SVEM: A Structural Variant Estimation Method using Multi-Mapped Reads on Breakpoints

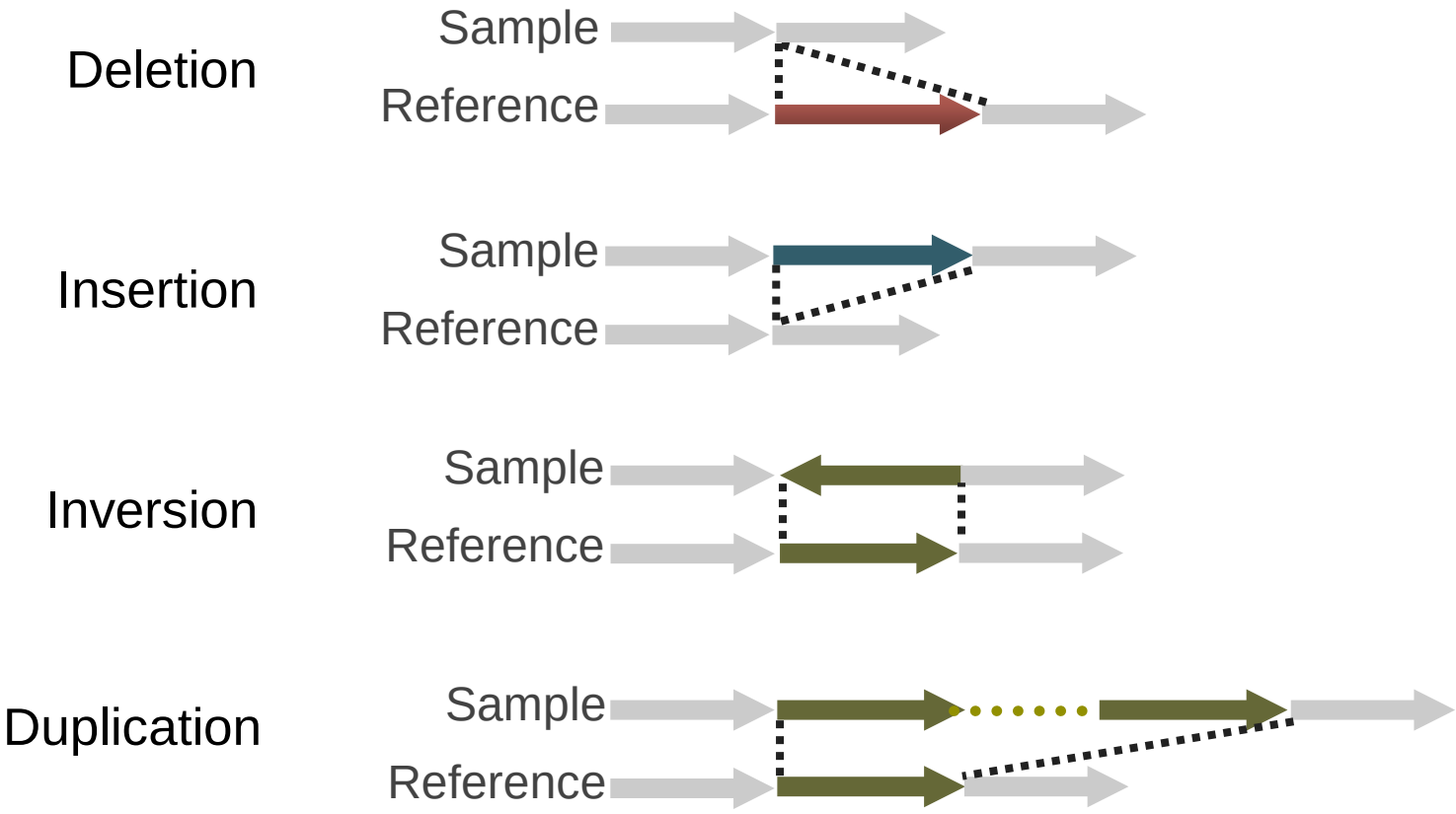
Tomohiko Ohtsuki, Naoki Nariai, Kaname Kojima, Takahiro Mimori, Yukuto Sato, Yosuke Kawai, Yumi Yamaguchi-Kabata, Testuo Shibuya and Masao Nagasaki

Tohoku Medical Megabank Organization, Tohoku University

1st International Conference on Algorithms for Computational Biology, AlCoB 2014

Structural Variants (SVs)

- Genomic alterations existing in human genome



- **SVs are related to human diseases (cancer, mental disorders, and etc.)**

Existing approaches for SV detection

- Microarray based approach
  - Based on intensity of cDNA probes for ref and alt alleles
  - Positional resolution is limited (about 1kb)
  - Difficult to identify high copy number

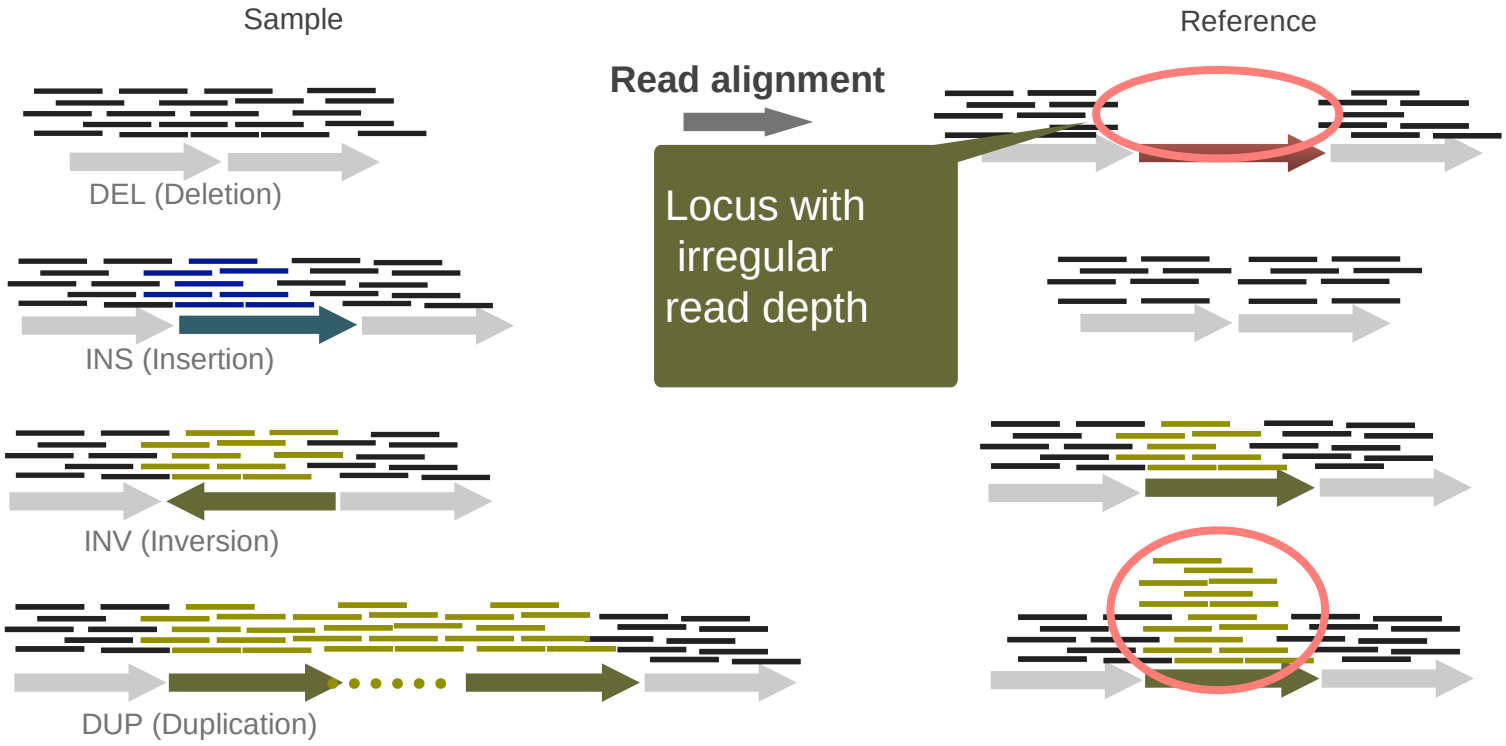
PennCNV: Wang et al., Genome Research, 2007

cnvPartition: www.illumina.com

Existing approaches for SV detection

- NGS based approaches
  - **RD approach**: irregular read depth
  - **RP approach**: discordant read pairs
  - **SR approach**: re-alignment of split read
  - **AS approach**: local de novo assembly

Read depth (RD) approach



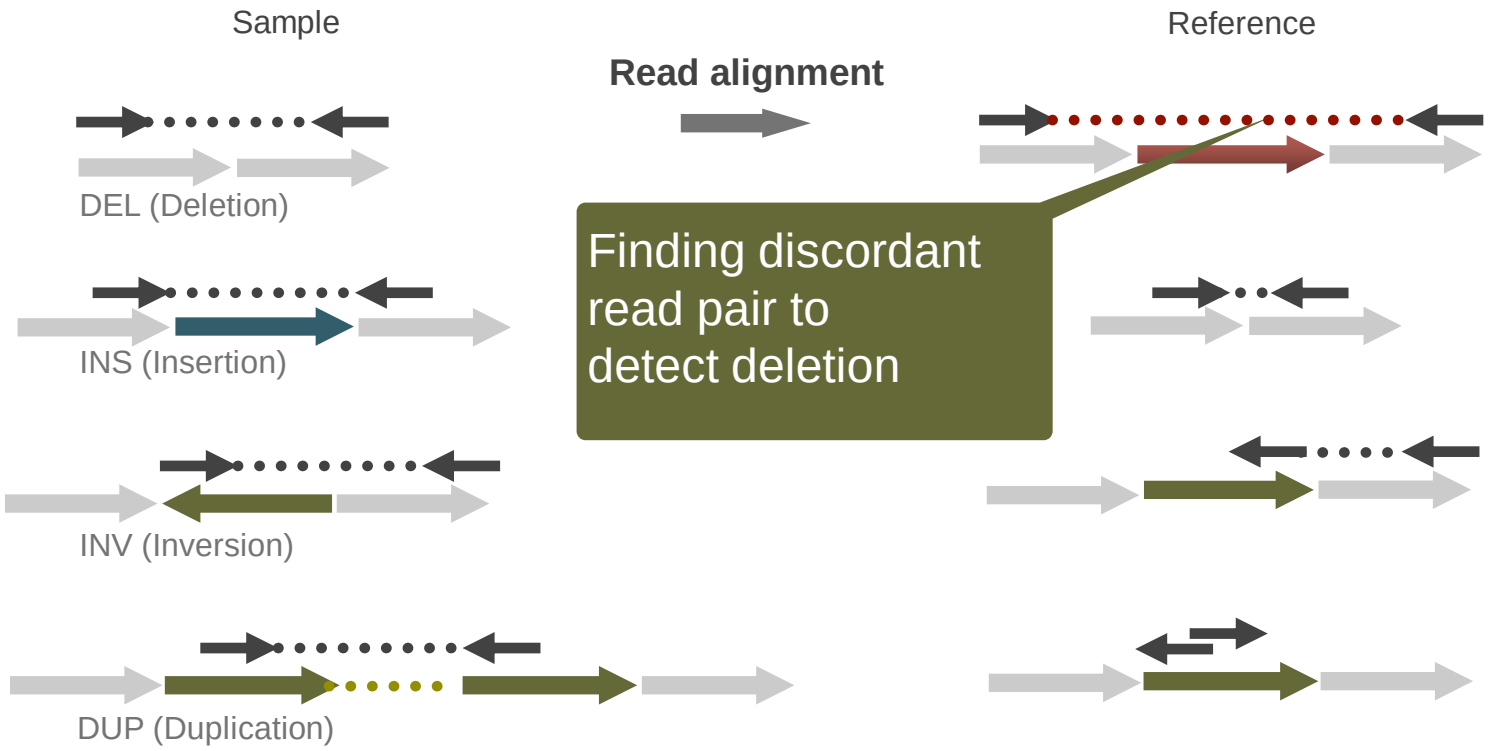
CNVnator : Abyzov *et al. Genome Res* (2011)

FREEC : Boeva *et al. Bioinformatics Application note* (2012)

PROS
- Can discover large DEL/DUP
- Fast

CONS
- Cannot detect INS/INV
- Low resolution

Read pair (RP) approach

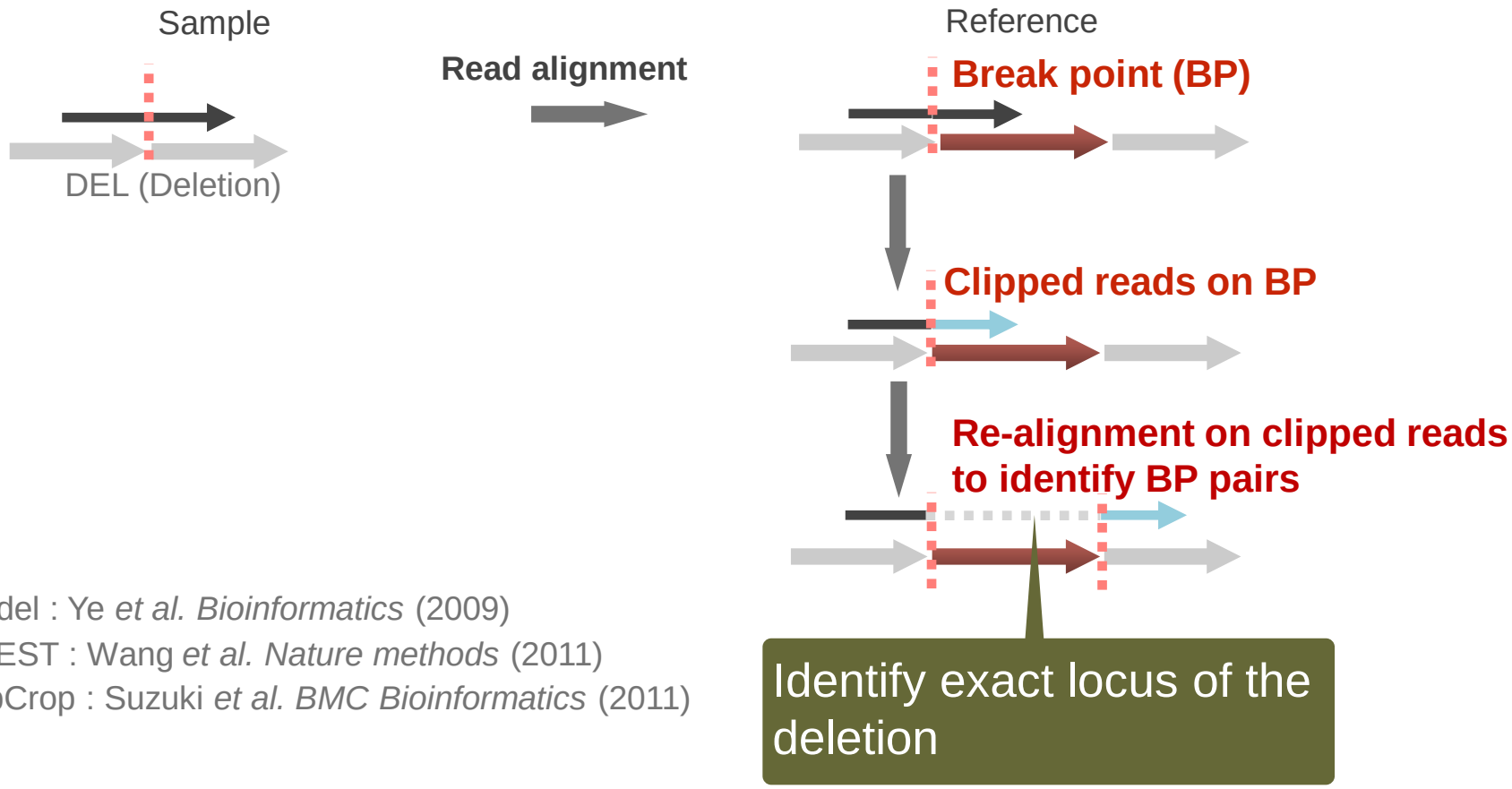


BreakDancer : Chen *et al. Nature methods* (2009)
DELLY : Rausch *et al. Bioinformatics* (2012)

PROS
- Can discover DEL/INS/INV/DUP
- Fast

CONS
- Low resolution
- Cannot determine copy number

Split read (SR) approach

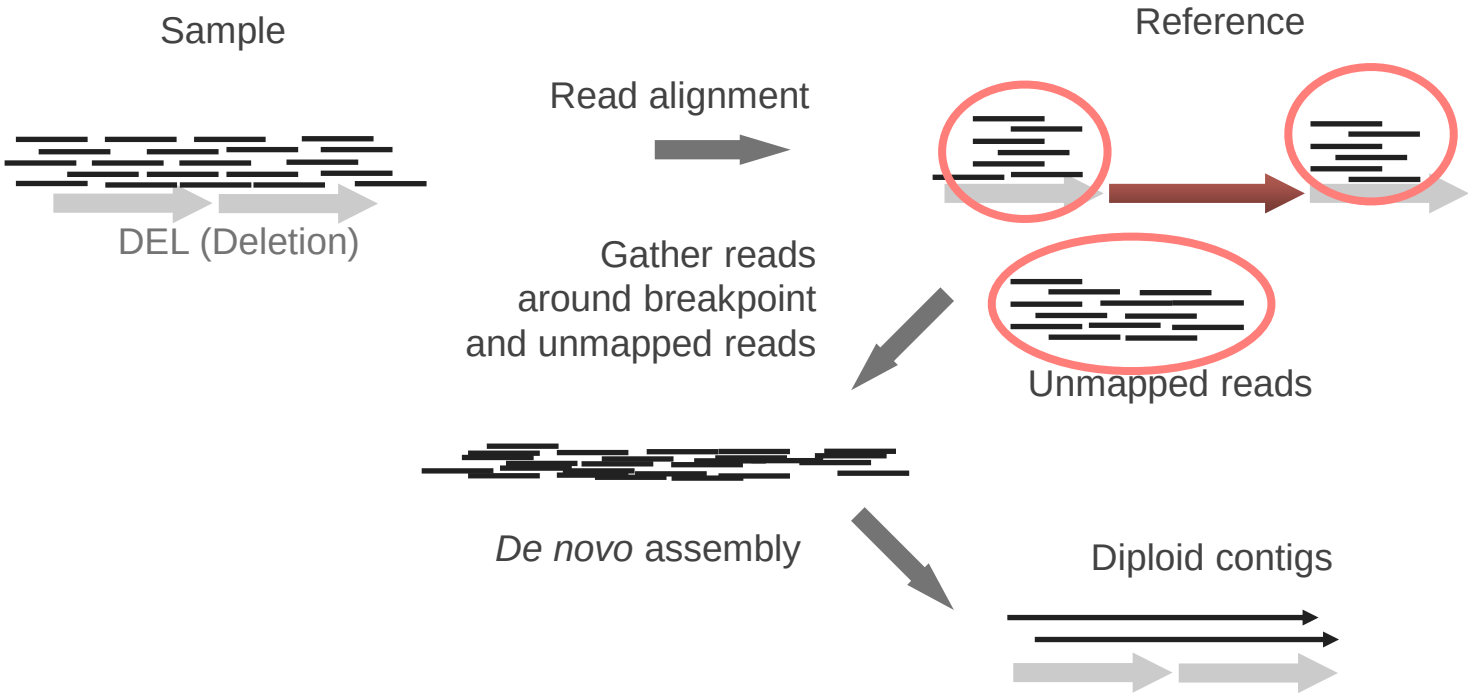


Pindel : Ye *et al. Bioinformatics* (2009)
CREST : Wang *et al. Nature methods* (2011)
ClipCrop : Suzuki *et al. BMC Bioinformatics* (2011)

PROS
- Can discover DEL, INS, INV, DUP
- Identify exact loci of SVs

CONS
- False positives
- Computationally high cost

Assembly (AS) approach



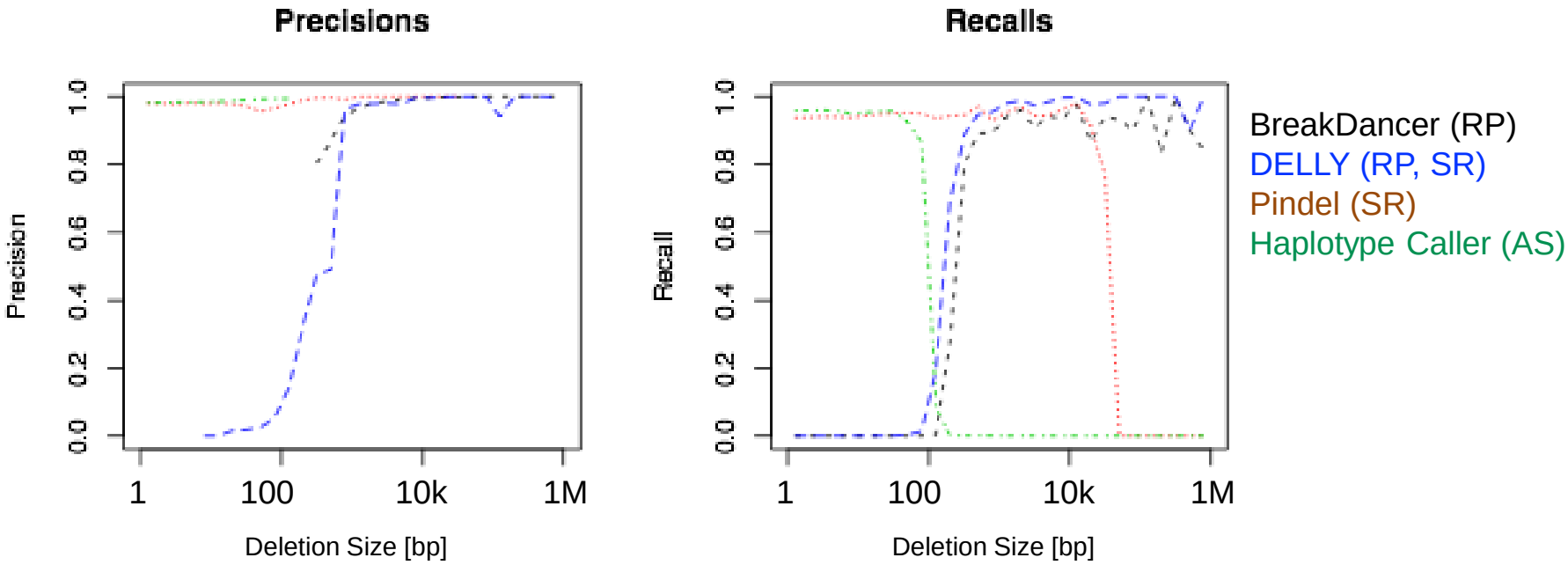
SOAPindel : Shengting *et al. Genome Res.* (2013)
Haplotype Caller : http://www.broadinstitute.org/gatk/

PROS
- Detects exact position of SVs
- Determines actual sequence directly

CONS
- Computational cost is even higher than SR method

Challenges in SV detection from NGS data

- Performance of SV detection algorithms depends on SV size



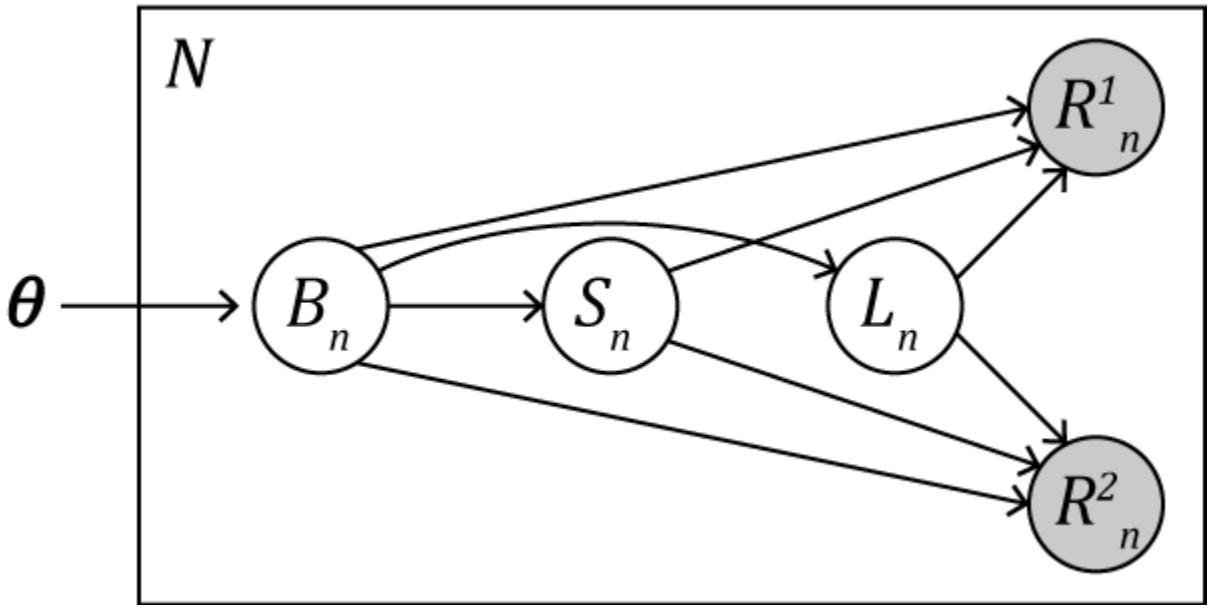
Mimori et al., BMC Systems Biology, 2013

- **Split read (SR) method** generally works well for 1-10kb size

SVEM

- Split read (SR) approach
- We use a **statistical framework to handle ambiguous mapping of reads on break points** (existing methods do not consider multi-mapped reads explicitly)
- Maximum likelihood **estimation of locations of break points and correct alignments of reads simultaneously** by the EM algorithm

Generative model of NGS reads



$\boldsymbol{\theta}$: Read abundance parameter
B_n: Breakpoint choice
S_n: Read start position
L_n: Fragment length
R_n^1, R_n^2: Reads

$$P(B_n, S_n, F_n, R_n^1, R_n^2 \mid \theta) = P(B_n \mid \theta)P(S_n \mid B_n)P(F_n \mid B_n) \cdot P(R_n^1 \mid B_n, S_n, F_n)P(R_n^2 \mid B_n, S_n, F_n).$$

Likelihood of data

$$P(B_n, S_n, F_n, R_n^1, R_n^2 \mid \theta) = P(B_n \mid \theta)P(S_n \mid B_n)P(F_n \mid B_n) \cdot P(R_n^1 \mid B_n, S_n, F_n)P(R_n^2 \mid B_n, S_n, F_n).$$

$$P(B_n = b \mid \theta) = \theta_b, \quad \text{where} \sum_b \theta_b = 1.$$

$$P(S_n \mid B_n) = \begin{cases} 1/(l-1) \text{ if the distance between } S_n \text{ and } B_n < l \\ 0 \qquad\qquad\qquad\qquad \text{otherwise} \end{cases}.$$

$$P(F_n = f_n \mid B_n) = \frac{\exp(-\frac{(f_n-\mu)^2}{2\sigma^2})}{\sum_{x=2l}^{f_{max}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})},$$

$$P(R_n^1 \mid B_n, S_n, F_n) = \prod_{i=1}^{l} r_{ni}^1,$$

where

$$r_{ni}^1 = \begin{cases} 1 - 10^{-Q_i/10} \text{ if } i\text{th nucleotide of read } n \text{ is the same as the reference} \\ 10^{-Q_i/10} \qquad\qquad\qquad\qquad\qquad\qquad \text{otherwise} \end{cases}.$$

EM algorithm

- E-step
  - Based on the current estimate of $\theta_{(old)}$, calculate the expected alignment of read *n* on each multi-mapped location:

$$E[Z_{nbsf}] = \begin{cases} \frac{\rho_{nbsf}}{\sum_{(b',s',f') \in \pi_n} \rho_{nb's'f'}} & \text{if } (b,s,f) \in \pi_n \\ 0 & \text{otherwise} \end{cases}.$$

where

$$\rho_{nbsf} = P(B_n \mid \theta_{(old)}) P(S_n \mid B_n) P(F_n \mid B_n) P(R_n^1 \mid Z_{nbsf}) P(R_n^2 \mid Z_{nbsf}).$$

EM algorithm

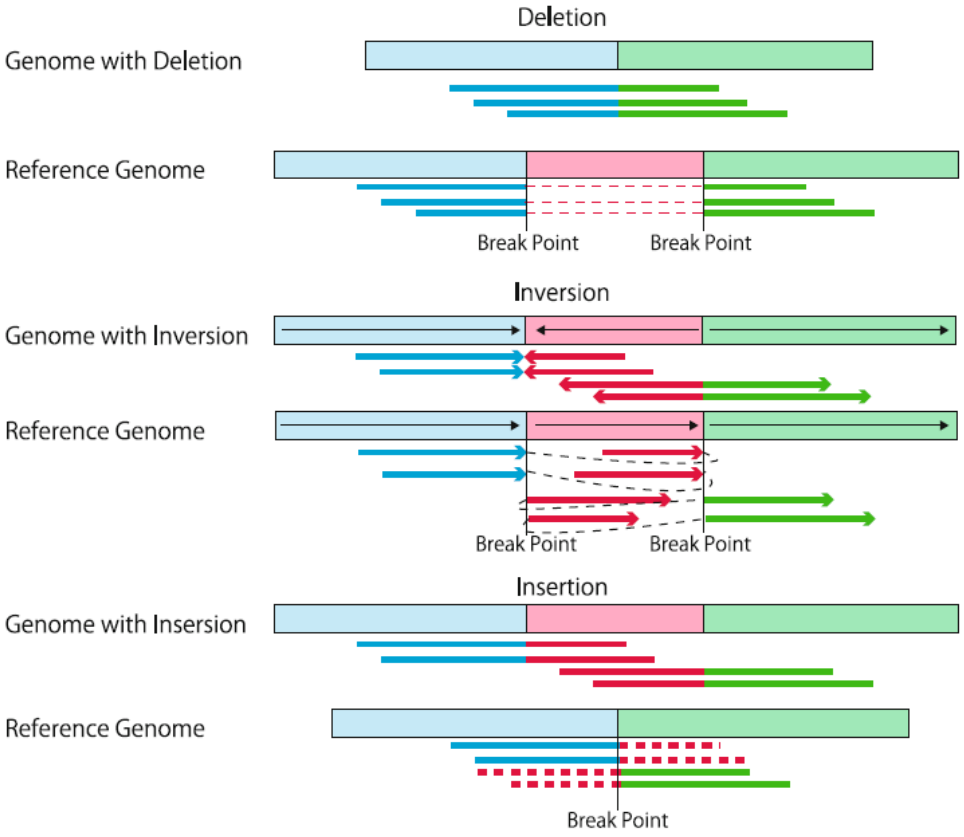
- M-step
  - Based on the current estimate of $E[Z_{nbsf}]$, update the parameter for each BP:

$$\theta_b = \frac{\sum_{n,s,f} E[Z_{nbsf}]}{\sum_{n,b',s,f} E[Z_{nb'sf}]}.$$

- E and M-steps are iterated until parameters are no longer updated

Identification of SVs from BPs

- Filter out spurious BPs (< three reads)
- Based on the estimated BP pairs, SVs are called



Simulation data analysis

- An artificial DNA sequences of chr21 with one SNP per 1kbp are prepared from GRCh37
- Both homozygous and heterozygous deletions, inversions, and insertions are incorporated
- 100 bp paired-end reads are generated (50x) with 0.1% substitution errors
- BWA-MEM is used for identifying clipped reads and re-alignment of clipped part of reads
- Performances are evaluated for midium-size SVs (500 bp) and larger-size SVs (1k bp)

Performance evaluation for the simulation data analysis

Midium-size SVs

Method	Predicted SVs	TP	FP	Precision	Recall	F-measure
SVEM	331	300	31	0.91	0.86	0.88
BreakDancer	506	3	503	0.0059	0.0086	0.0070
DELLY	894	50	844	0.056	0.14	0.080
Pindel	375	298	77	0.79	0.85	0.82

Larger-size SVs

Method	Predicted SVs	TP	FP	Precision	Recall	F-measure
SVEM	123	121	2	0.98	0.86	0.92
BreakDancer	119	10	109	0.084	0.071	0.076
DELLY	314	105	209	0.33	0.75	0.46
Pindel	123	119	4	0.97	0.85	0.91

$$\text{Precision} = \frac{\#\ \text{TP}}{\#\ \text{detected SV}} \qquad \text{Recall} = \frac{\#\ \text{TP}}{\#\ \text{true SV}} \qquad \text{F-measure} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

Performance evaluation for the simulation data analysis

Midium-size SVs

Method	Predicted SVs	TP	FP	Precision	Recall	F-measure
SVEM	331	300	31	0.91	0.86	0.88
BreakDancer	506	3	503	0.0059	0.0086	0.0070
DELLY	894	50	844	0.056	0.14	0.080
Pindel	375	298	77	0.79	0.85	0.82

Larger-size SVs

Method	Predicted SVs	TP	FP	Precision	Recall	F-measure
SVEM	123	121	2	0.98	0.86	0.92
BreakDancer	119	10	109	0.084	0.071	0.076
DELLY	314	105	209	0.33	0.75	0.46
Pindel	123	119	4	0.97	0.85	0.91

SVEM performed better compared to other comparable SV detection methods.

Real data analysis

- 100 bp paired-end sequencing data (45x) of NA12878, a CEU sample in the 1000 Genomes Project
- SVEM predicted 7,082 SVs in total, among which 214 were experimentally validated by other groups (Mills et al., Nature 2011)

Method	Predicted SVs	TP	Recall	Estimated TP ratio
SVEM	7081	214	0.35	0.030
BreakDancer	3213	21	0.034	0.0065
DELLY	206968	43	0.070	0.00021
Pindel	288783	205	0.33	0.0007

Computational resources

- All the experiments were performed on Intel Xeon CPU E5-2670 processor (2.60GHz)

Method	CPU Time(minutes)	Memory (GBytes)
SVEM	2.25	48
BreakDancer	1.5	5.6
DELLY	90	47.7
Pindel	10	1

SVEM was faster then DELLY and Pindel, but required 48 GB memory.

Conclusions

- SVEM is a statistical method to handle multi-mapped reads on BPs for SV detection
- SVEM performed better than Pindel, BreakDancer, and DELLY for the simulation data
- SVEM was faster than Pindel and DELLY with practical memory requirement for our experiments

Future works

- Incorporation of other approaches, such as read depth (RD), read pair (RP), and assembly (AS) algorithms.
- Incorporation of pedigree information or population genetics information
- Efficient implementation for shorter CPU time and memory usage

Acknowledgements

- The super-computing resource was provided by Human Genome Center, Institute of Medical Science, University of Tokyo
- Funding: This work was supported (in part) by MEXT Tohoku Medical Megabank Project