CS681: Advanced Topics in Computational Biology

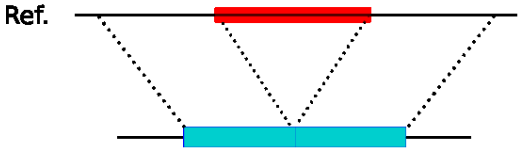
Week 6 Lectures 2-3

Can Alkan
EA224
calkan@cs.bilkent.edu.tr

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

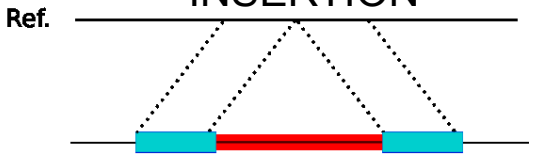
Structural Variation Classes

DELETION

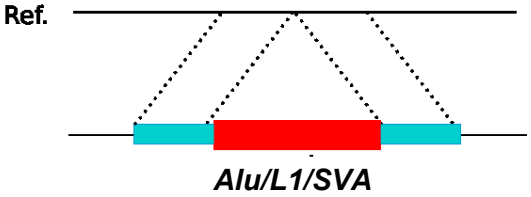


Autism, mental retardation, Crohn's

NOVEL SEQUENCE INSERTION

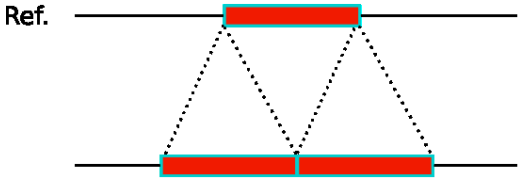


MOBILE ELEMENT INSERTION

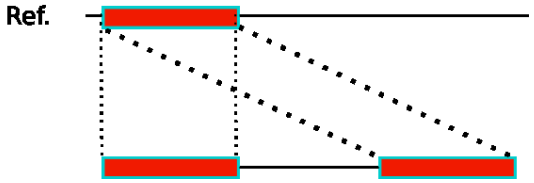


Haemophilia

TANDEM DUPLICATION



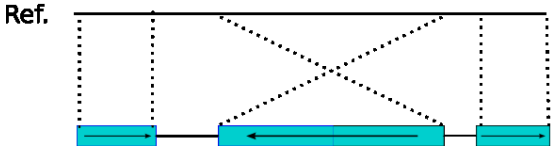
INTERSPERSED DUPLICATION



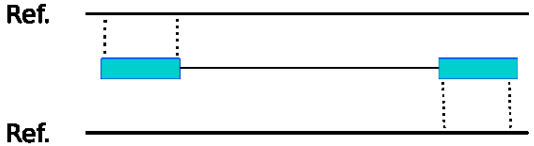
Schizophrenia, psoriasis

CNV: Copy number variants

INVERSION



TRANSLOCATION

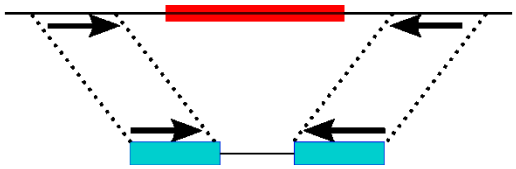


Chronic myelogenous leukemia

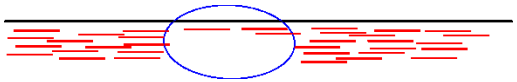
Balanced rearrangements

Sequence signatures of structural variation

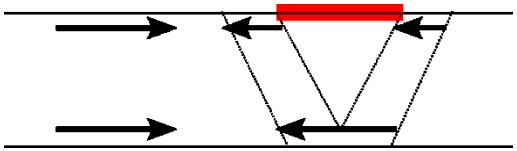
- Read pair analysis
  - Deletions, small novel insertions, inversions, transposons
  - Size and breakpoint resolution dependent to insert size



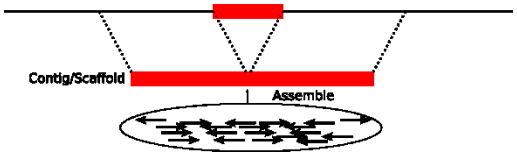
- Read depth analysis
  - Deletions and duplications only
  - Relatively poor breakpoint resolution



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

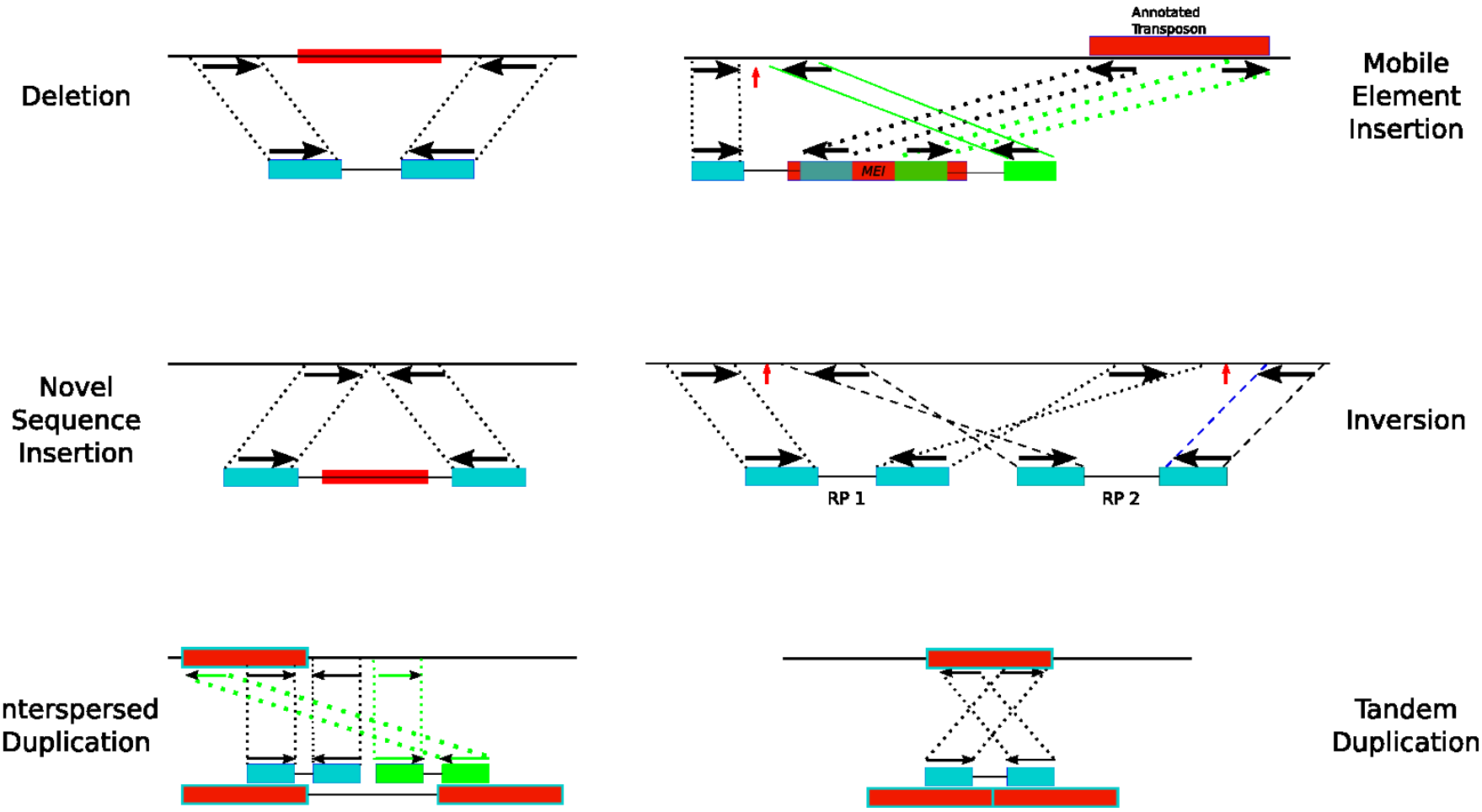


- Local and *de novo* assembly
  - SV in unique segments
  - 1bp breakpoint resolution

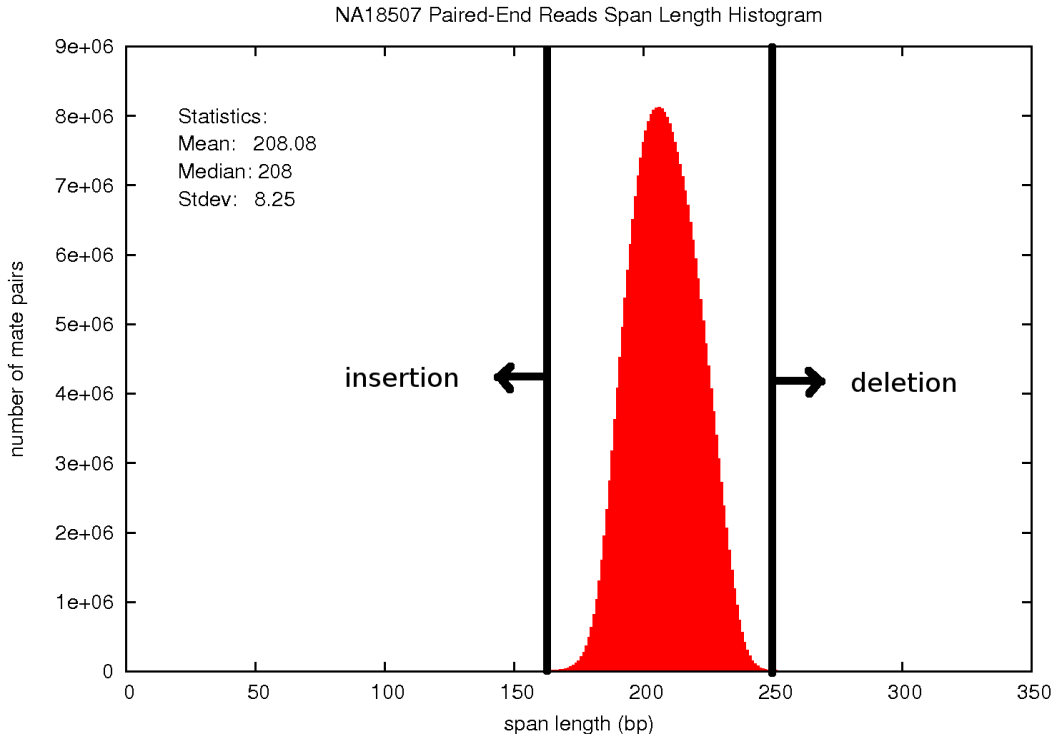


READ PAIR

Read Pair analysis



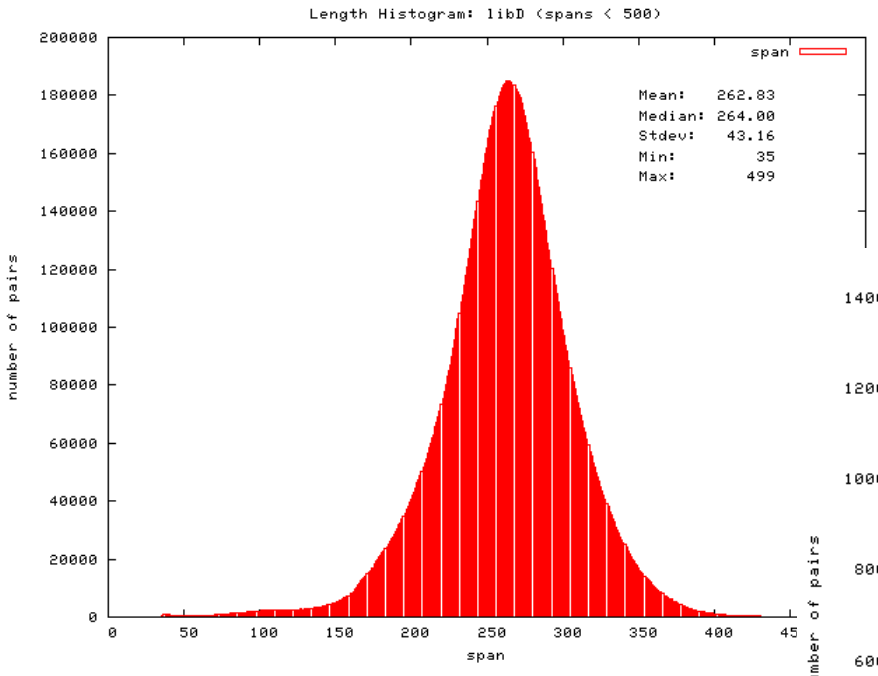
Span size distribution

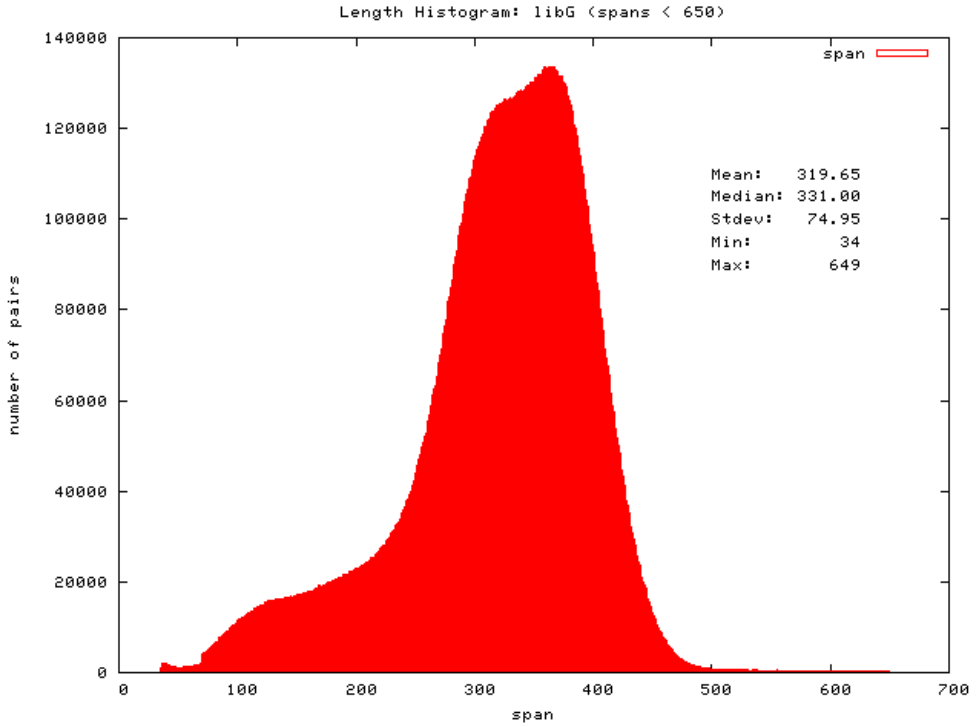


Span size = fragment length = insert size

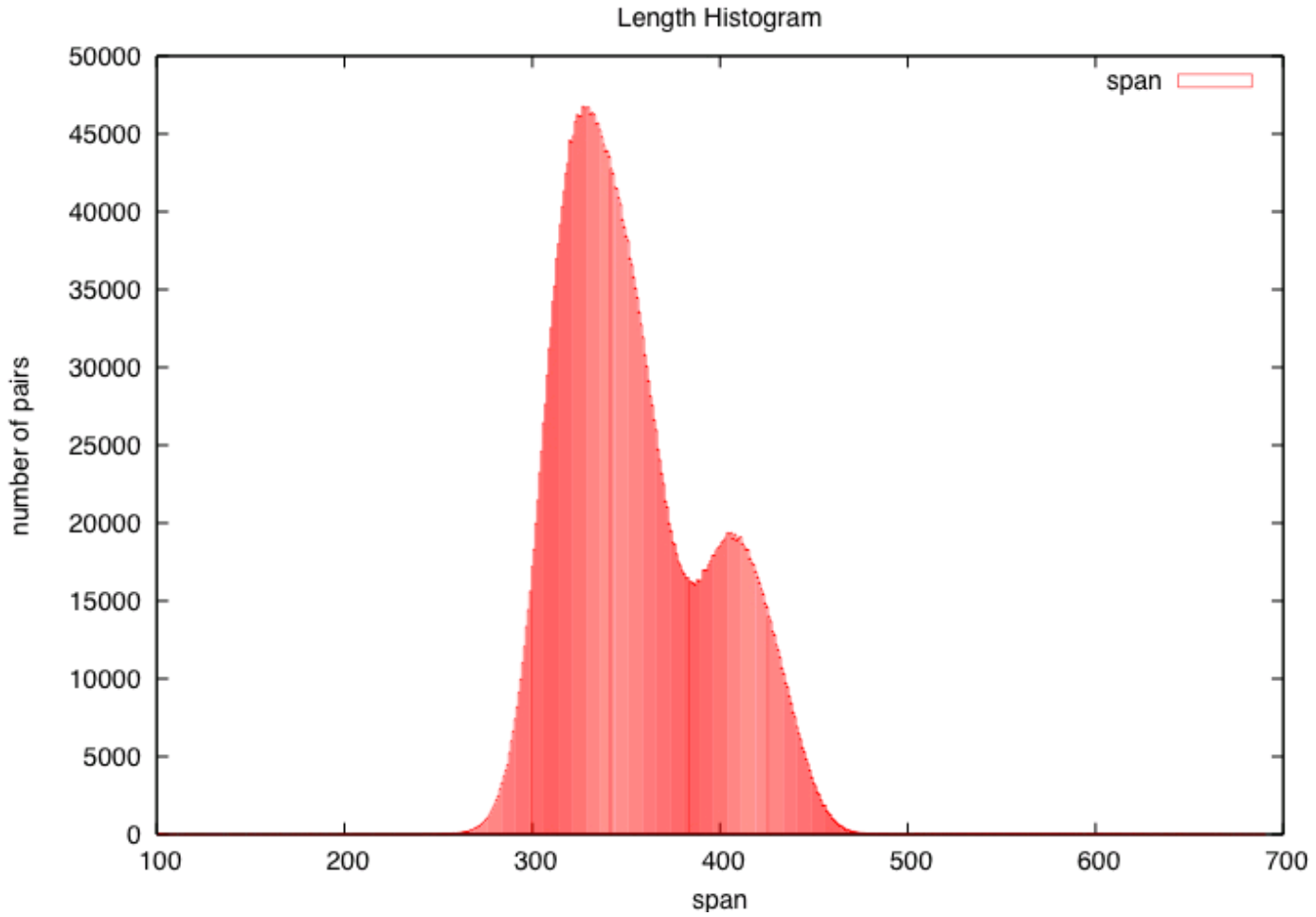
Concordant = read pairs that map in expected orientation & size
Discordant = read pairs that map different than what is expected

Span size distribution: not-so-good

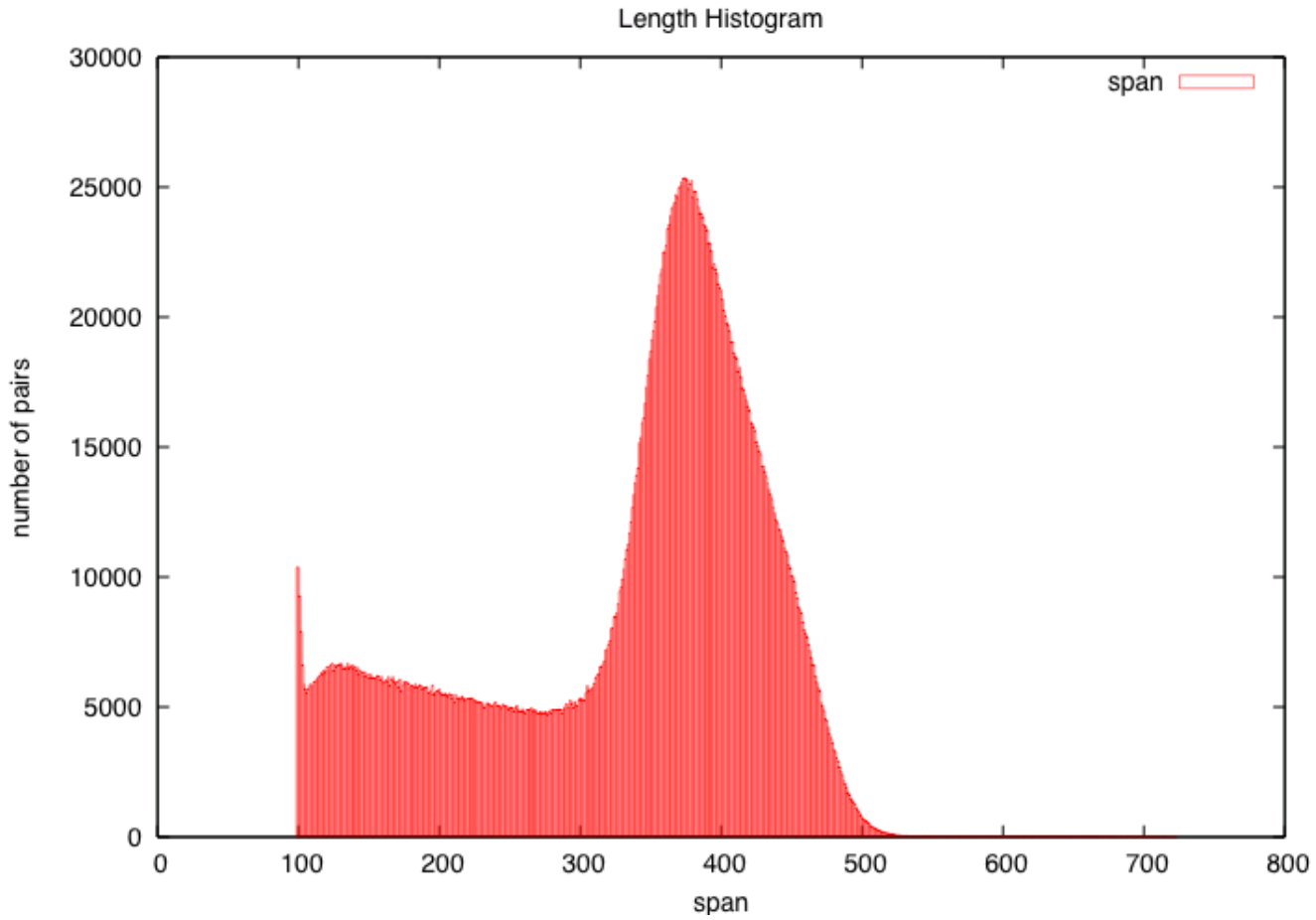




Span size distribution: bad



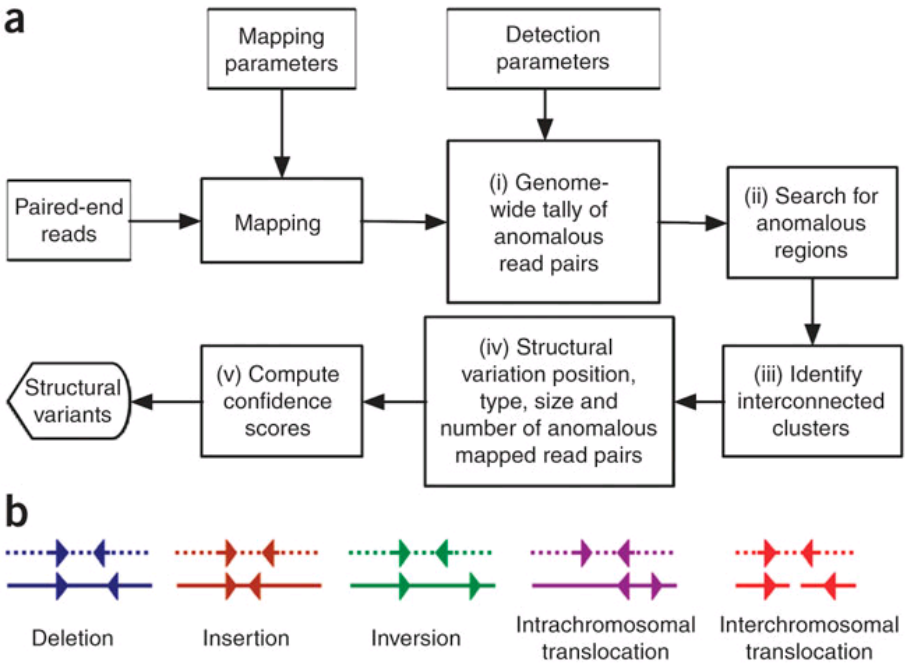
Span size distribution: bad



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



- 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

Chen et al., Nature Methods, 2009

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

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

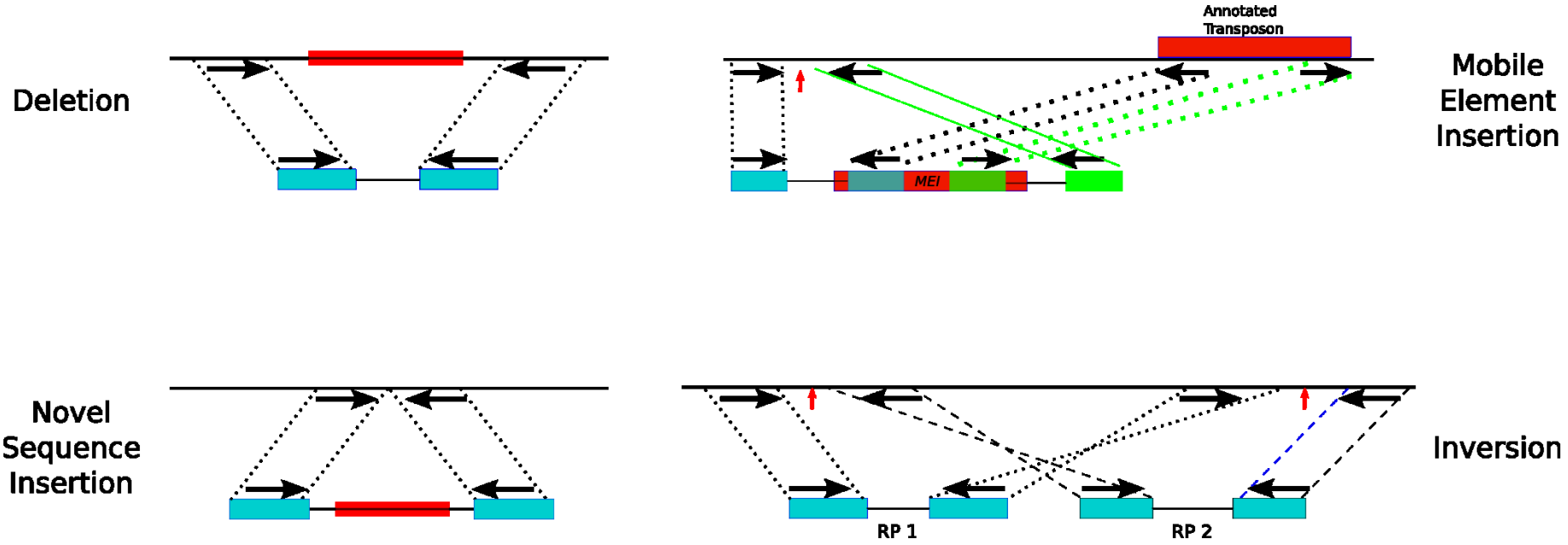
s: size of the region ARPs are anchored
N_i: total number or ARPs of type *i* in the data
G: length of the reference genome

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

Chen et al., Nature Methods, 2009

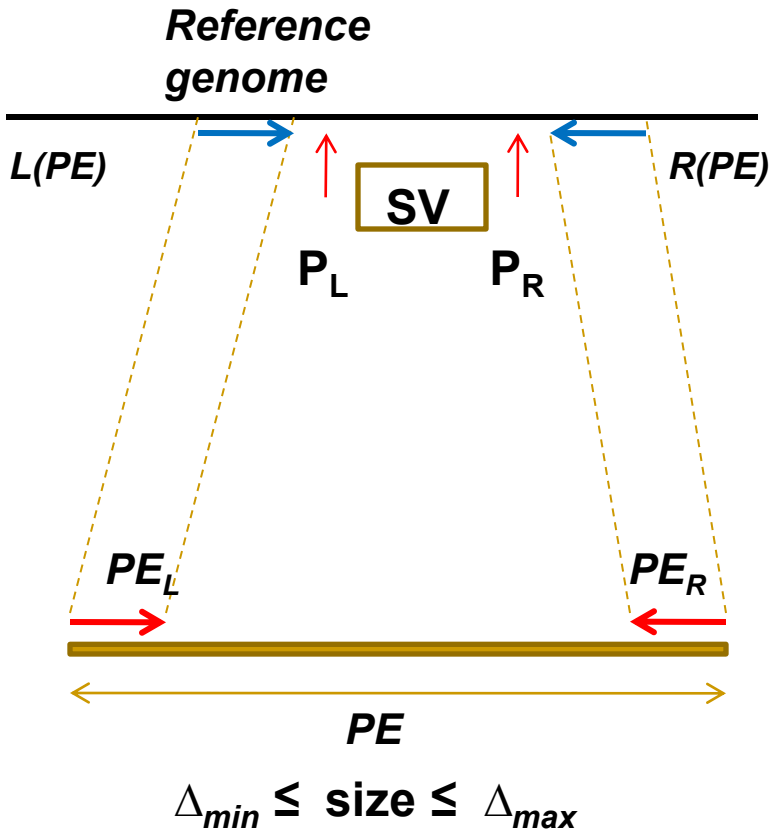
VariationHunter

- **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

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

Hormozdiari, Alkan, et al, Genome Res. 2009

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:

$$L(PE) \leq P_L$$

$$R(PE) \geq P_R$$

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

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

Hormozdiari, Alkan, et al, Genome Res. 2009

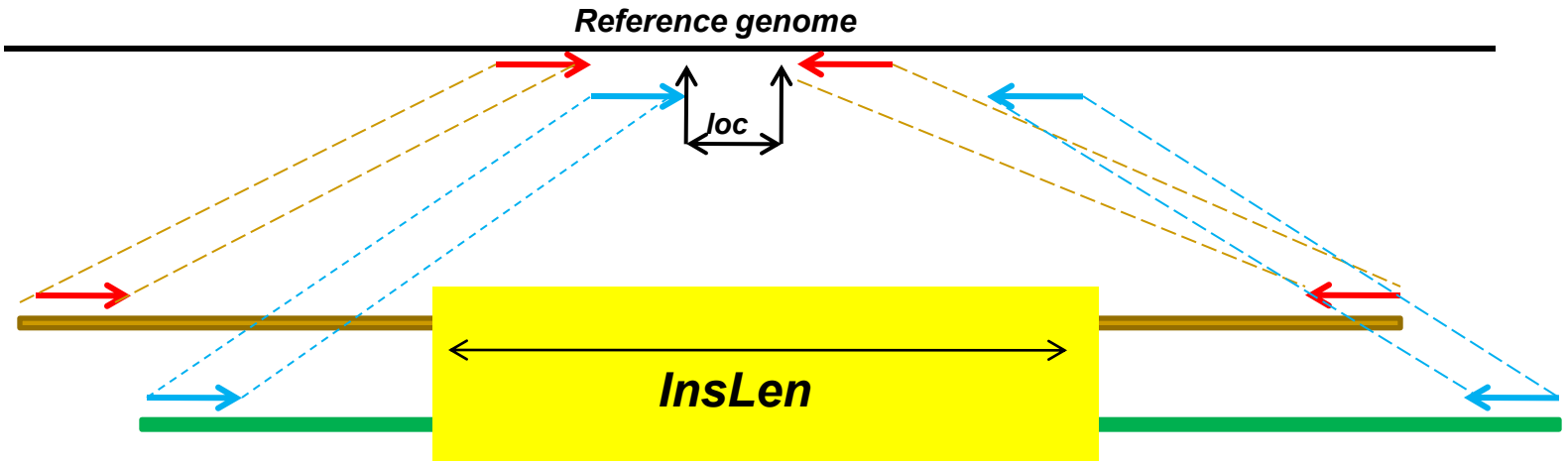
Valid clusters

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

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

$$\exists\ oc, \forall\ PE \in\ ' : L(APE) < oc < R(APE)$$

$$\exists\ nsLen, \forall\ PE \in\ ' : \Delta_{\text{in}} - R(APE) - L(APE)) < nsLen < \Delta_{\text{ax}} - R(APE) - L(APE))$$



Hormozdiari, Alkan, et al, Genome Res. 2009

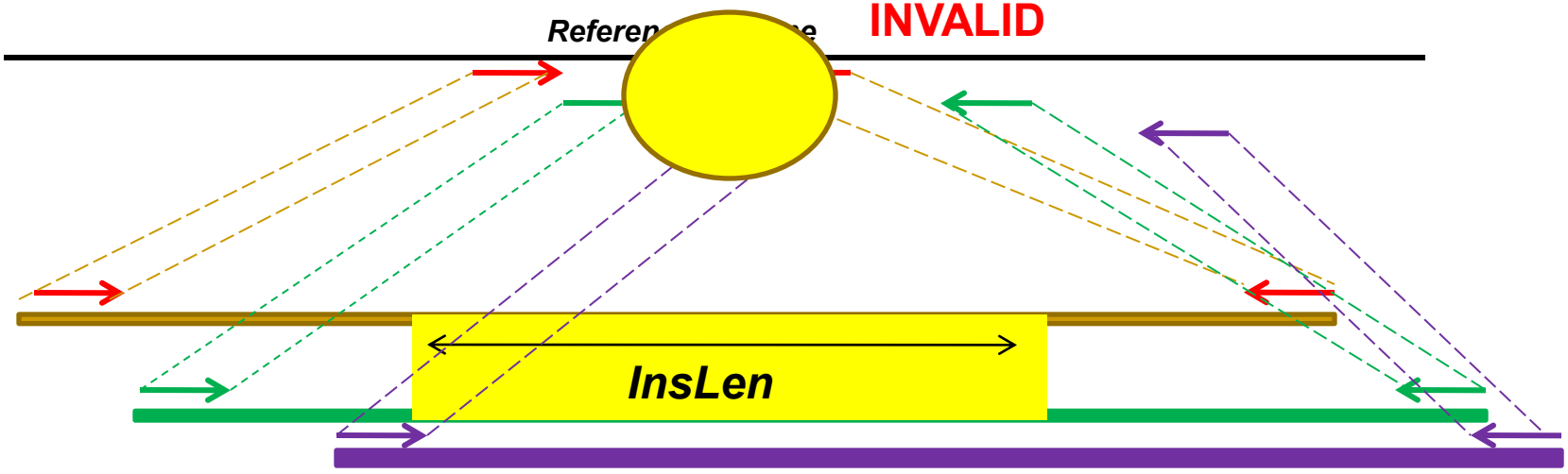
Valid clusters

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

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

$$\exists\ oc, \forall\ PE \in\ ' : L(APE) < oc < R(APE)$$

$$\exists\ nsLen, \forall\ PE \in\ ' : \Delta\ _{in} - R(APE) + L(APE) < nsLen < \Delta\ _{ax} - R(APE) + L(APE)$$



Hormozdiari, Alkan, et al, Genome Res. 2009

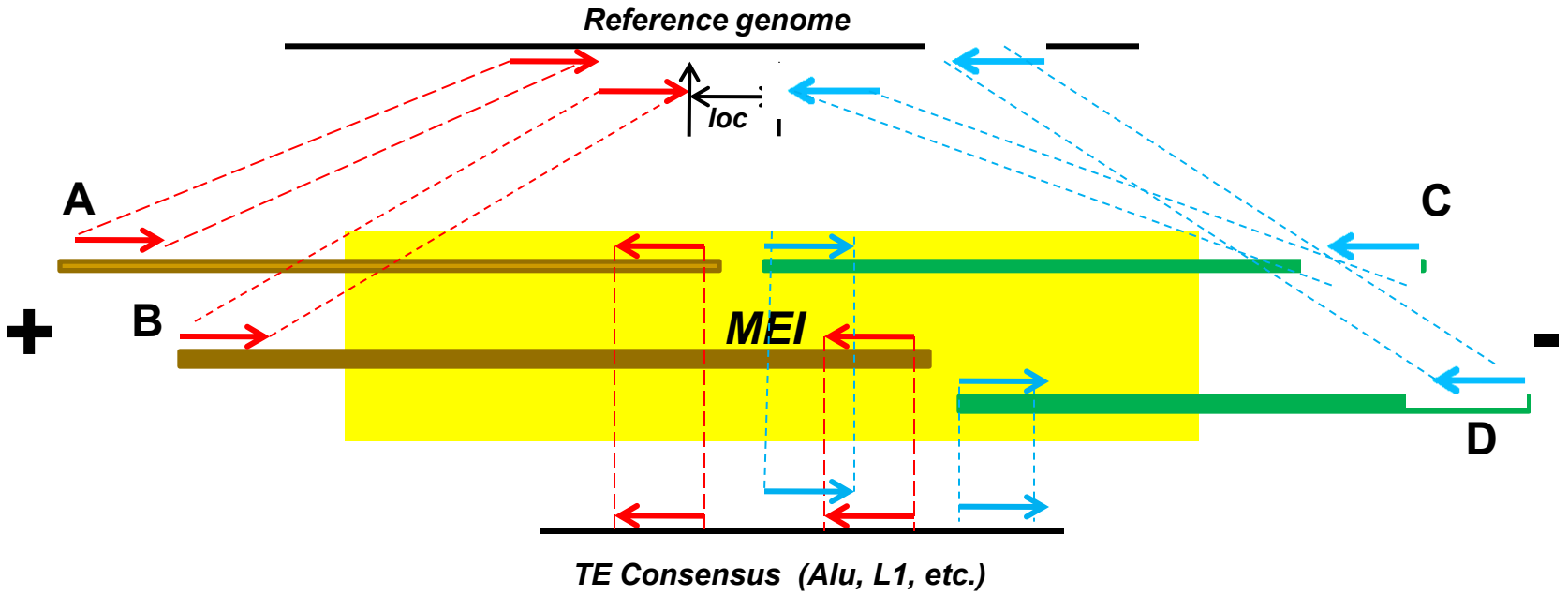
Maximal Valid Clusters for Insertions

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

Hormozdiari, Alkan, et al, Genome Res. 2009

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.

$$\Pr(SV_j) = \Gamma(\forall \quad e \in \quad E : \Pr(pe \text{ supports } SV_j); L_{\min}; L_{\max})$$

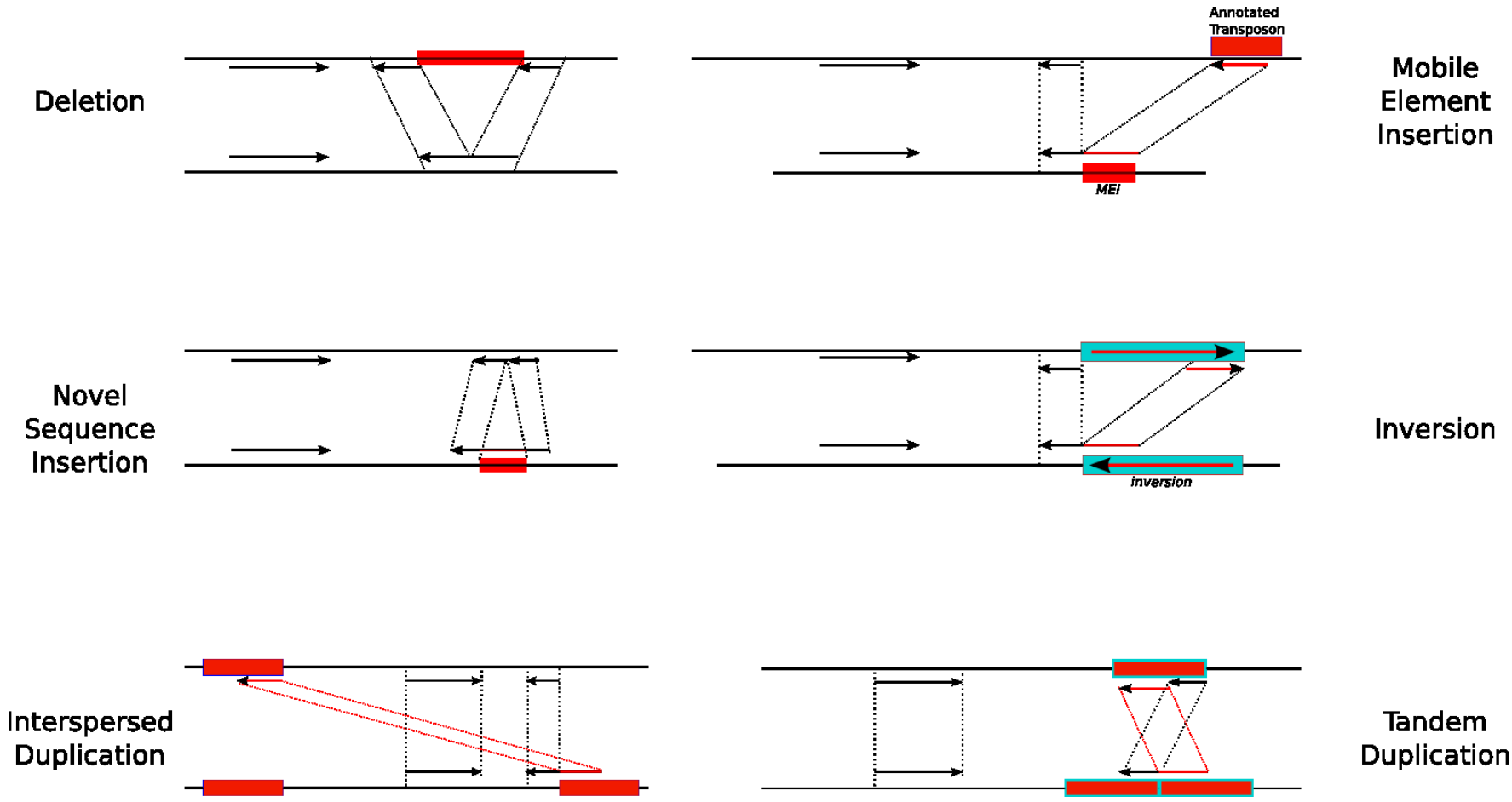
$$\Pr(pe \text{ supports } SV_j) = \Im(SeqSim(pe, SVj); \forall \quad V : \Pr(SV)$$

  - Iterative heuristic method to find a solution

Hormozdiari, Alkan, et al, Genome Res. 2009

SPLIT READ

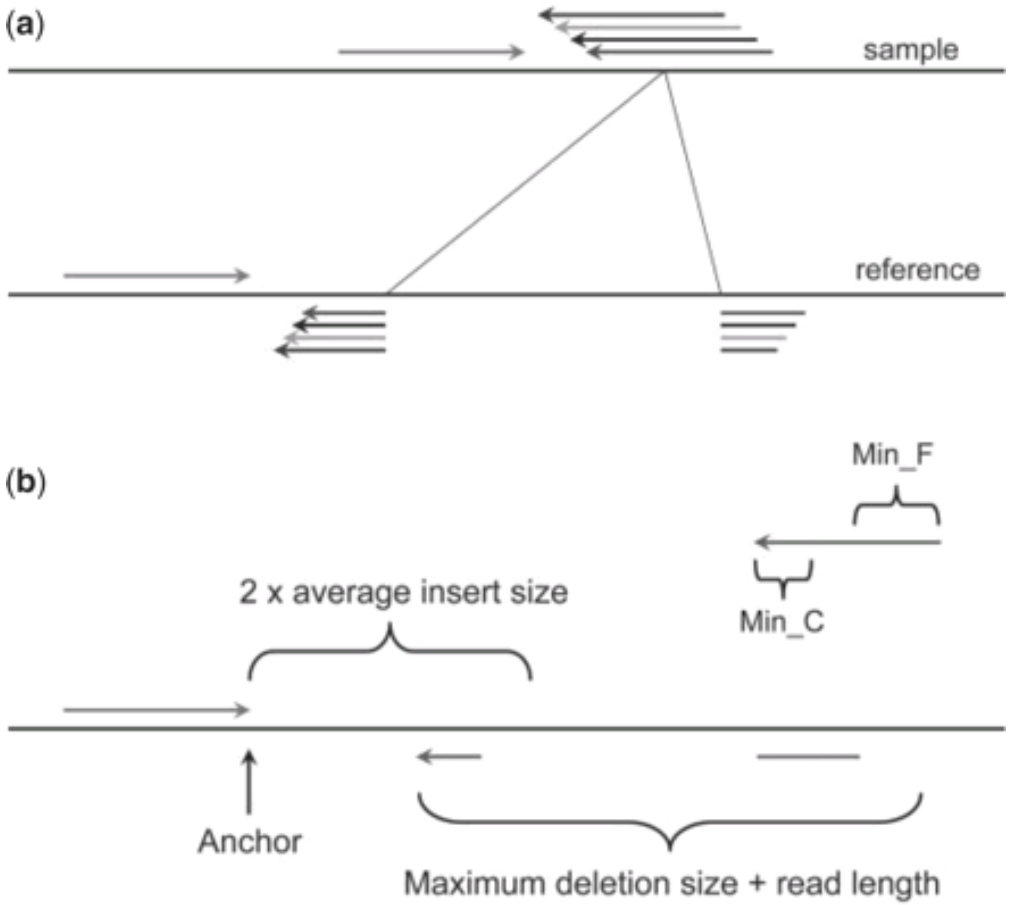
Split Read analysis



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

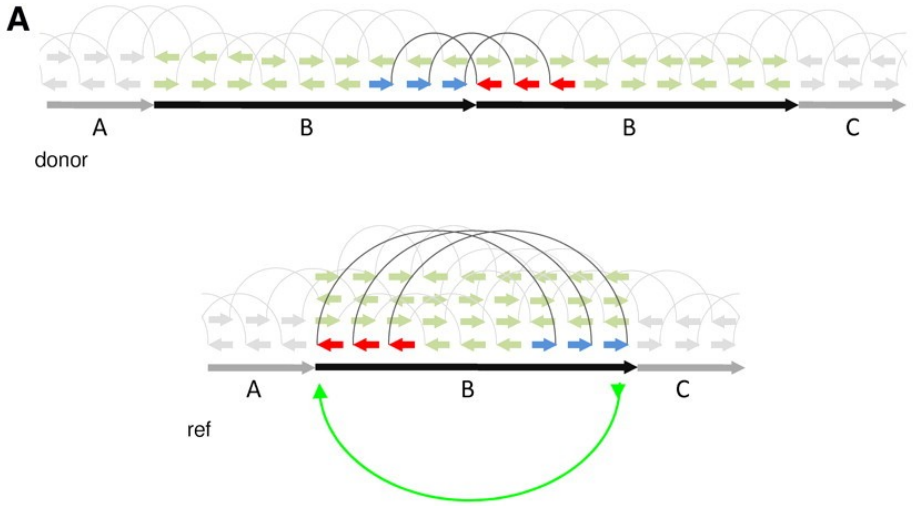
Ye et al. Bioinformatics, 2009

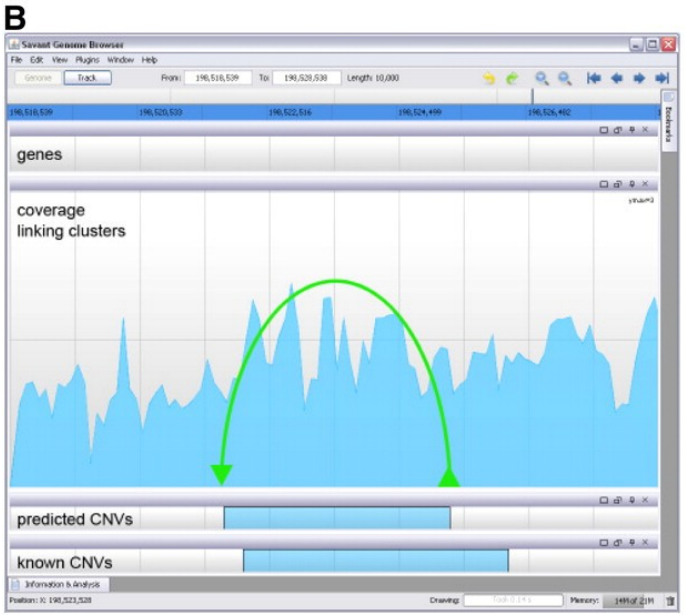
MULTIPLE SIGNATURE

Multiple signature algoritms

- SPANNER (Stewart et al., unpublished)
  - Find candidates with RP
  - Filter with RD
- Genome STRiP (Handsaker et al., Nat Genet, 2011)
  - Discovery: as above; also integrate multiple genomes in a population
  - Genotyping also includes SR
- CNVer (Medvedev et al., Genome Res, 2010)
  - Build a graph with RP; edge weights by RD
  - Solve minimum-cost-flow

CNVer

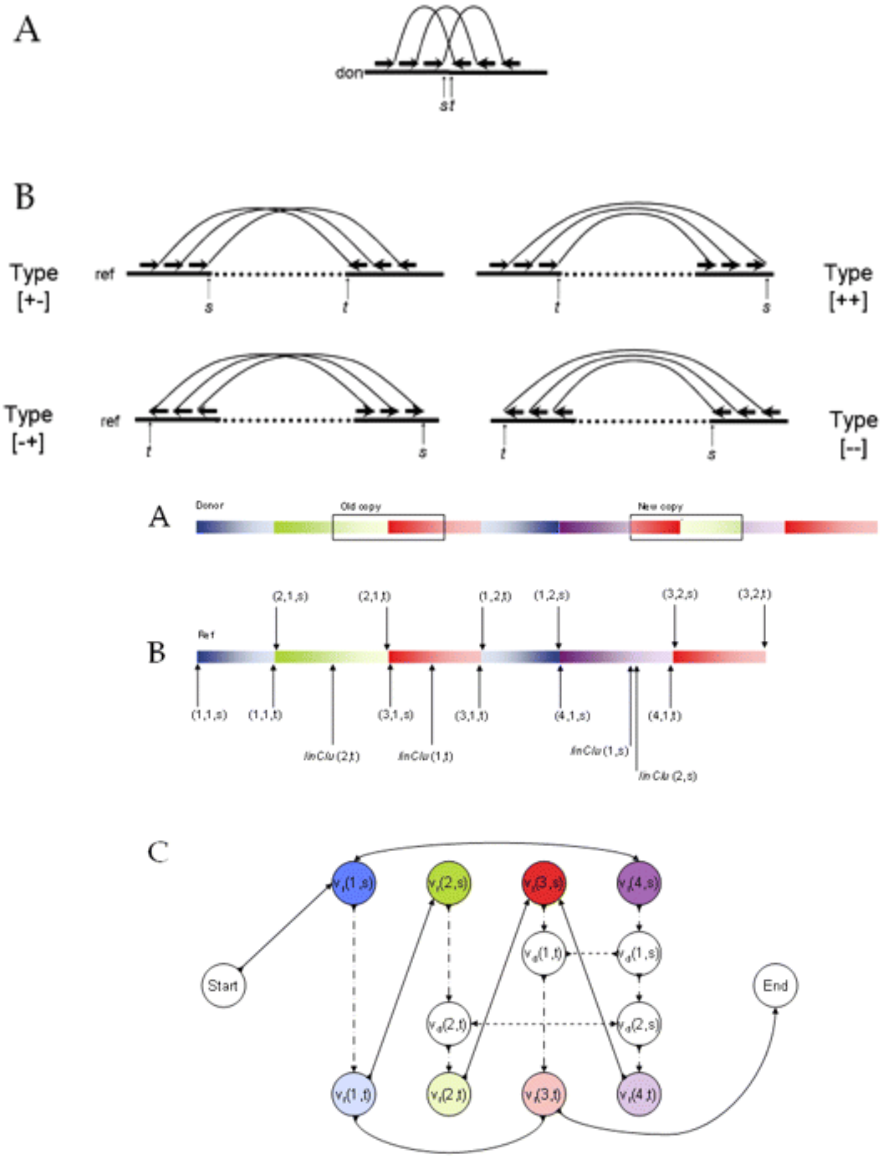




Medvedev et al., Genome Res, 2010

CNVer

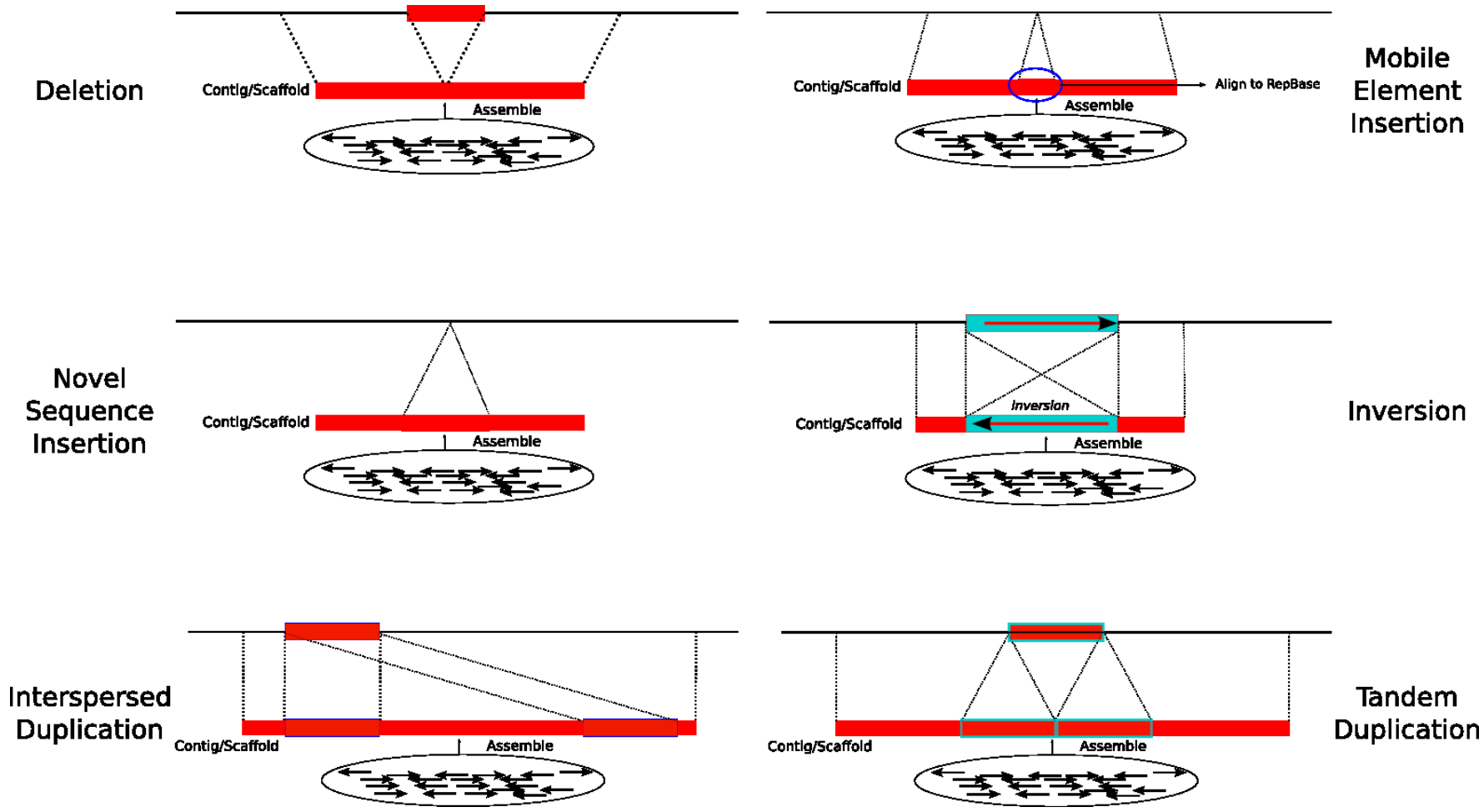
- Build “donor graph” from RP data
- Partition reference genome (self-alignment)
- Probabilistic score to flows in donor graph
  - Length, copy count (unknown variable f_e), and depth (RD data)
- Find minimum cost flow
- Where flow is divergent from reference: CNVs



Medvedev et al., Genome Res, 2010

ASSEMBLY

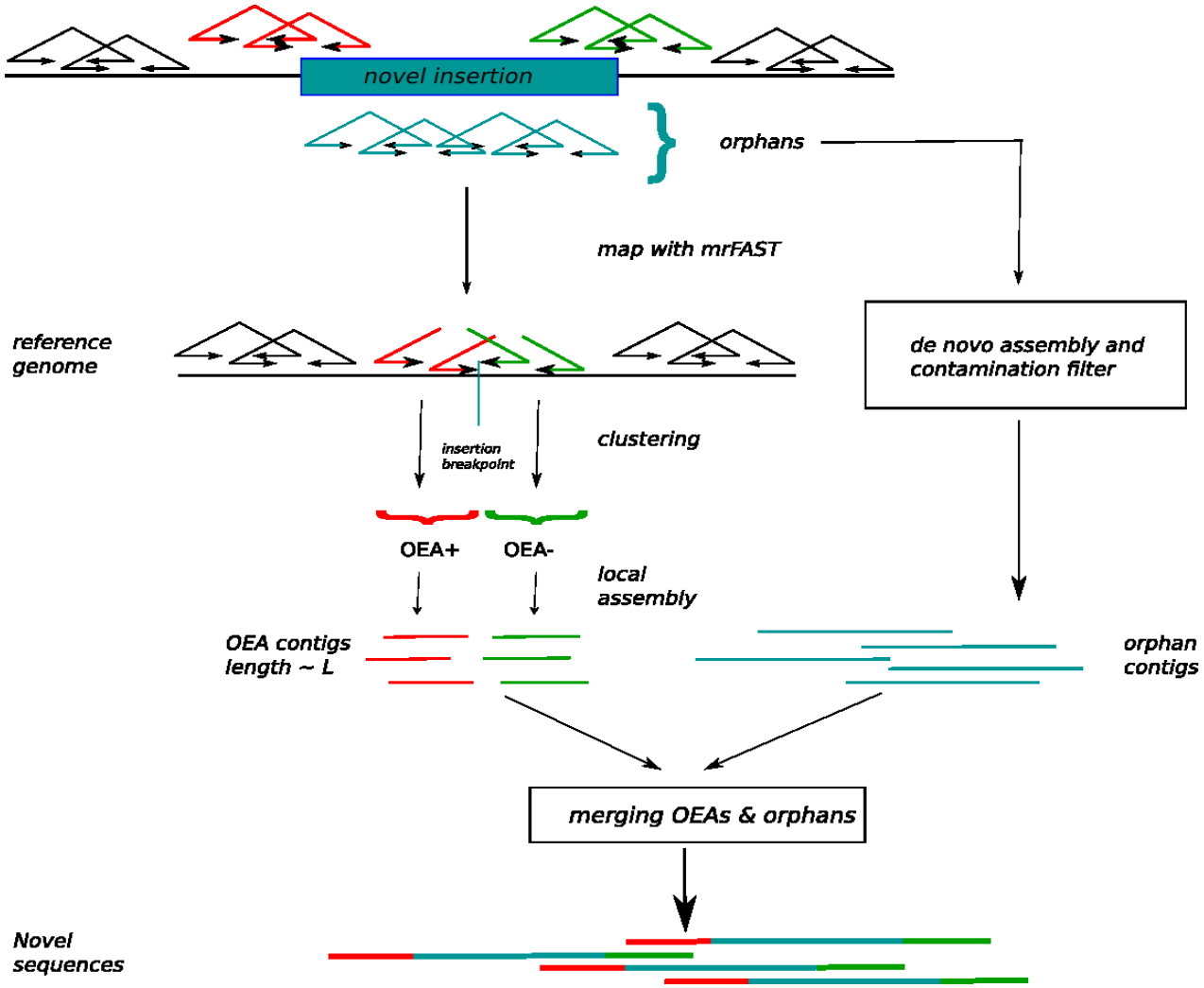
Assembly analysis



Assembly analysis

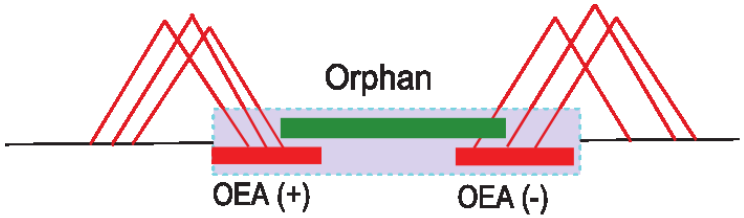
- Collect all reads; and assemble into contigs/scaffolds using:
  - Velvet, EULER, ABySS, Cortex, SOAPdenovo, ALLPATHS-LG, etc.
- Align to reference, and find SV
- SV-specific framework:
  - *NovelSeq: Poor man's method: Going through the trash that the mapper left*

NovelSeq

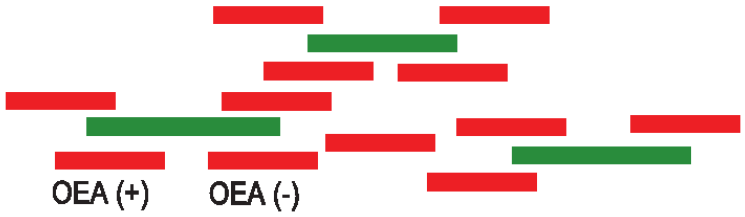


Hajirasouliha et al., Bioinformatics, 2010

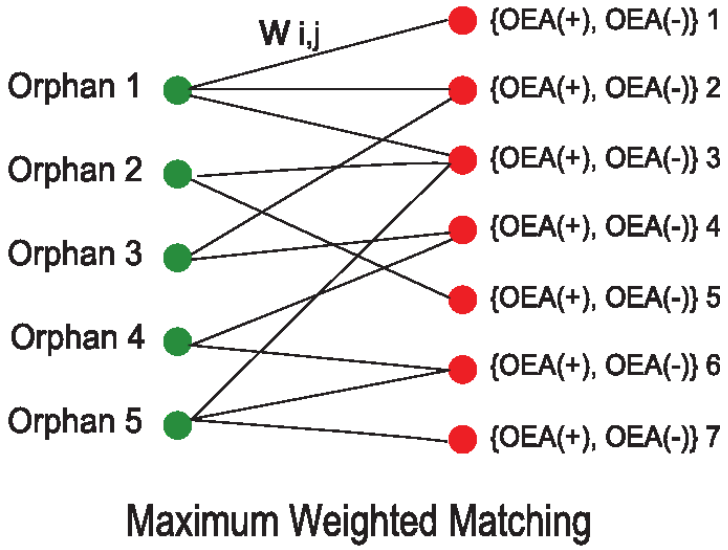
NovelSeq: merging OEA+orphan



Novel insertion {OEA(+), orphan, OEA (-)}



Overlaps between {OEA(+), OEA(-)} and orphan contigs



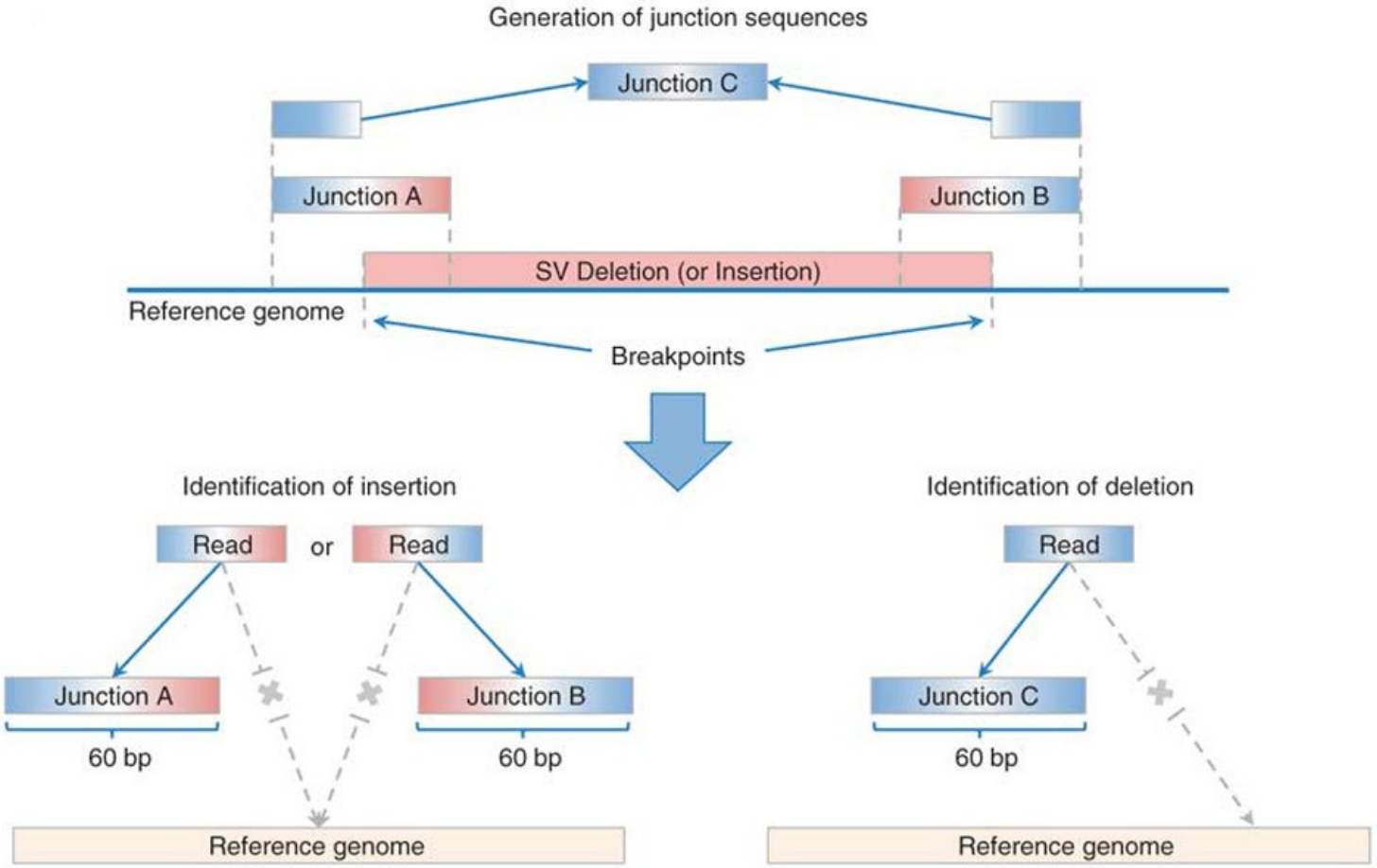
Maximum Weighted Matching

Hungarian Method

Hajirasouliha et al., Bioinformatics, 2010

GENOTYPING SV

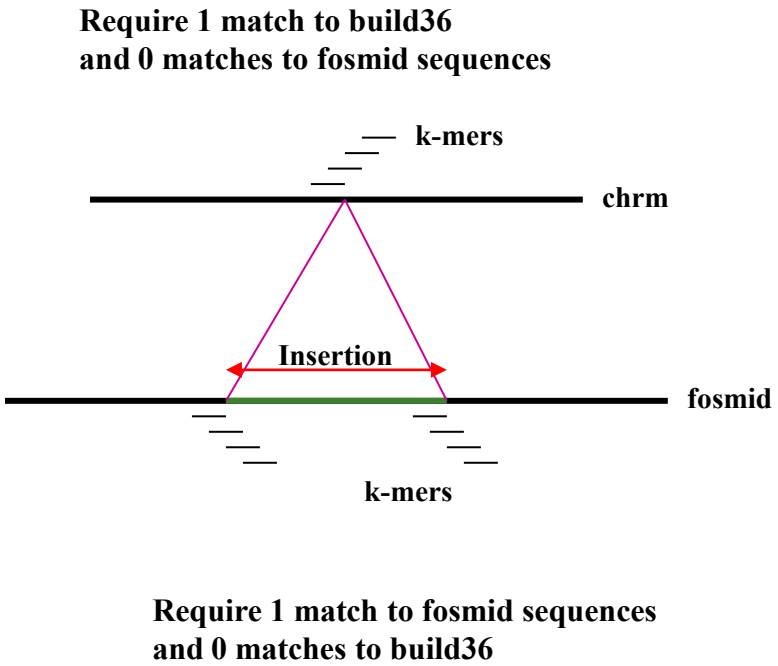
BreakSeq

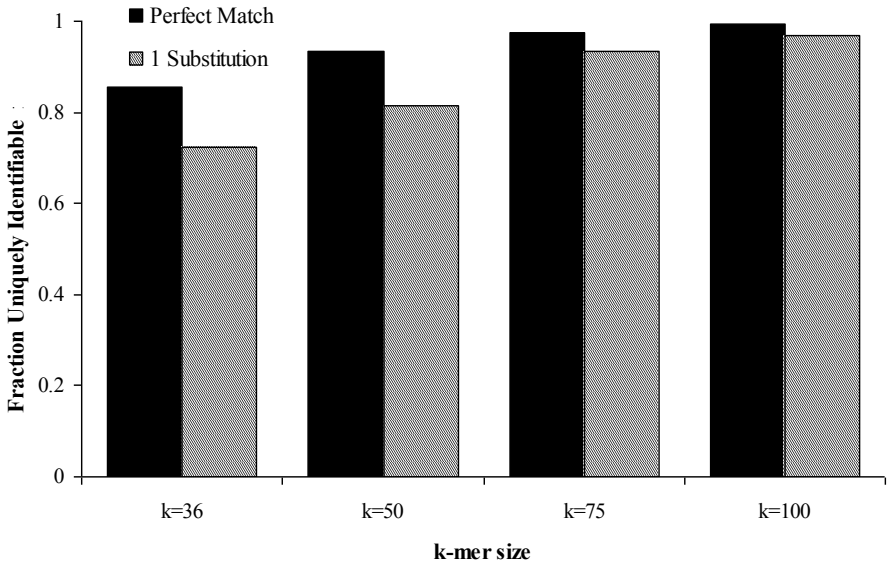


Read overlaps <10 bp to one side of the breakpoint is discarded and read matches also to the reference genome is classified as non-unique match

Lam et al., Nature Biotechnology, 2010

Diagnostic k-mer genotyping





- To be genotyped a variant must be represented by at least 1 insertion and at least 1 deletion k-mer
- 72% (110/152) of targeted variants are uniquely identifiable with k=36 and match criteria that permit 1 substitution

Kidd et al., Nature Methods, 2010

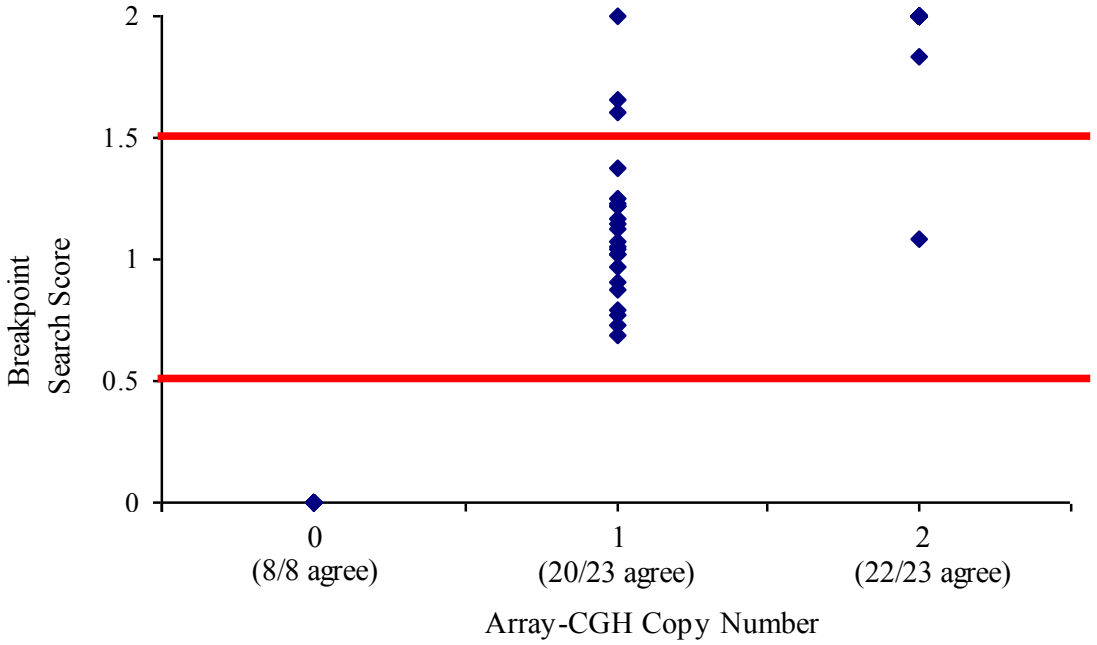
Genotyping insertions with NGS

T_I, T_D: number of diagnostic k-mers for the insertion and deletion alleles

R_I, R_D are the number of matching reads

$$I = \frac{R_I}{T_I} \qquad D = \frac{R_D}{T_D}$$

$$\text{breakpoint search score} = 2\left(\frac{I}{I+D}\right)$$



Kidd et al., Nature Methods, 2010

RESULTS & OPEN PROBLEMS

SV calling in 1000 Genomes

Low coverage data

Approach	Algorithm name	Plat-form	Genomes analyzed	SV types discovered (size-range of validated SVs in basepairs)	SV calls made	SVs validated	FDR (PCR)	FDR (array)	FDR (hierarch.)
RD	**N/A**	Illumina	8	DEL (200 - 77,700)	10,965	1,049	-	0.535	0.535*
	Event-wise testing	Illumina	162	DEL (200 - 67,500)	10,019	3,436	-	0.234	0.234*
	CNVnator	Illumina	65	DEL (200 - 402,150)	5,507	402	-	0.695	0.695*
PE	**Spanner**	Illumina	138	TEINS (56 - 6,049)	3,276	182	0.052	-	0.052
	Spanner	Illumina	138	DEL (53 - 195,139)	5,555	4,615	0.054	0.067	0.059
	PEMer	SOLiD	25	DEL (773 - 184,792)	2,177	1,188	0.258	0.434	0.380
	BreakDancer	Illumina	138	DEL (51 - 959,495)	7,643	4,425	0.337	0.271	0.320
	N/A	Illumina	144	DEL (210 - 959,499)	8,011	5,541	0.214	0.245	0.227
SR	**Mosaik**	454	22	TEINS (300 - 6,000)	2,833	172	0.044	-	0.044*
	Pindel	Illumina	145	DEL (51 - 47,040)	11,189	5,400	0.211	0.309	0.229
	SriC	454	5	DEL (54 - 6,047); INS (51 - 268	10,697	74	0.575	-	0.575*
IN	**Spanner**	Illumina	138	TANDUP (55 – 64,230)	407	55	0.125	-	0.125*
	Genome STRiP	Illumina	168	DEL (100 - 471,351)	7,015	5,852	0.057	0.019	0.037

1000 Genomes Consortium, Nature, 2010

SV calling in 1000 Genomes: sensitivity

Low coverage data

Supplementary Table 6A. Sensitivity in discovering deletions for different methods, assessed in NA12156(*)

Approach	Callset Origin	Algorithm	Sequencing platform	Kidd (n=54)	Conrad (n=353)	McCarroll (n=118)	Mills (n=151)
RD	SD	Event-wise testing	Illumina	0.46	0.65	0.70	0.06
	YL	CNVnator	Illumina	0.20	0.19	0.31	0.09
RP	BC	Spanner	Illumina	0.26	0.19	0.17	0.21
	SI	N/A	Illumina	0.30	0.28	0.25	0.21
	YL	PEMer	SOLiD	0.11	0.28	0.09	0.03
	WU	BreakDancer	Illumina	0.20	0.20	0.18	0.17
	LN	Pindel	Illumina	0.13	0.08	0.13	0.10
PD	BI	Genome STRiP	Illumina	0.63	0.50	0.40	0.21

Mills et al., Nature, 2011

SV calling in 1000 Genomes

High coverage data

Approach	Algorithm name	Platform	Genomes	SV types discovered (size-range of validated SVs in basepairs)	SV calls	validated	FDR (PCR)	FDR (array)	FDR (hierarch.)
RD	**Event-wise testing**	Illumina	6	DEL (200 - 221,800); DUP (200 - 415,700)	5,762	1,952	0	0.230	0.230
	CNVnator	Illumina	6	DEL (100 - 412,475)	17,036	2,361	-	0.142	0.142
PE	**AB large indel tool**	SOLiD	1	DEL (67 - 83,391)	1,138	480	0.188	0.084	0.143
	AB large indel tool	SOLiD	1	INS (448 - 2,213)	632	42	0.176	-	0.176
	Spanner	Illumina	6	TEINS (51 - 6,012)	2,013	179	0.022	-	0.022
	Spanner	Illumina	6	DEL (50-192,167)	4,718	3,619	0.100	0.033	0.087
	PEMer	454	1	DEL (941 - 960,004)	1,062	483	0.095	0.363	0.363
	VariationHunter	Illumina	6	DEL (52 - 498,738)	11,028	4,231	0.103	0.419	0.190
	BreakDancer	Illumina	6	DEL (51 - 1,035,808)	5,973	3,587	0.115	0.145	0.121
	N/A	Illumina	6	DEL (276 - 959,518)	3,419	2,584	0.136	0.085	0.121
SR	**Mosaik**	454	2	TEINS (300 - 6,000)	1,463	172	0.055	-	0.055
	Pindel	Illumina	6	DEL (51 - 46,384)	3,879	2,960	0.201	0.127	0.189
	N/A	454	1	DEL (51 - 703,404); INS (52 - 295)	32,187	3,845	0.545	0.519	0.543
AS	**SOAPdenovo**	Illumina	6	DEL (64 - 3,907)	160	55	0.531	0.531	0.497
	SOAPdenovo	Illumina	6	INS (55 – 4,116)	3,894	22	0.810	-	0.810
	Cortex	Illumina	1	DEL(52-39,512);DUP(83-2,090)	2,787	896	0.415	0.415	0.410
	Cortex	Illumina	1	INS(50-828)	389	84	0.398	-	0.398
	NovelSeq	Illumina	6	INS (200 - 8,224)	657	30	0.791	-	0.791
IN	**Spanner**	Illumina	6	TANDUP (55-64,230)	256	88	0.049	-	0.049

1000 Genomes Consortium, Nature, 2010

SV calling in 1000 Genomes: sensitivity

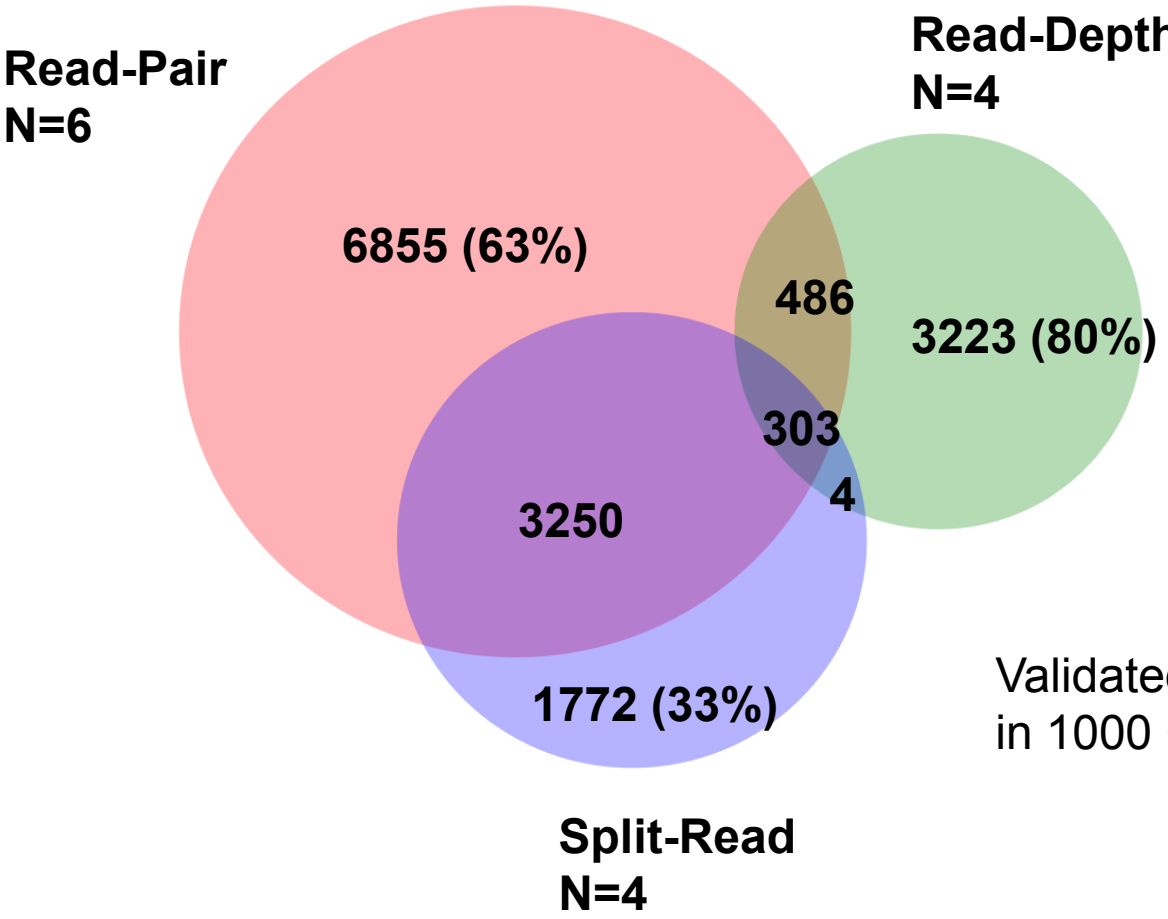
High coverage data

Supplementary Table 6B. Sensitivity in discovering deletions for different methods, assessed in NA12878(*)

Approach	Callset Origin	Algorithm name	Sequencing platform	Kidd (n=58)	Conrad (n=373)	McCarroll (n=130)	Mills (n=81)
RD	SD	Event-wise testing	Illumina	0.67	0.56	0.80	0.05
	UW	mrFAST	Illumina	0.16	0.07	0.22	0.00
	YL	CNVnator	Illumina	0.91	0.84	0.88	0.24
RP	BC	Spanner	Illumina	0.45	0.50	0.32	0.44
	SI	N/A	Illumina	0.50	0.55	0.42	0.24
	UW	VariationHunter	Illumina	0.55	0.53	0.50	0.30
	WU	BreakDancer	Illumina	0.50	0.55	0.44	0.40
	YL	PEMer	454	0.91	0.45	0.72	0.10
SR	LN	Pindel	Illumina	0.28	0.38	0.25	0.28
	YL	N/A	454	0.55	0.54	0.44	0.52

Mills et al., Nature, 2011

No method is comprehensive



Validated deletions
in 1000 Genomes data

Kidd et al., Cell, 2010

Open problems

- Identify ***inversions*** and ***translocations***
- Discover SVs in repeat- and duplication-rich regions
- Accurate & comprehensive detection of CNVs with a *single* algorithm
  - High sensitivity
  - High specificity