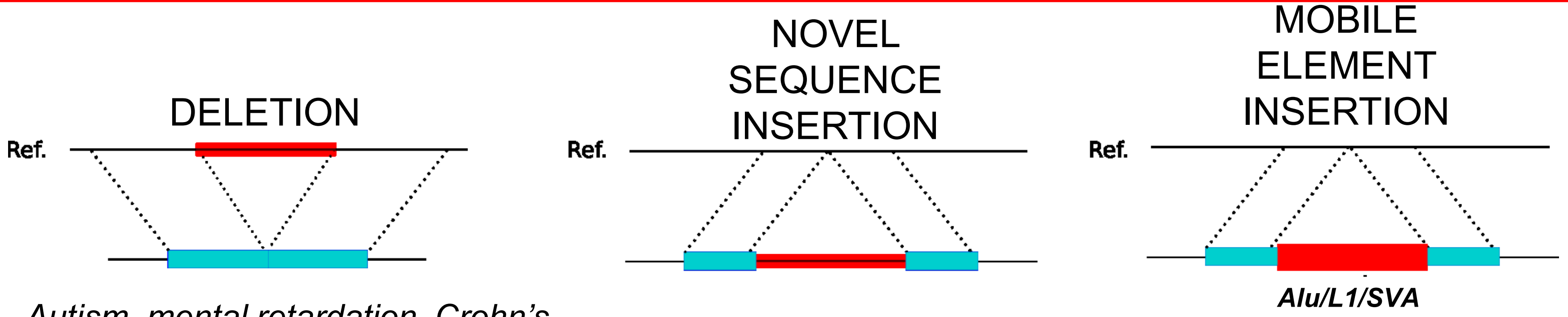
CS681: Advanced Topics in Computational Biology

Week 6 Lectures 2-3

Can Alkan
EA509
calkan@cs.bilkent.edu.tr

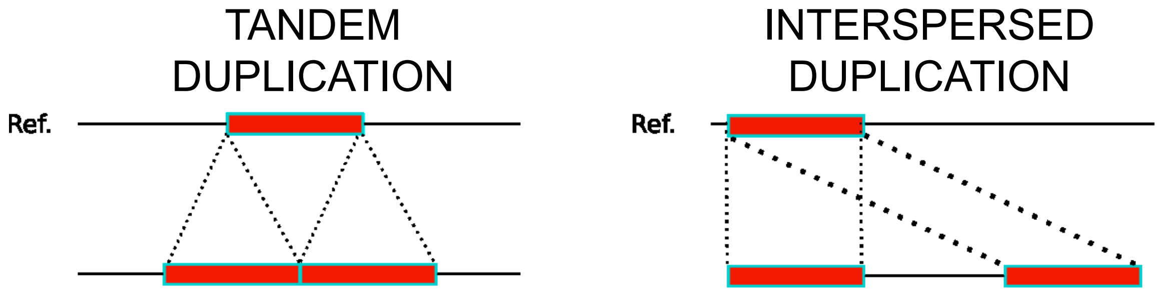
http://www.cs.bilkent.edu.tr/~calkan/teaching/cs681/

Structural Variation Classes



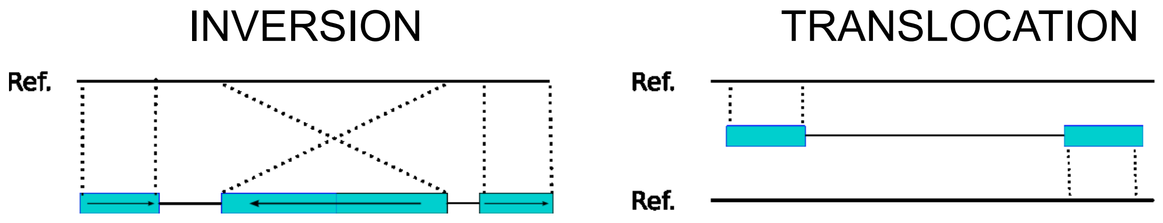
Autism, mental retardation, Crohn's

Haemophilia



Schizophrenia, psoriasis

CNV: Copy number variants

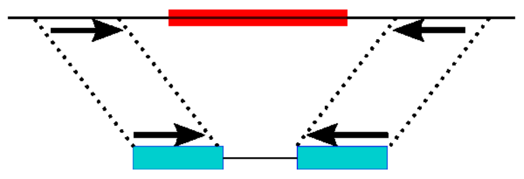


Chronic myelogenous leukemia

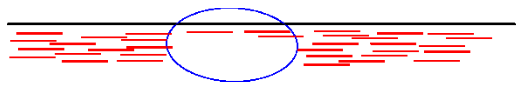
Balanced rearrangements

Sequence signatures of structural variation

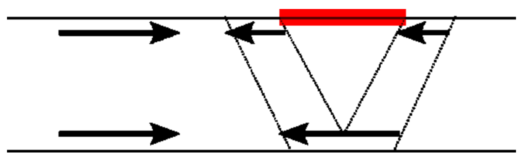
- Read pair analysis
  - Deletions, small novel insertions, inversions, transposons
  - Size and breakpoint resolution dependent to insert size



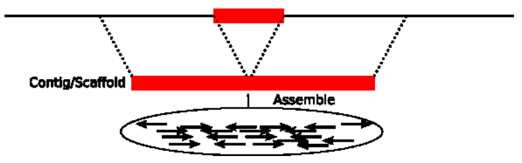
- Read depth analysis
  - Deletions and duplications only
  - Relatively poor breakpoint resolution



- Split read analysis
  - Small novel insertions/deletions, and mobile element insertions
  - 1bp breakpoint resolution

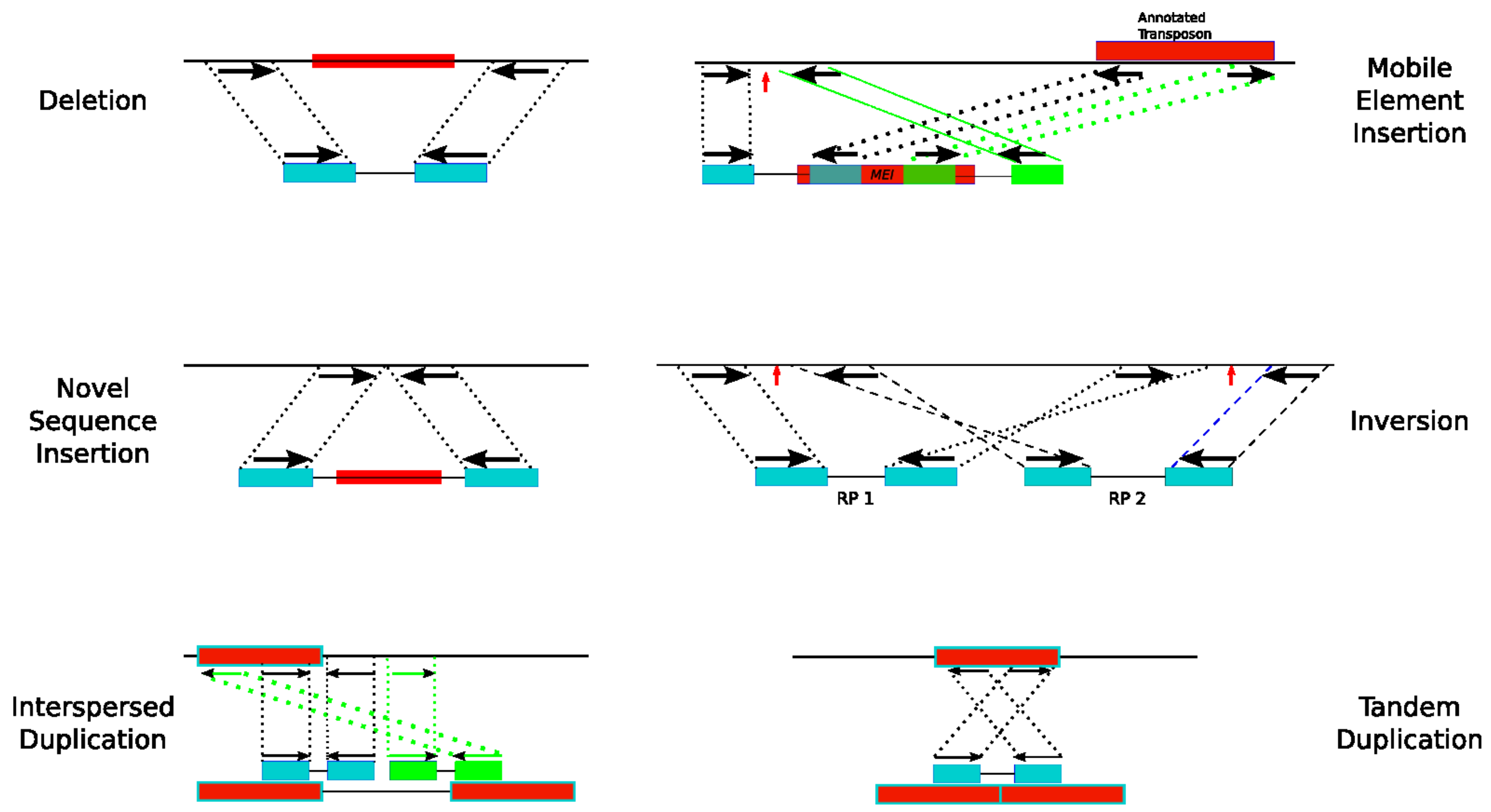


- Local and *de novo* assembly
  - SV in unique segments
  - 1bp breakpoint resolution

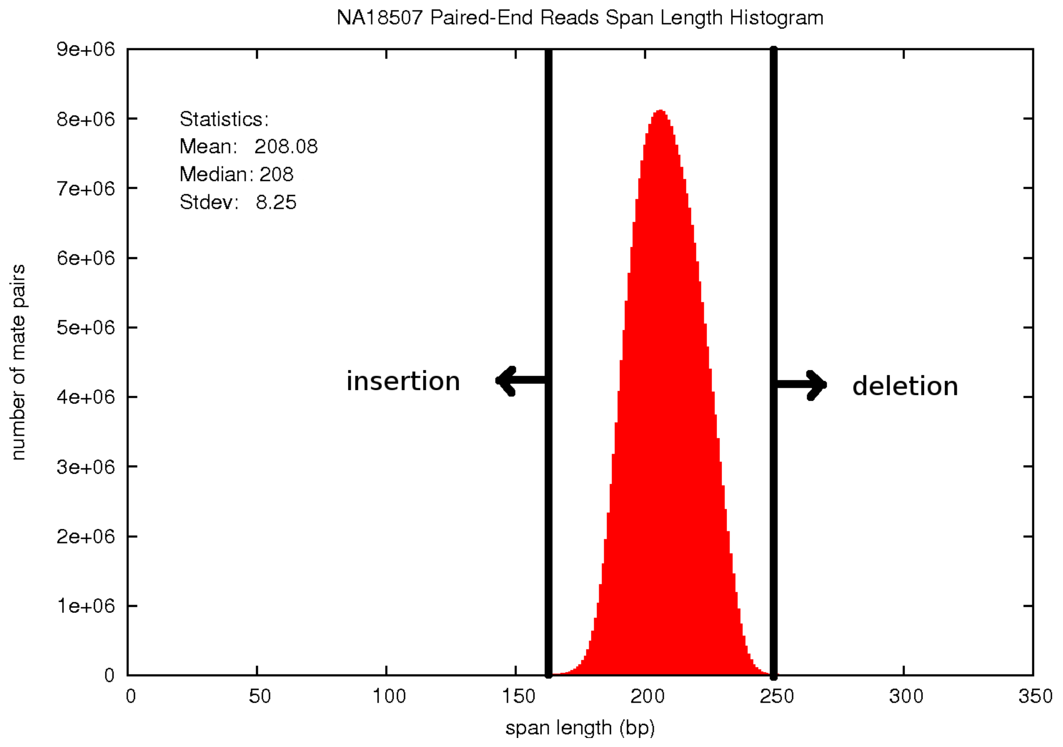


READ PAIR

Read Pair analysis



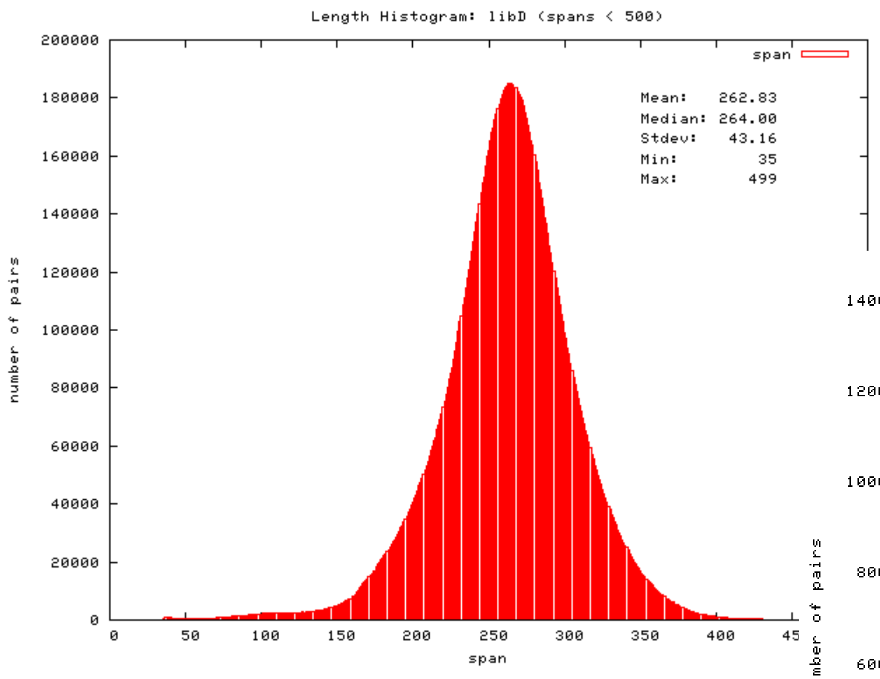
Span size distribution

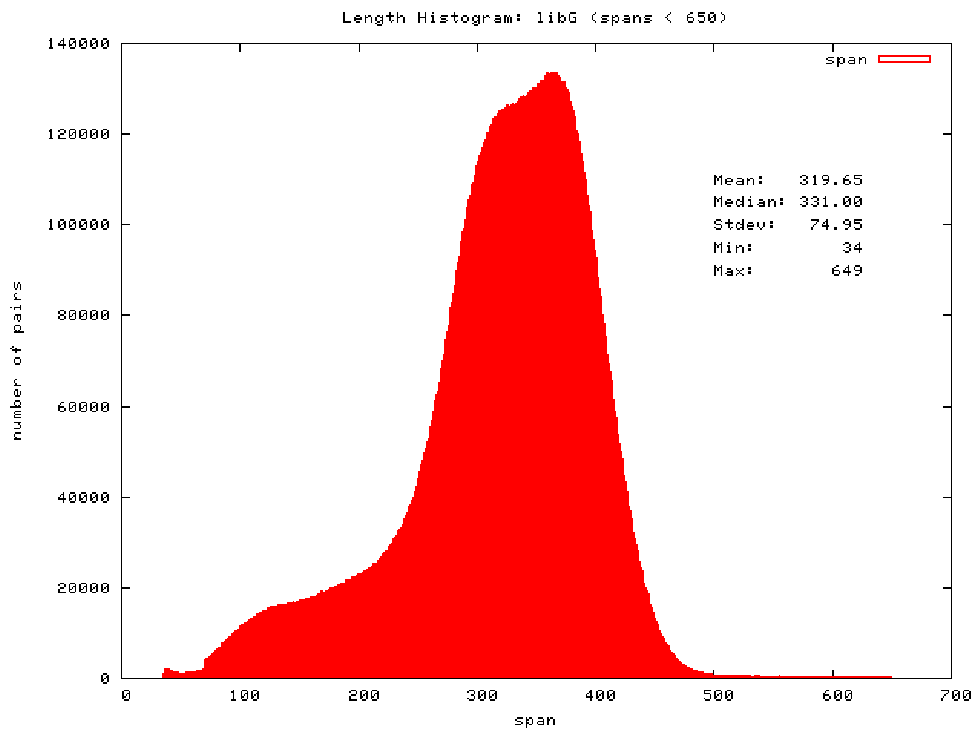


Span size = fragment length = insert size

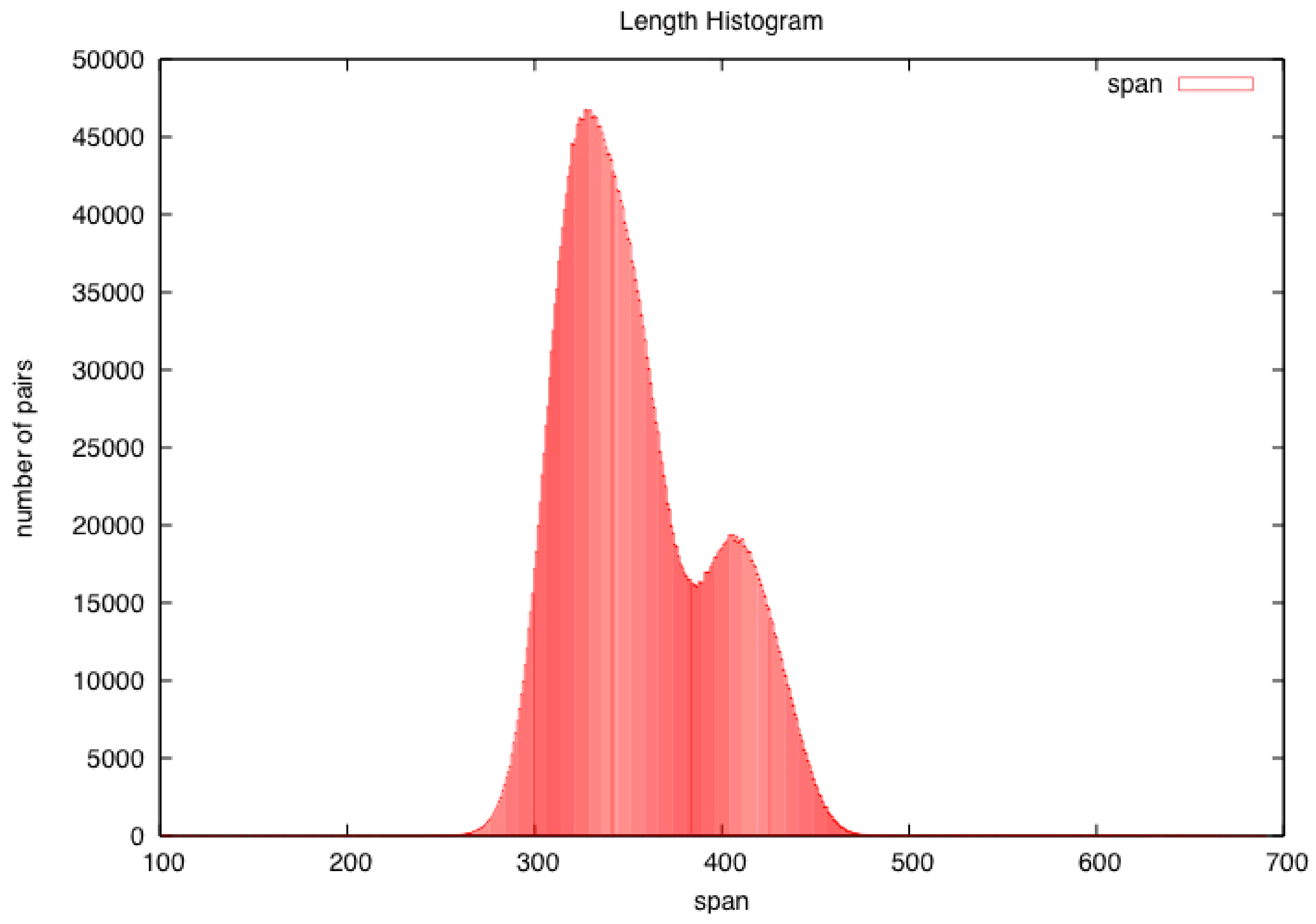
Concordant = read pairs that map in expected orientation & size
Discordant = read pairs that map different than what is expected

Span size distribution: not-so-good

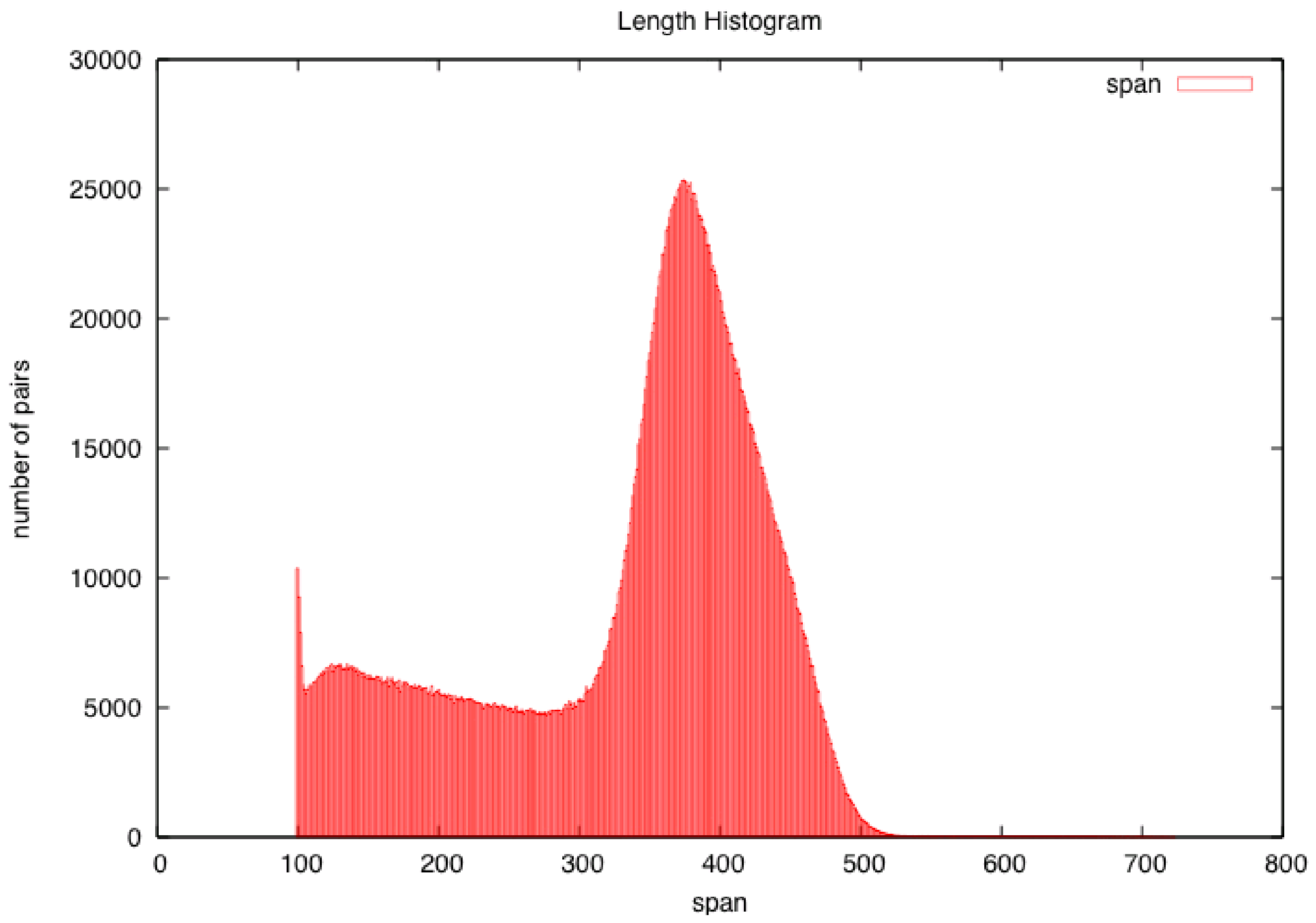




Span size distribution: bad



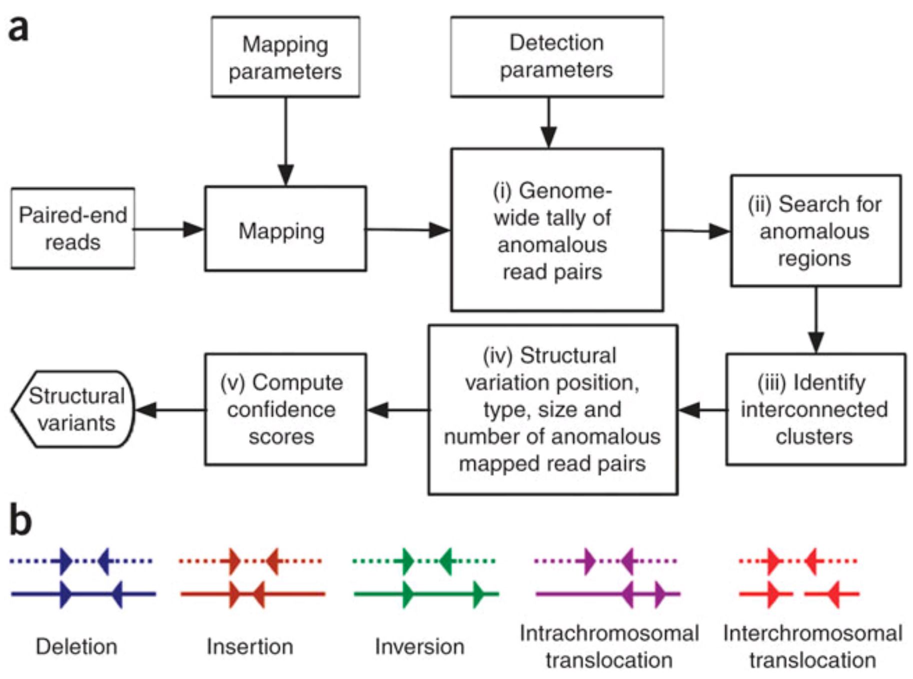
Span size distribution: bad



Read pair based SV callers

- Unique mapping:
  - BreakDancer, GenomeSTRiP, SPANNER, PEMer (454), Corona (SOLiD), etc.
- Multiple mapping:
  - VariationHunter, CommonLAW, MoDIL, MoGUL, HYDRA
- Multi-genome callers (pooled)
  - GenomeSTRiP, MoGUL, CommonLAW

BreakDancer


a
Mapping parameters
Detection parameters
Paired-end reads
Mapping
(i) Genome-wide tally of anomalous read pairs
(ii) Search for anomalous regions
(iii) Identify interconnected clusters
(iv) Structural variation position, type, size and number of anomalous mapped read pairs
(v) Compute confidence scores
Structural variants
b
Deletion
Insertion
Inversion
Intrachromosomal translocation
Interchromosomal translocation

- Unique mapping from MAQ/BWA, etc.
- Two versions:
  - BreakDancerMax
    - >100bp
  - BreakDancerMini
    - 10 – 100 bp

Chen et al., Nature Methods, 2009

BreakDancerMax

- Unique mapping only; filter low MAPQ
- Classify inserts as:
  - Normal, deletion, insertion, inversion, intra-translocation, inter-translocation
  - If not “normal”, name as ARP (anomalous read pair)
- Call SV if at least 2 ARPs are at the same location
- Assign confidence score

BreakDancerMax Confidence Score

Degree of clustering: Probability of having more than the observed number of inserts in a given region

$$P(n_i \geq k_i)$$

i: type of insert
n_i: Poisson random variable with mean λ_i
k_i: number of observed type *i* inserts

Estimation of λ_i

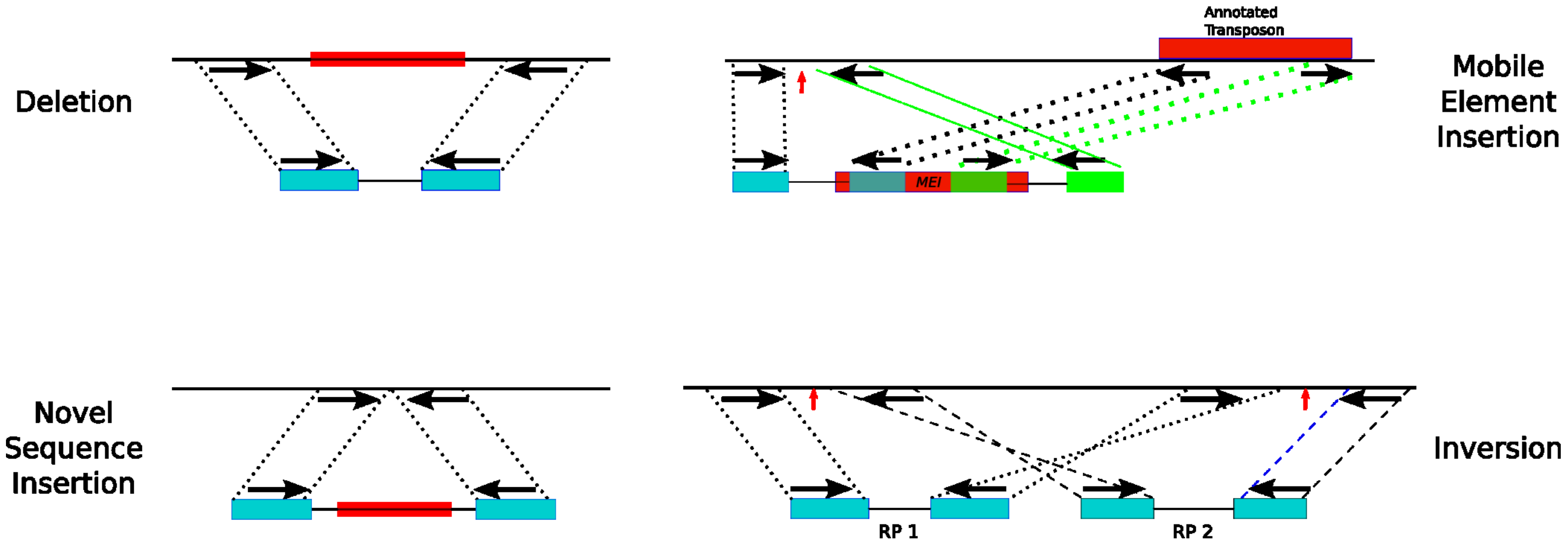
$$\lambda_i = \frac{sN_i}{G}$$

s: size of the region ARPs are anchored
N_i: total number or ARPs of type *i* in the data
G: length of the reference genome

Aim: find statistically significant SVs; i.e. $p<0.0001$

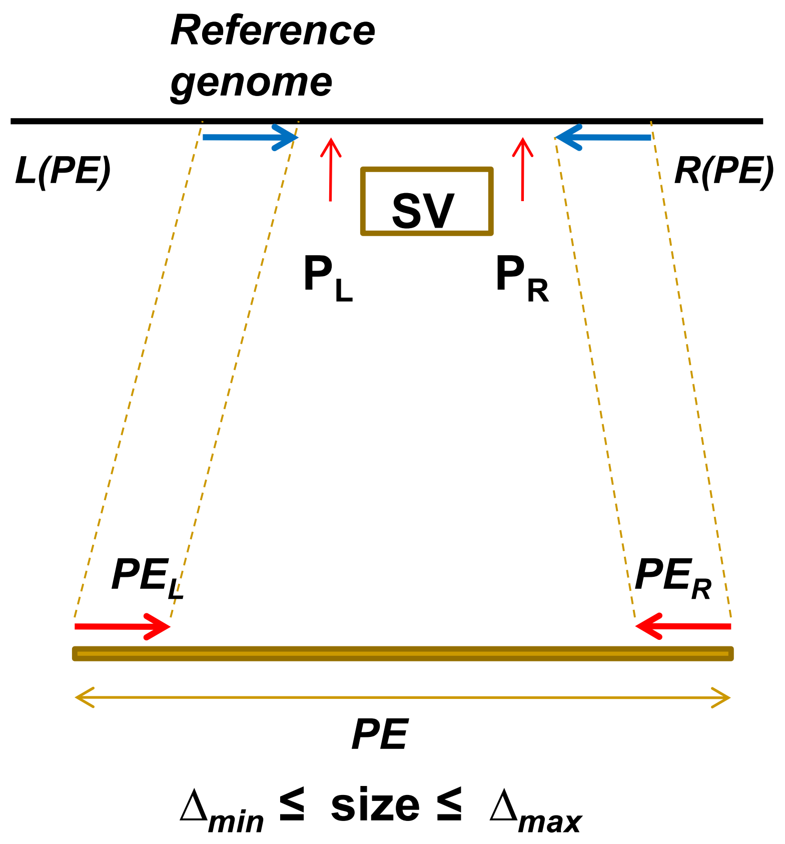
VariationHunter

- **VariationHunter-SC: Maximum parsimony approach; using all discordant map locations; finds an optimal set of SVs through a combinatorial algorithm based on *set-cover***
- *VariationHunter-Pr: Probabilistic version; tries to maximize the probability score of detected SVs*



Hormozdiari, Alkan, et al, Genome Res. 2009

Definitions



Paired-end read

PE:= (PE_L, PE_R)

PE-Alignment

(PE, L(PE), R(PE), O(PE))

O(PE): mapping orientation:

- "+/-": normal
- "+/+" or "-/-": inversion
- "-/+": tandem duplication

SV = (P_L, P_R, L_{min}, L_{max})

Mathematical model

Let L_{min}, L_{max} be *minimum* and *maximum* size of the predicted variant

A Structural Variation is defined by event:

$$SV = (P_L, P_R, L_{min}, L_{max})$$

A PE-Alignment APE=(PE, L(PE), R(PE), O(PE)) supports an insertion $SV = (P_L, P_R, L_{min}, L_{max})$ if:

$$L(PE) \leq P_L$$

$$R(PE) \geq P_R$$

$$L_{min} \geq \Delta_{min} - (R(PE) - L(PE))$$

$$L_{max} \leq \Delta_{max} - (R(PE) - L(PE))$$

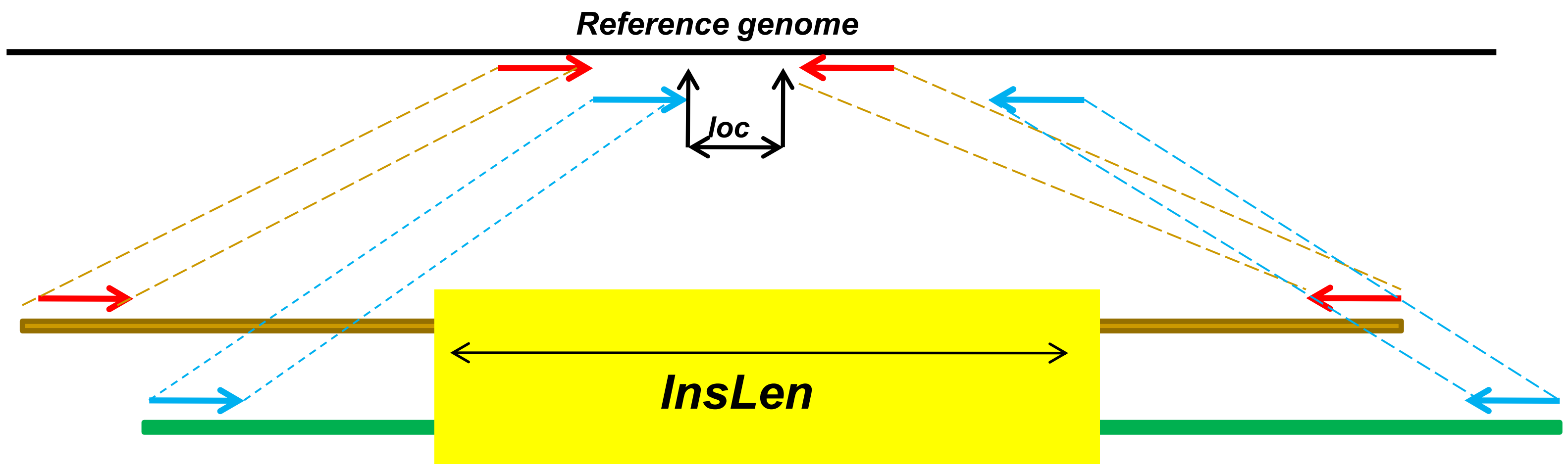
Valid clusters

A set of PE-Alignments that support the same structural variation event SV

A cluster C is a *valid cluster* supporting insertions if:

$$\exists loc, \forall APE \in C : L(APE) < loc < R(APE)$$

$$\exists InsLen, \forall APE \in C : \Delta_{\min} - (R(APE) - L(APE)) < InsLen < \Delta_{\max} - (R(APE) - L(APE))$$



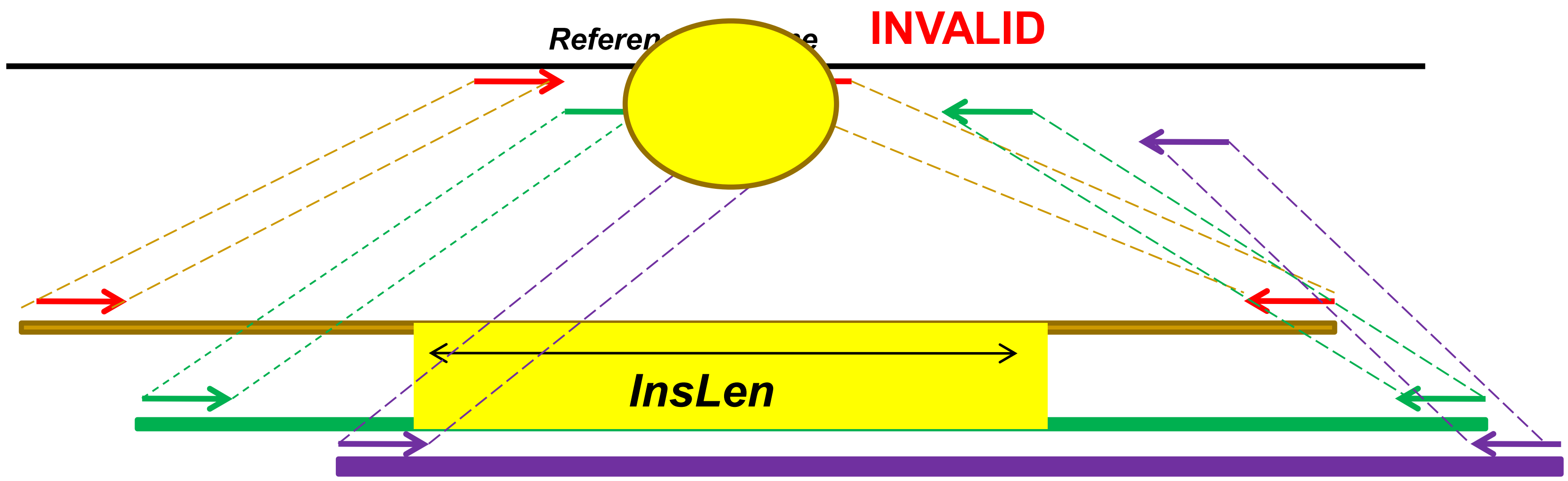
Valid clusters

A set of PE-Alignments that support the same structural variation event SV

A cluster C is a *valid cluster* supporting insertions if:

$$\exists loc, \forall APE \in C : L(APE) < loc < R(APE)$$

$$\exists InsLen, \forall APE \in C : \Delta_{\min} - R(APE) + L(APE) < InsLen < \Delta_{\max} - R(APE) + L(APE)$$

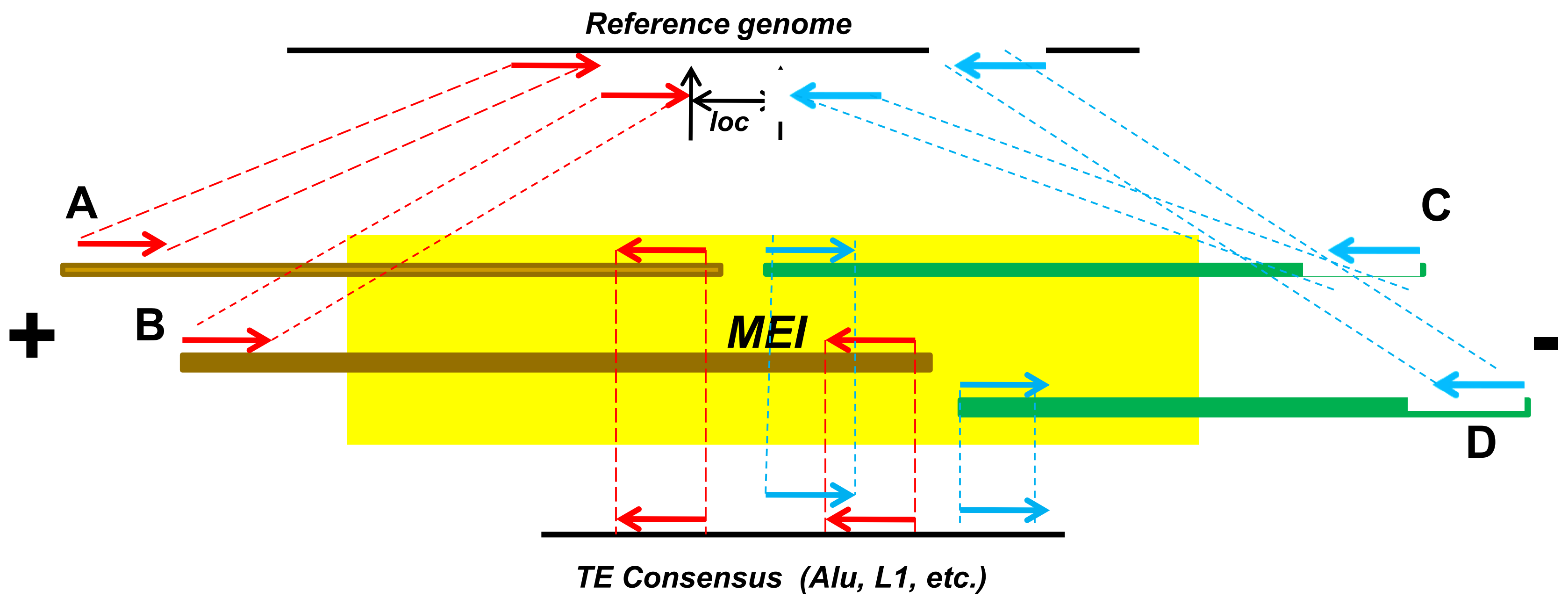


Maximal Valid Clusters for Insertions

A *Maximal Valid Cluster* is a valid cluster that no additional APE can be added without violating the validity of the cluster

1. Find all the Maximal sets of overlapping paired-end alignments
2. For each maximal set S_k found in Step 1, find all the maximal subsets s_i in S_k that the insertion size (*InsLen*) they suggest is overlapping
3. Among all the sets s_i found in Step 2, remove any set which is a proper subset of another chosen set

MEI sequence signature



- Strand rules: MEI-mapping "+" reads and MEI mapping "-" reads should be in different orientations:
  - +/- and -/+ clusters; or +/+ and -/- clusters (inverted MEI)
- Span rules: A=(A1, A2); B=(B1, B2); C=(C1, C2); D=(D1, D2)
  - |A1-B1| ~ |A2-B2| and |C1-D1| ~ |C2-D2| (simplified; we have 8 rules)
- Location and 2-breakpoint rule:

$$\exists loc, \forall PE : RightMost(+) < loc < LeftMost(-)$$

Hormozdiari et al., Bioinformatics 2010

Problem and Solutions

Problem: Among all the maximal valid clusters, which ones are correct?
Aim: Assign a single PE-Alignment to all paired-end reads

- Maximum Parsimony Structural Variation
  - Find a *minimum* number of *SVs* such that all the paired-end reads are covered
    - Similar to SET-COVER problem
    - Greedy algorithm. Approximation factor ***O(log(n))***
- Calculating the probabilities of each potential structural variation.

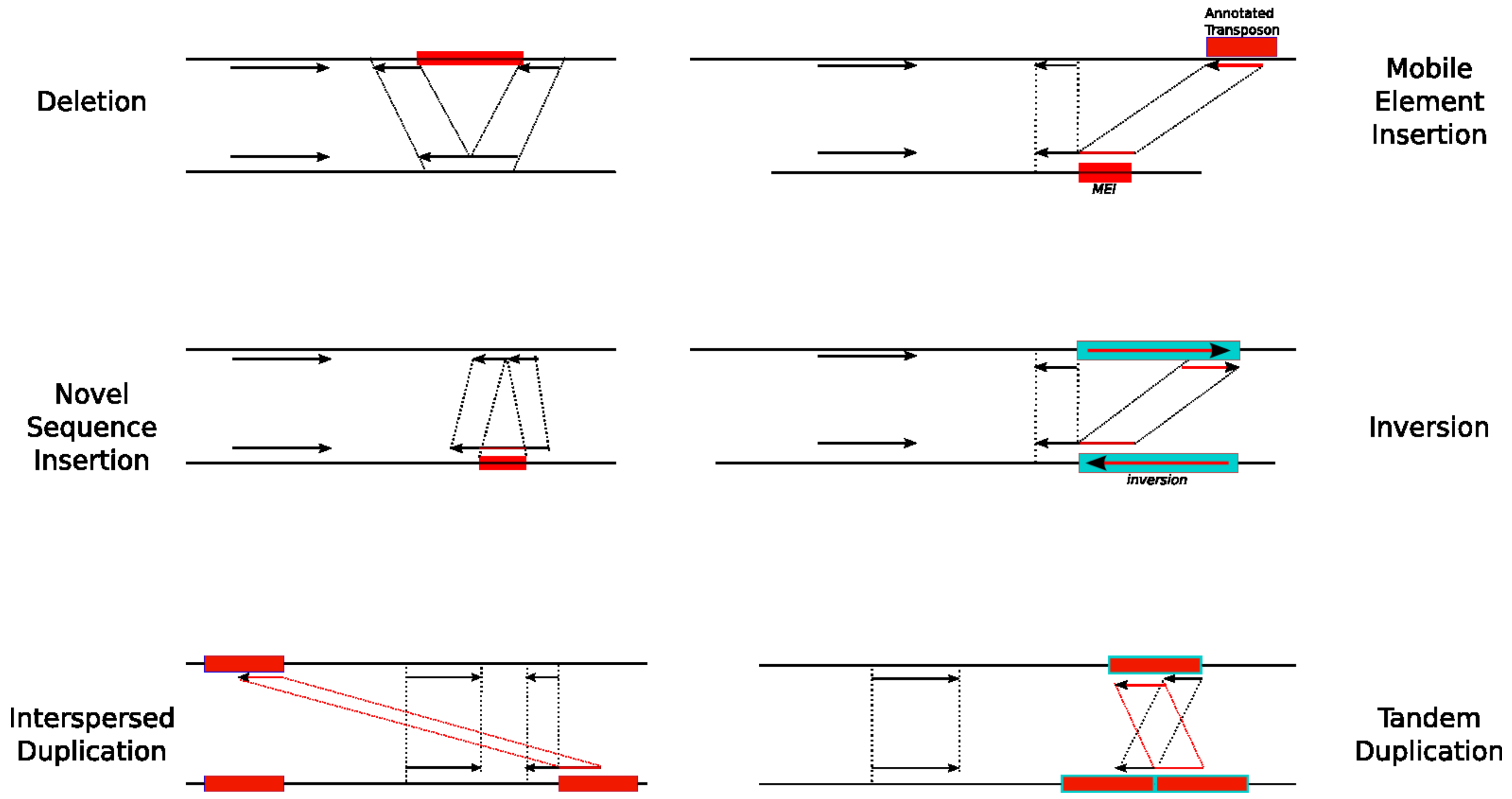
$$\Pr(SV_j) = F(\forall pe \in PE : \Pr(pe \text{ supports } SV_j); L_{\min}; L_{\max})$$

$$\Pr(pe \text{ supports } SV_j) = G(SeqSim(pe, SVj); \forall SV : \Pr(SV))$$

  - Iterative heuristic method to find a solution

SPLIT READ

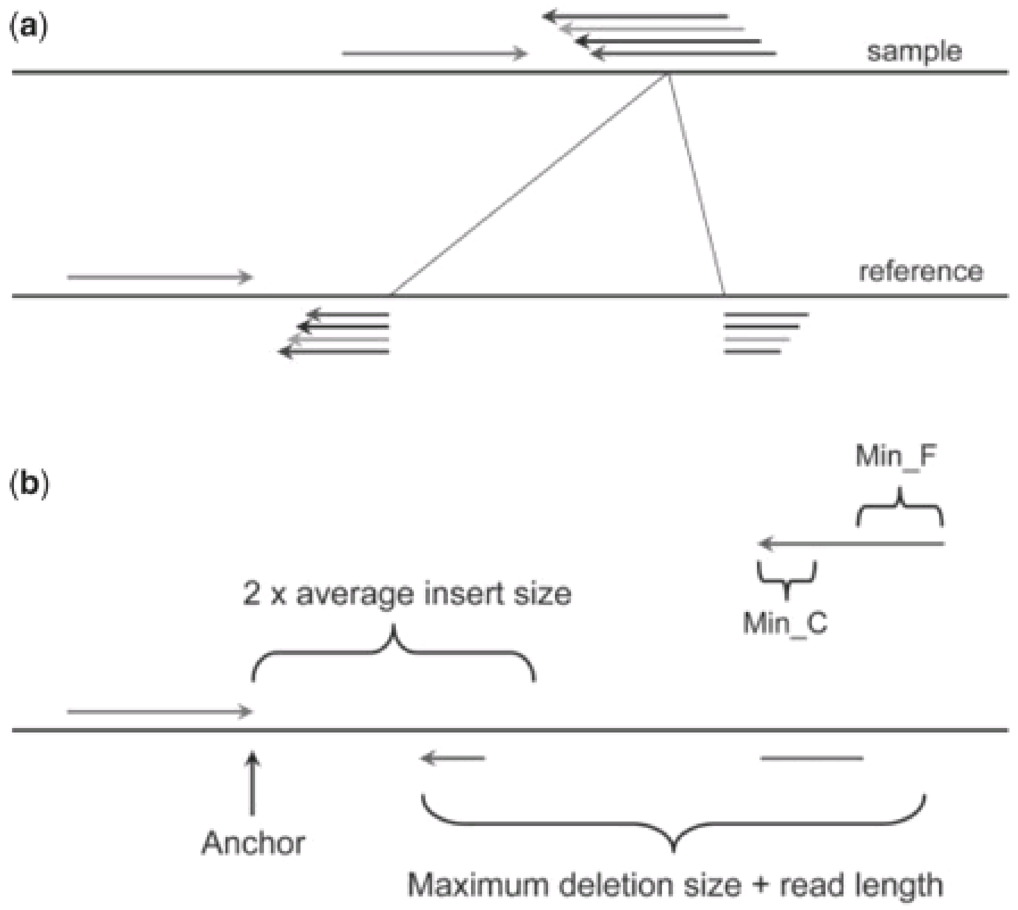
Split Read analysis



Split Read based algorithms

- Unique mapping:
  - Pindel (Ye et al. Bioinformatics, 2009)
  - SRiC (for the 454 platform; Zhang et al., BMC Bioinformatics, 2011)
- Multiple mapping:
  - SPLITREAD (Karakoc et al., Nature Methods, 2012)
- Specialized for RNA alternative splicing:
  - TopHat (Trapnell et al., Bioinformatics, 2009)

Pindel: pattern growth approach



Ye et al. Bioinformatics, 2009

Pattern growth

S = ATCAAGTATGCTTAGC
P = ATGCA

Search **A:**
ATCAAGTATGCTTAGC

Projected database of **A:**
1,4,5,8,14

Search **T** in Projected Database of **A**:
ATCAAGTATGCTTAGC

Projected database of **AT**:
1,8

Search **G** in Projected Database of **AT**:
ATCAAGTATGCTTAGC

Projected database of **ATG**:
8

ATG appears only once: minimum unique substring of pattern P

Search **C** in Projected Database of **ATG**:
ATCAAGTATGCTTAGC

Projected database of **ATGC**:
8

No **ATGCA**. Therefore, ATGC is the maximum unique substring of pattern P

Ye et al. Bioinformatics, 2009

Pindel

1. Read in the location and the direction of the mapped read from the mapping result obtained in the preprocessing step;
2. Define the 3′ end of the mapped read as anchor point;
3. Use pattern growth algorithm to search for minimum and maximum unique substrings from the 3′ end of the unmapped read within the range of two times of the insert size from the anchor point;
4. Use pattern growth to search for minimum and maximum unique substrings from the 5′ end of the unmapped read within the range of read length+*Max_D_Size* starting from the already mapped 3′ end of the unmapped read obtained in step 3;
5. Check whether a complete unmapped read can be reconstructed combining the unique substrings from 5′ and 3′ ends found in steps 3 and 4. If yes, store it in the database *U*. Note that exact matches and complete reconstruction of the unmapped read are required so that neither gap nor substitution is allowed.

- Large Max_D_Size -> slow execution

Ye et al. Bioinformatics, 2009