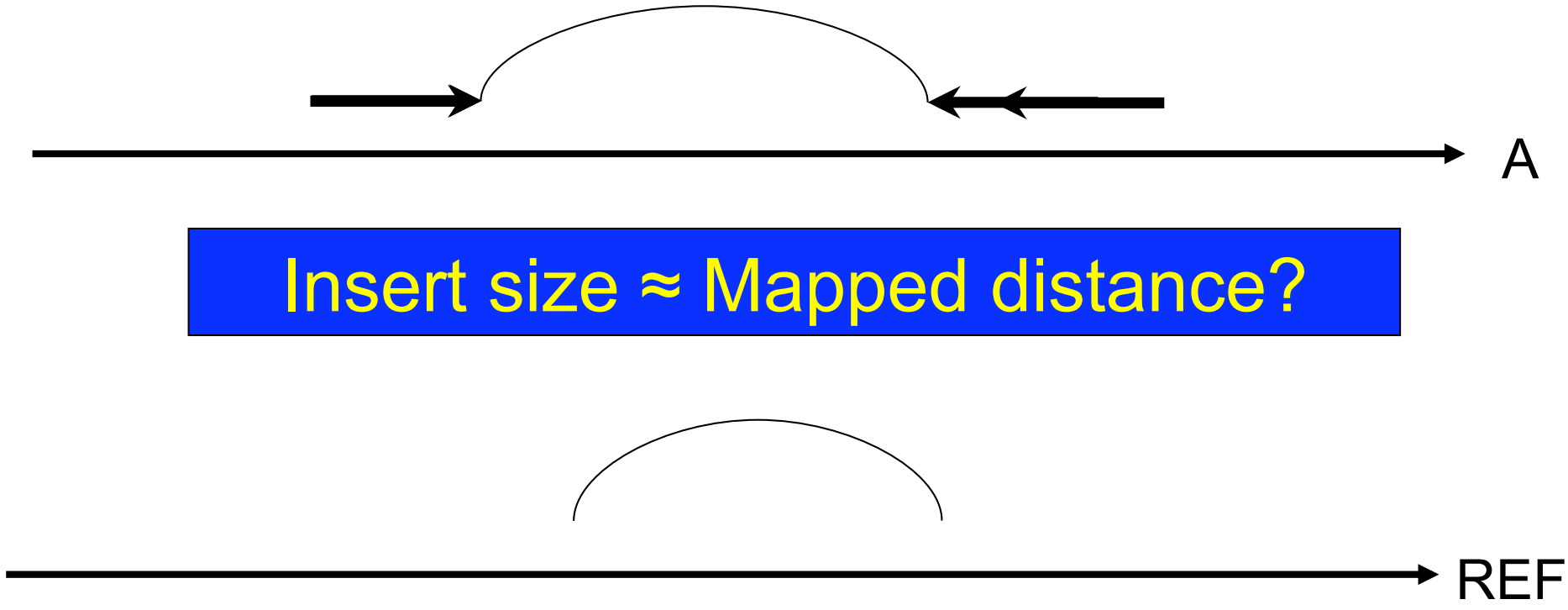


Distribution of Insert Sizes

- In reality, insert sizes of matepairs are not perfect

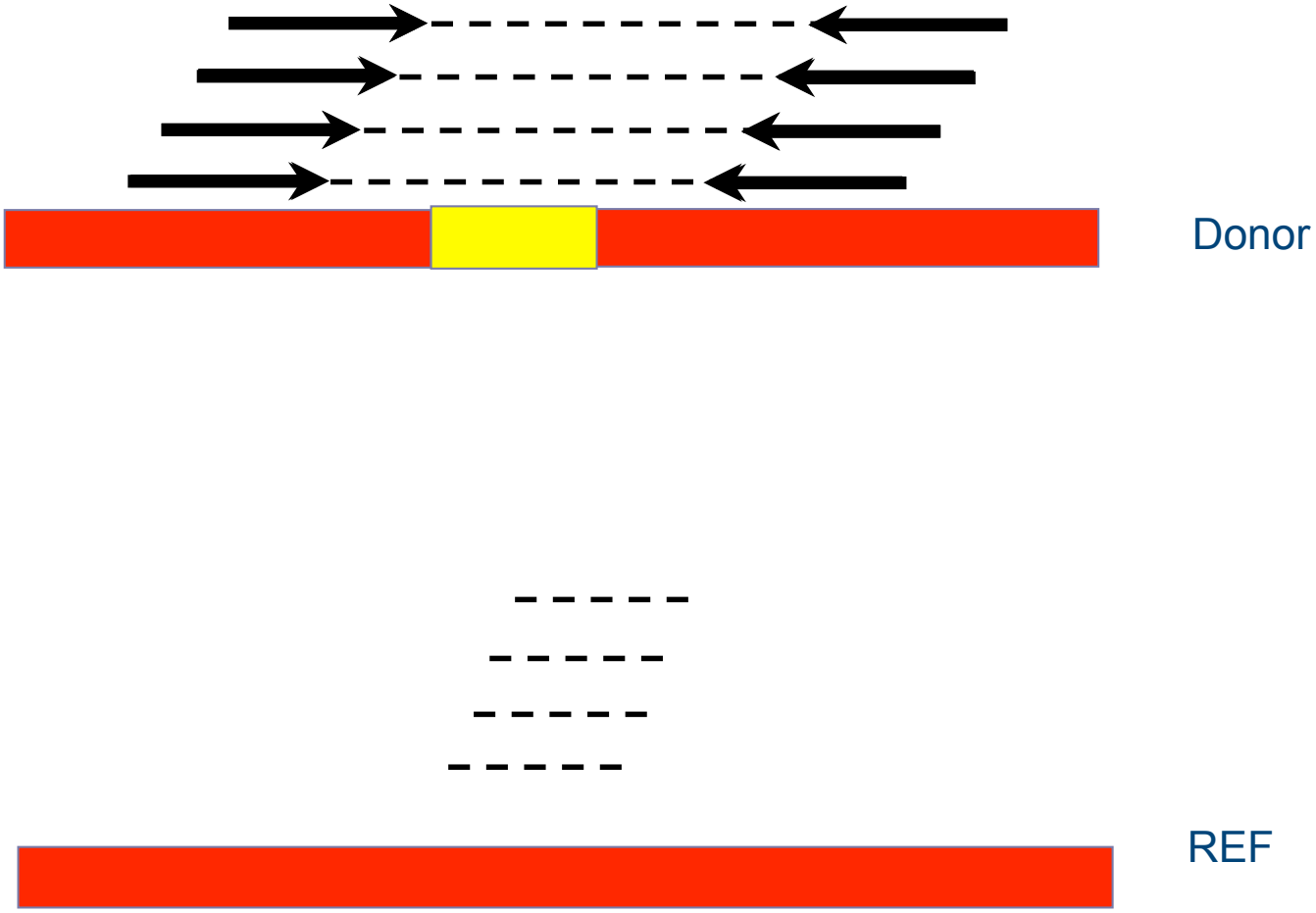


Detecting Smaller INDELS



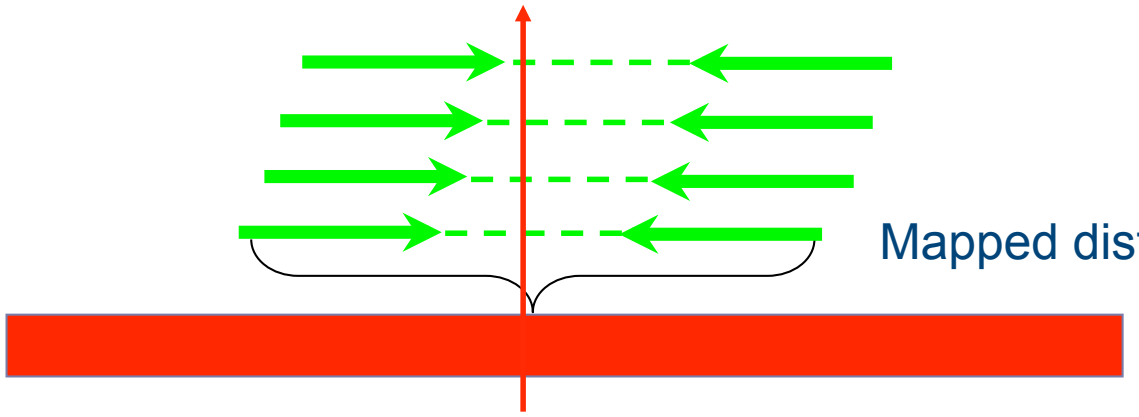
Small Insertion… or noise

Haploid Case – Alignment



Haploid Case – Alignment

Donor



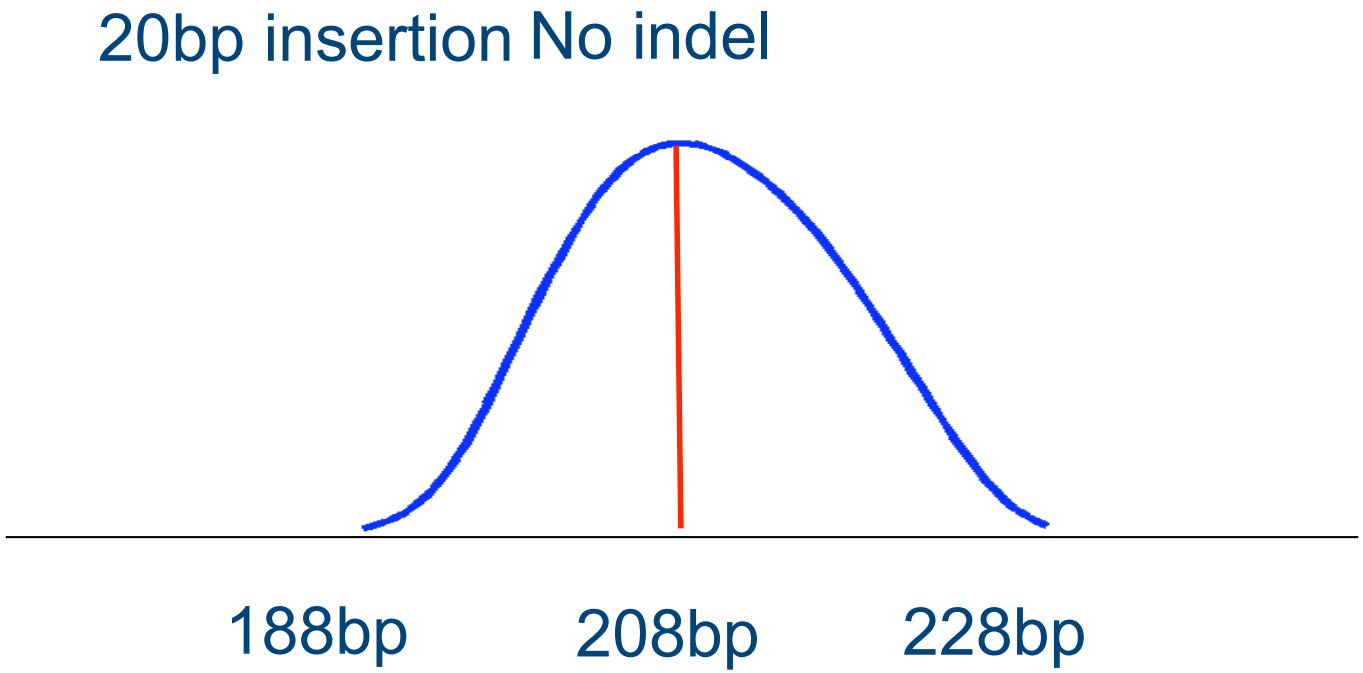
Mapped distance

REF

Cluster

Haploid Case – Distribution

Make a distribution of mapped distances in each cluster
=> The distribution shifts if there is an INDEL

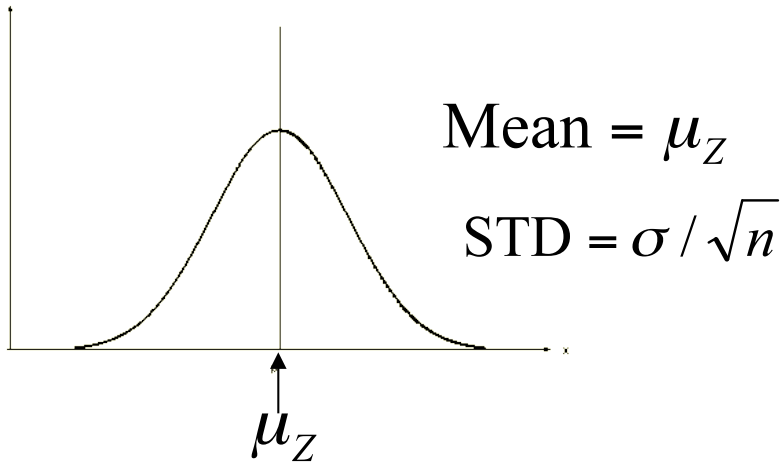


Accuracy of INDEL Estimation

Central limit theorem

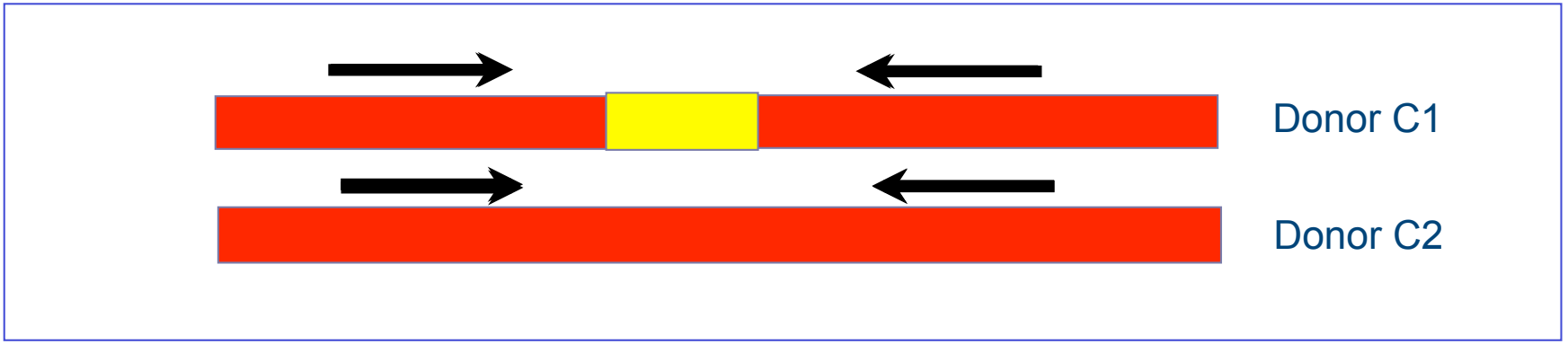
Mean of N independent random variables with finite mean μ and variance σ^2 follows Gaussian with mean μ and standard deviation $\sigma / \sqrt{N}$

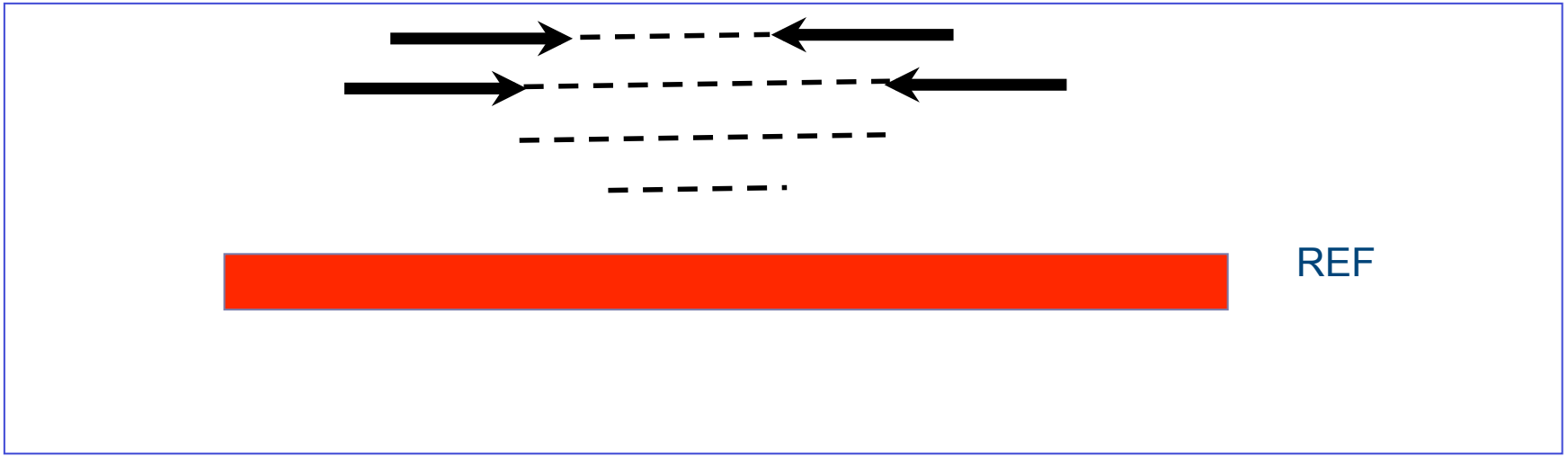
$\mathbf{Z} = \{Z_1 \ldots Z_n\}$: random vars for size of indels from each pair



Diploid Case – Alignment & Clustering

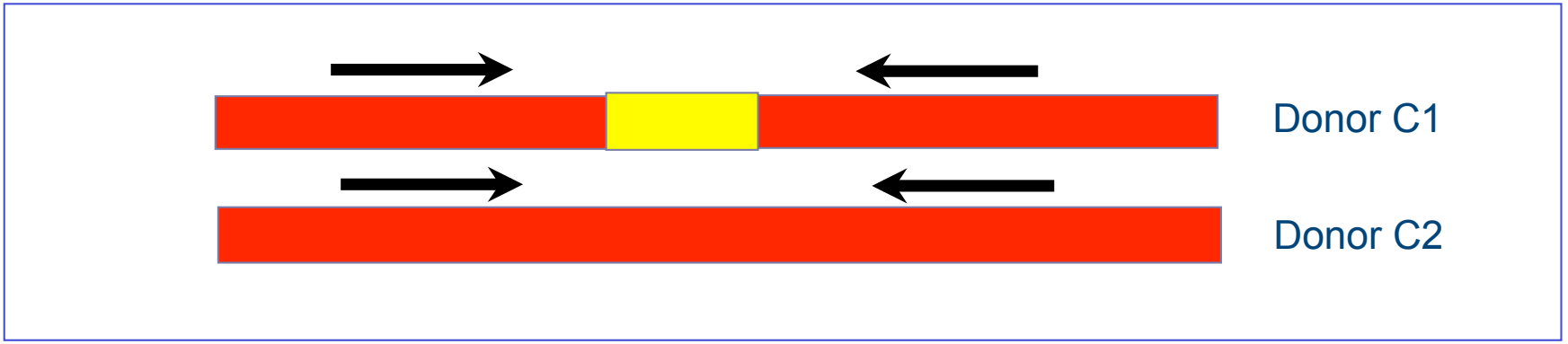
Heterozygous insertion

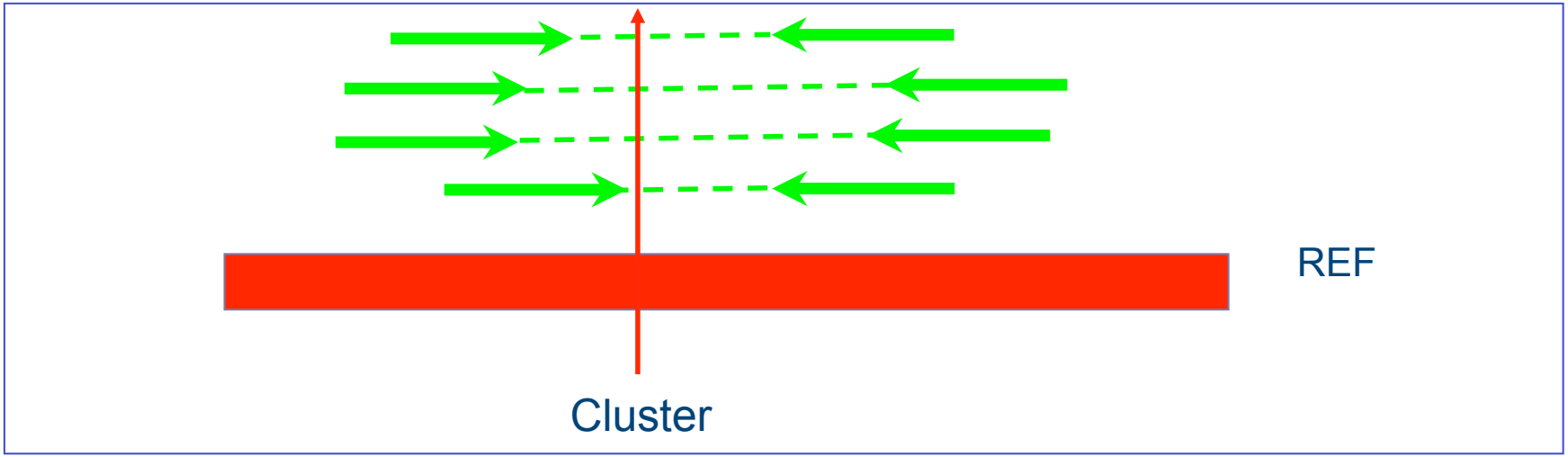




Diploid Case – Alignment & Clustering

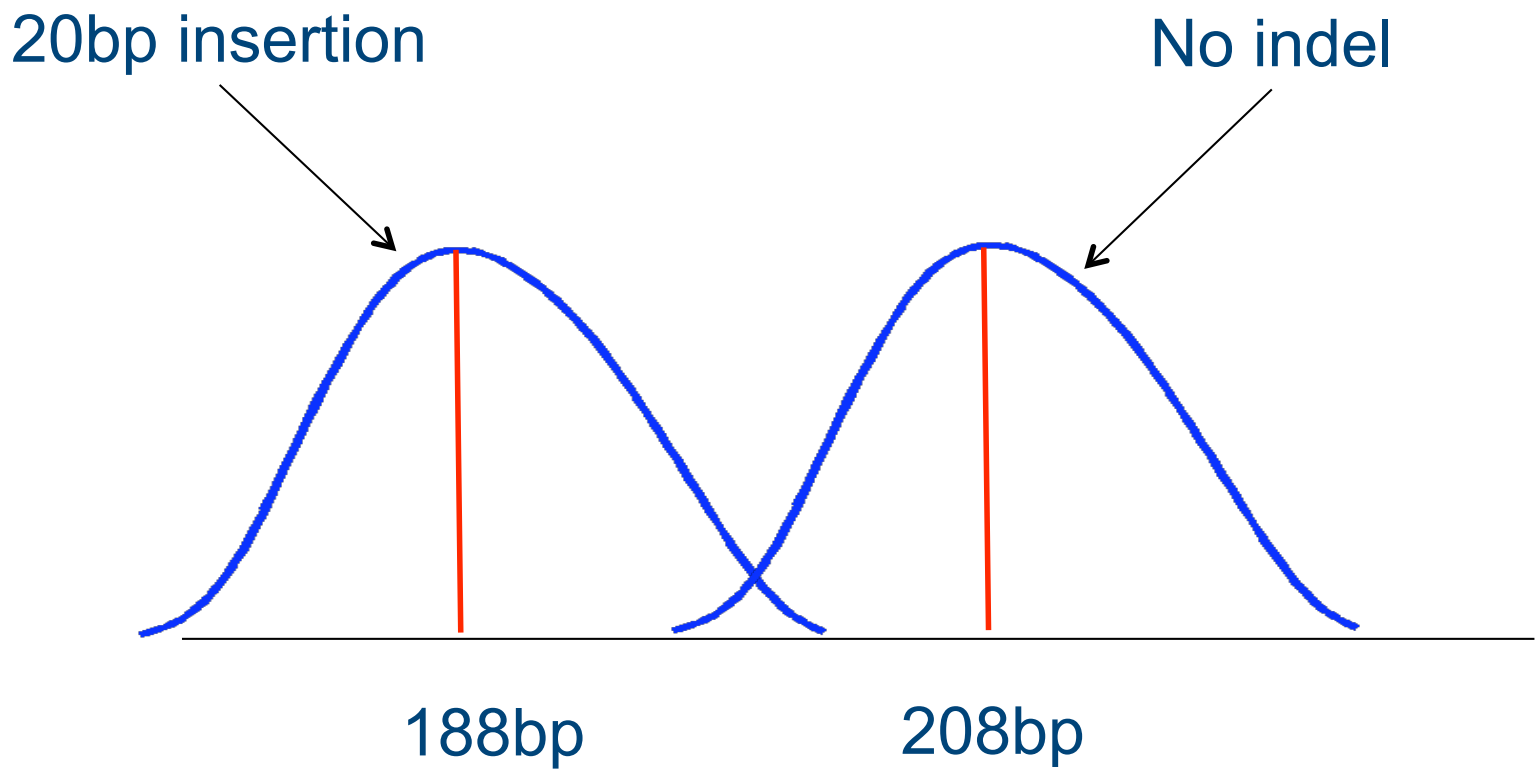
Heterozygous insertion





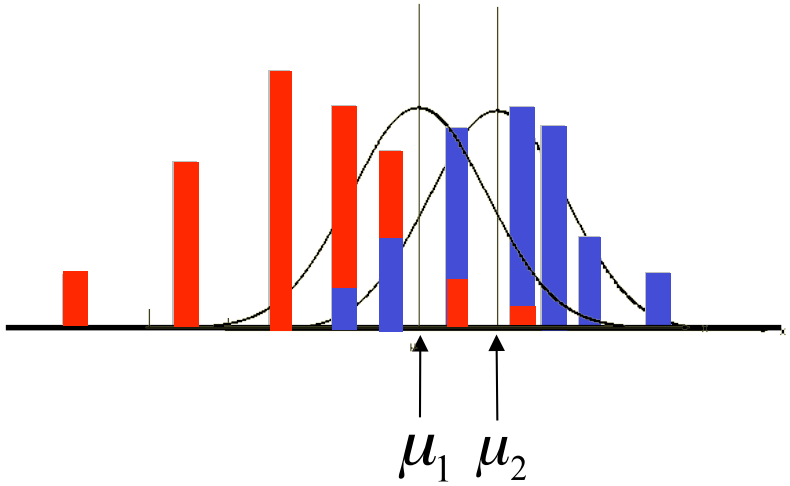
Diploid Case – Distributions

You expect to see matepairs from two distributions.



MoDIL EM Algorithm

1. Randomly initialize μ_1 and μ_2



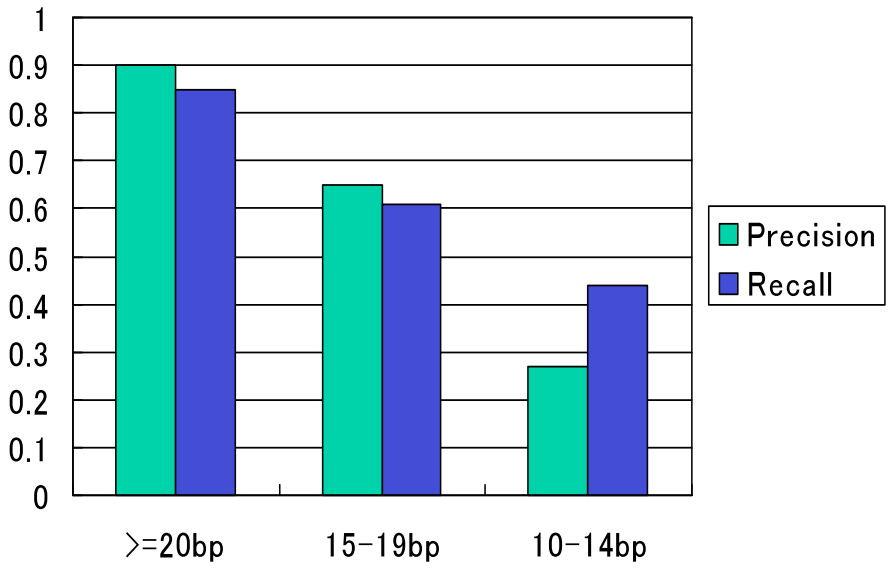
2. E step: Assign each matepair, M_i, to one of two distributions

Assign M_i to p_1 *with* probability $\frac{p_1(M_i)}{p_1(M_i)+p_2(M_i)}$, p_2 with $1-\frac{p_1(M_i)}{p_1(M_i)+p_2(M_i)}$

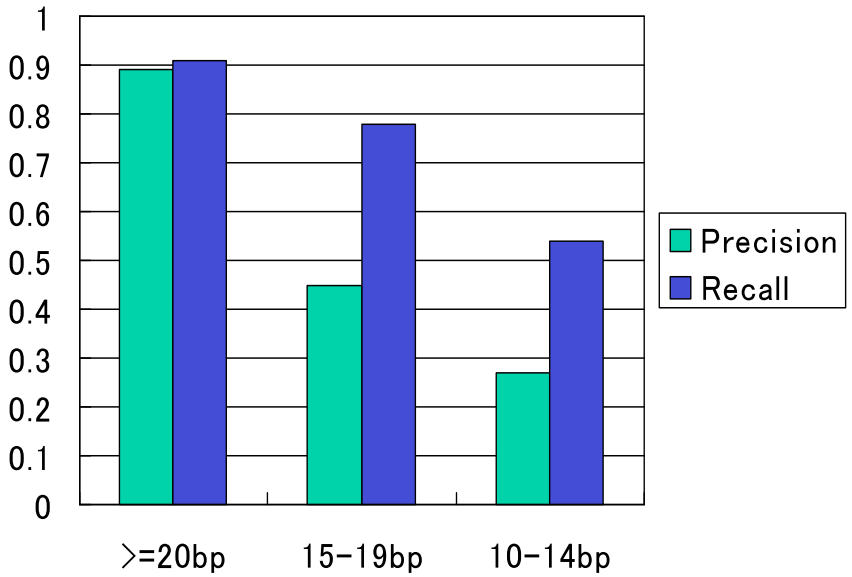
3. M step: Update μ_1 and μ_2 by searching the optimal μ_1 and μ_2 which minimizes Kolmogorov–Smirnov statistic $D = \sum_{t=1}^{2} l_t \sup\left|F_t^o(z) - F_t(z)\right|$

Simulation Results

- Implanted all indels from Mills et al. into chromosome 1 and generated ~51 million matepairs
- Run MoDIL on simulated data and compute precision & recall.



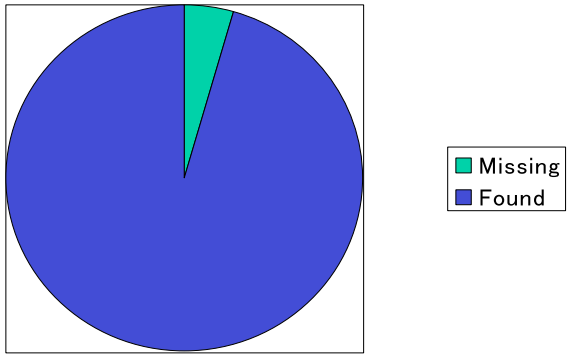
Insertion



Deletion

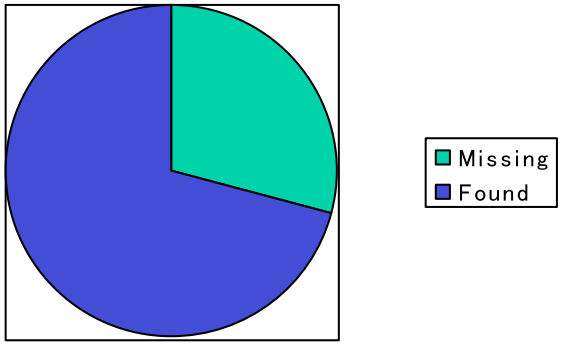
Analysis of NA18507

- NA 18507 (40x Illumina coverage, 208±13bp pairs)
- Kidd et al. found small fraction of INDELs using Sanger style reads (0.3x coverage)
- Computed False Negative Rate (FNR) by taking into account Kidd et al. indels covered by >=20 matepairs



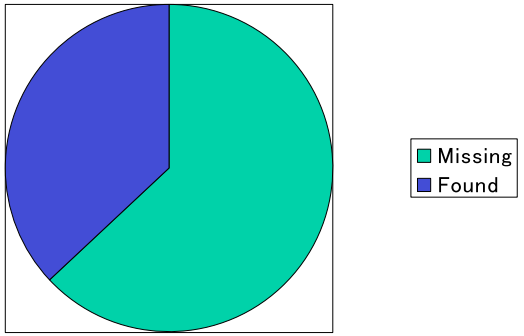
>=20bp INDELs

FNR=0.05



15-19bp INDELs

FNR=0.3

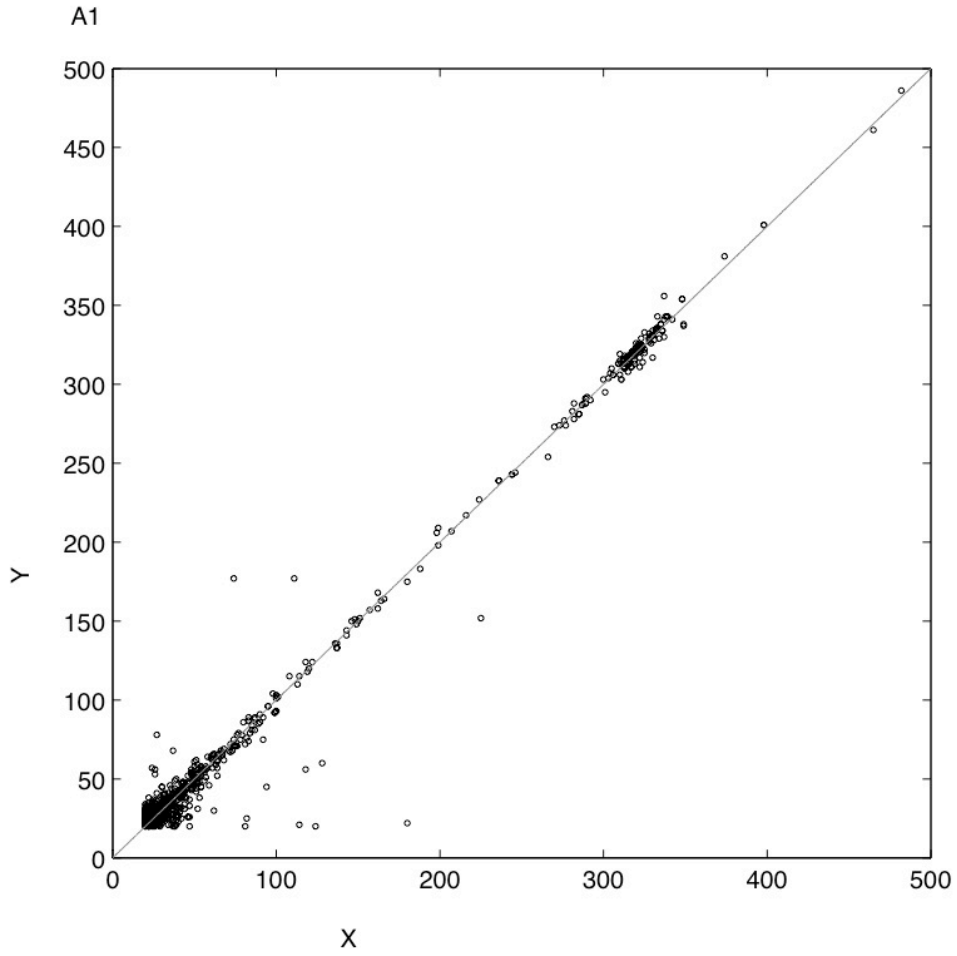


10-14bp INDELs

FNR=0.65

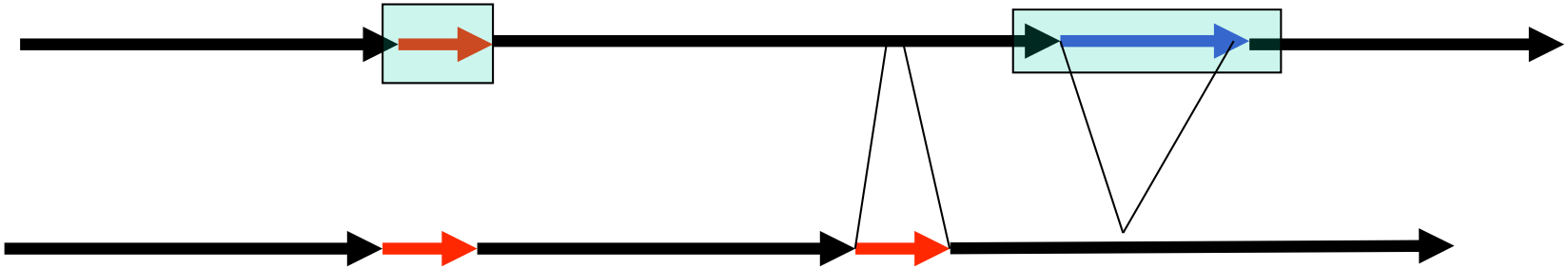
Analysis of NA18507

- NA 18507 (40x Illumina coverage, 208±13bp pairs)

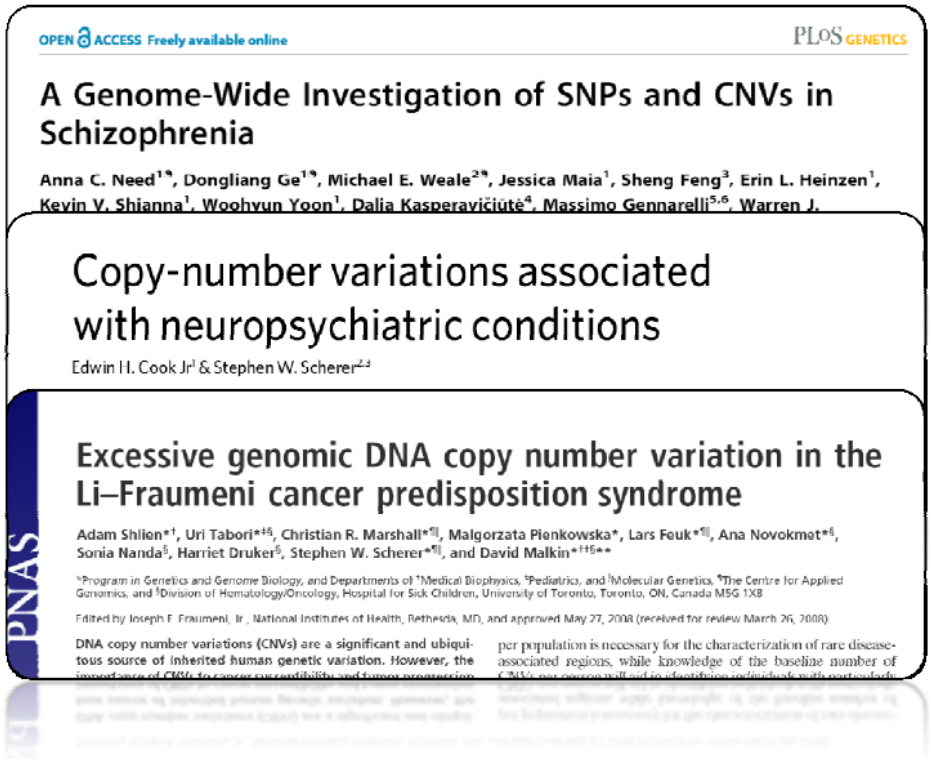


Copy Number Variants (CNVs)

- Large regions that appear a different number of times within different indiv.



- CNVs are associated with a number of diseases
- Input
  - reference human genome
  - sequenced donor genome
- Output
  - CNV annotations in ref



OPEN ACCESS Freely available online

PLoS GENETICS

A Genome-Wide Investigation of SNPs and CNVs in Schizophrenia

Anna C. Need[1¶], Dongliang Ge[1¶], Michael E. Weale[2¶], Jessica Maia[1], Sheng Feng[3], Erin L. Heinzen[1], Kevin V. Shianna[1], Woohyun Yoon[1], Dalia Kasperaviciute[4], Massimo Gennarelli[5,6], Warren J.

Copy-number variations associated with neuropsychiatric conditions

Edwin H. Cook Jr[1] & Stephen W. Scherer[2,3]

PNAS

Excessive genomic DNA copy number variation in the Li–Fraumeni cancer predisposition syndrome

Adam Shlien*†, Uri Tabori*‡§, Christian R. Marshall*¶‖, Malgorzata Pienkowska*, Lars Feuk*¶‖, Ana Novokmet*§, Sonia Nanda§, Harriet Druker§, Stephen W. Scherer*¶‖, and David Malkin*†‡§**

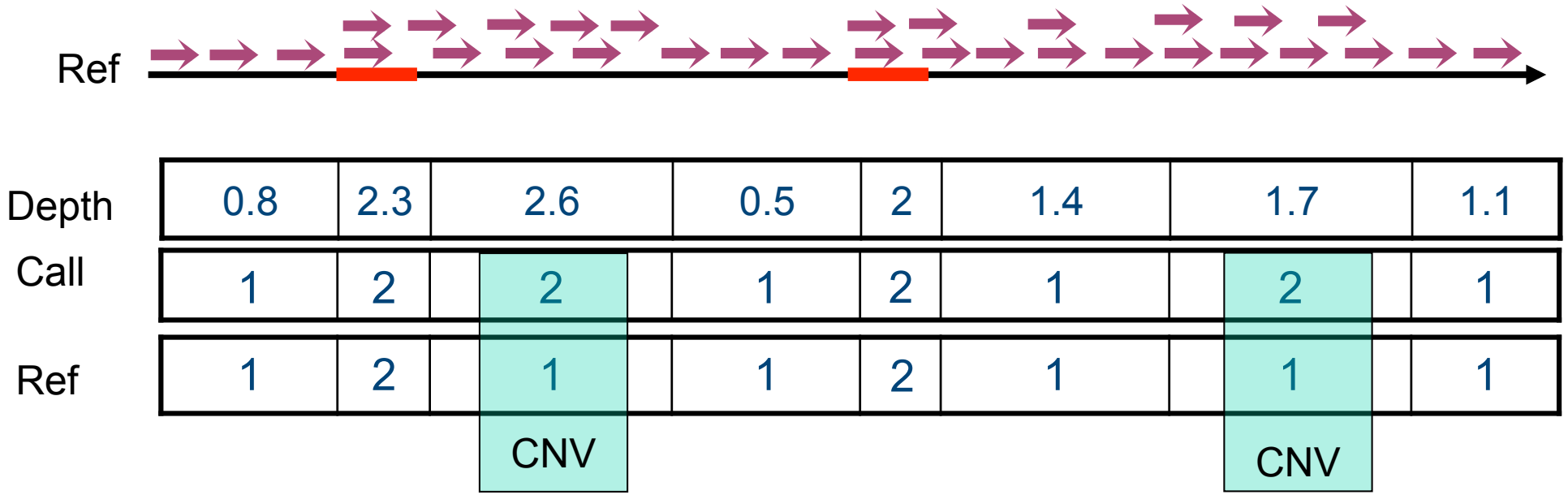
*Program in Genetics and Genome Biology, and Departments of †Medical Biophysics, ‡Pediatrics, and ‖Molecular Genetics, ¶The Centre for Applied Genomics, and §Division of Hematology/Oncology, Hospital for Sick Children, University of Toronto, Toronto, ON, Canada M5G 1X8

Edited by Joseph F. Fraumeni, Jr., National Institutes of Health, Bethesda, MD, and approved May 27, 2008 (received for review March 26, 2008)

DNA copy number variations (CNVs) are a significant and ubiquitous source of inherited human genetic variation. However, the

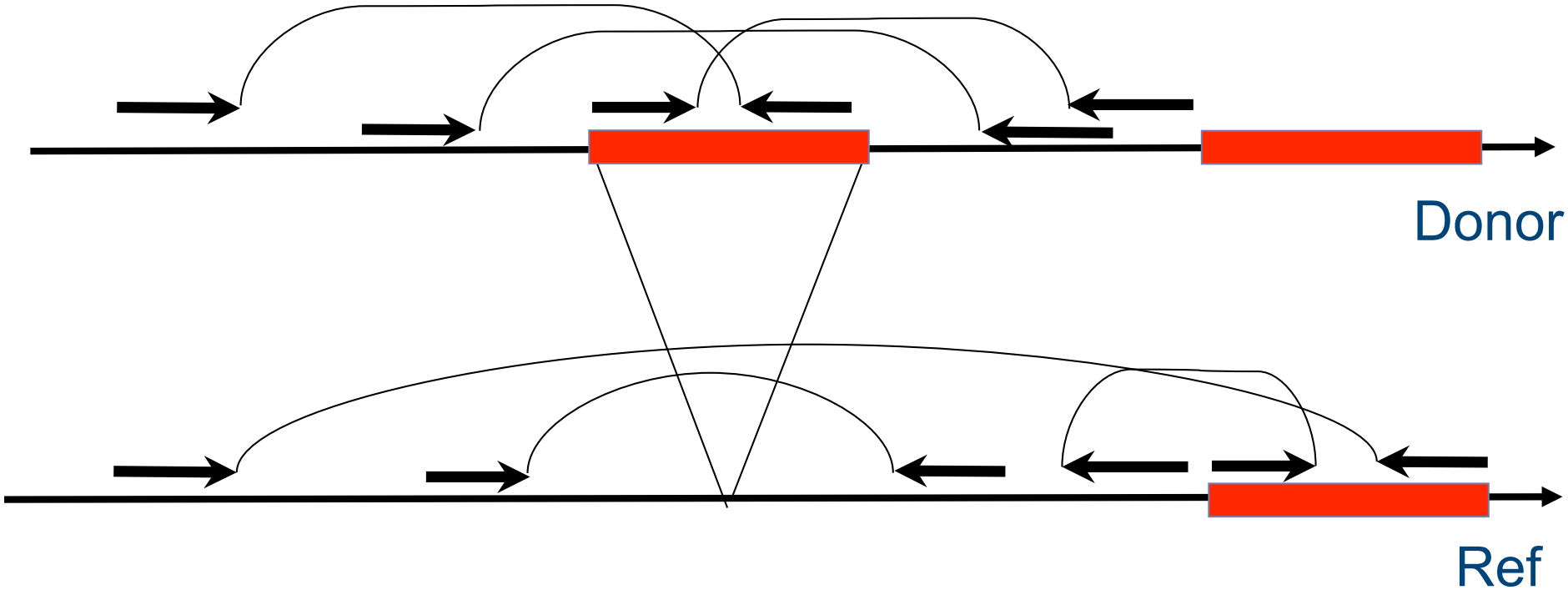
per population is necessary for the characterization of rare disease-associated regions, while knowledge of the baseline number of

Calling CNVs

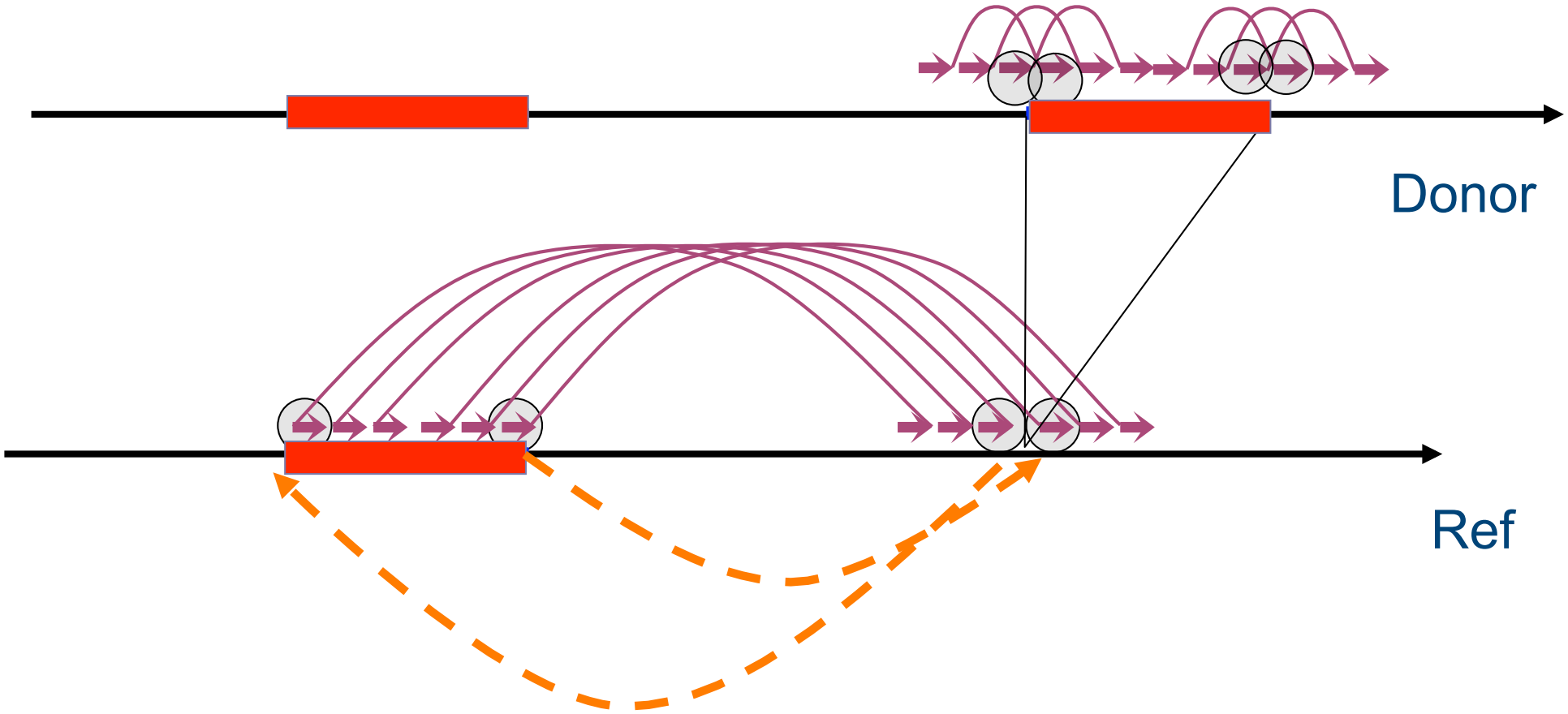


Back to… Structural Variants

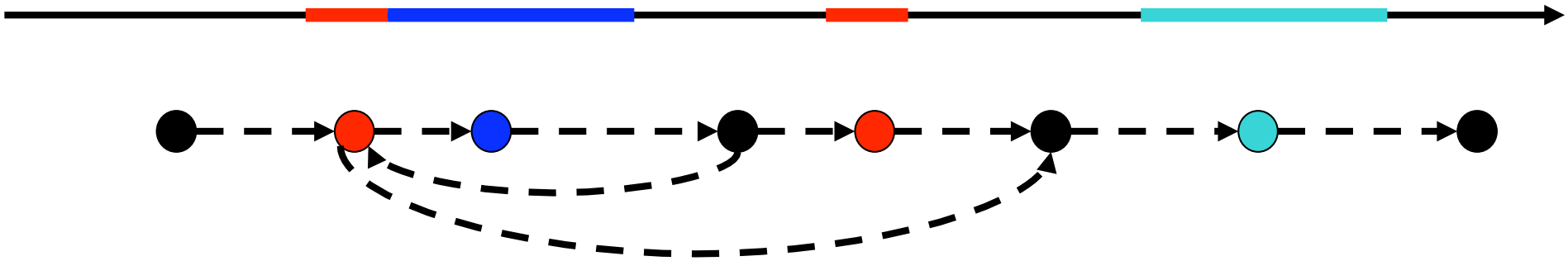
- What if the inserted segment is present elsewhere?



The Linking Signature

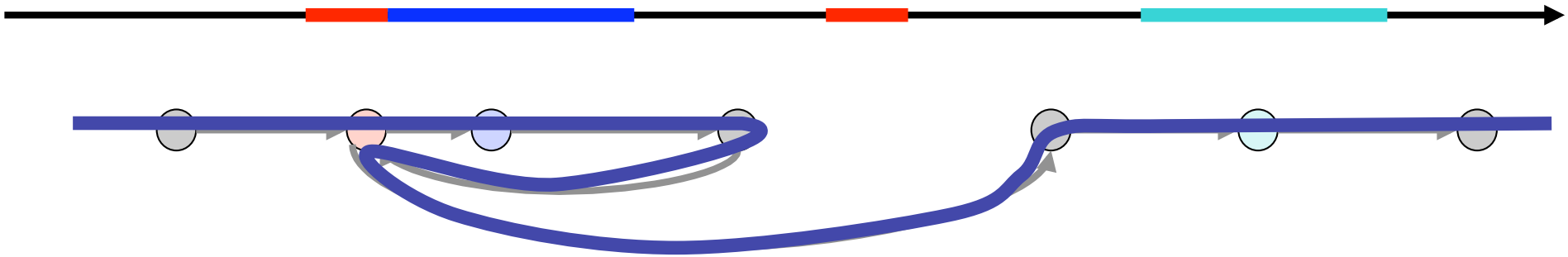


Step 1 – Build Repeat Graph



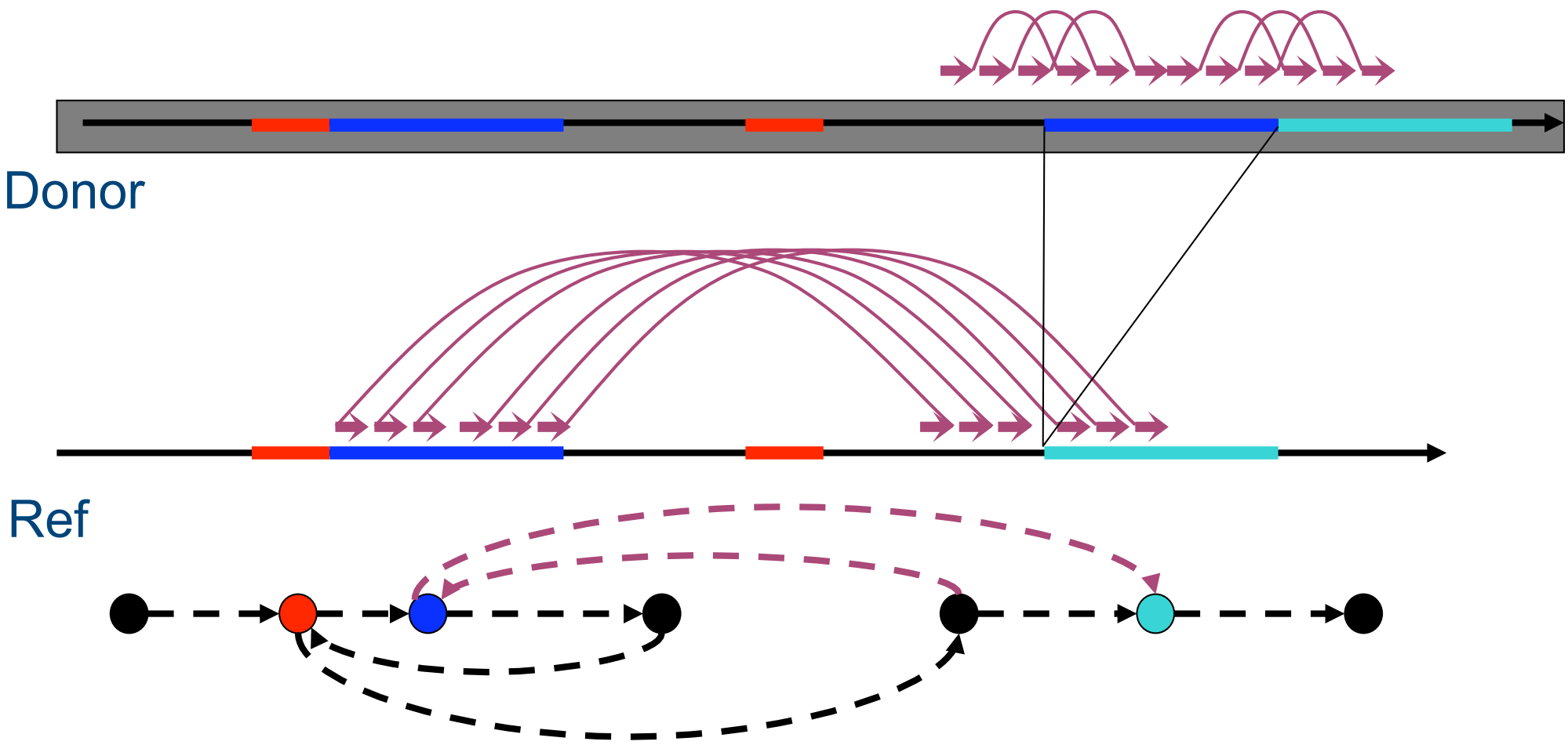
Pevzner, Tang, Tesler (2004)

Step 1 – Build Repeat Graph

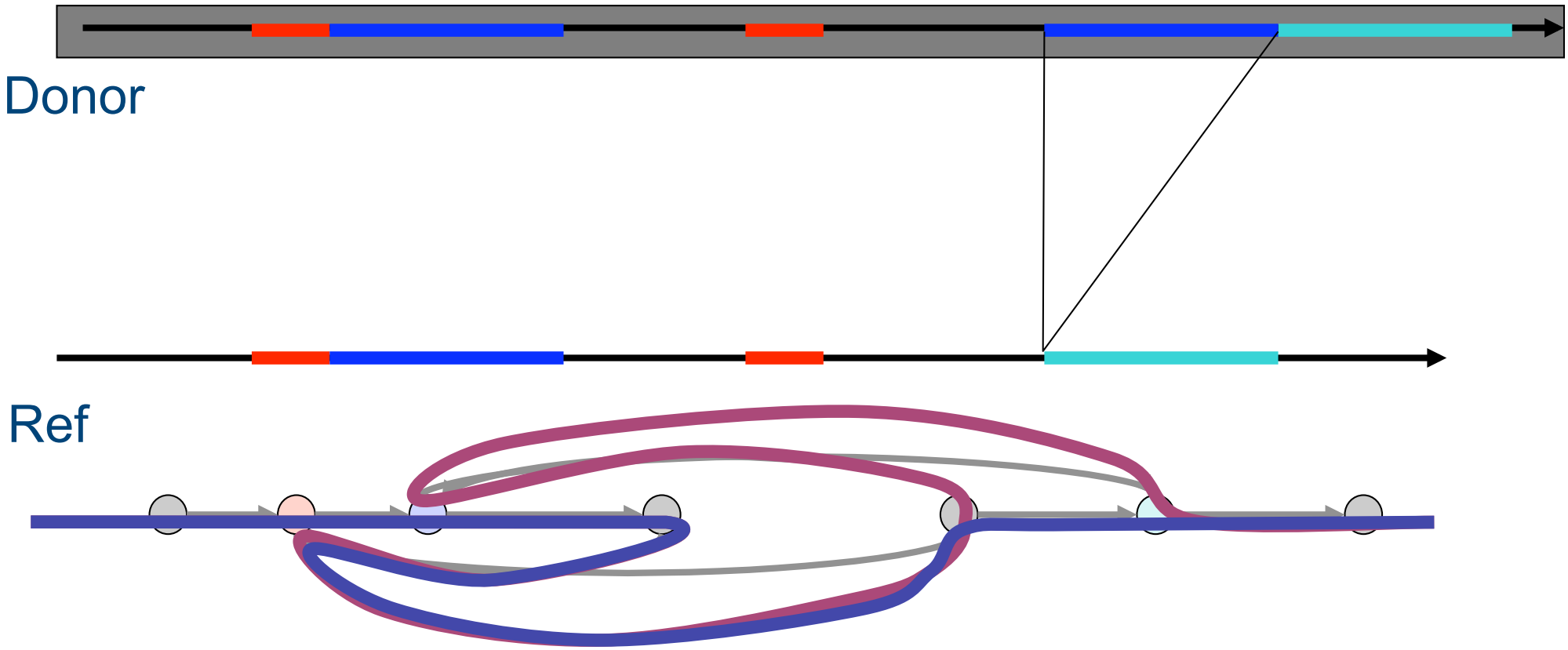


Pevzner, Tang, Tesler (2004)

Step 2 – Capture Donor Adjacencies



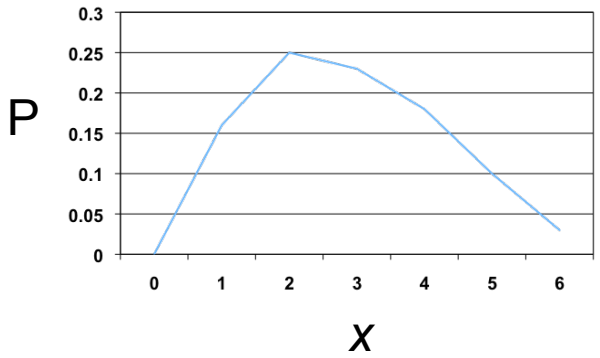
Step 2 – Capture Donor Adjacencies

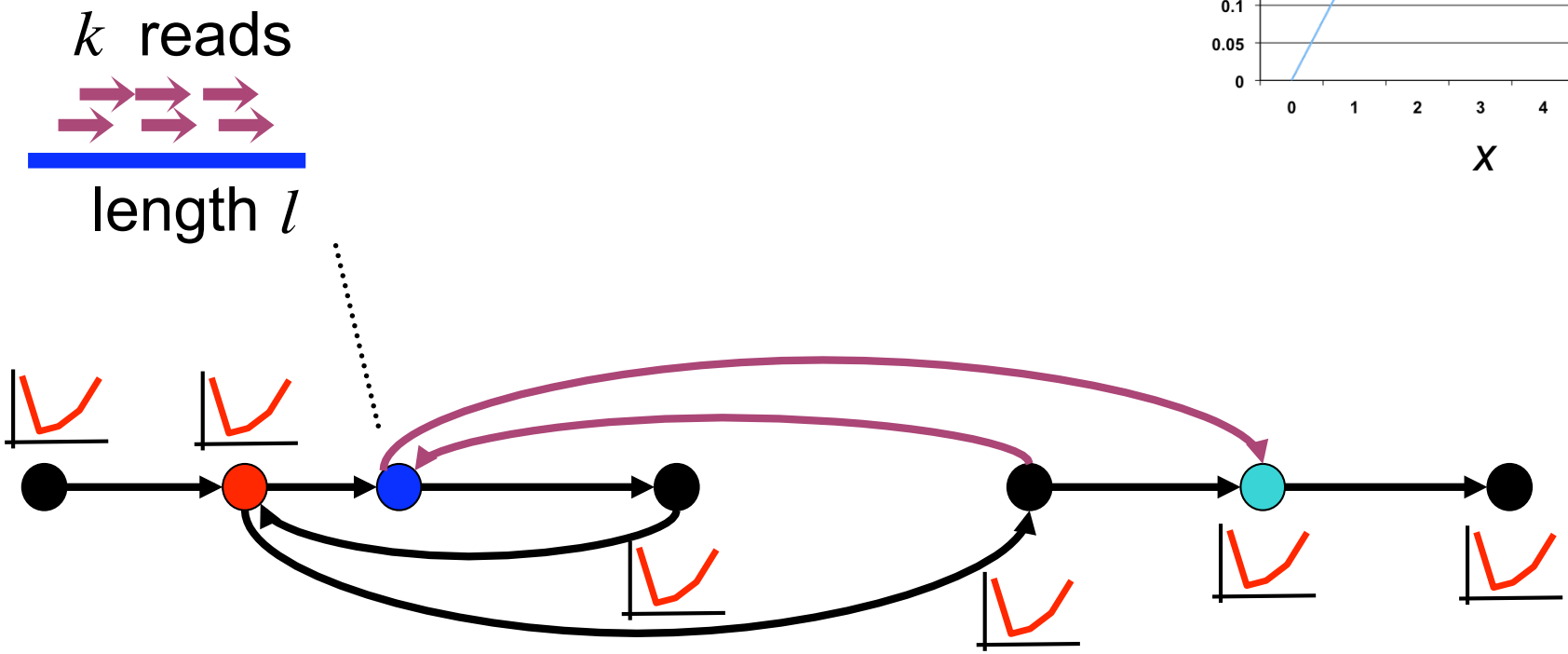


Step 3– Defining Walk Costs

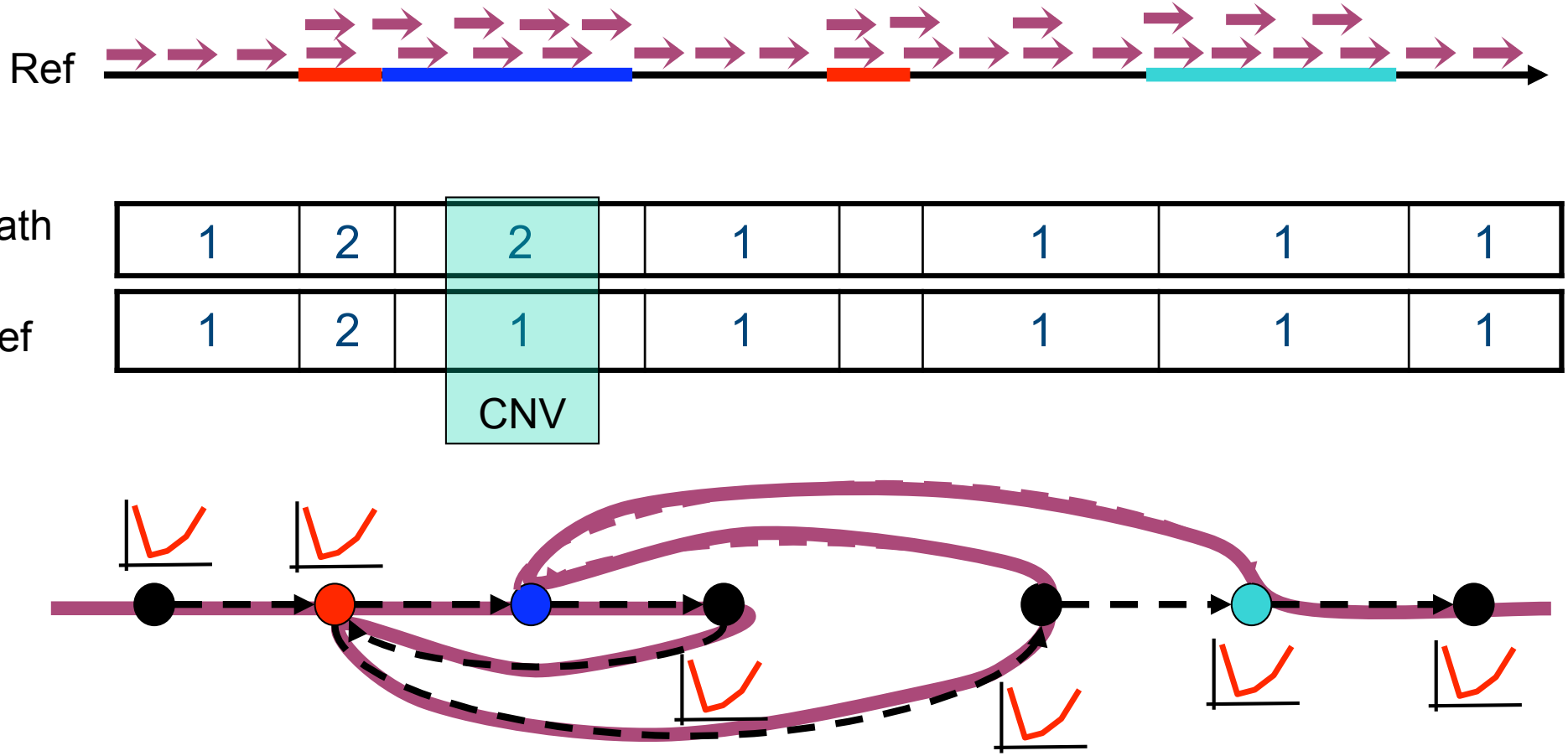
Each function represents the probability that the segment of length *l* appears *x* times in the donor given that there are *k* reads mapped to that segment

$$P(k \mid xl) = \frac{(x\lambda)^k e^{-(x\lambda)}}{k!} \quad \text{where} \quad \lambda = N * l / G$$





Calling CNVs



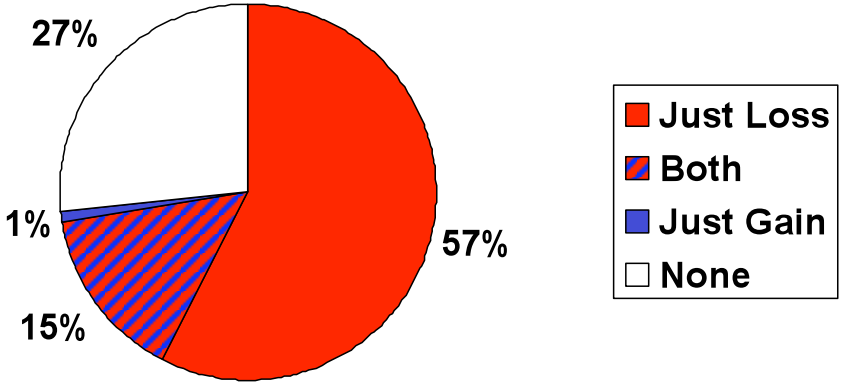
Finds the path “most faithful” to the DOC (Network Flow)

- Probabilistic model to score “faithfulness”

Preliminary Results

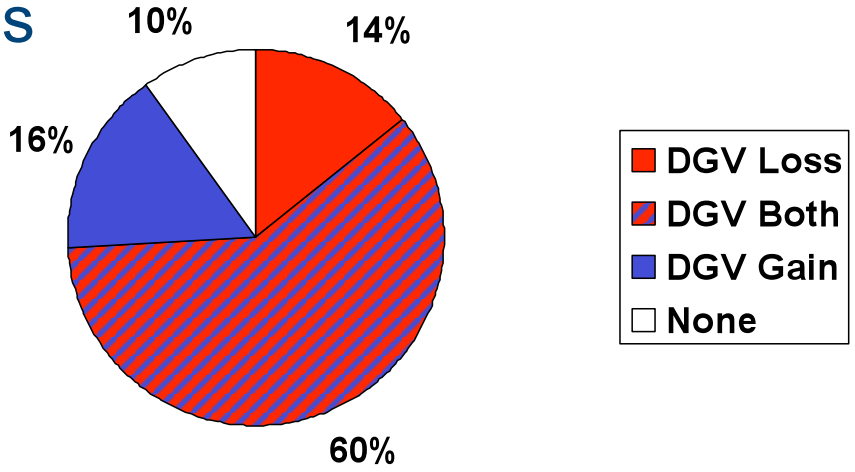
- Total 9909 CNV calls (>1k; 2.5%) – 5795 losses, 4114 gains

Kidd et al's variants detected (out of 146; Sensitivity)

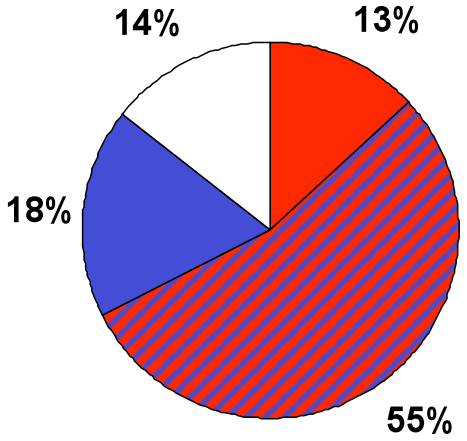


DGV Overlap (Specificity)

Gains



Losses



Take-home points

- MoDIL
  - Take advantage of high clone coverage to find smaller INDELs with high accuracy
  - ~90% accuracy and recall for INDELs ≥ 20bp.
- CNVs
  - Combine pair-end and arrival information to find CNVs
  - Good Concordance with previous results
- Matepairs are key
  - Length & distribution of insert sizes key
  - Read length (sometimes) less so

Acknowledgments

Seunghak Lee
Can Alkan (UW)
Fereydoun Hormozdiari (SFU)

Paul Medvedev
Marc Fiume
Misko Dzamba
Tim Smith

http://compbio.cs.toronto.edu/modil

brudno@cs.toronto.edu

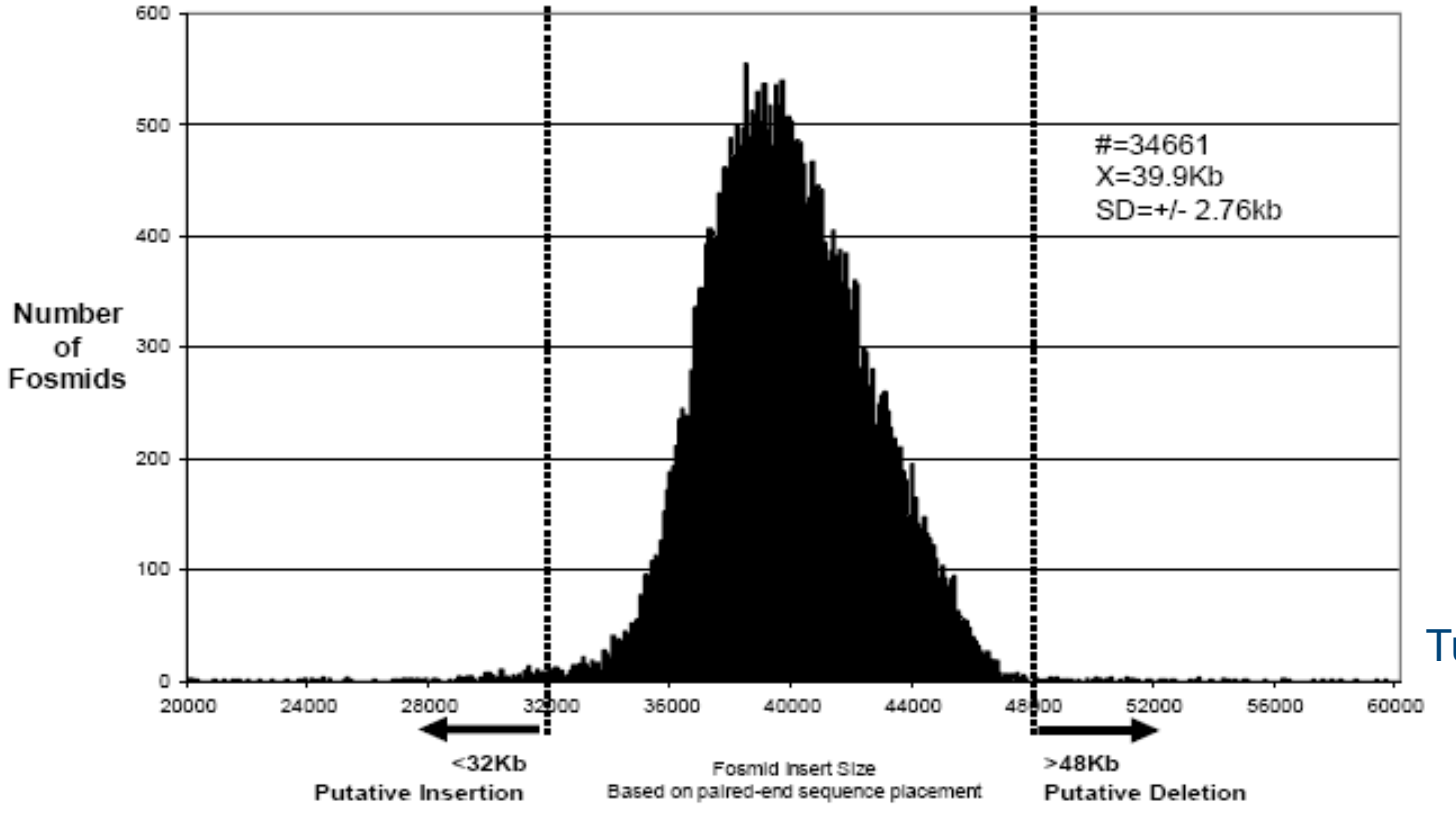
Outline

1. MoDIL: Detecting INDELs with Mixtures of Distributions
   - Haploid case - Detecting INDELs with a distribution
   - Diploid case – Detecting INDELs with mixtures of distributions
   - Results

2. Finding CNVs with Matepairs and Depth-of-coverage

Difficulties in Small INDEL Detection

- In reality, insert sizes of matepairs are not perfect
  - Unable to detect small indels (e.g. < 3STD)



Tuzun et al. 2005

Outline: MoDIL

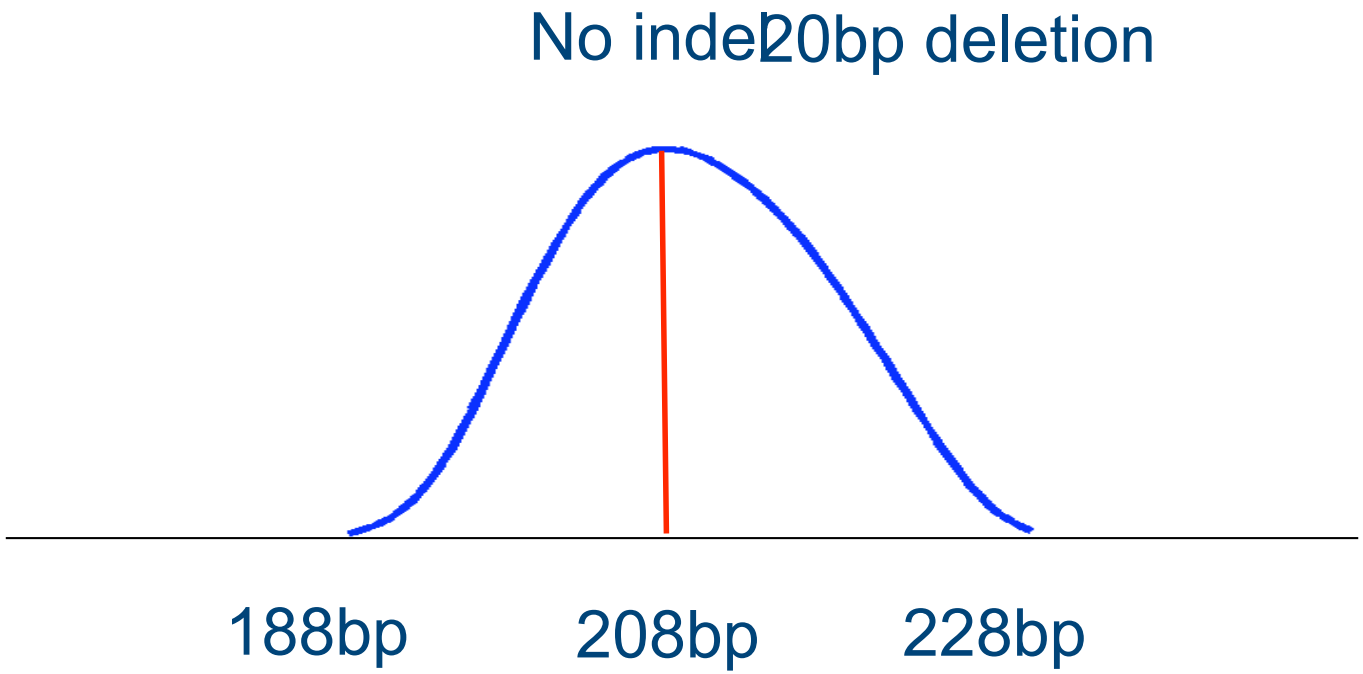
1. Haploid case - Detecting INDELs with a distribution

2. Diploid case – Detecting INDELs with mixtures of distributions

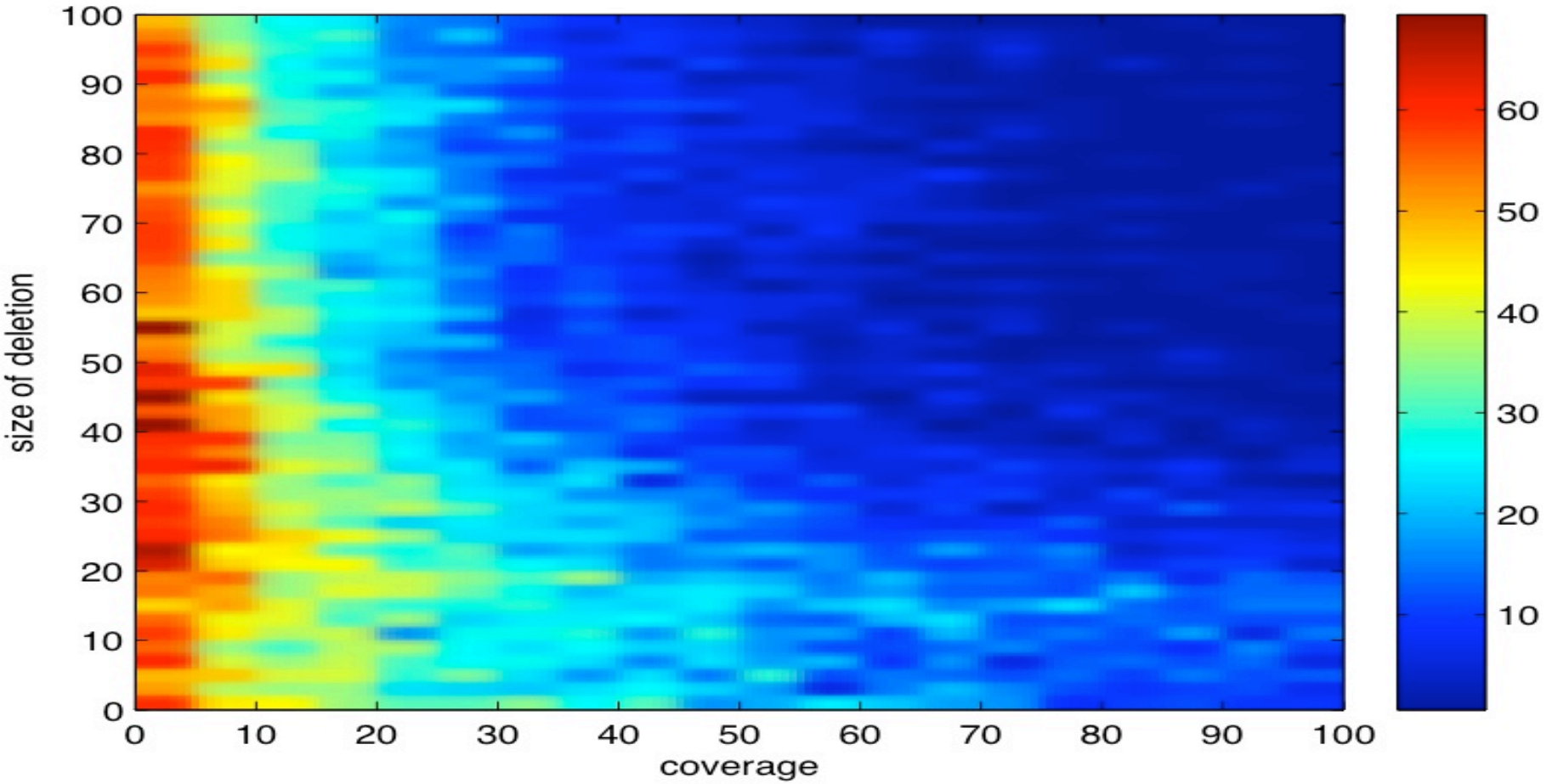
3. Results

Haploid Case – Distribution

Make a distribution of mapped distances in each cluster
=> The distribution shifts from distribution of insert size
if there is an INDEL



EM Algorithm Sensitivity



Percent error (> 5bp off)

Probabilistic Framework

- S_{ACAT} = 2
- S_{GGCA} = 1

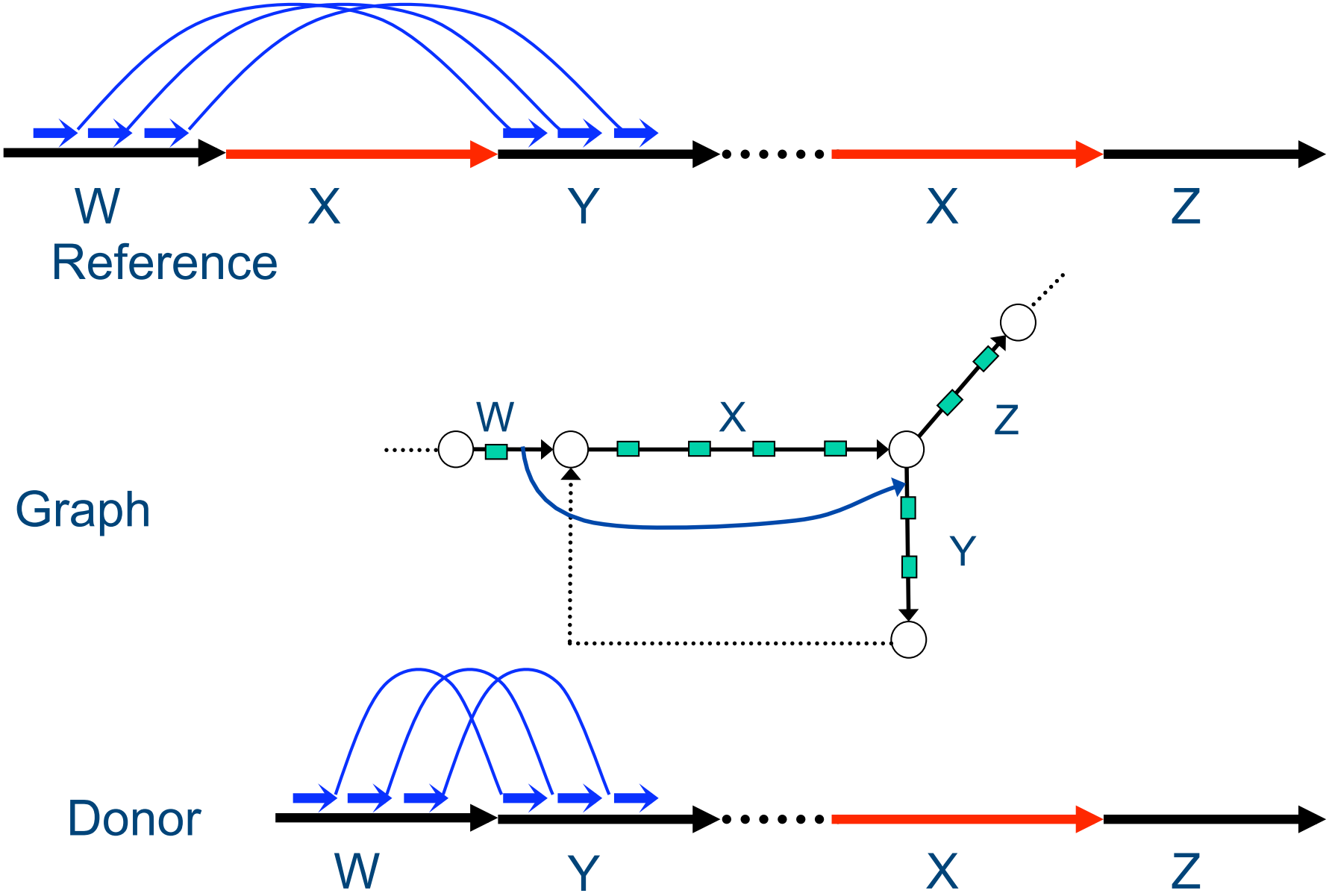
- Example:
  - 50x coverage
  - 25-long reads

- Every 25-long window of the genome is sampled 2 times, on average.

- Let s_i=4
- $g_i = s_i / 2$, so $g_i \approx 2$

ACAT	ACAT	GGCA

Discordant Matepairs



MoDIL: Detecting INDEL Variation with Clone-end Sequencing

Seunghak Lee,Fereydoun Hormozdiari,
Can Alkan, Michael Brudno

This paper is in pressat Nature Methods

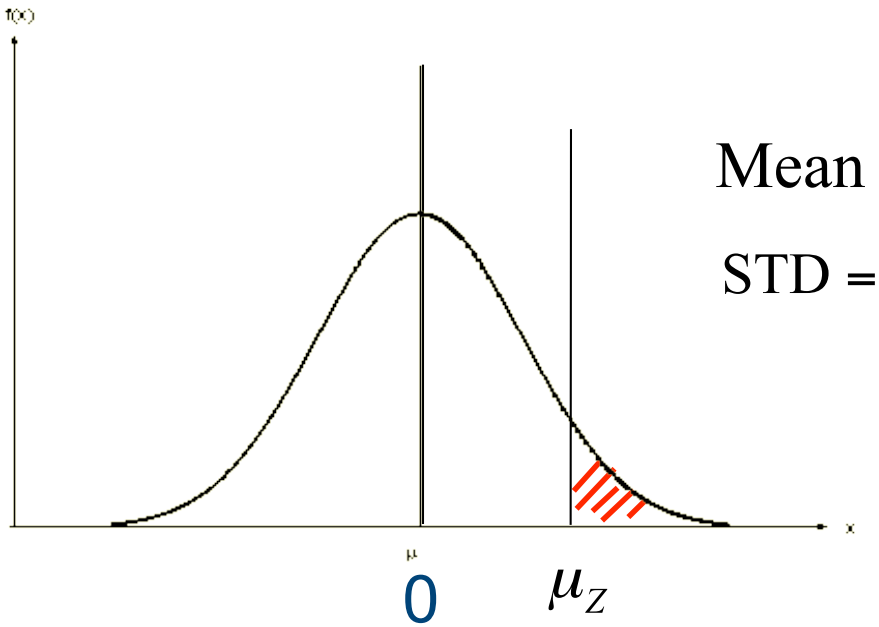
http://compbio.cs.toronto.edu/modil/

P-value (assigning a confidence)

P-value

Probability that a cluster is generated from a region without an indel

$$\text{P - value} = \sum_{\mu_Z}^{\infty} p(Z' = z \mid 0)$$



P-heterozygosity

P-heterozygosity

posterior probability that an indel is hetozygous

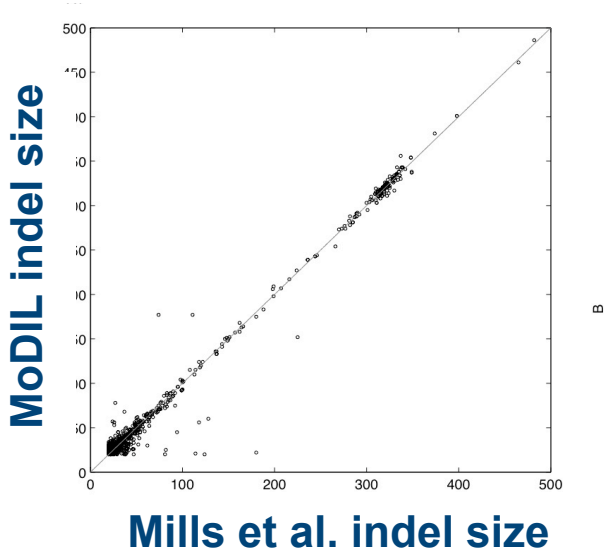
$$\mathrm{P-het} = P(hetero \mid X_1, \ldots X_N) = \frac{P(X_1, \ldots X_N \mid hetero)}{P(X_1, \ldots X_N \mid herero) + P(X_1, \ldots X_N \mid \mathrm{hom})}$$

$$P(X_1, \ldots X_N \mid hetero) = \prod_{i=1}^{N} P(X_i \mid hetero)$$

$$= \prod_{i=1}^{N} \left\{ 0.5P(X_i \mid \mu) + 0.5P(X_i \mid 0) \right\}$$

Accuracy of Size Estimation of MoDIL

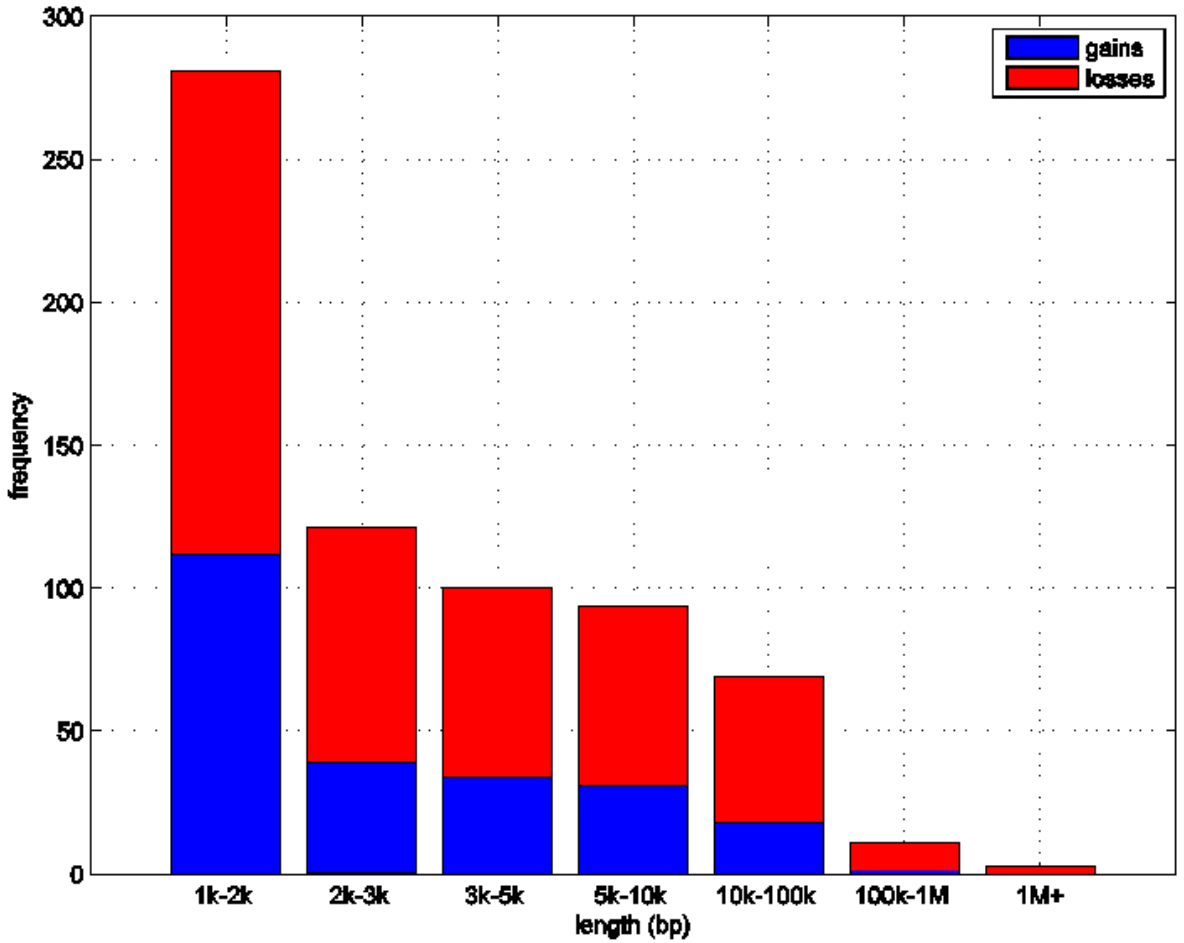
- Large # of indels (~32%) overlapped with Mills et al. results (>=20bp)
- Compared sizes of Mills et al. and MoDIL
  - Pearson's correlation coefficient, r^2=0.96
  - (Mills et al. minus MoDIL) overlaps with Gaussian with STD=4 (expected STD for a cluster with 20 matepairs)



Preliminary Results

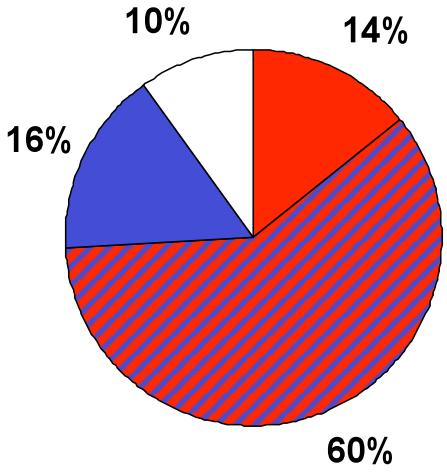
- NA18507 individual sampled with Illumina

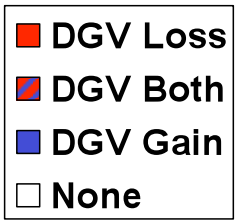
- Total of 9909 CNV calls
- 5795 losses, 4114 gains



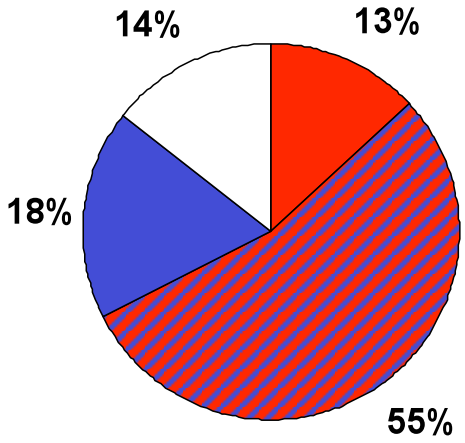
Preliminary Results (Specificity)

Percent of our GAIN calls that overlap with DGV:

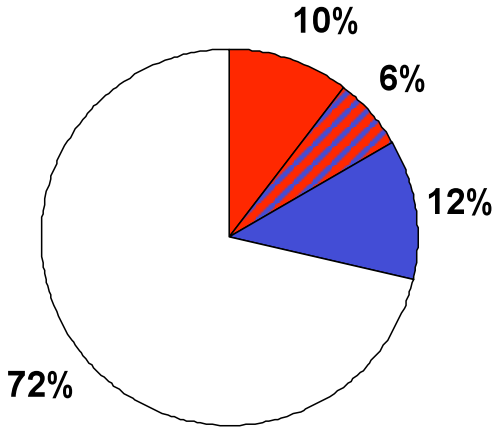


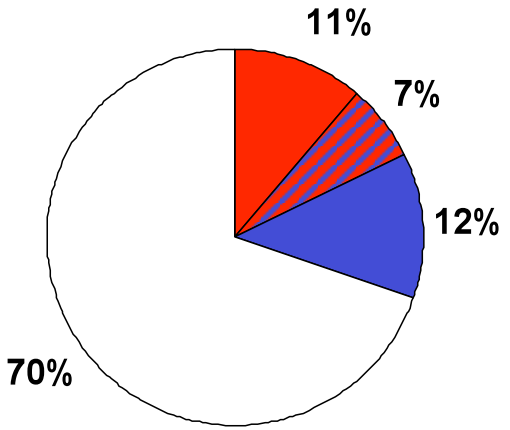


Percent of our LOSS calls that overlap with DGV:



After shuffle:





Diversity of Humans

- Humans are diverse
  - Genomic Variation

- Single Nucleotide Polymorphisms
  - SNPs occur ~1/1000 positions
  - Find by comparing reads from one individual to the reference human genome

```
G: 798   GAACCCCTTACAACTGAACCCCTTAC
         |||||||||| |||||||||||||||
R:       GAACCCCTTATAACTGAACCCCTTAC
```

- Structural variations are large scale genomic alterations
  - Insertions, deletions, inversions, translocations and changes in copy numbers