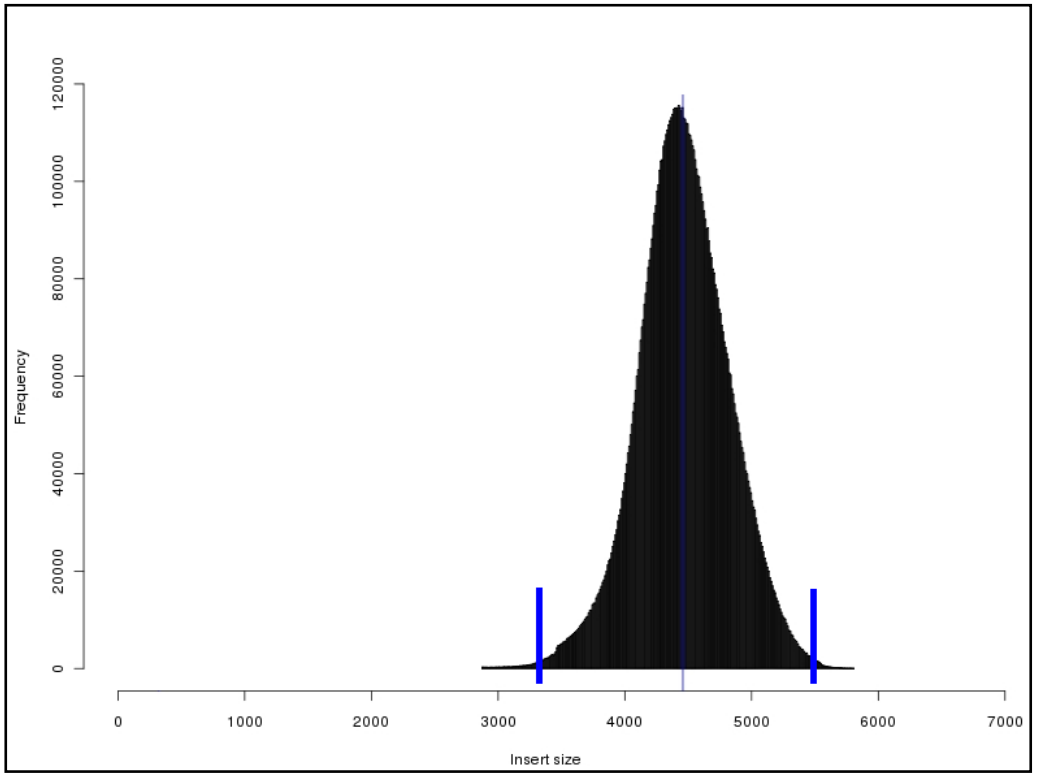
Structural Variations

02-715 Advanced Topics in Computational Genomics

Challenges

- Insert sizes of each mate-pair are unknown and can vary
- Only the approximate distribution of insert sizes is available



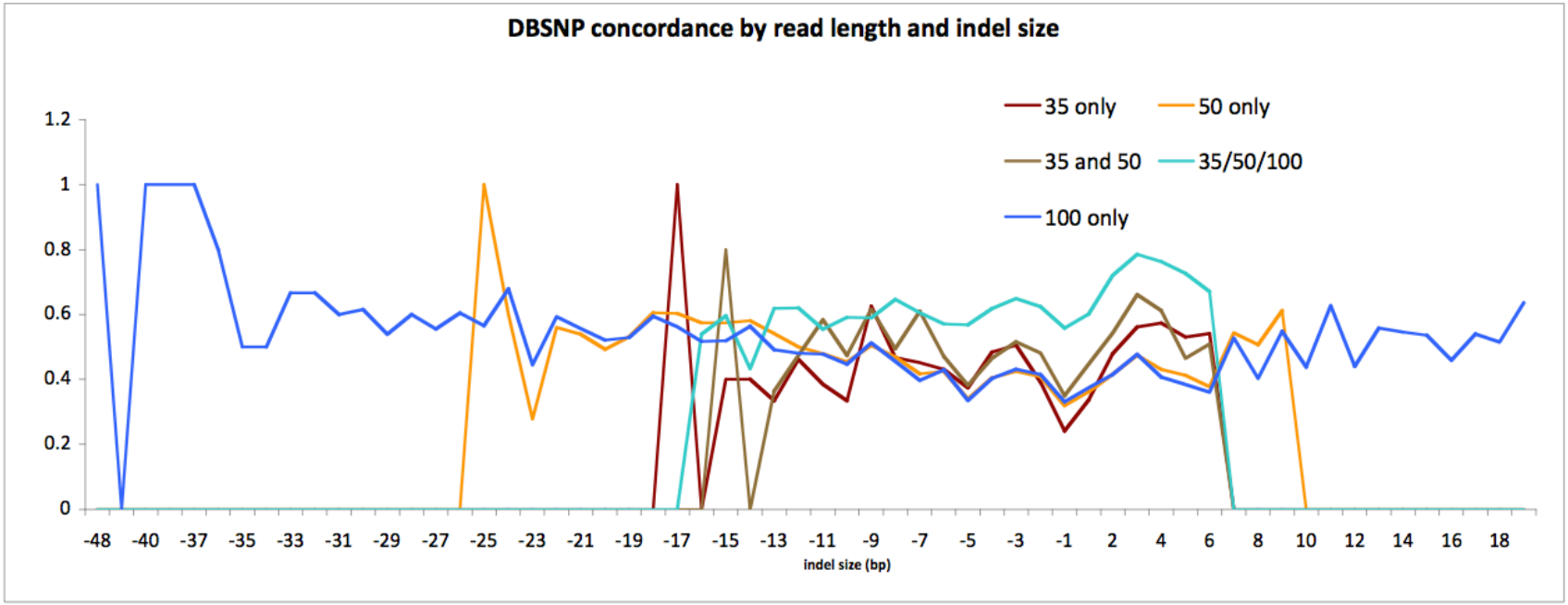
Indels (Insertions/Deletions)

- Small-size indels with size<10bp: readily identifiable with Smith-Waterman algorithm
- Large-size indels with size>50bp: relatively easy to identify
- Medium-size indels?
  - Difficult to distinguish between true indels and insert-size variations
  - Needs methods for better resolution

Paired-End Smith-Waterman Alignment Algorithm

- Detects short indels
  - First, align the reads without gaps
  - For those read pairs, where only one read is aligned and the other is not, apply gapped alignment for the unaligned read

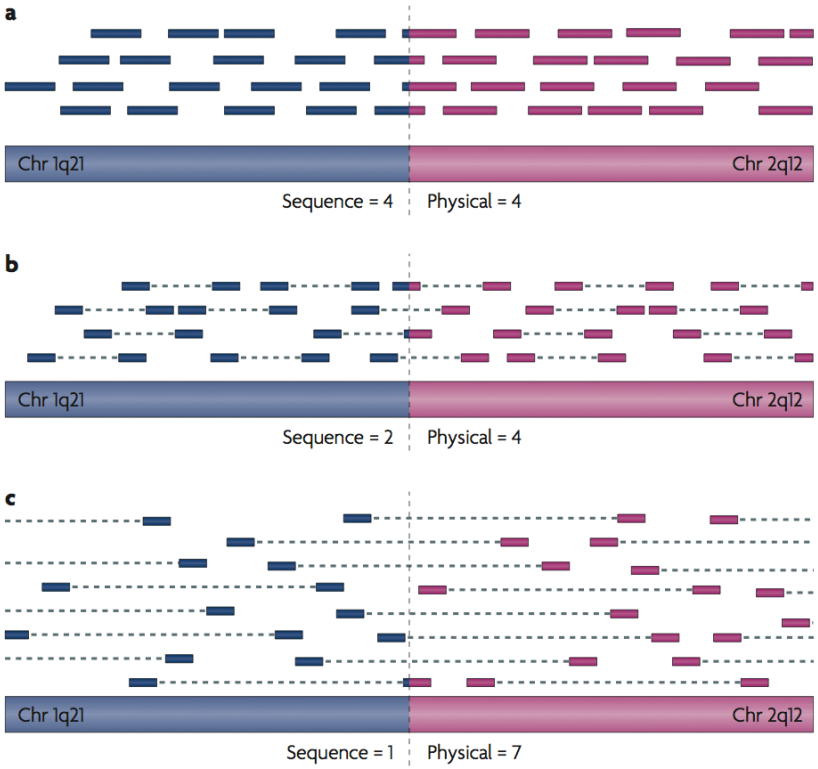
Indels Detected by Paired-End Smith-Waterman Alignment Algorithm



Other Factors

- Accuracies can depend on
  - Coverage
  - Insert sizes
  - Read length
  - Read alignment accuracy

Depth of Coverage and Physical Coverage



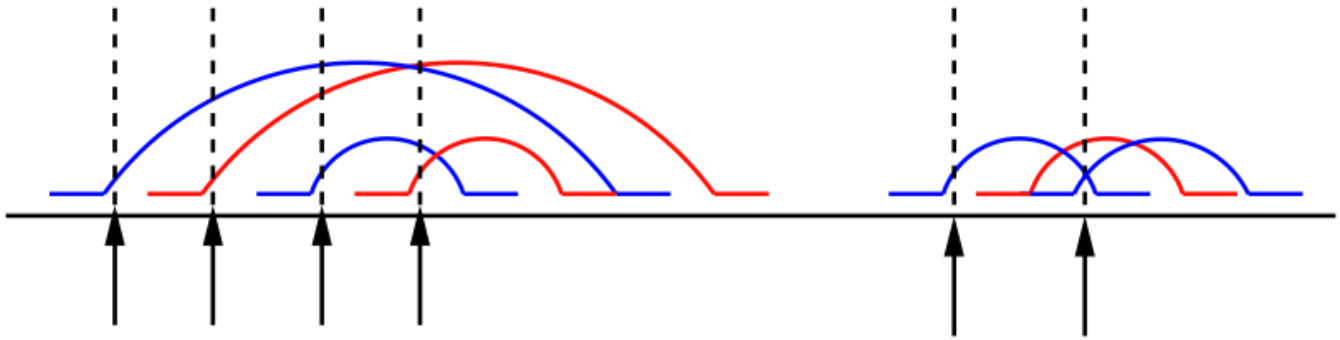
- Single-end sequencing
- Paired-end sequencing
- Paired-end sequencing

Statistical Methods for Detecting Structural Variants

- MODIL: small indels
- MOGUL: small indels, low coverage, many individuals
- BreakDancer
  - BreakDancerMax: detects different types of structural variations
  - BreakDancerMini: small indels
- All methods are based on mixture modeling

Mate-pair Clusters

- Many methods are applied to mate-pair clusters after mapping the mate-pairs to reference genome
- Clustering of mate pairs



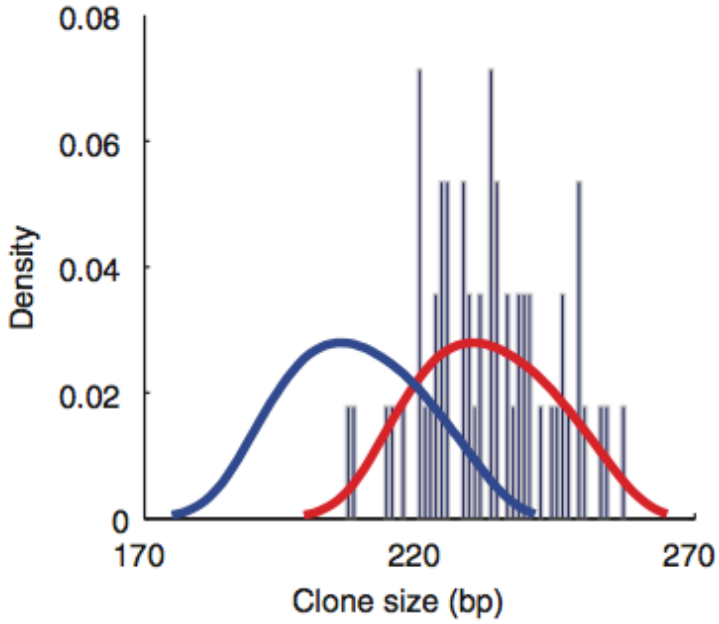
MODIL

- Mixture of distributions indel locators

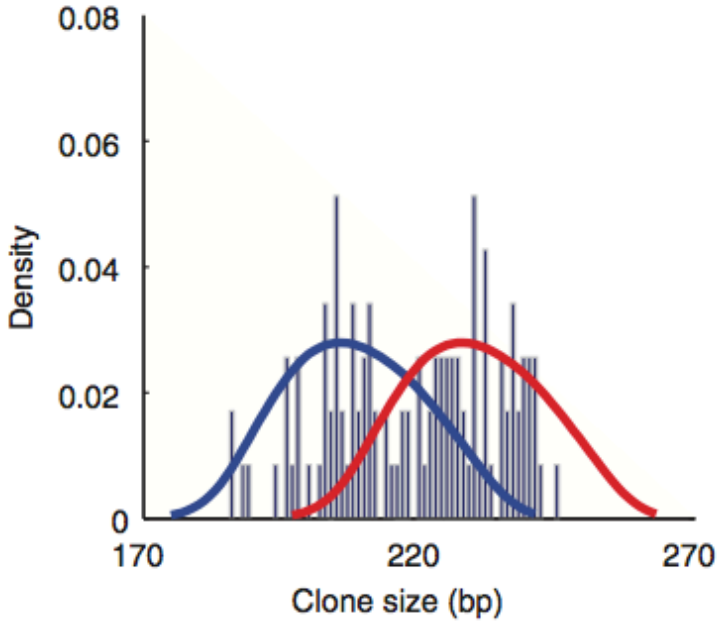
- Model insert-size distributions at each locus *i*
  - Mixture component 1 $P(Y)$: mapped distances for no indels
  - Mixture component 2 $P(C_i)$: mapped distances for indels

MODIL

- Blue: insert-size distribution with no deletions
- Red: insert-size distribution with deletions



Homozygous deletions



Heterozygous deletions

MODIL Algorithm

- Map the mate-pairs to reference genome
- Estimate $P(Y)$ from mapped distances across the whole genome
- For each locus i, estimate $P(C_i)$ from the mate-pairs that span the locus
  - Location-shifted distribution of $P(Y)$
  - Expected size of indels D_1, D_2 for each of two haplotypes
  - EM algorithm
- Expected indel size:

$$\mu = \mu_{p(Y)} - \mu_{D_k} \text{ and } \sigma = \sigma_{p(Y)} / \sqrt{n}$$

Applying MODIL to Illumina Dataset

- 40-fold read coverage
- Observed insert size: mean 208, standard deviation 13
- To determine whether there is an insertion/deletion at each locus, find a cluster of mate pairs spanning that locus. Each cluster is required to have 20 mate pairs

MODIL: Performance

			Overlap with known indels[7]			Simulation	
Size	Type	MoDIL	Total	Found	FNR	Recall	Precision
≥20 bp	Insertion	1,336	78	75	0.04	0.85	0.90
	Deletion	3,799	196	187	0.05	0.91	0.89
15–19 bp	Insertion	1,601	119	84	0.29	0.61	0.65
	Deletion	5,333	178	126	0.29	0.78	0.45
10–14 bp	Insertion	936	370	130	0.65	0.44	0.37
	Deletion	3,682	593	227	0.62	0.54	0.27

Number of insertions/deletions discovered by MODIL

Mixture of Genotypes Variant Locator (MOGUL): Motivation

- Higher coverage leads to more accurate results
  - MODiL works for more than 20 mate-pairs covering each locus
  - 1000 genome project: less than 4 mate-pairs covering each locus for each individual
- What if we have many individuals with each having lower coverage
  - How can we combine the statistical strength across multiple individuals?
  - Can we estimate variant locations/sizes and allele frequencies?
  - MOGUL (Mixture of Genotypes Variant Locator) (Lee et al., 2010)
  - Allele frequency/coverage/number of individuals can influence the performance

1000 Genome Project

(The 1000 Genome Project Consortium, Nature 2010)

The **goal** is to characterize over **95% of variants** that are in genomic regions accessible to current high-throughput sequencing technologies and that have **allele frequency of 1% or higher** (the classical definition of polymorphism) in each of **five major population groups** (populations in or with ancestry from Europe, East Asia, South Asia, West Africa and the Americas)

Pilot project:

- 179 individuals from four populations
  (low coverage: 2-6x)
- 6 individuals in two trios
  (deep sequencing: average 42x)
- 697 individuals from seven populations
  (exon sequencing of 8,140 exons: average 50x)

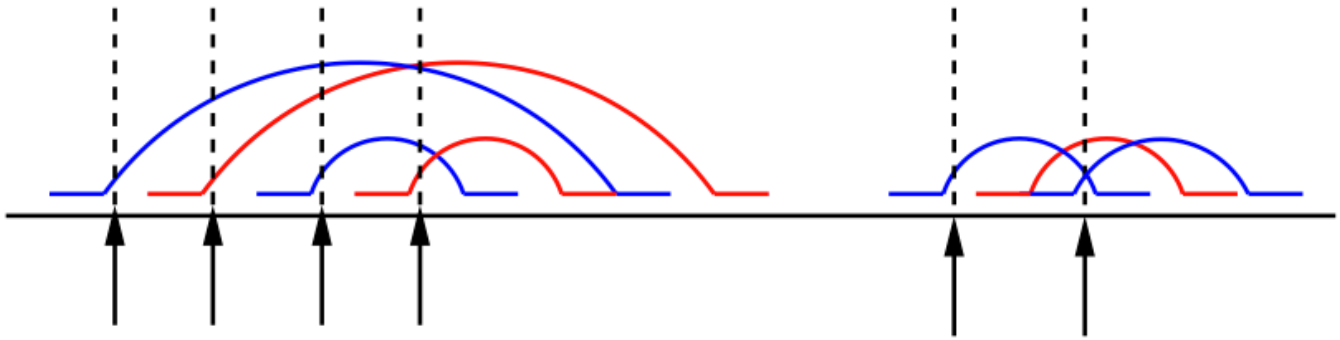
Main project: sequence 2500 genomes at 4x coverage

MoGUL

- A Bayesian approach for discovering indels from a large number of individuals sequenced at a low coverage
  - Explicitly models each individual as homozygous or heterozygous at each locus
  - Computes expected minor allele frequency (MAF) at each locus
  - Allows identification of indels > 30 bases for MAF > 0.04

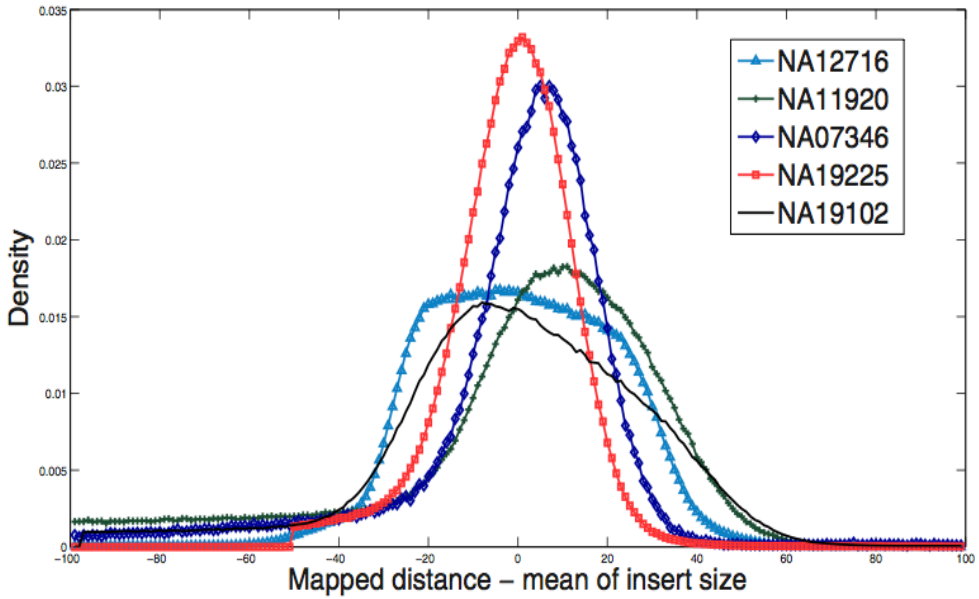
Mate-pair Clusters

- Clustering of mate pairs
- Blue/red for each of two individuals

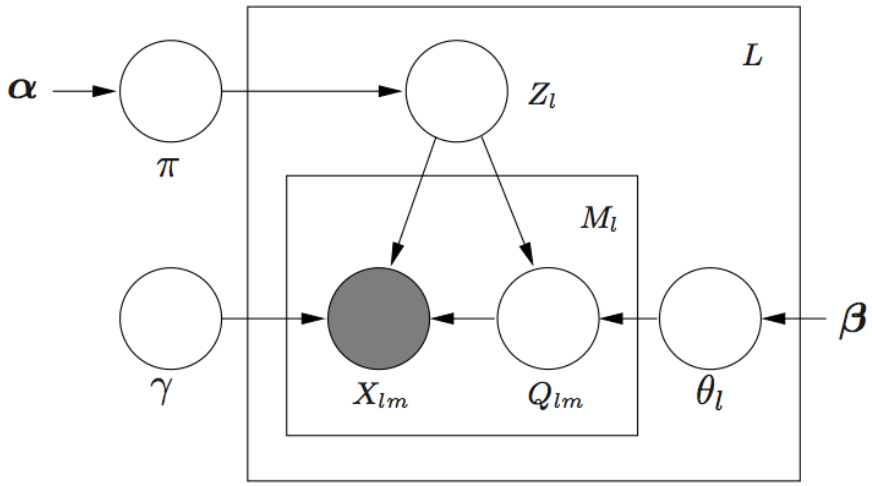


Insert-size Distributions

- The insert size distribution varies across libraries and individuals
- Insert sizes for each individual need to be modeled as separate random variables



MOGUL

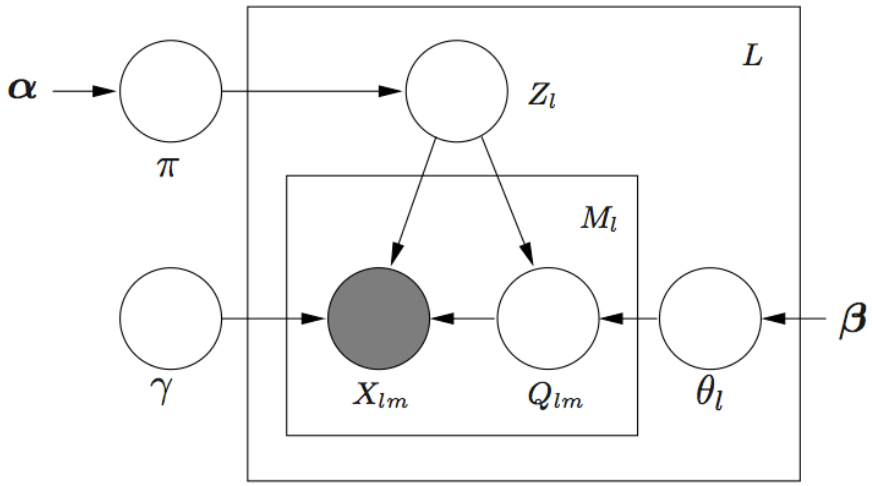


- For a given locus
  - X_{lm}: insert size for individual l, mate pair m

$$X_{lm} = D_{lm} - \mu_{Y_l}$$

  - D_{lm}: mapped distance for l-th individual, m-th mate pair
  - μ_{Yi}: mean of the insert size in the case of no indels

MOGUL



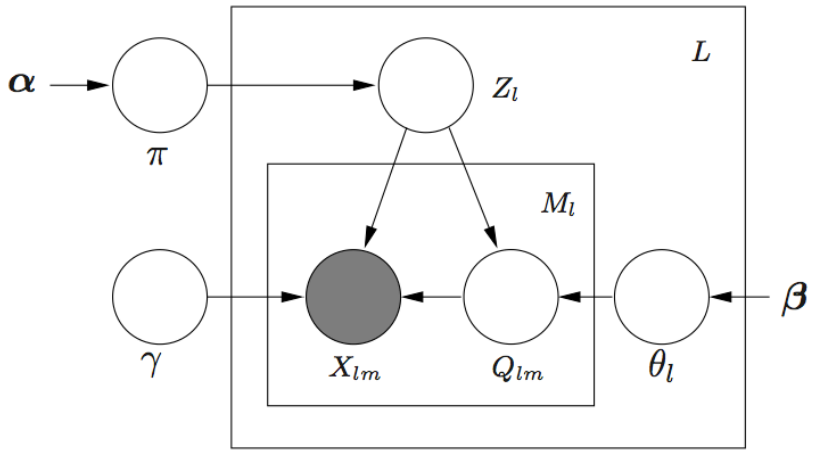
- For a given locus
  - L: Number of individuals
  - M_l: Number of mate-pairs for individual l
  - Z_l: 0/1 for no indels/indels

$$Z_l = \begin{cases} 0 \text{ if individual } l \text{ has no indel} \\ 1 \text{ if individual } l \text{ has an indel.} \end{cases}$$

  - X_{lm}: insert size for individual l, mate pair m
  - Q_{lm} : Two copies of chromosomes

$$Q_{lm} = \begin{cases} 0 \text{ if } Z_l = 1 \text{ and chromosome contains no indel} \\ 1 \text{ if } Z_l = 1 \text{ and chromosome contains an indel} \\ 2 \text{ if } Z_l = 0. \end{cases}$$

MoGUL



- Prior distributions

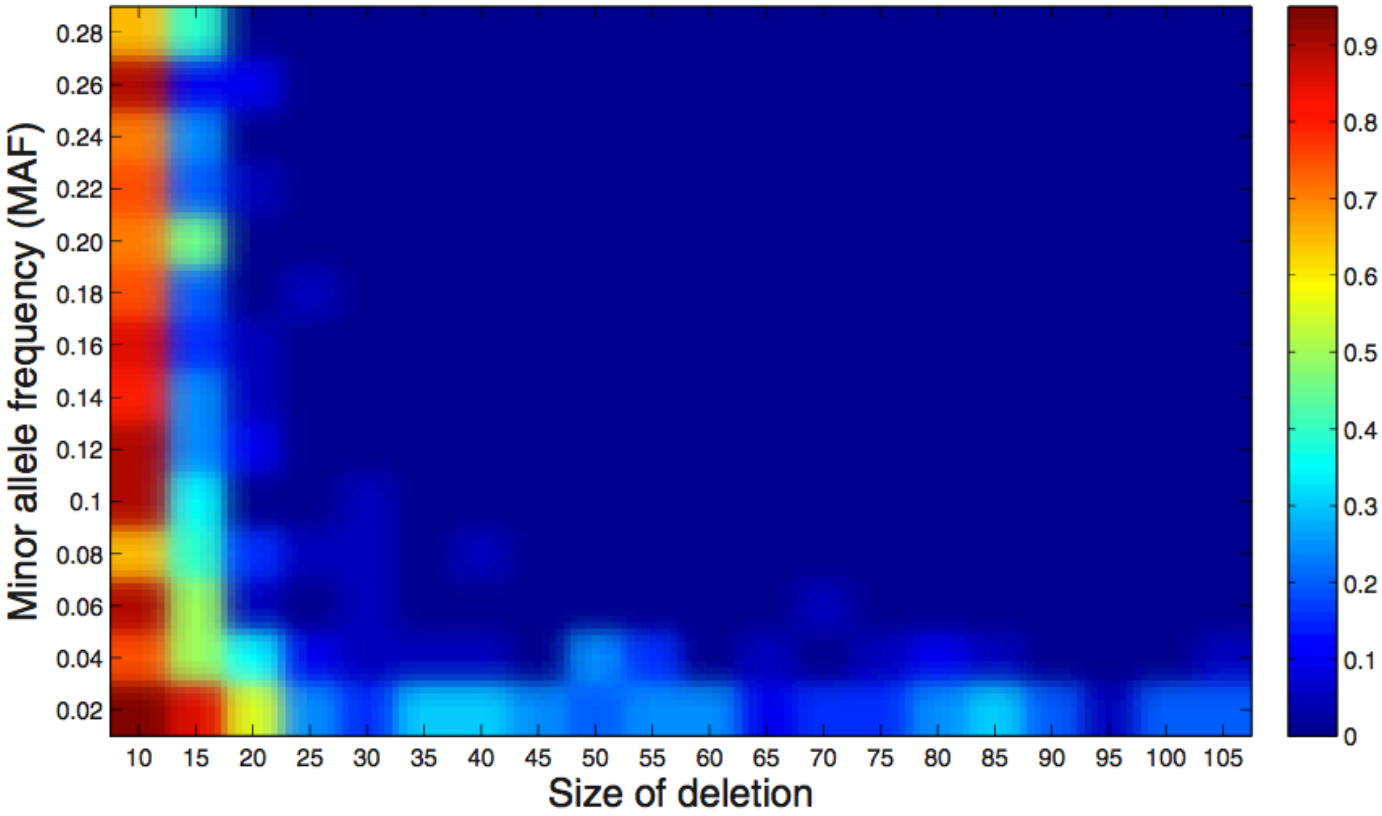
$$p(Z_l = z|\pi) = \pi^z(1-\pi)^{1-z}$$

$$p(Q_{lm} = q|Z_l = 1, \theta_l) = \theta_l^q(1-\theta_l)^{1-q}$$

- Find a MAP (maximum a posteriori) estimate of the unknown parameters

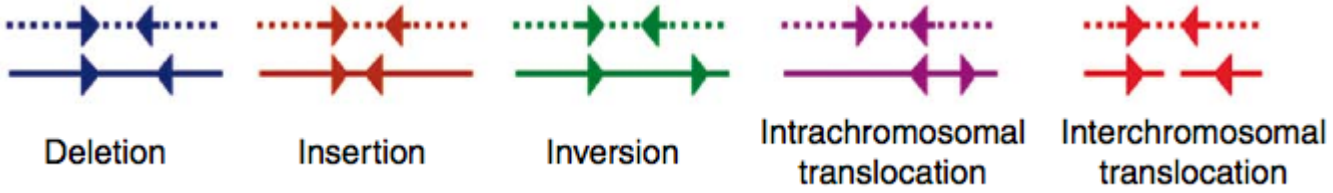
MOGUL: Simulation Study

- Heatmap for average error rates of 20 MOGUL simulations



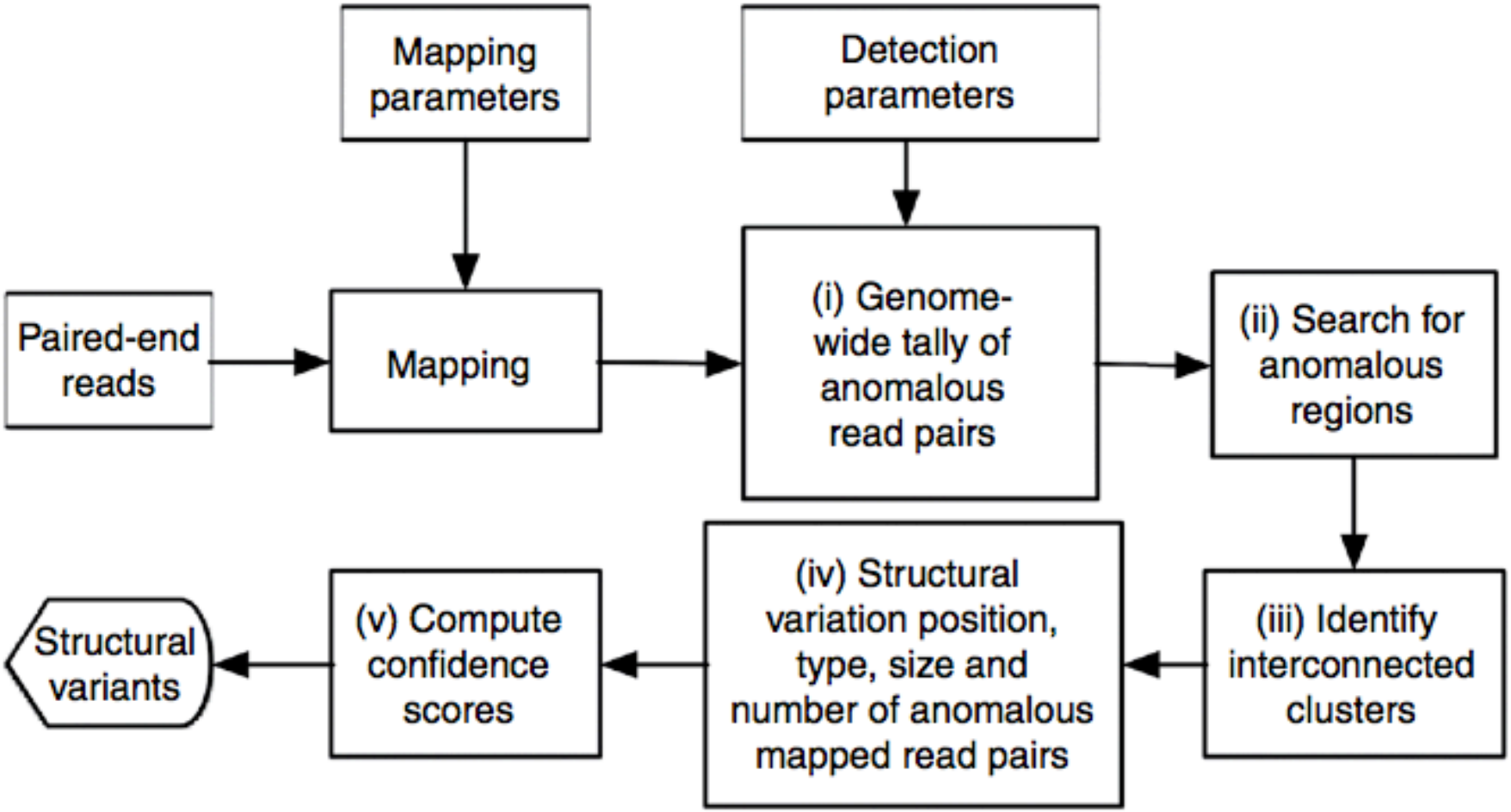
BreakDancer

- BreakDancerMax
  - Detects deletions, insertions, inversions, intrachromosomal and interchromosomal translocations



- BreakDancerMini
  - Focuses on small indels (10-100bp) that are often missed by BreakDancerMax

BreakDancer



BreakDancerMax

- Detects normal, deletion, insertion, inversion, intrachromosomal translocation and interchromosomal translocation

- Focuses on relatively large insertions/deletions

- Poisson mixture model with a mixture component for each type of structural variant

BreakDancerMax Algorithm

- Align mate-pairs to reference genome
- Assign each mate-pair to categories of normal/deletion/ insertion/inversion/translocation
- Select those regions spanned by two or more anomalous read pairs as candidate structural variants
- Confidence score based on Poisson mixture model is assigned to each candidate structural variant

BreakDancerMax: Confidence Score

- Under the null hypothesis of no variant:
  - the anomalous read pairs are uniformly distributed across the genome for each type of structural variations

BreakDancerMax: Confidence Score

- Given the Poisson mixture model, confidence score for structural variant type i is given as

$$P(n_i \geq k_i)$$

  - k_i: the number of observed type i structural variants
  - n_i: Poisson random variables with mean λ_i
  - The probability of having more than the expected number of inserts of a structural variant of each type
- Poisson mixture model for structural variant type i under the null hypothesis has mean

$$\lambda_i = \frac{sN_i}{G}$$

  - G: the length of the reference genome
  - s: the cumulative size of the regions that the anomalous read pairs anchor to
  - N_i: the number of type i structural variants in the entire dataset. Simply count the number of variants in the dataset

BreakDancerMini

- Detects small indels that are often missed by BreakDancerMax
- Makes use of the normally mapped read pairs that were ignored by BreakDancerMax
  - Regions of normal insert size could be small indels
- Sliding window test
  - Null hypothesis: the distribution of insert sizes in the given window is the same as the distribution of insert sizes in the whole genome

Summary

- MoDIL
  - Statistical method for detecting small indels from paired-end sequencing data
- MoGUL
  - Probabilistic model for detecting structural variants from a large number of individuals sequenced at a low coverage
- BreakDancer
  - Genome-wide detection of structural variants