Technologie w skali genomowej 2/ Algorytmiczne i statystyczne aspekty sekwencjonowania DNA

Structural variation discovery

Ewa Szczurek
szczurek@mimuw.edu.pl

Instytut Informatyki
Uniwersytet Warszawski

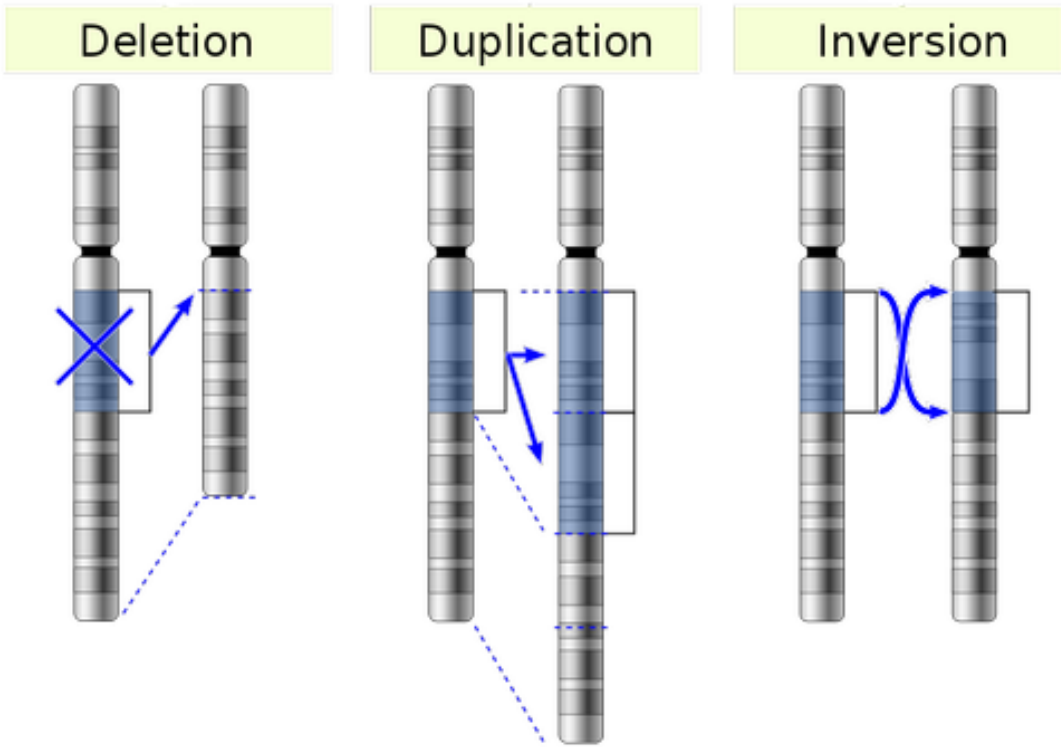
Structural variation

Change of the structure of the genome, including all insertions, deletions and inversions.
Structural variants are generally categorized into

- copy-number variants (CNVs), affecting the copy count of any genomic region, e.g. insertions and deletions (indels)
- copy-count invariant events, e.g. inversions

We will consider methods for SV detection from paired-end NGS read data.

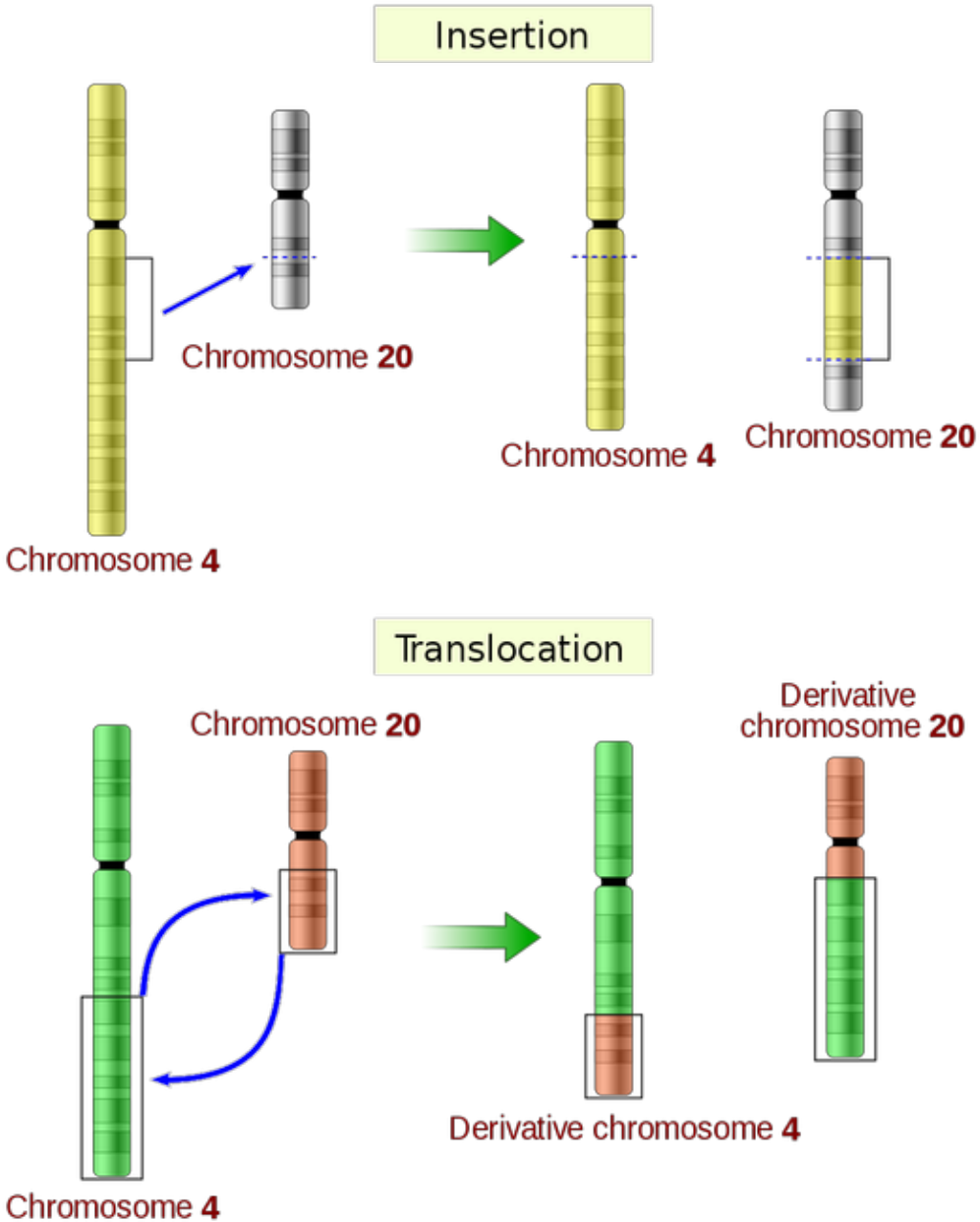
Single chromosome structural variations



Note: If the size of insertion > insert size of the sequenced fragment

- the basic insertion signature does not appear
- the inserted sequence cannot be identified invariant events, e.g. inversions

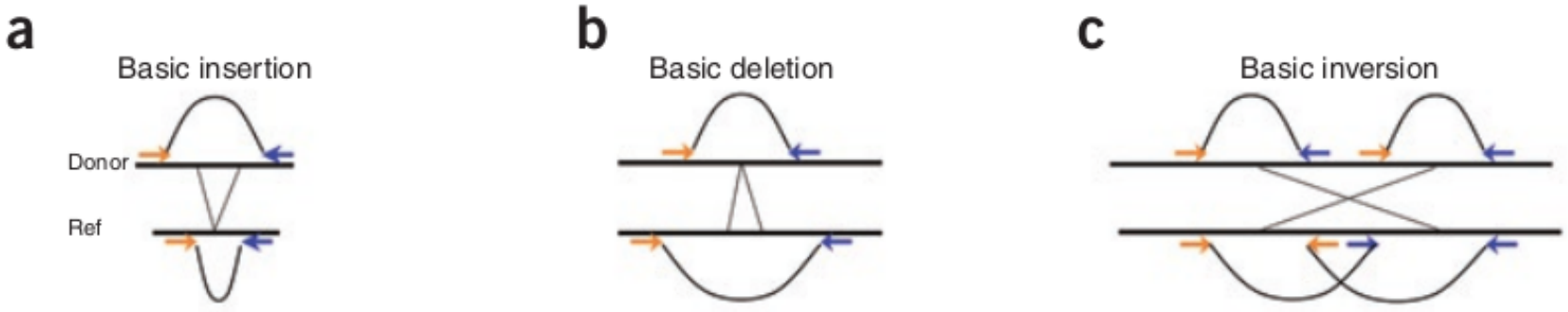
Inter-chromosome structural variations



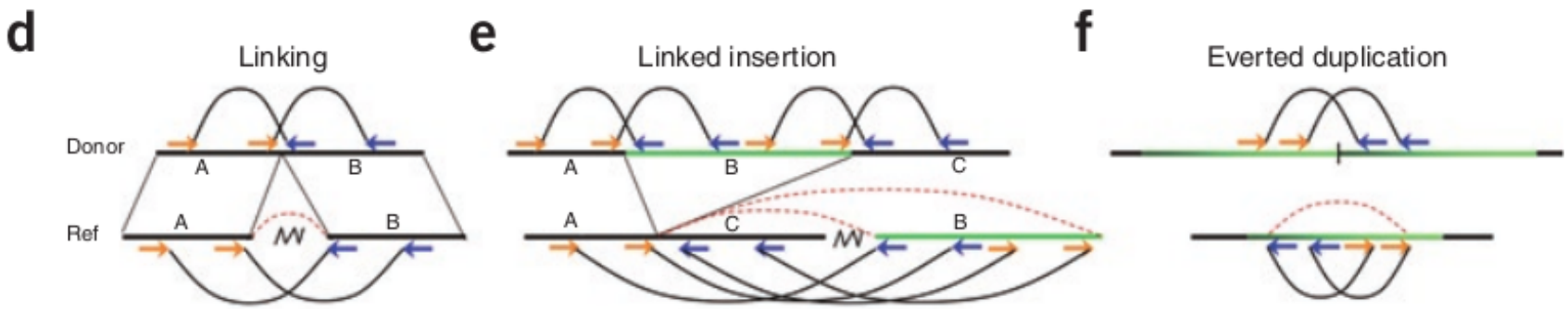
Approaches to structural variation discovery

- Read Pair patterns
- Split Read patterns
- Read Depth patterns

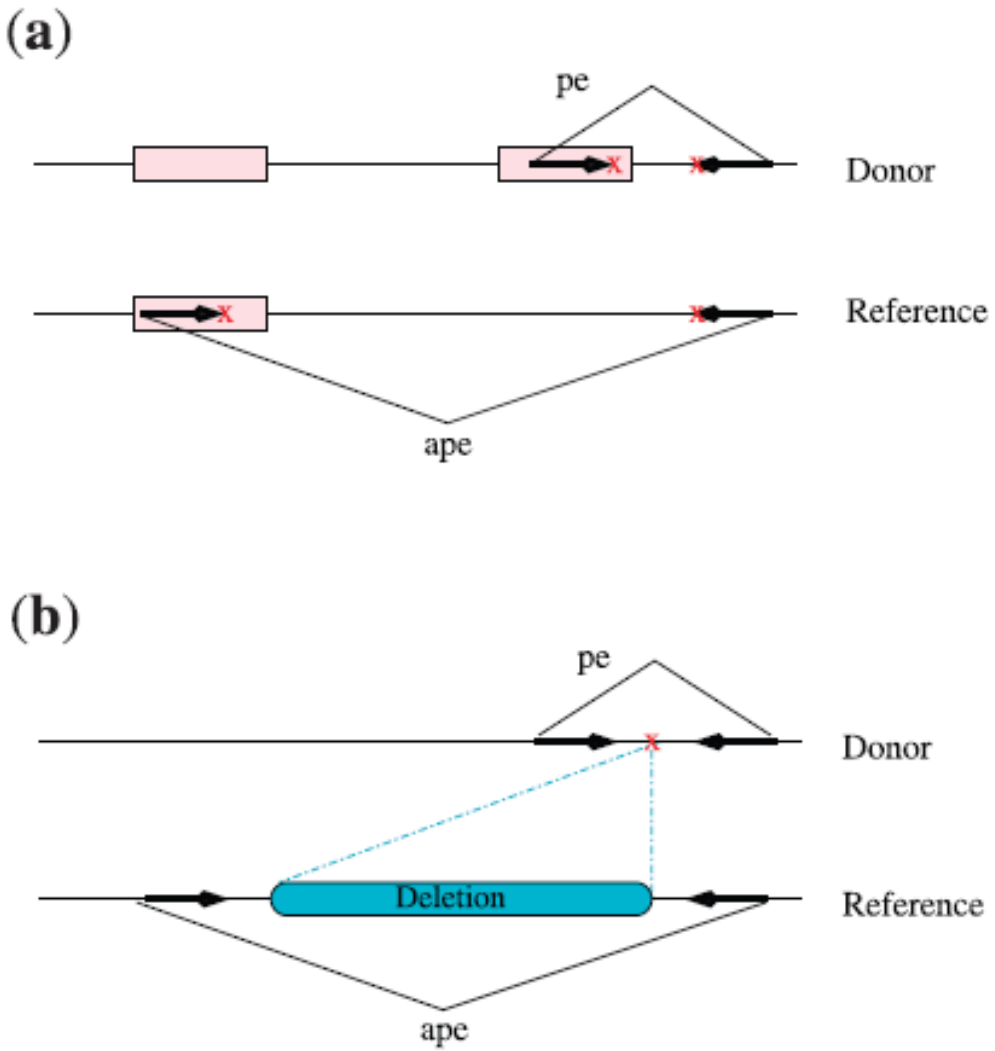
Read Pair patterns: insertions, deletions and inversions



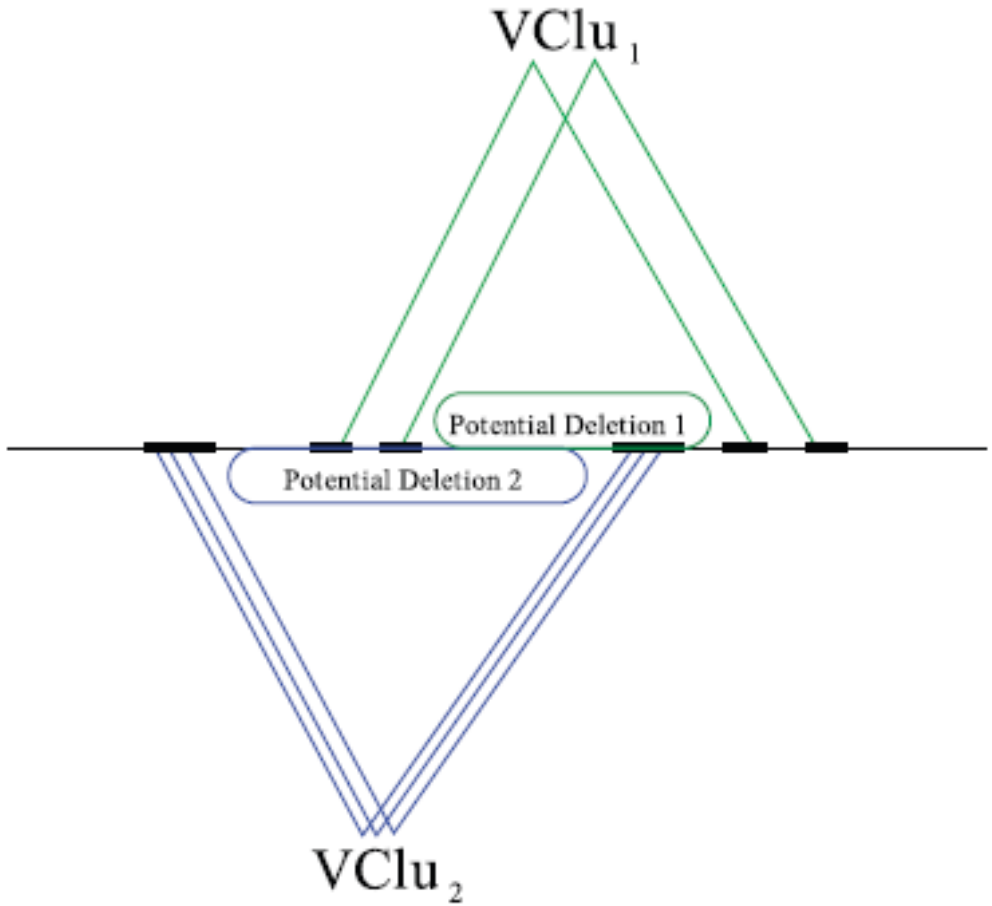
Read Pair patterns: linking and duplications



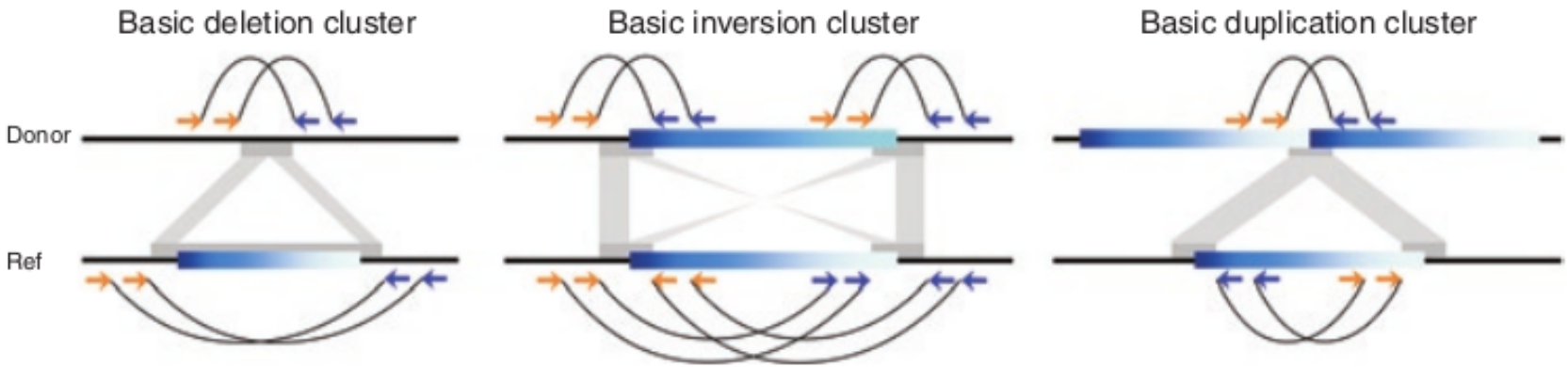
Ambigous Read Pair predictions



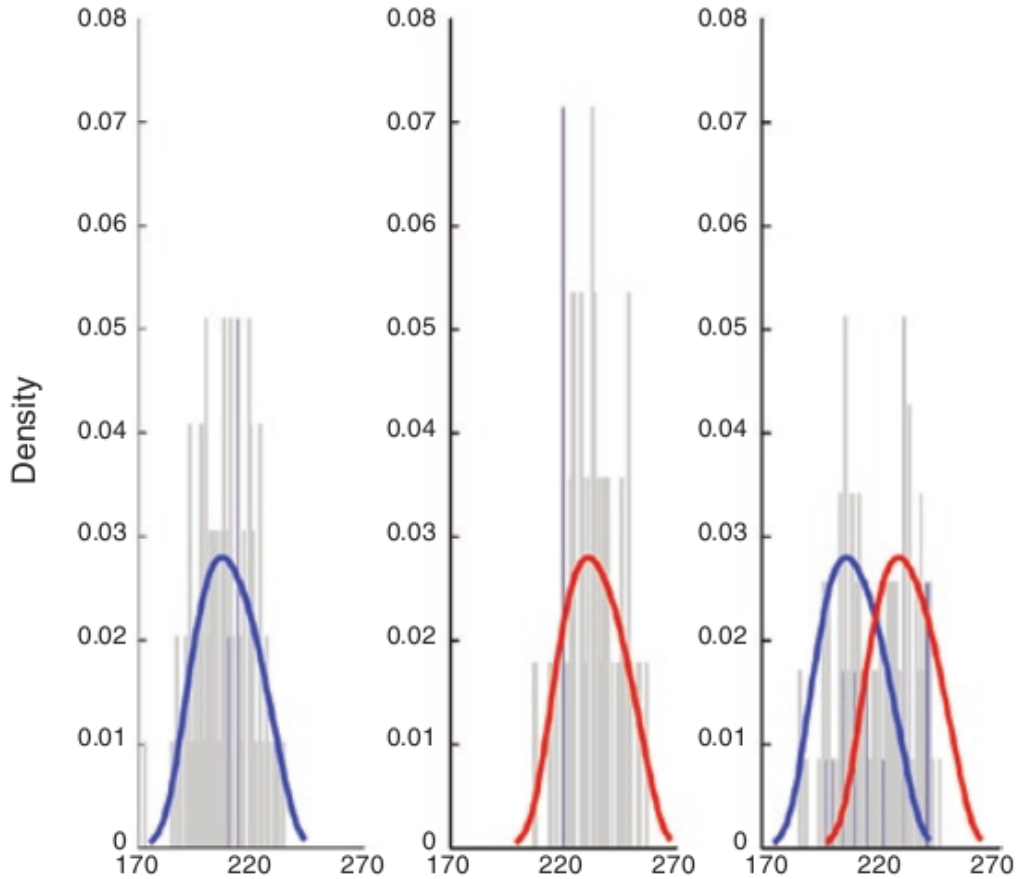
Diploid genome conflicts



Indels, inversions and duplications – improving resolution with clustering



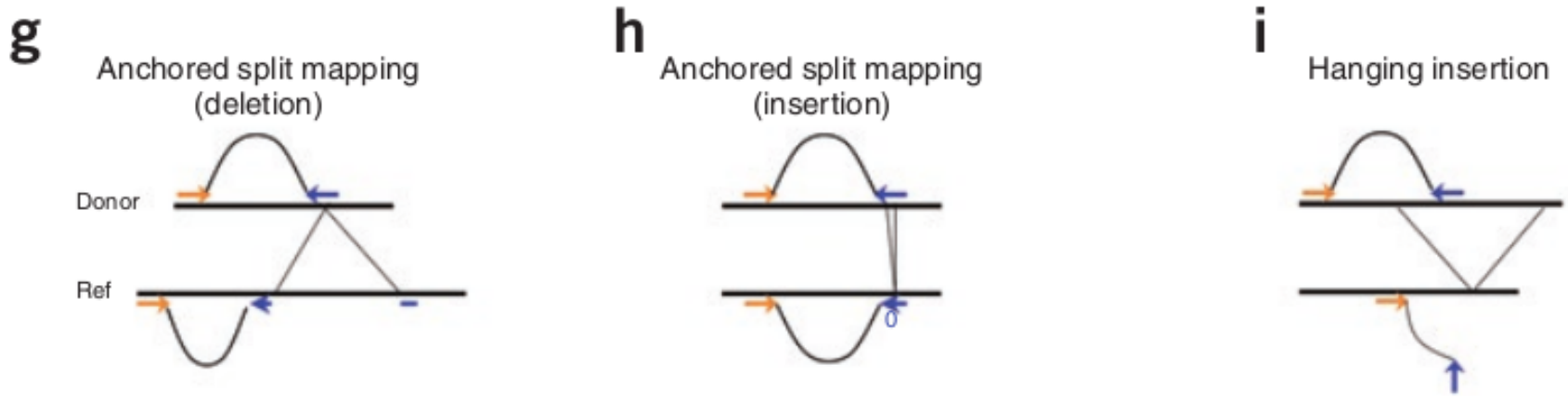
Distribution-based clustering



Insert length distribution in

- area with no variation (insert size is 208bp)
- area with homozygous deletion of length 24bp
- area with hemizygous deletion of length 22bp

Split Read patterns



Depth-of-coverage patterns

- Assumption: sequencing process is uniform

Depth-of-coverage patterns

- Assumption: sequencing process is uniform

$\Longrightarrow$ the number of reads covering a region

  - follows a Poisson distribution
  - with expected value proportional to its copy number

Depth-of-coverage patterns

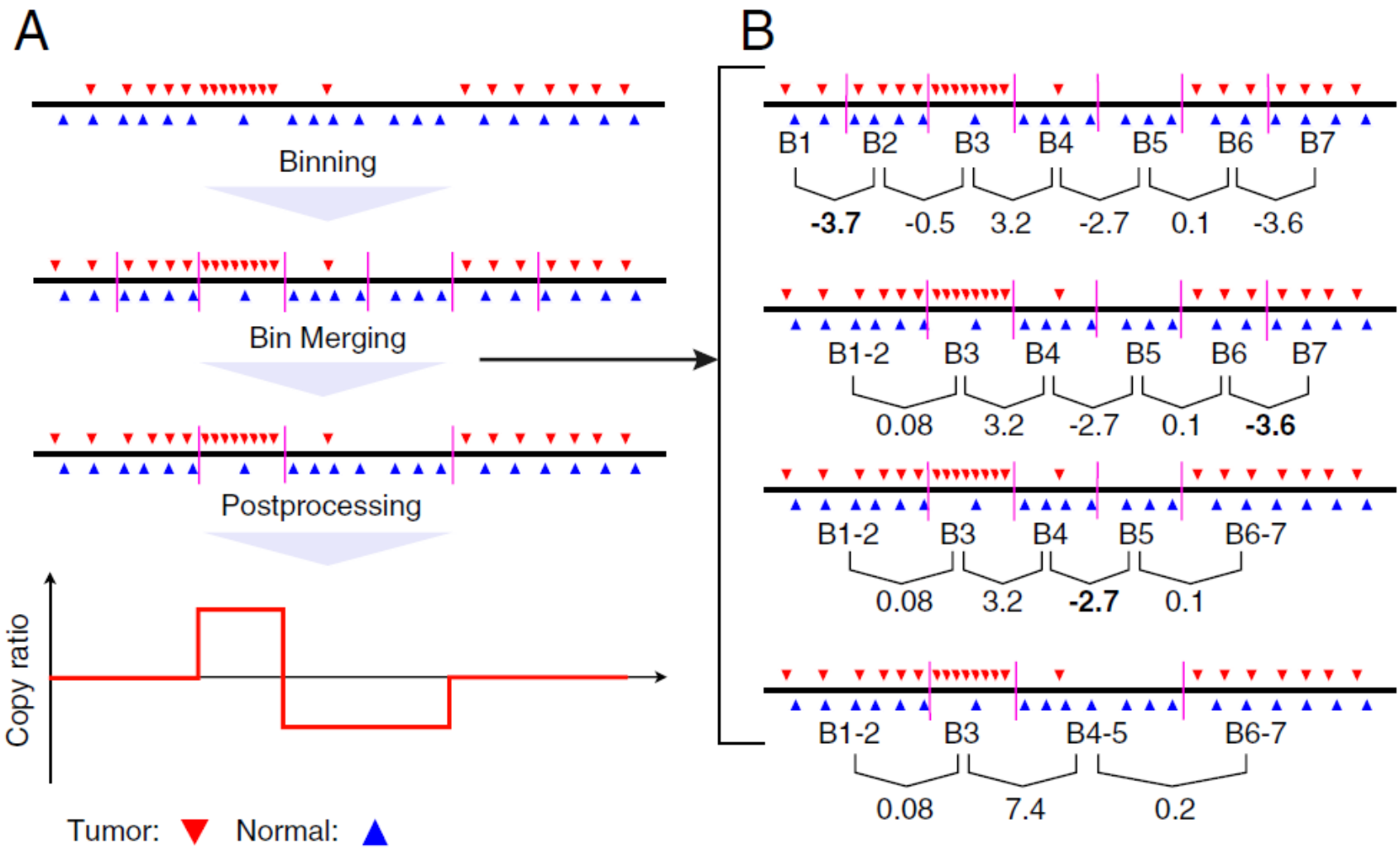
- Assumption: sequencing process is uniform

$\Longrightarrow$ the number of reads covering a region

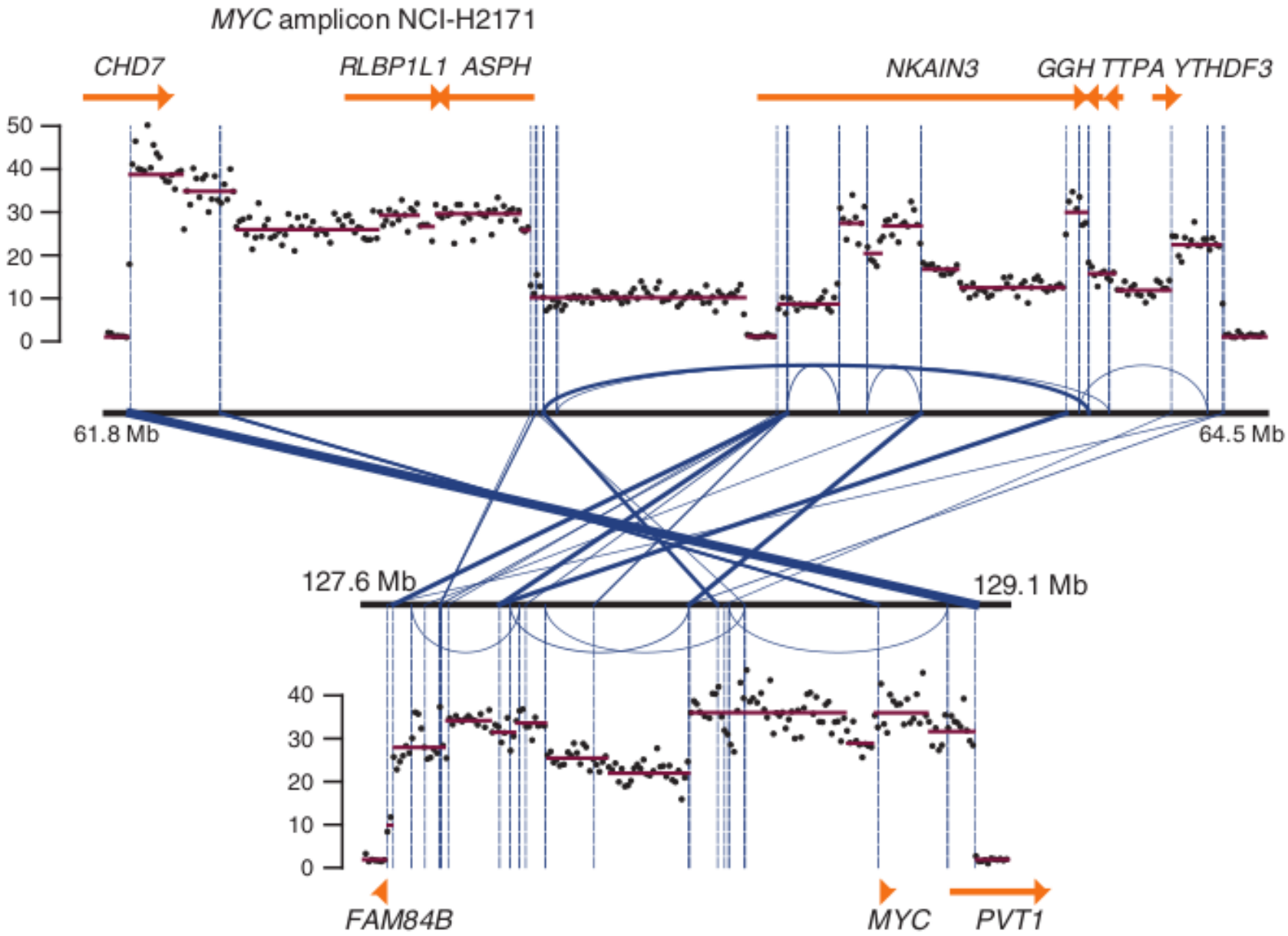
  - follows a Poisson distribution
  - with expected value proportional to its copy number

- regions must be large enough to support statistically significant signal from the distribution of coverage

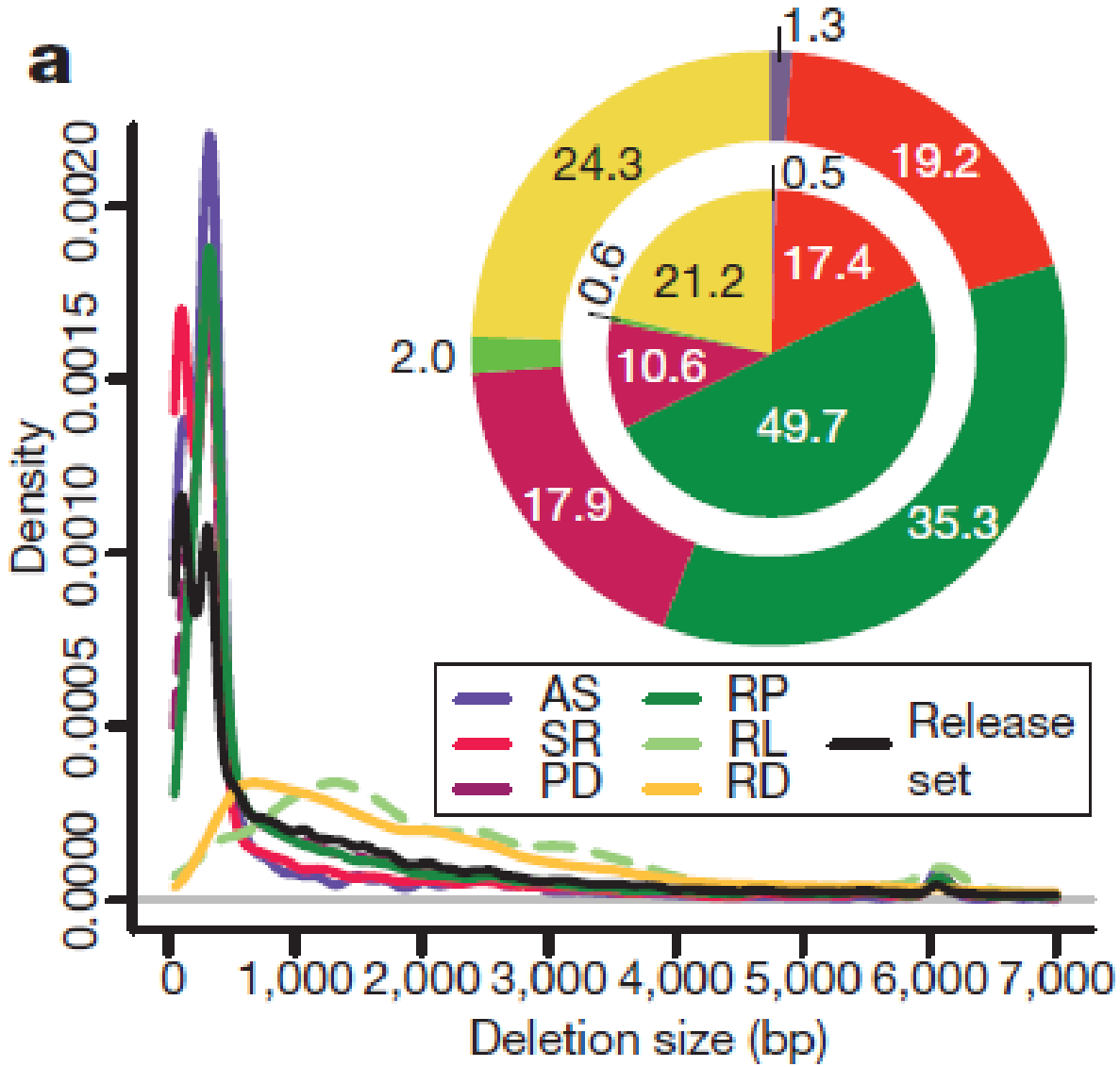
Depth-of-coverage patterns – BIC-seq



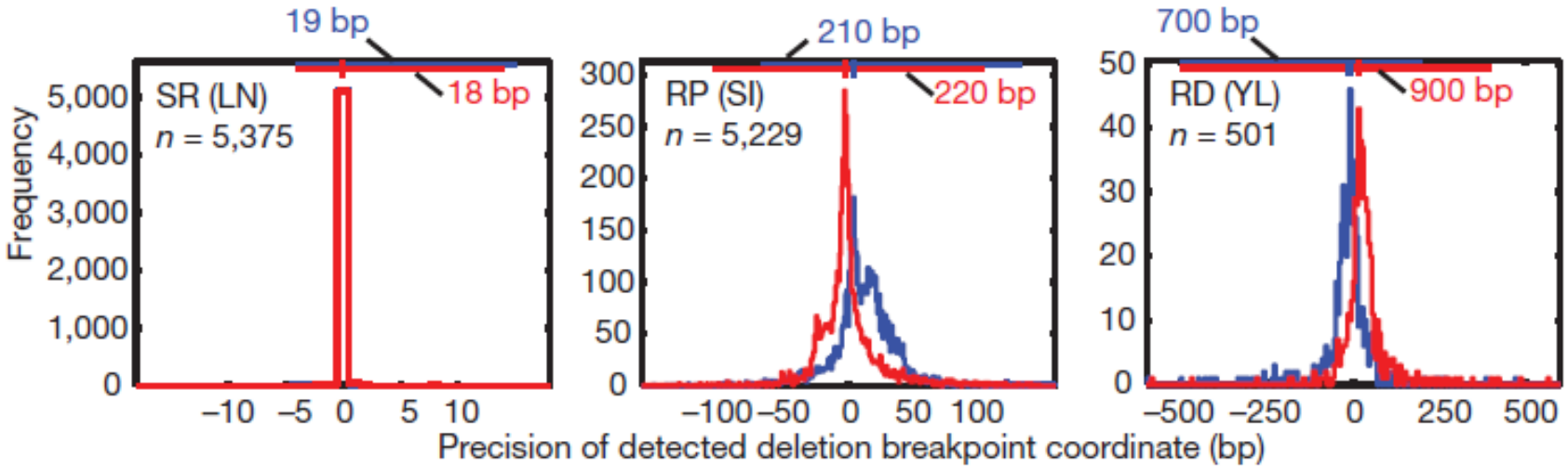
Depth-of-coverage patterns



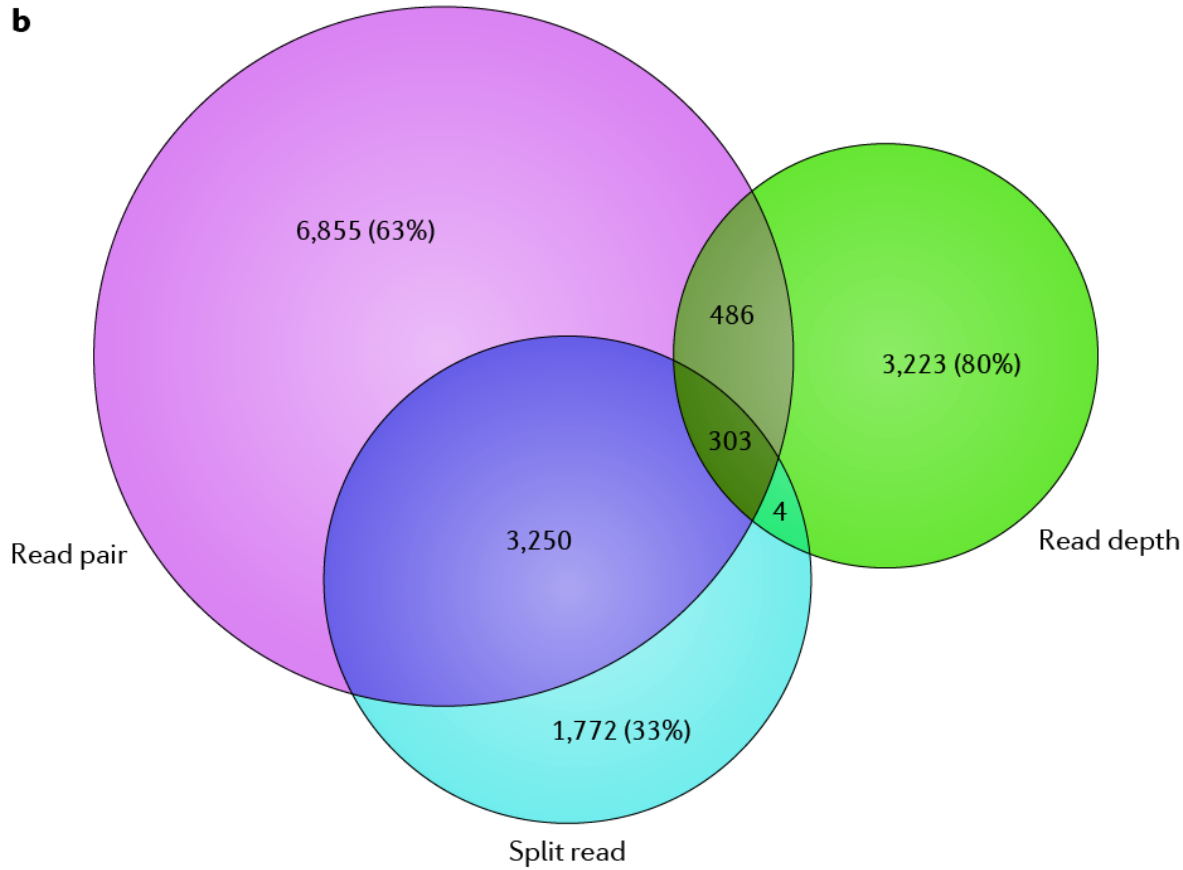
Approach comparison – size of detected variations



Approach comparison – precision of breakpoint prediction



Approach comparison – unique and shared predictions



Tools for structural variation discovery

			Signatures detected									
Refs.	**Name**	**Availability**	**Basic deletion**	**Basic insertion**	**Basic inversion**	**Linking**	**Linked insertion**	**Hanging insertion**	**Anchored split mapping**	**Everted duplication**	**Gain/loss**	**Clustering and/or windowing strategies**
3, 41	PEMer	Downloadable	•	•	•	•	•					Standard
34		None				•					•	Binary circular segmentation
44	SegSeq	Downloadable									•	Local change-point analysis
9		In the future	•	•				•			•	Standard
10	VariationHunter	Downloadable	•	•	•					•		Soft
11	MoDIL	Downloadable	•	•								Soft, distribution-based
36	Pindel	Downloadable							•			Standard
43	BreakDancer	Downloadable	•	•	•			•				Standard, distribution-based
42	ABI Tools	Downloadable	•	•	•						•	Standard, distribution-based, binary circular segmentation

Tools for structural variation discovery

Refs.	Technology	Individual or cell line	Read length	Mean insert size	Coverage	Detectable events	Mean breakpoint resolution	Range of calls
3	454	NA15510 NA18505	109 bp	~3,000 bp	×2.1 ×4.3[a,b]	Ins, del, inv	644 bp	>3 kbp
34	Illumina	NCI-H2171 NCI-H1770	29–36 bp	~400 bp ~90 bp	2.4 Gb 1.8 Gb[a,c]	Ins, del	500 bp	>30 kbp
44	Illumina	HCC1954 HCC1143 HCI-H2347	32–36 bp	Unpaired	637 Mb 541 Mb 503 Mb[d]	Ins, del	440 bp	10–500 kbp
9						Ins, del	Not available	50 bp–35 kbp (del) 60–160 bp (ins)
10						Ins, del, inv	Not available	<500 kbp (del) <137 bp (ins) <10 Mb (inv)
11	Illumina	NA18507	~36 bp	~200 bp	~×42[e]	Ins, del	<100 bp	>20 bp (del) 20–120 bp (ins)
36						Ins, del	1 bp	<10 kbp (del) <20 bp (ins)
43						Ins, del, inv	Not available	>10 bp (del) 10–130 bp (ins)
42	ABI SOLiD	NA18507	25–50 bp	600–3,500 bp	~×15[e]	Ins, del, inv	Not available	>80 bp (del) 30–1,300 bp (ins)

Ins, insertion; del, deletion; inv, inversion.
[a]Total sequence generated by reads that were part of a mate pair that had a mapping that was not rejected by the algorithm. [b]With respect to the diploid genome. [c]Clone coverage. [d]Total sequence generated that had a high-quality alignment. [e]Total sequence generated with respect to the haploid genome.

Bibliography

- P. Medvedev et al., *Computational methods for detecting structural variation with next generation sequencing*. Nat. Methods 2009.
- R. E. Mills et al., *Mapping copy number variation by population-scale genome sequencing*. Nature 2011.
- F. Hormozdiari et al., *Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery*. Bioinformatics 2010.
- R. Xi et al., *Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion*. PNAS 2011.