CS681: Advanced Topics in Computational Biology

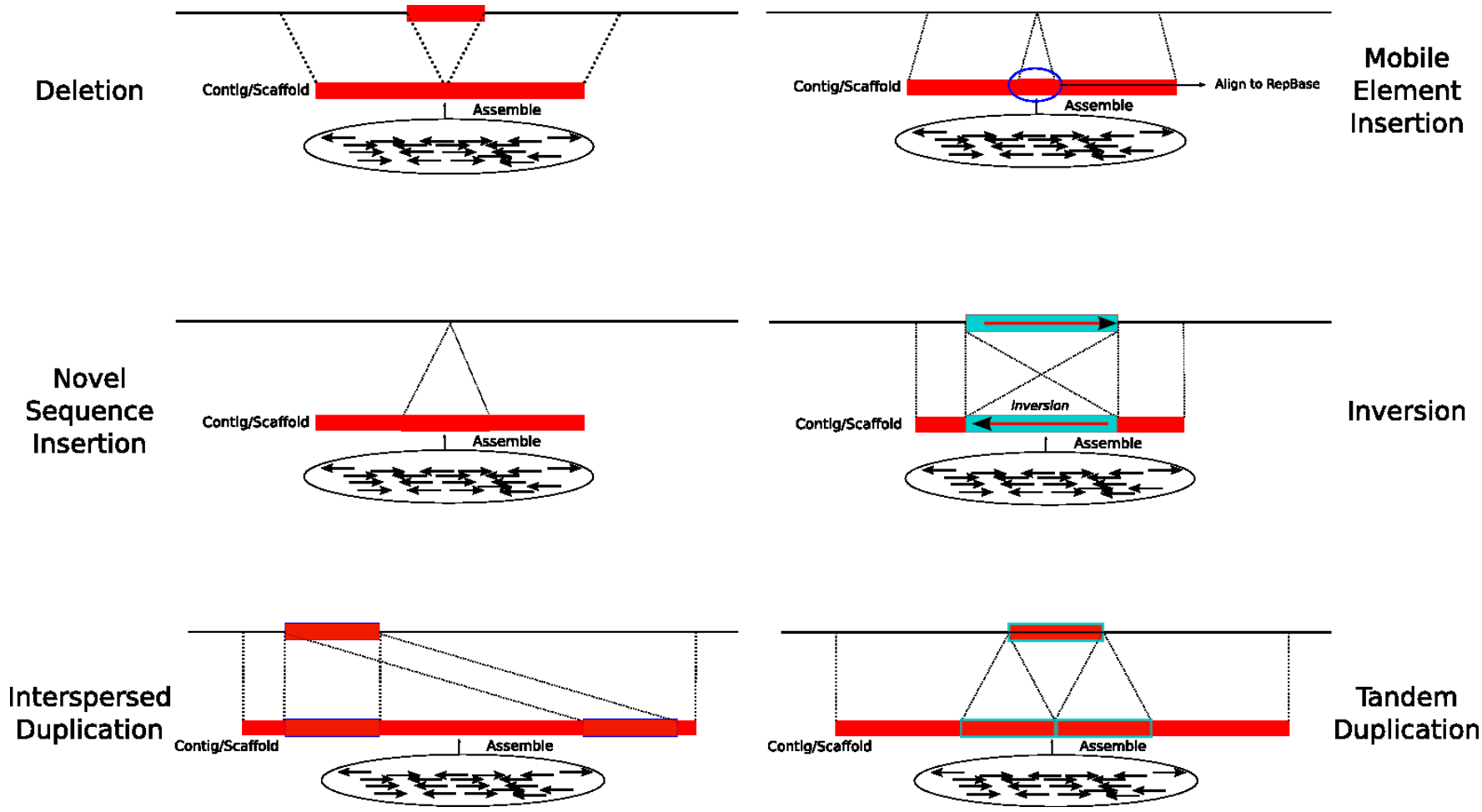
Week 6 Lectures 2-3

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

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

ASSEMBLY & SV CALLING

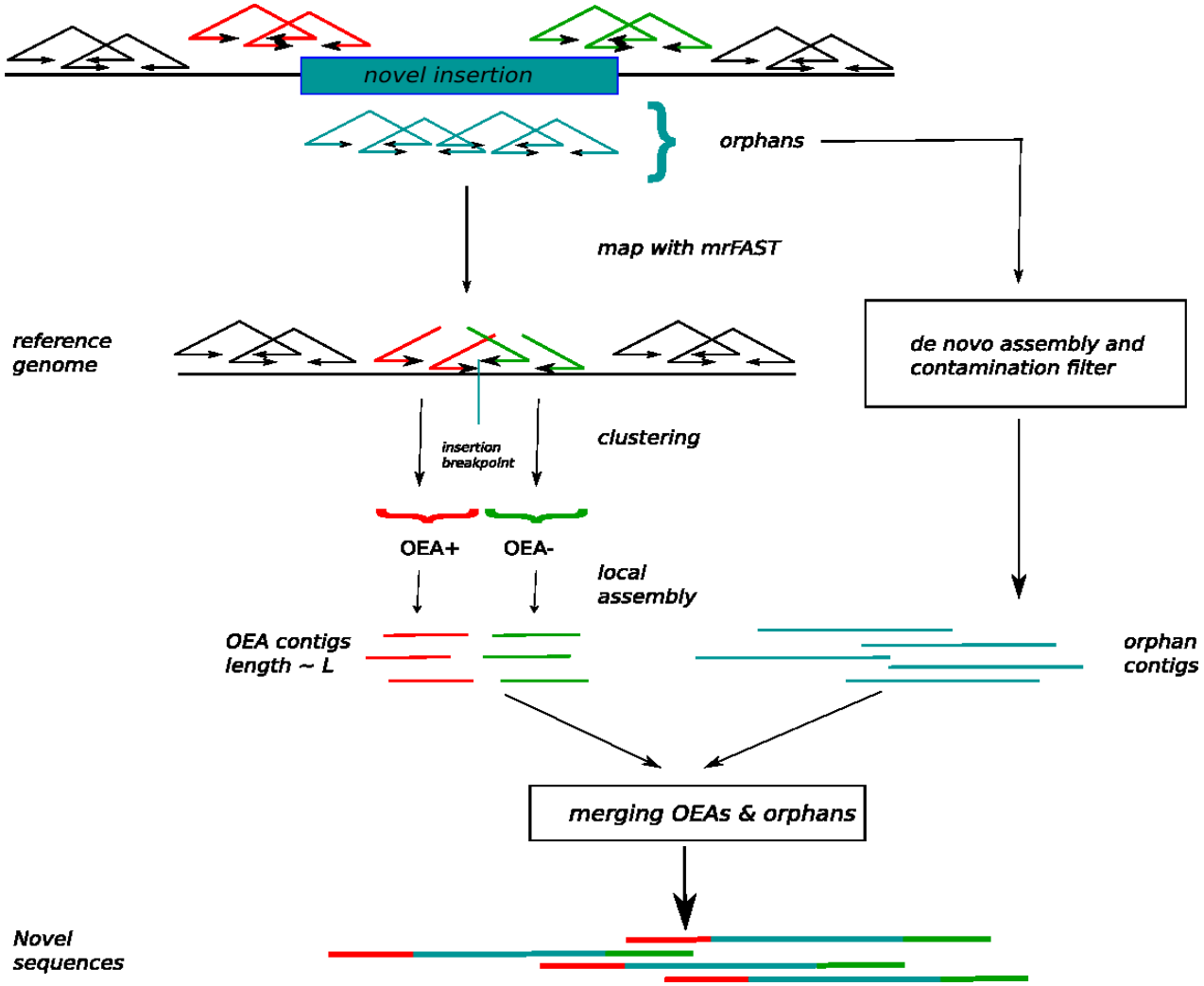
Assembly analysis



Assembly analysis

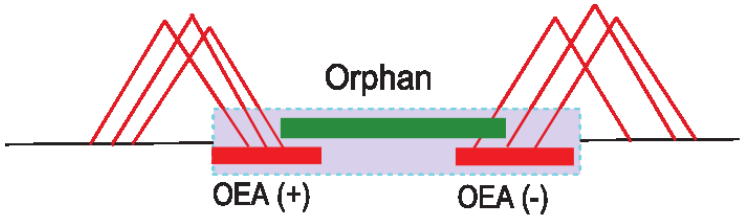
- Collect all reads; and assemble into contigs/scaffolds using:
  - Velvet, EULER, ABySS, Cortex, SOAPdenovo, ALLPATHS-LG, etc.
- Align to reference, and find SV
- SV-specific framework:
  - *NovelSeq (Hajirasouliha et al., 2010)*
  - *Pamir (Kavak et al., 2017)*
  - *PopIns (Kehr et al., 2016)*

NovelSeq

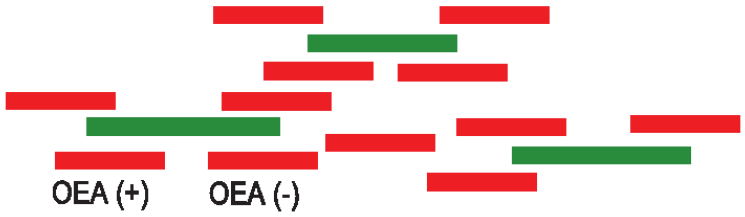


Hajirasouliha et al., Bioinformatics, 2010

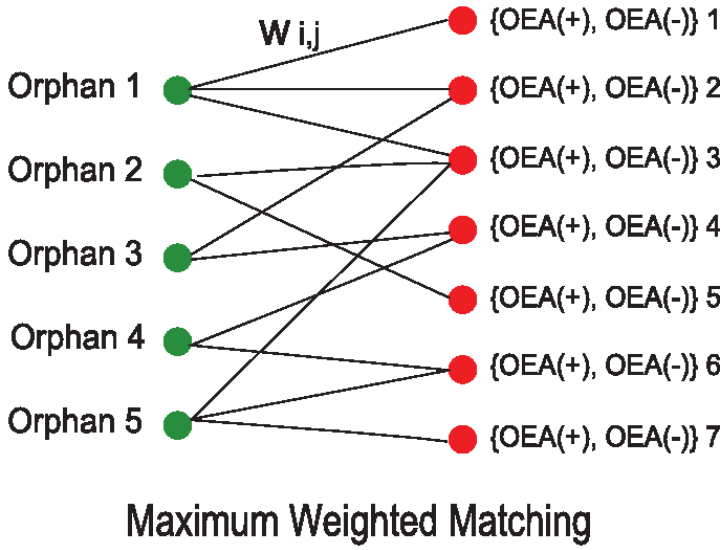
NovelSeq: merging OEA+orphan



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



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



Maximum Weighted Matching

Hungarian Method

Hajirasouliha et al., Bioinformatics, 2010

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	)EL (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

1000 GENOMES SV

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

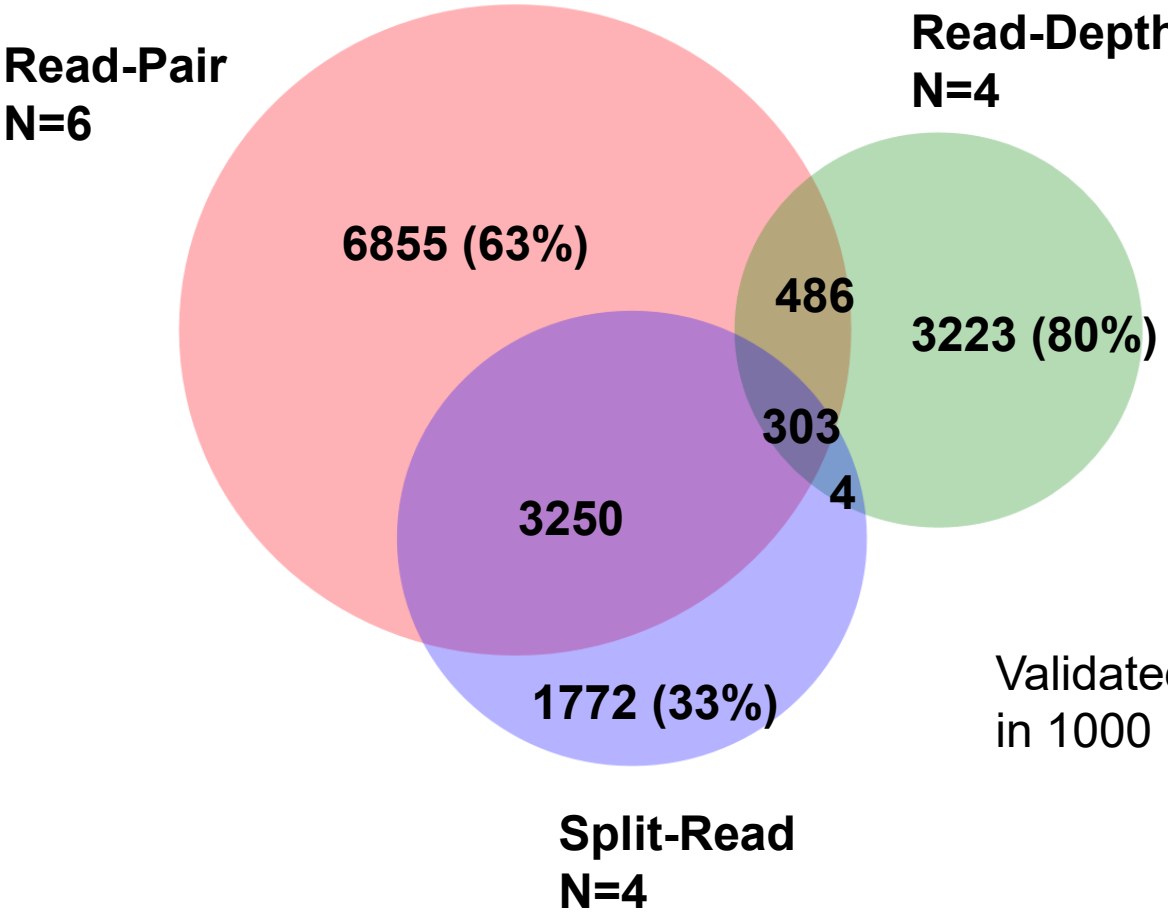
High coverage data

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

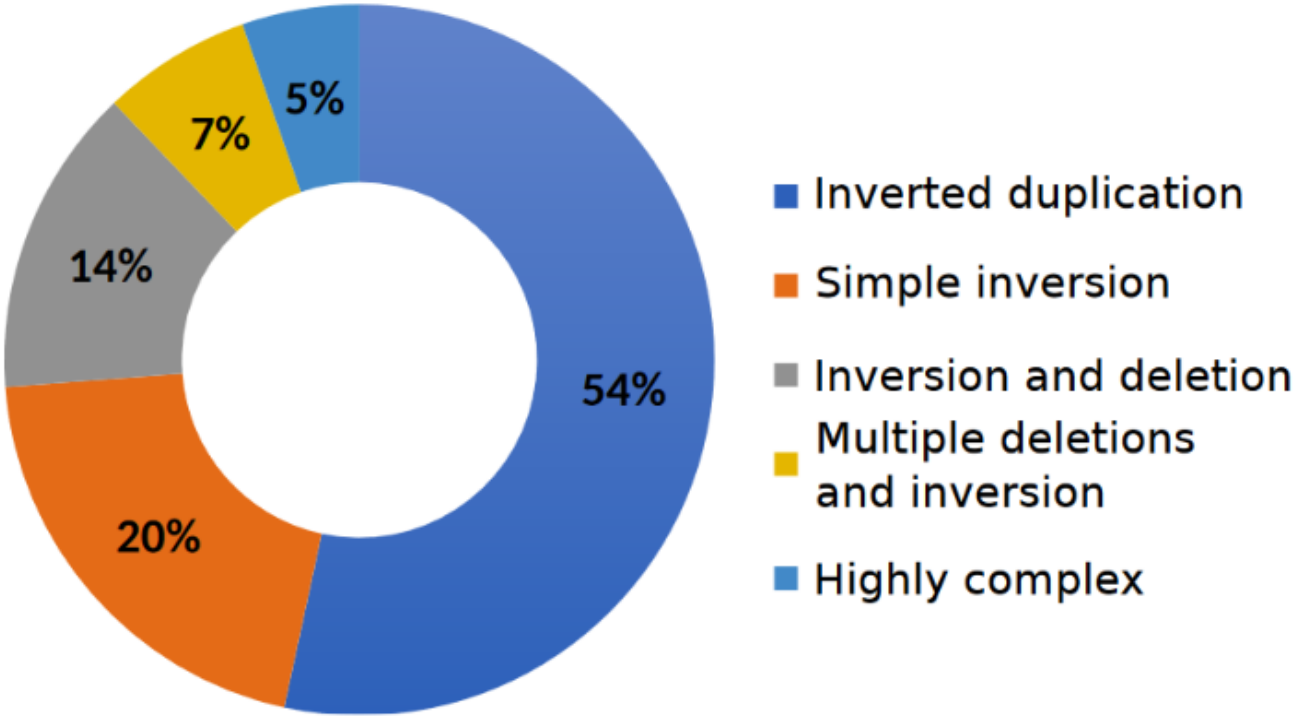


Validated deletions
in 1000 Genomes data

Kidd et al., Cell, 2010

SV: MULTIPLE SIGNATURES

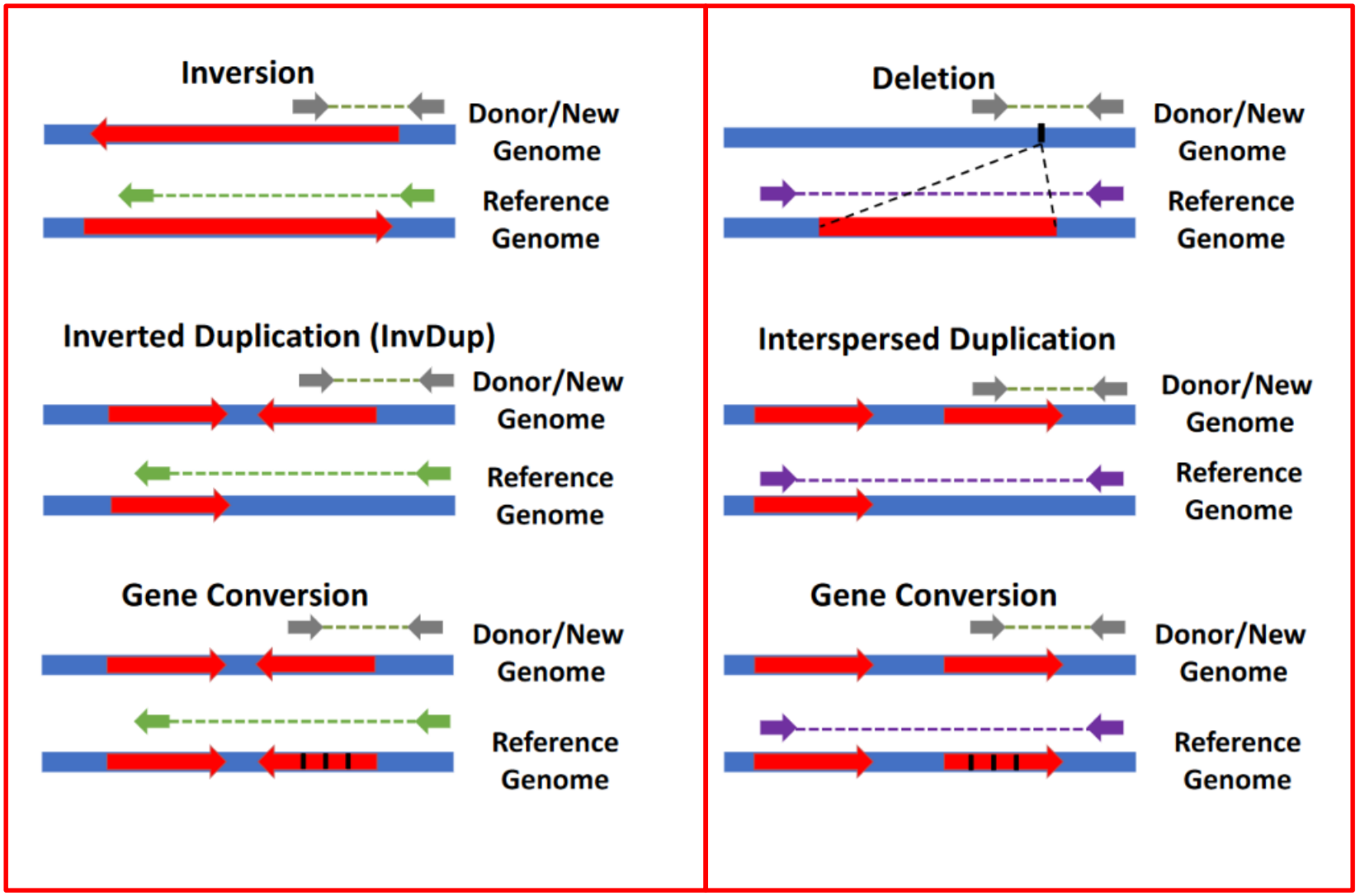
Complex SVs are hard(er)



Reported as inversions by the 1000 Genomes Projects

Sudmant et al, Nature 2015

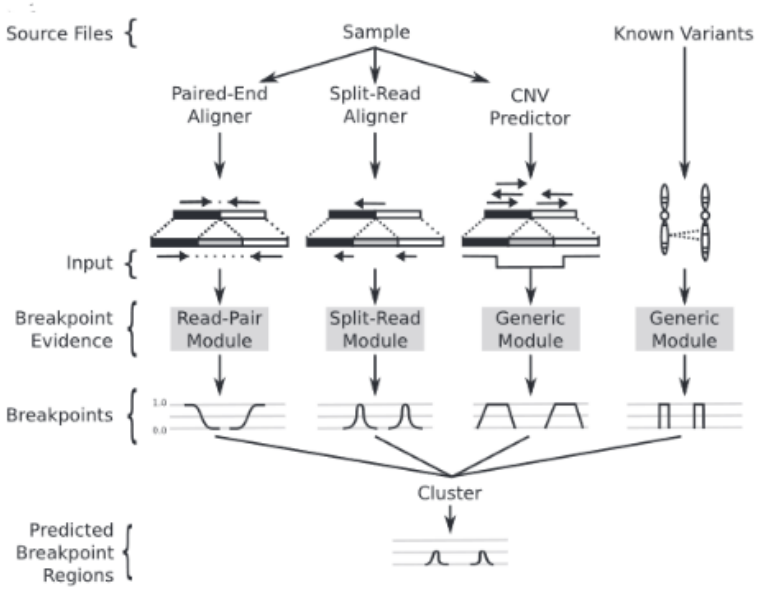
Inversions & inverted duplications



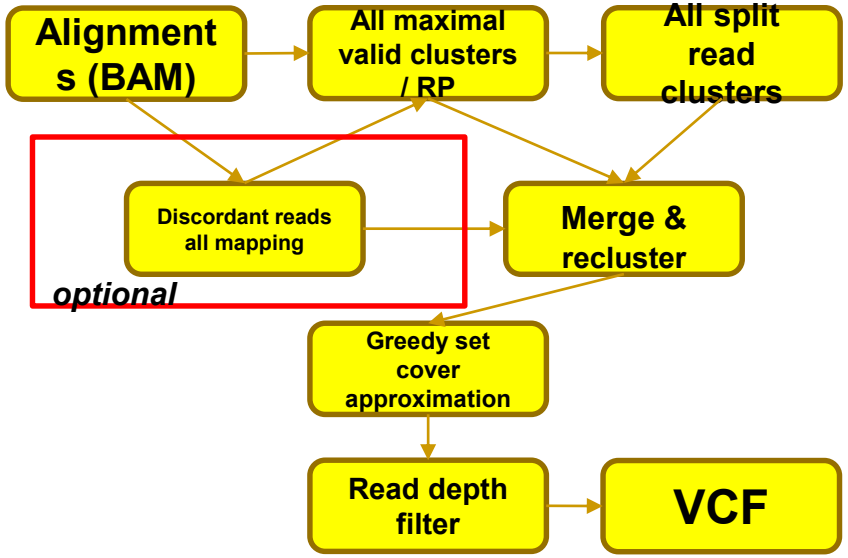
Soylev et al, Bioinformatics 2019

Multi-signature SV callers

- Integrate (combinations of) read pair, read depth, split read, local assembly



LUMPY (Layer, 2014)



TARDIS (Soylev, 2019)

Multi-signature SV callers

- RP+RD:
  - Genome STRiP (Handsaker 2011)
  - SV-Bay (Iakovishina 2016)
- RP+SR:
  - DELLY (Rausch 2012)
- RP+AS:
  - SvABA (Wala 2018) [+RD for dels < 300 bp)
- RP+RD+SR:
  - LUMPY (Layer 2014)
  - Wham (Kronenberg 2015)
  - TIDDIT (Eisfeldt 2017)
  - TARDIS (Soylev 2019)
- RP+SR+AS:
  - Manta (Chen 2016)
  - GRIDDS (Cameron 2017)

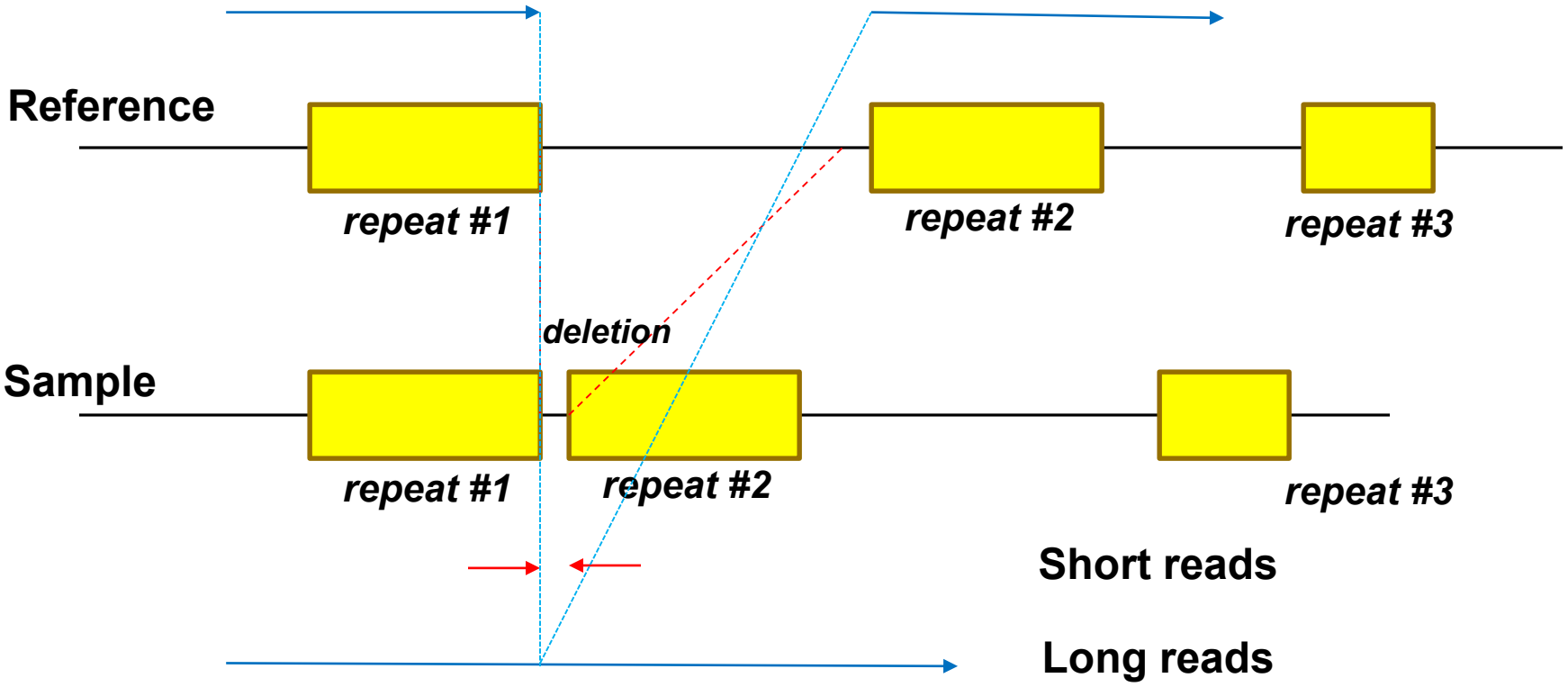
STRUCTURAL VARIATION – ENSEMBLE ALGORITHMS

Ensemble algorithms

- Aim to combine and integrate SV call sets generated by multiple algorithms
- Different approaches:
  - Intersection: overlap or union
  - Validation: split reads, local assembly, or signature prioritization
  - Data mining / machine learning methods using truth sets
- Tools:
  - SpeedSeq (Chiang et al., 2015), svtools (Larson et al., 2018)
  - HugeSeq (Lam et al., 2012), SVMerge (Wong et al., 2010)
  - Parliament2 (Zarate et al., 2018), FusorSV (Becker et al., 2018)
- For exomes: CN-Learn (Pounraja et al., 2019)

STRUCTURAL VARIATION USING LONG READS

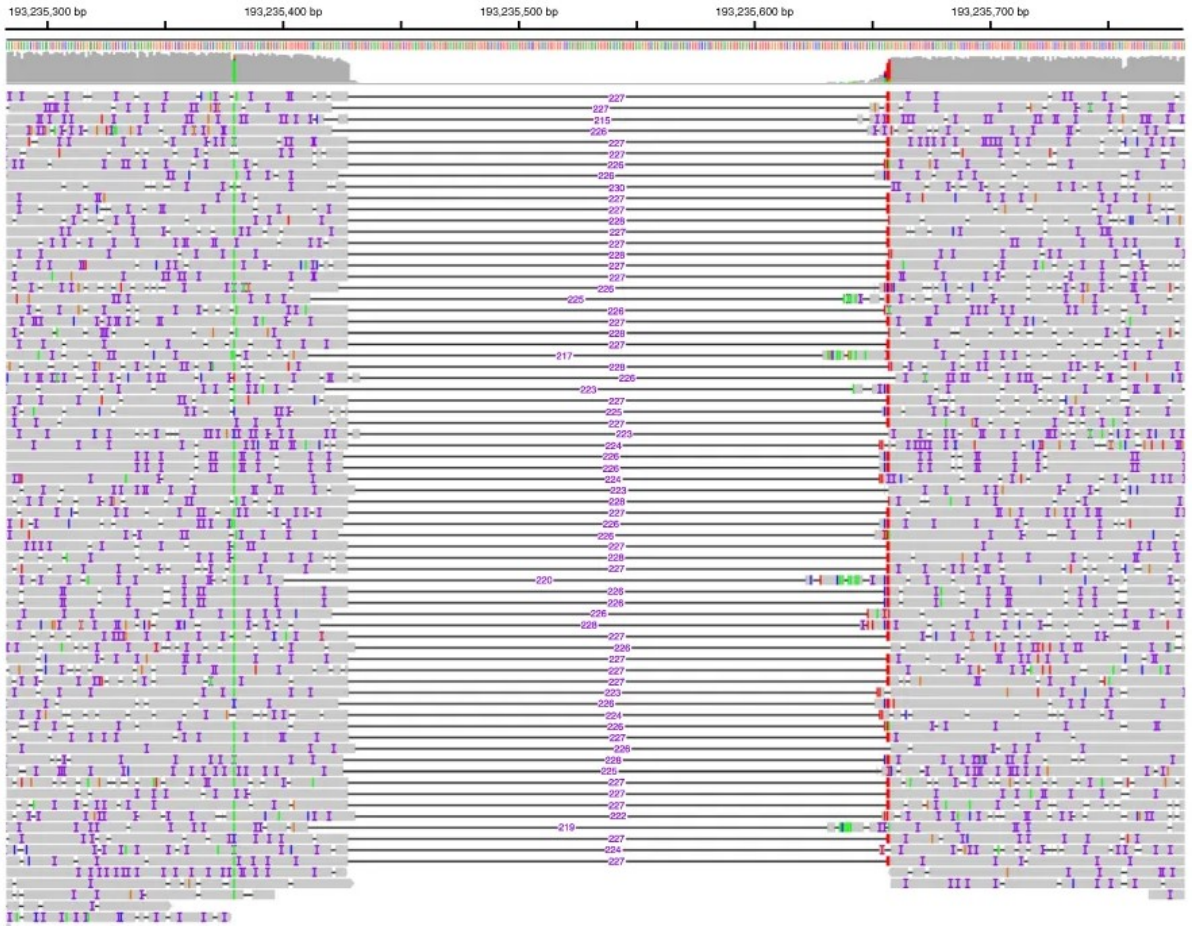
Mapping short vs. long reads



SV callers using long reads

SR/AS based

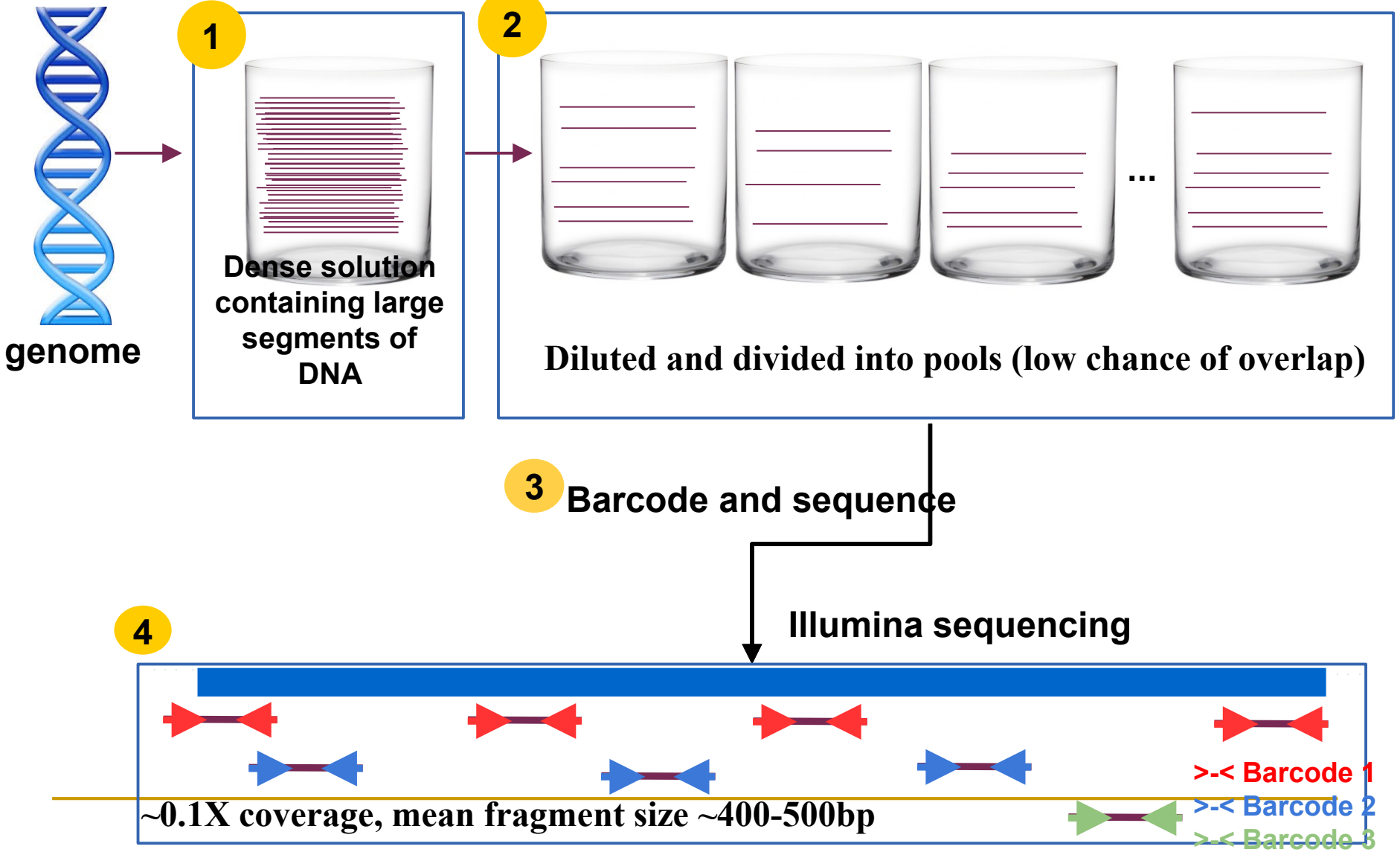
- Sniffles
- SMRT-SV
- NanoSV
- NextSV
- CORGi
- SVIM
- Picky



Sedlazeck et al., 2018

STRUCTURAL VARIATION USING LONG RANGE INFORMATION

Long Range Information: Linked-Reads

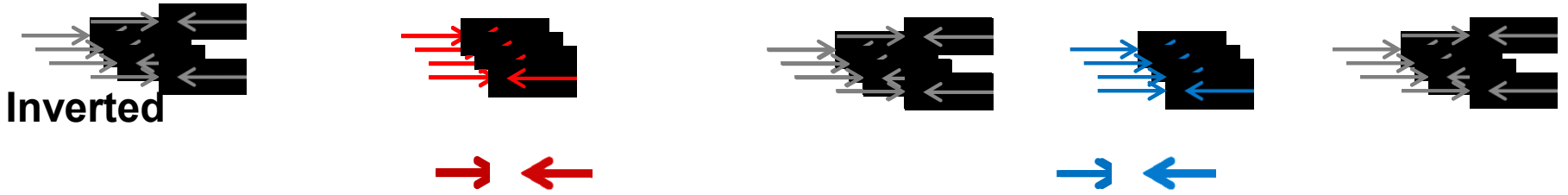


10x Genomics Linked-Reads

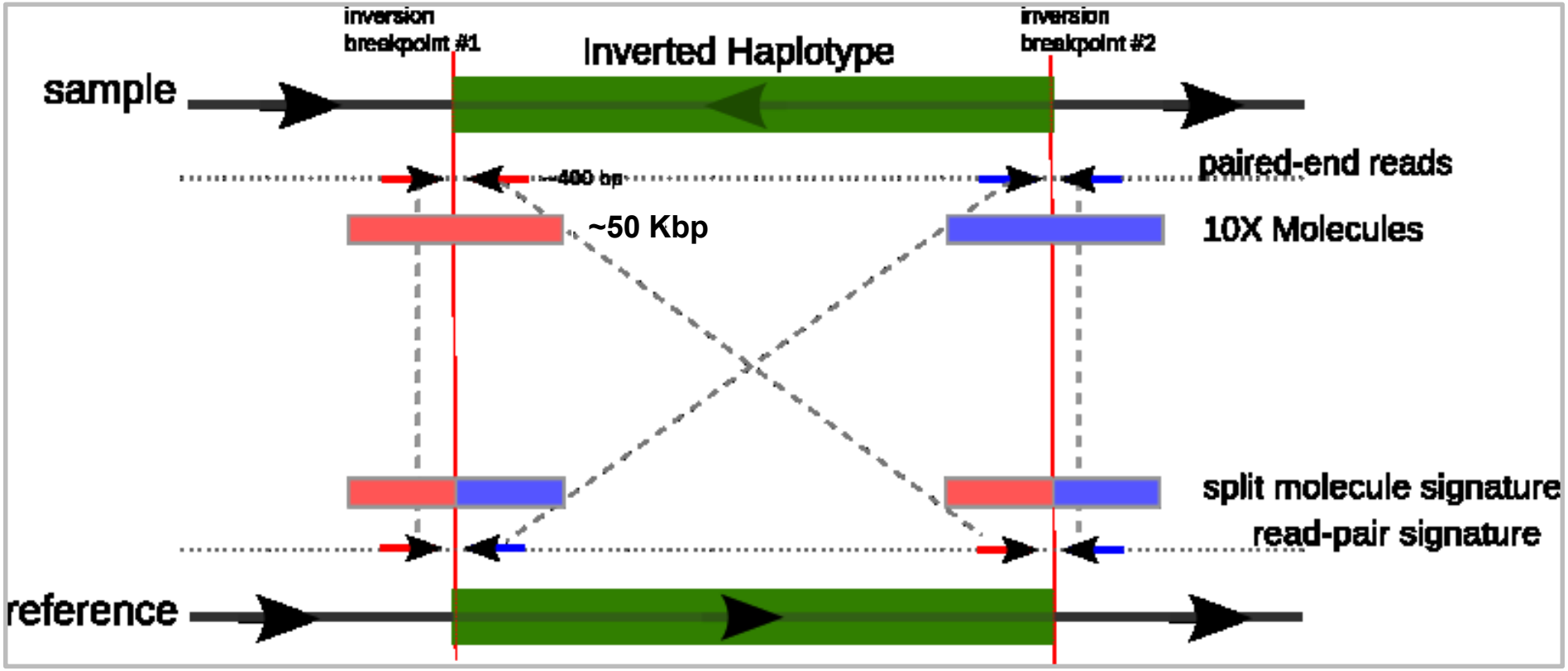
- ~50 Kb (average) molecules
- No cloning required
  - Automated process
  - No cloning bias, but not tight size distribution
- ~0.1x coverage per molecule
- Up to 4M pools
  - ~20 molecules per pool
- SV callers:
  - Long Ranger, **VALOR**, NAIBR, GROC-SVs, Novel-X, ZoomX

Example: inversion signature

Referenc
e

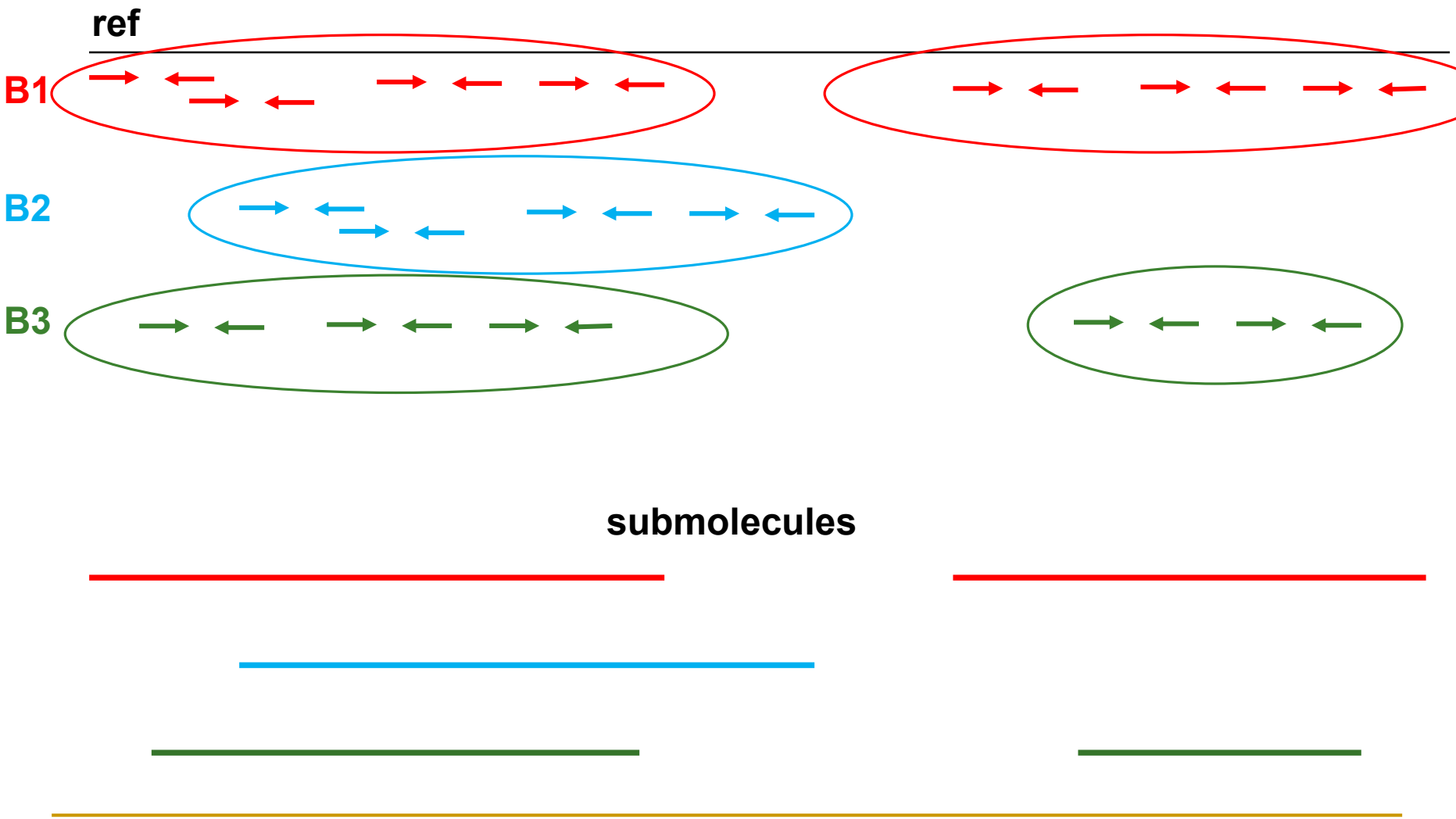


Split molecule signature for inversions



Eslami Rasekh et al, 2017

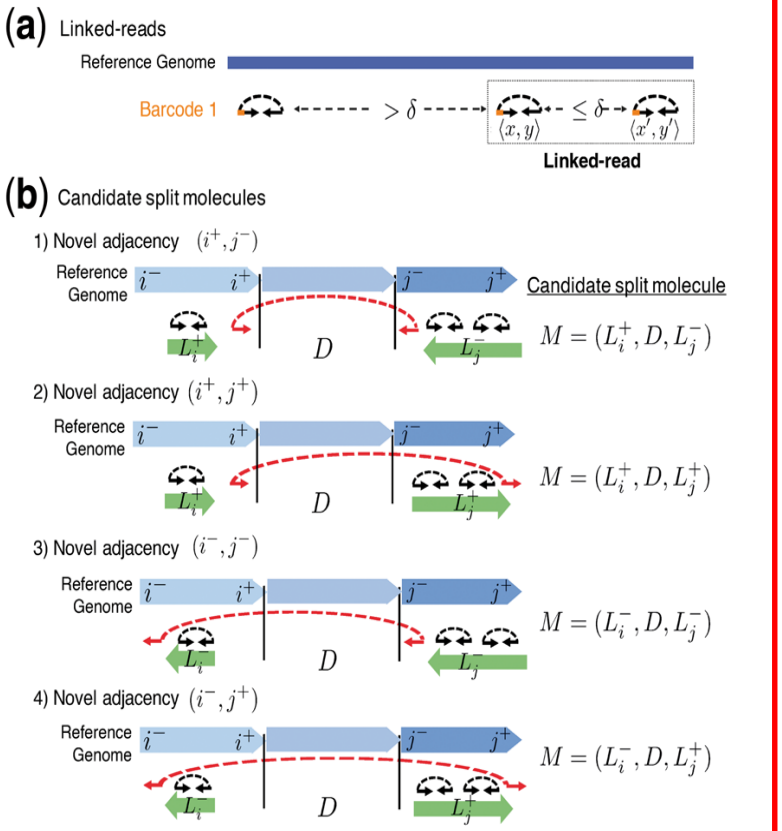
Identifying submolecules



Similar method in NAIBR (Elyanow, 2018)

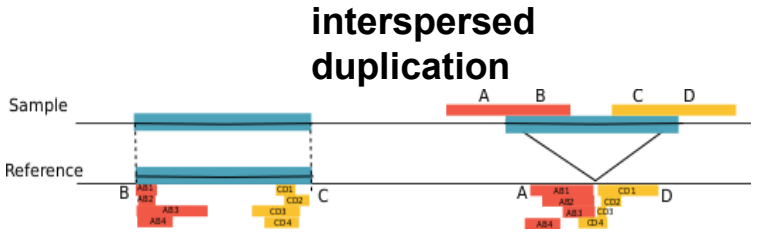
Eslami Rasekh et al., 2017, Karaoglanoglu et al., 2018

SV detection from split molecules

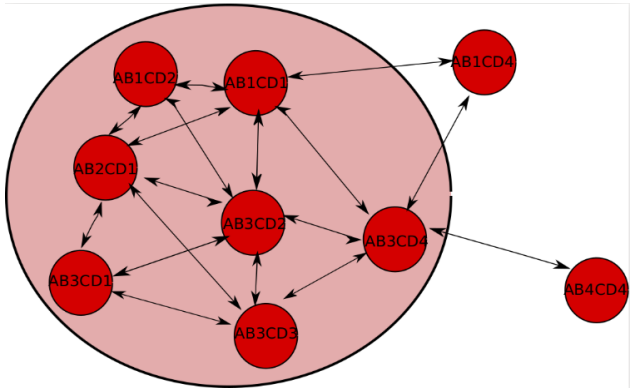


NAIBR

Elyanow et al., 2018



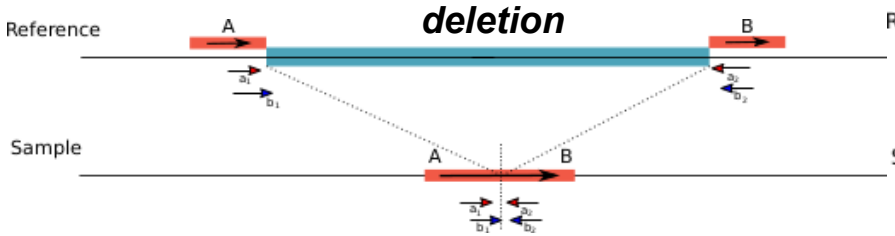
SV Graph: ***Max quasi-clique***

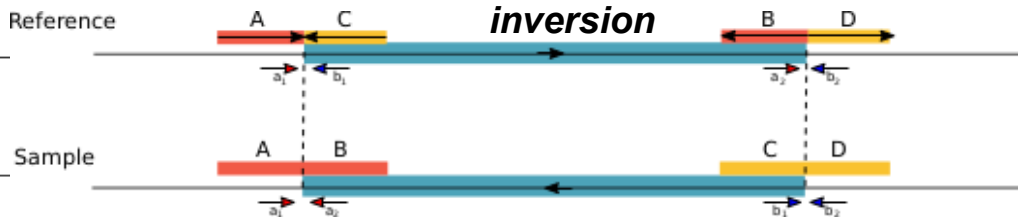


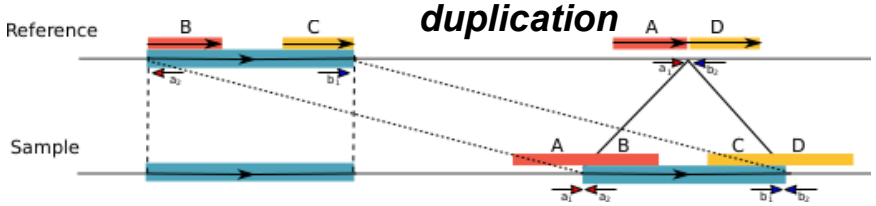
VALOR & VALOR2

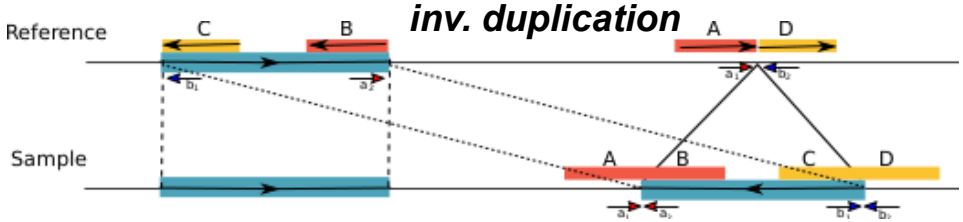
Eslami Rasekh et al., 2017, Karaoglanoglu et al., 2018

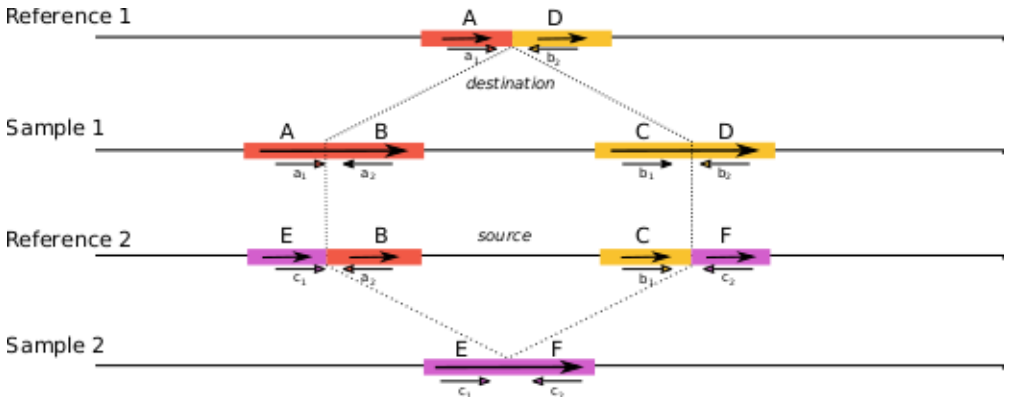
Identifiable SV types









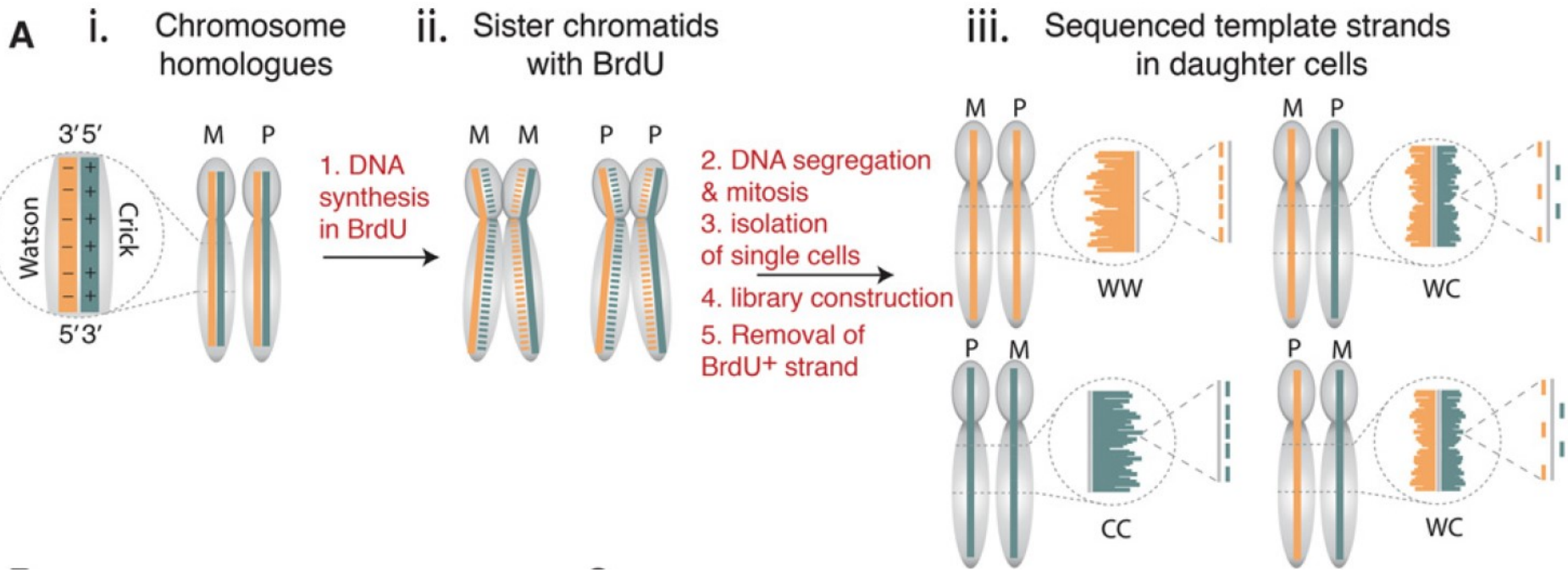


translocation

Karaoglanoglu et al., 2018

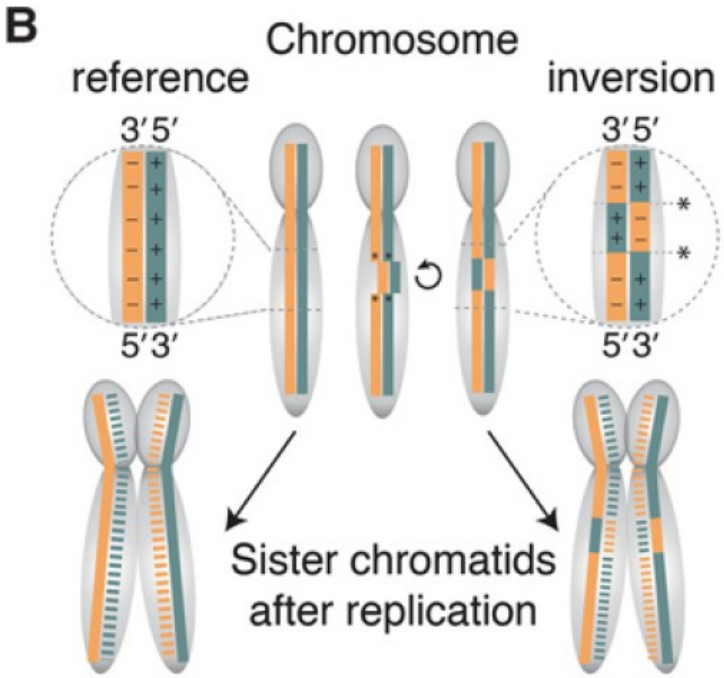
EMERGING TECHNOLOGIES

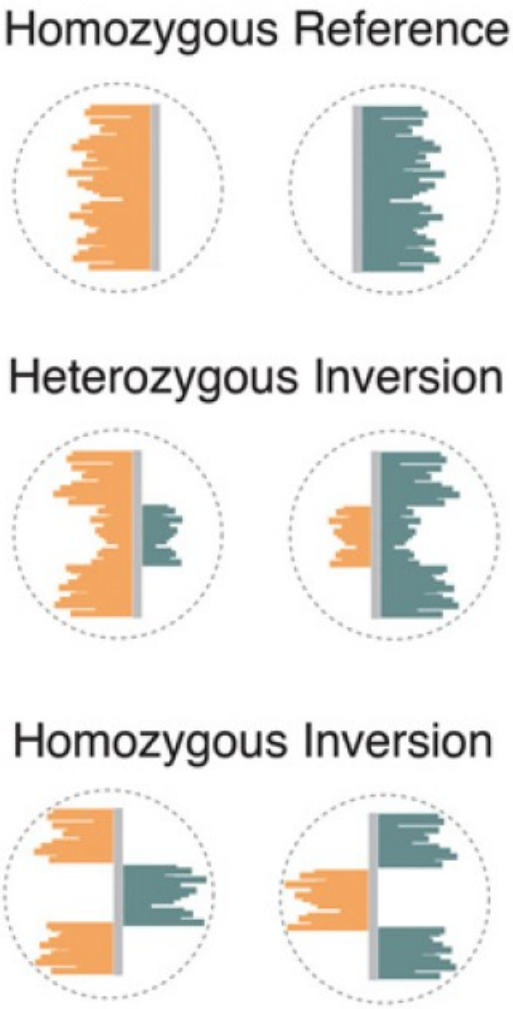
Strand-Seq



Sanders et al., Genome Res 2016

Strand-Seq

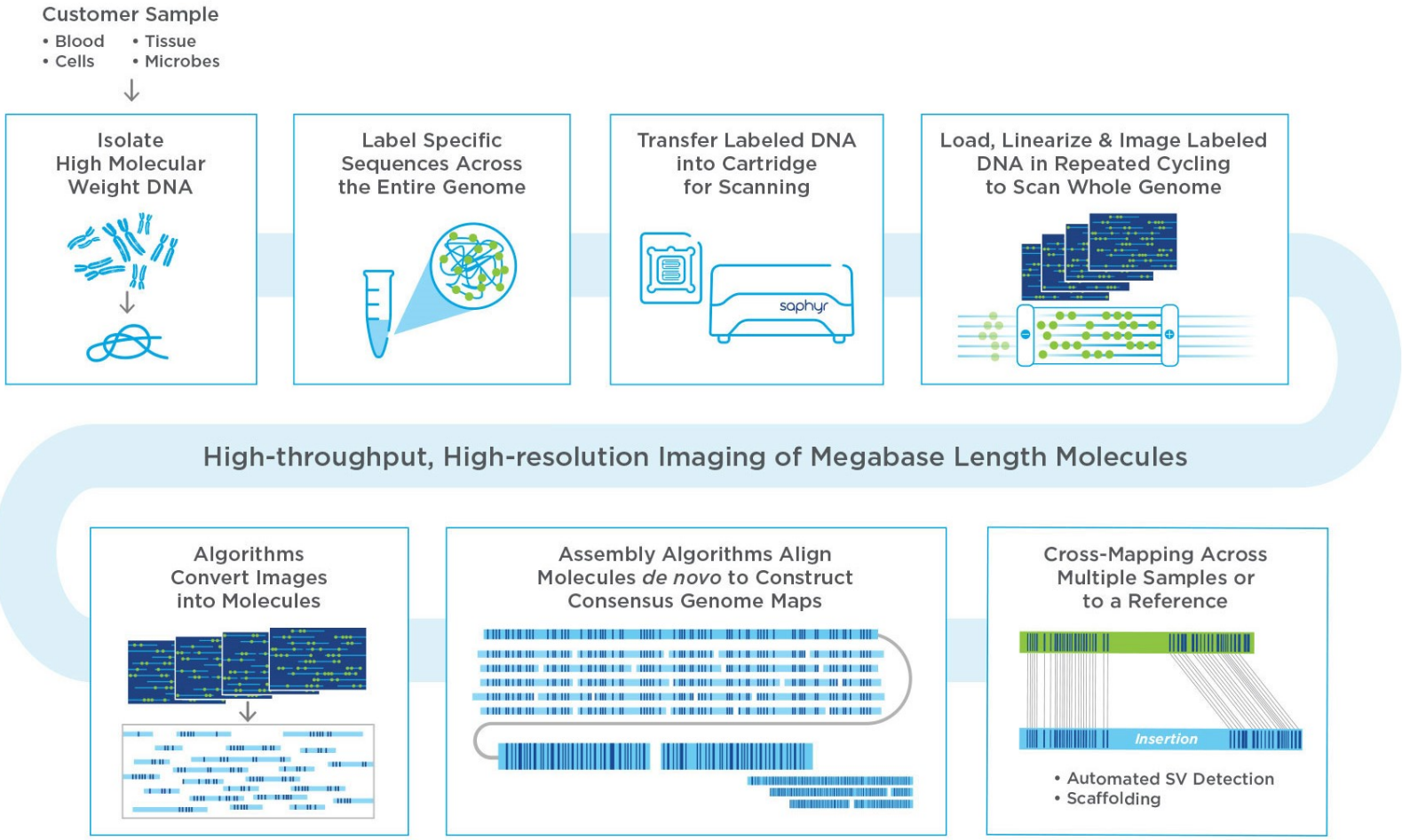




BAIT: inversions only

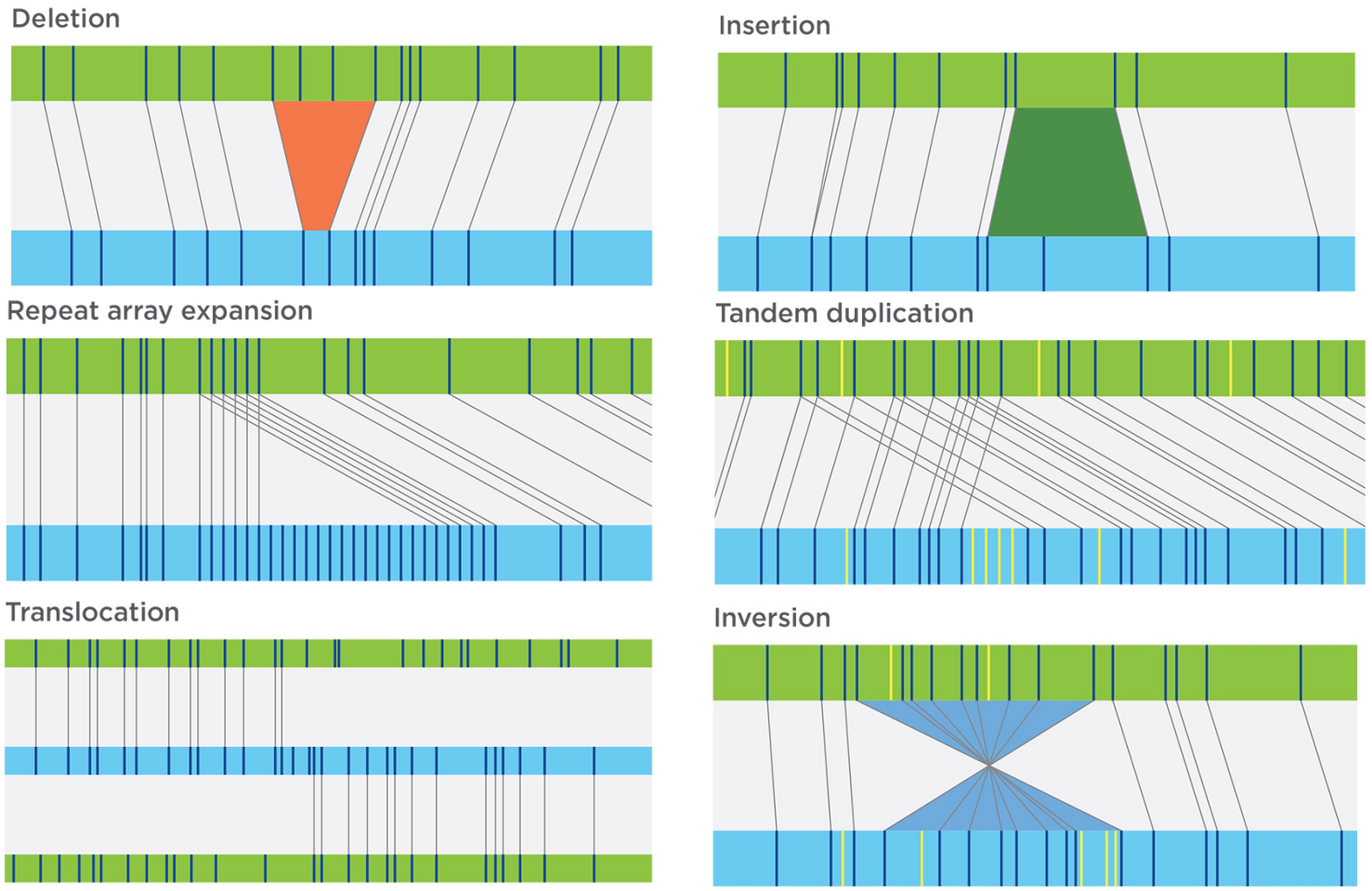
Hills et al., Genome Med 2013Sanders et al., Genome Res 2016

Optical Mapping



https://bionanogenomics.com

Optical Mapping: SV discovery

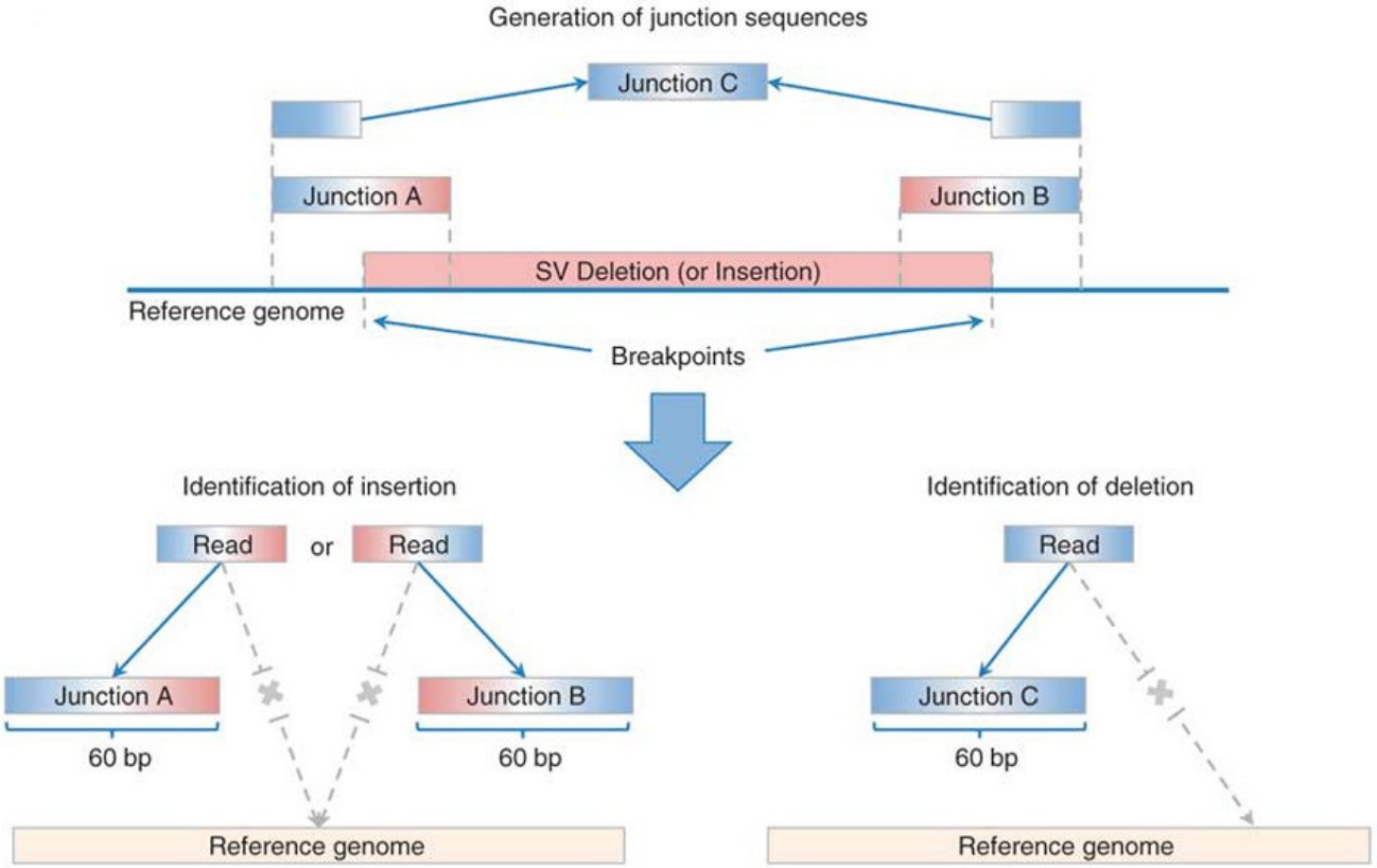


https://bionanogenomics.com

OMSV (Li et al., 2017)
Bionano Solve (Hastie et al., 2017)

GENOTYPING SV

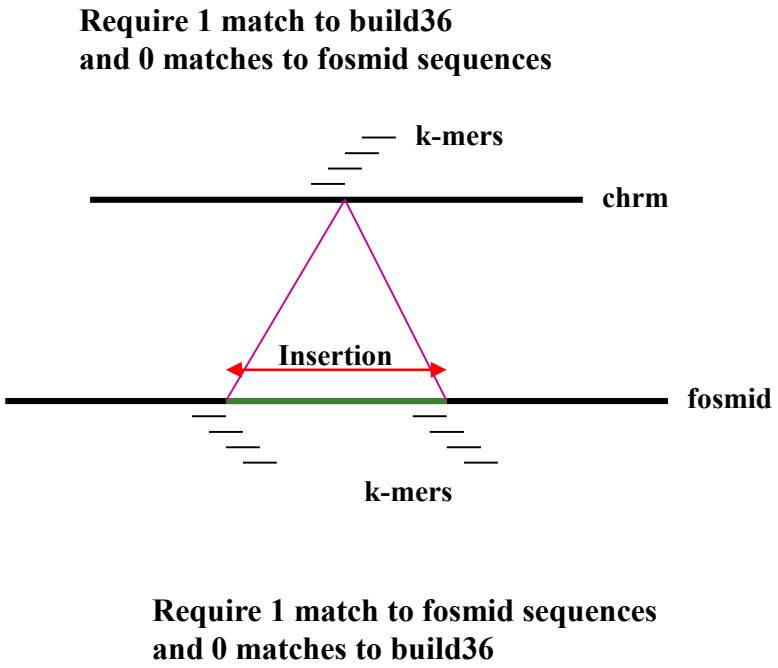
BreakSeq

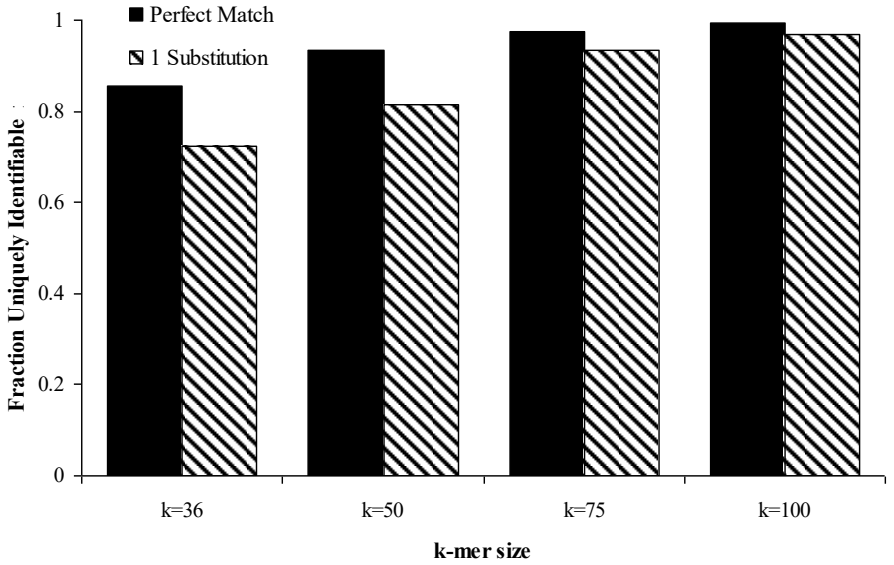


Read overlaps <10 bp to one side of the breakpoint is discarded and read matches also to the reference genome is classified as non-unique match

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

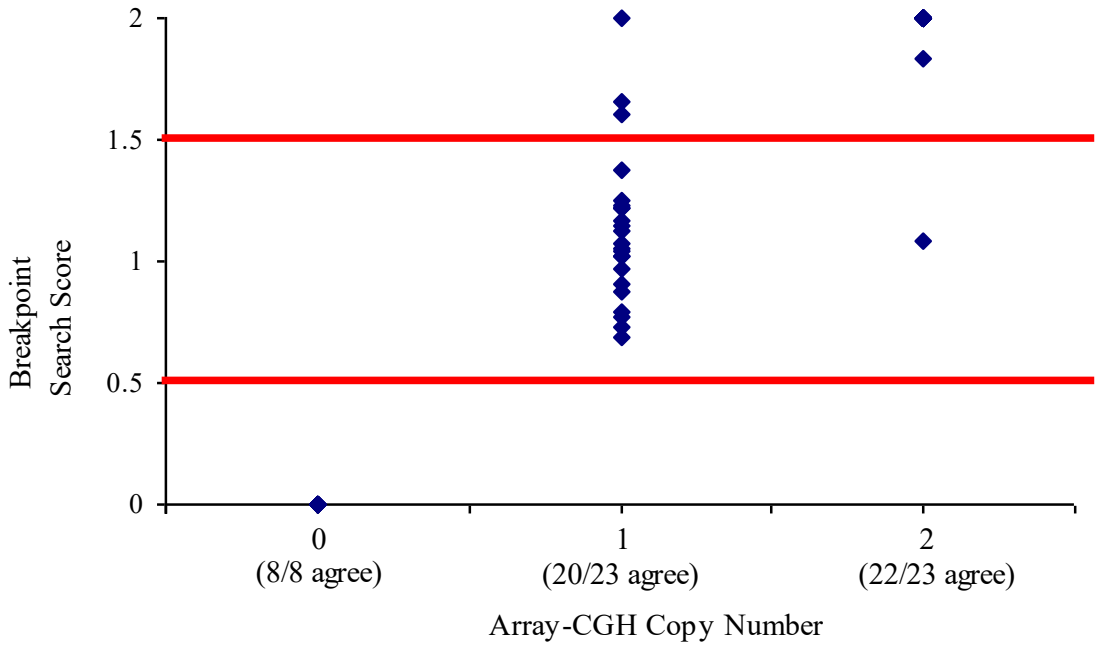
Genotyping insertions with NGS

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

Further reading

- Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Kosugi et al., Genome Biol. 2019*

- A robust benchmark for germline structural variant detection. *Zook et al. bioRxiv, June 2019*

Open problems

- Identify ***inversions***, ***translocations,*** *and* ***complex rearrangements***
- Reference-free STR typing
- Discover SVs in repeat- and duplication-rich regions
- Accurate & comprehensive detection of SVs with a *single* algorithm
  - High sensitivity
  - High specificity