Short-read alignment, variant detection in NGS

27/11/2012

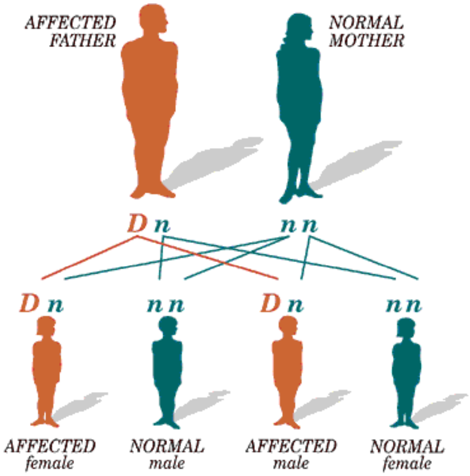
Gézsi András
gezsi@mit.bme.hu

Agenda

- Sequence alignment
  - Scoring schemes
  - Local sequence alignment
  - Short-read sequence alignment
- Variant calling
  - SNPs
  - Structural variations

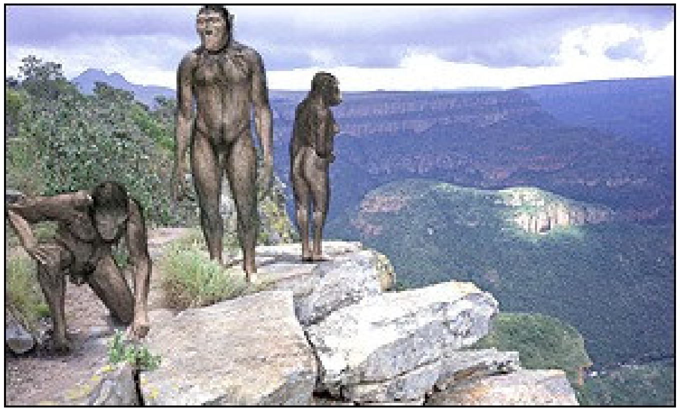
Why do we care about variations?

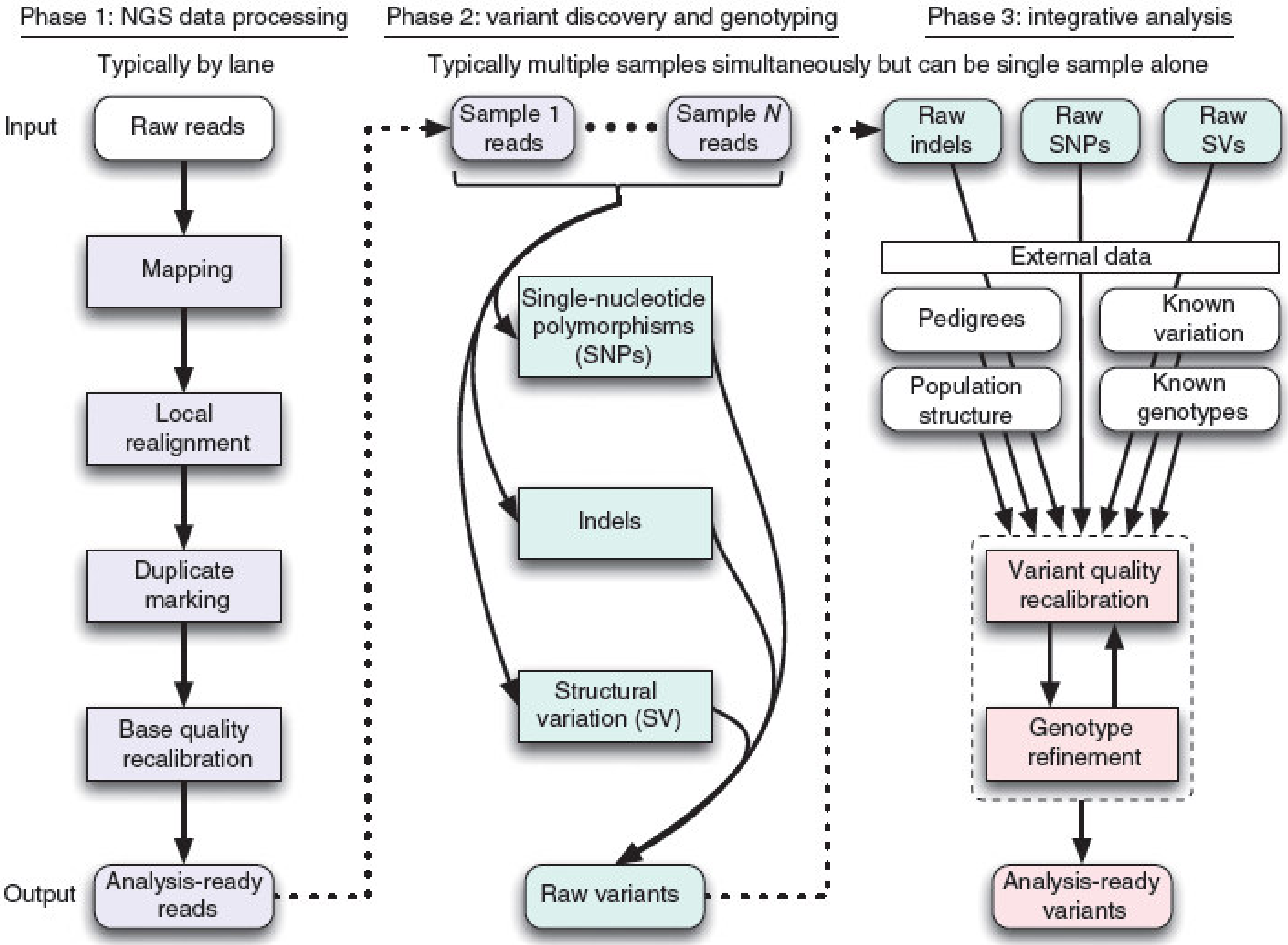
underlie phenotypic
differences



cause inherited
diseases

allow tracking ancestral
human history




Phase 1: NGS data processing
Typically by lane
Input
Raw reads
Mapping
Local realignment
Duplicate marking
Base quality recalibration
Output
Analysis-ready reads
Phase 2: variant discovery and genotyping
Typically multiple samples simultaneously but can be single sample alone
Sample 1 reads
Sample N reads
Single-nucleotide polymorphisms (SNPs)
Indels
Structural variation (SV)
Raw variants
Phase 3: integrative analysis
Raw indels
Raw SNPs
Raw SVs
External data
Pedigrees
Known variation
Population structure
Known genotypes
Variant quality recalibration
Genotype refinement
Analysis-ready variants

Global sequence alignment

- The best alignment over the entire length of two sequences
- Suitable when the two sequences are of similar length, with a significant degree of similarity throughout.
- Example:

```
SIMILARITY
 |  |||
PIL-LAR---
```

Local sequence alignment

- Involving stretches that are shorter than the entire sequences, possibly more than one.
- Suitable when comparing substantially different sequences, which possibly differ significantly in length, and have only a short patches of similarity.
- For example, the local alignment of `SIMILARITY` and `PILLAR`:

```
SIMI-LARITY
   | |||
  PILLAR
```

Alignment "by eye"

- Consider the "best" alignment of ATGGCGT and ATGAGT

```
ATGGCGT
||| :||
ATG-AGT
```

- Intuitively we seek an alignment to maximize the number of residue-to-residue matches.

A mathematical framework

- *Sequence alignment is the establishment of residue-to-residue correspondence between two or more sequences such that the order of residues in each sequence is preserved.*
- A gap, which indicates a residue-to-nothing match, may be introduced in either sequence.
- A gap-to-gap match is meaningless and is not allowed.

The scoring scheme

- Given two sequences we need a number to associate with each possible alignment (i.e. the **alignment score** = goodness of alignment).
- The **scoring scheme** is a set of rules which assigns the alignment score to any given alignment of two sequences.

1. The scoring scheme is residue based: it consists of **residue substitution scores** (i.e. score for each possible residue alignment), plus penalties for gaps.
2. The alignment score is *the sum of substitution scores and gap penalties*.

A simple scoring scheme

- **Use +1 as a reward for a match, -1 as the penalty for a mismatch, and ignore gaps**

- The best alignment "by eye" from before:

```
ATGGCGT
ATG–AGT    score: +1 + 1 + 1 + 0 – 1 + 1 + 1 = 4
```

- An alternative alignment:

```
ATGGCGT
A–TGAGT    score: +1 + 0 – 1 + 1 – 1 + 1 + 1 = 2
```

Gaps

- So far we ignored gaps (amounts to gap penalty of 0)
- A gap corresponds to an insertion or a deletion of a residue
- A conventional wisdom dictates that the penalty for a gap must be several times greater than the penalty for a mutation. That is because a gap/extra residue
  - Interrupts the entire polymer chain
  - In DNA shifts the reading frame

Gap initiation and extension

- The conventional wisdom: the creation of a new gap should be strongly disfavored.
- However, once created insertions/deletions of chunks of more than one residue should be much less expensive (i.e. insertion of domains often occurs).
- A simple yet effective solution is **affine gap penalties**:

$$\gamma(n) = -o - (n-1)e$$

Affine gaps: a physical insight

- Affine gaps favor the alignment:

```
ATGTAGTGTATAGTACATGCA
ATGTAG-------TACATGCA
```

  Over the alignment:

```
ATGTAGTGTATAGTACATGCA
ATGTA--G--TA---CATGCA
```

- Exactly what we want from the biological viewpoint.

How do we find the best alignment?

- Brute-force approach:
  - Generate the list all possible alignments between two sequences, score them
  - Select the alignment with the best score
- The number of possible global alignments between two sequences of length *N* is

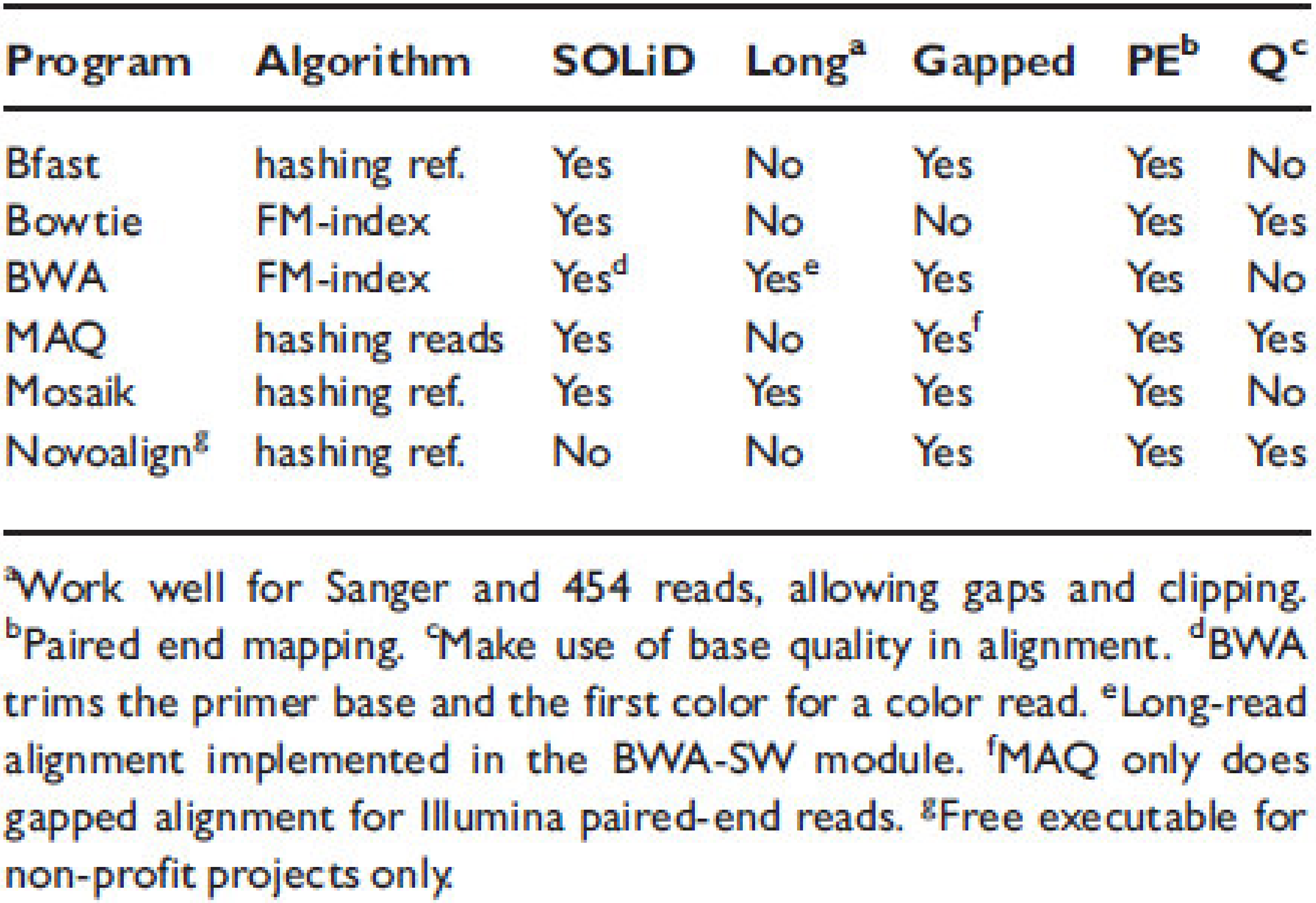
$$\frac{2^{2N}}{\sqrt{\pi N}}$$

- For two sequences of 250 residues this is ~10^{149}

How do we find the best alignment?

- Global sequence alignment: The **Needleman & Wunsch algorithm**
- Local sequence alignment: The **Smith & Waterman algorithm**
- Features
  - Both are based on dynamic programming
    - Break the problem into smaller subproblems.
    - Solve the smaller problems optimally.
    - Use the sub-problem solutions to construct an optimal solution for the original problem.
  - Optimal; guaranteed to find the best solution
  - $O(NM)$ running time and memory usage

Short-read alignment in NGS

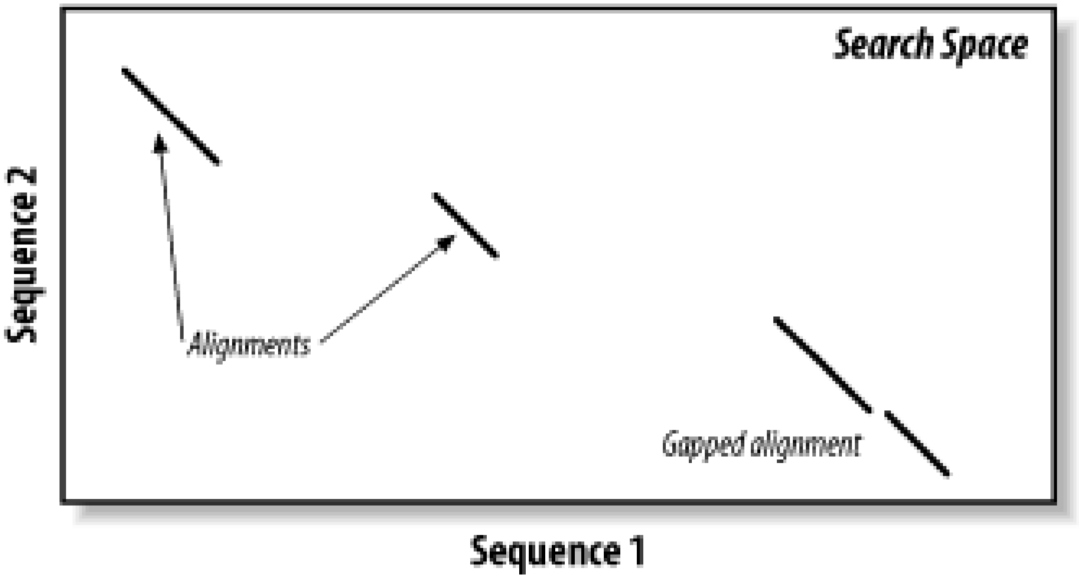


Program	Algorithm	SOLiD	Long[a]	Gapped	PE[b]	Q[c]
Bfast	hashing ref.	Yes	No	Yes	Yes	No
Bowtie	FM-index	Yes	No	No	Yes	Yes
BWA	FM-index	Yes[d]	Yes[e]	Yes	Yes	No
MAQ	hashing reads	Yes	No	Yes[f]	Yes	Yes
Mosaik	hashing ref.	Yes	Yes	Yes	Yes	No
Novoalign[g]	hashing ref.	No	No	Yes	Yes	Yes

[a] Work well for Sanger and 454 reads, allowing gaps and clipping. [b] Paired end mapping. [c] Make use of base quality in alignment. [d] BWA trims the primer base and the first color for a color read. [e] Long-read alignment implemented in the BWA-SW module. [f] MAQ only does gapped alignment for Illumina paired-end reads. [g] Free executable for non-profit projects only.

Algorithms based on hash tables

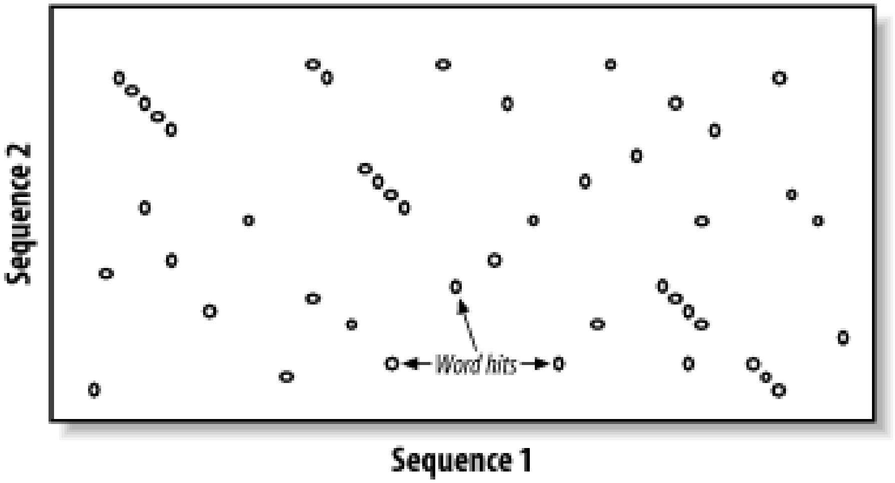
- BLAST

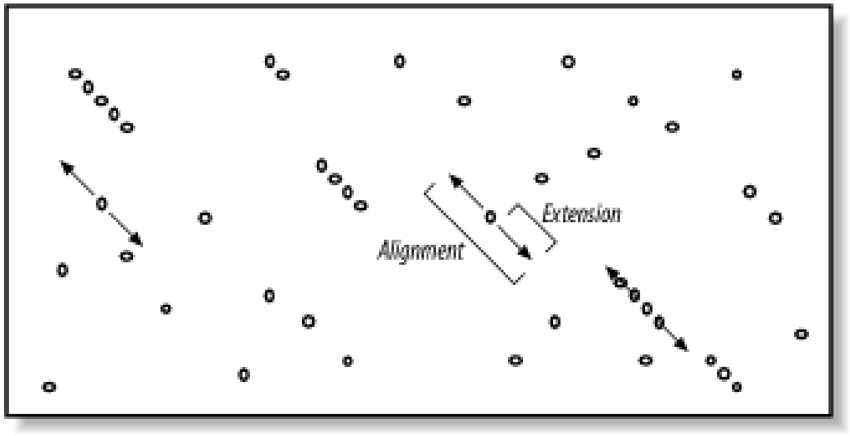


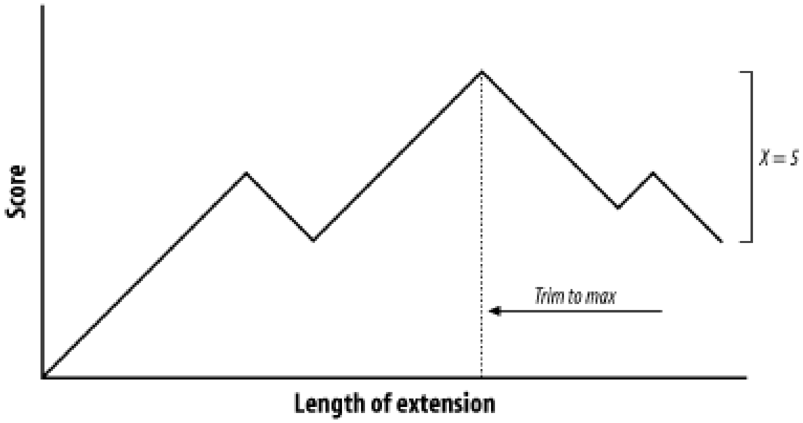
Three steps:

1. Seeding
2. Extension
3. Evaluation

BLAST


Sequence 2
Word hits
Sequence 1


Extension
Alignment


Score
X = 5
Trim to max
Length of extension

Algorithms based on hash tables

- Improvement on seeding: **spaced seed**

  seed: `111010010100110111`

  11 matches, 55% more sensitivity
  for two sequences of 70% similarity

Eland (reads, 2 mismatches)

SOAP (ref., 2 mismatches)

MAQ (reads, *k* mismatches)

Algorithms based on hash tables

- Improvement on seeding: **q-gram filter, multiple seed hits**

  The occurrence of a *w*-long query string with at most *k* differences, the query and the *w*-long database substring share at least (*w*+1)-(*k*+1)*q* common substring of length *q*.

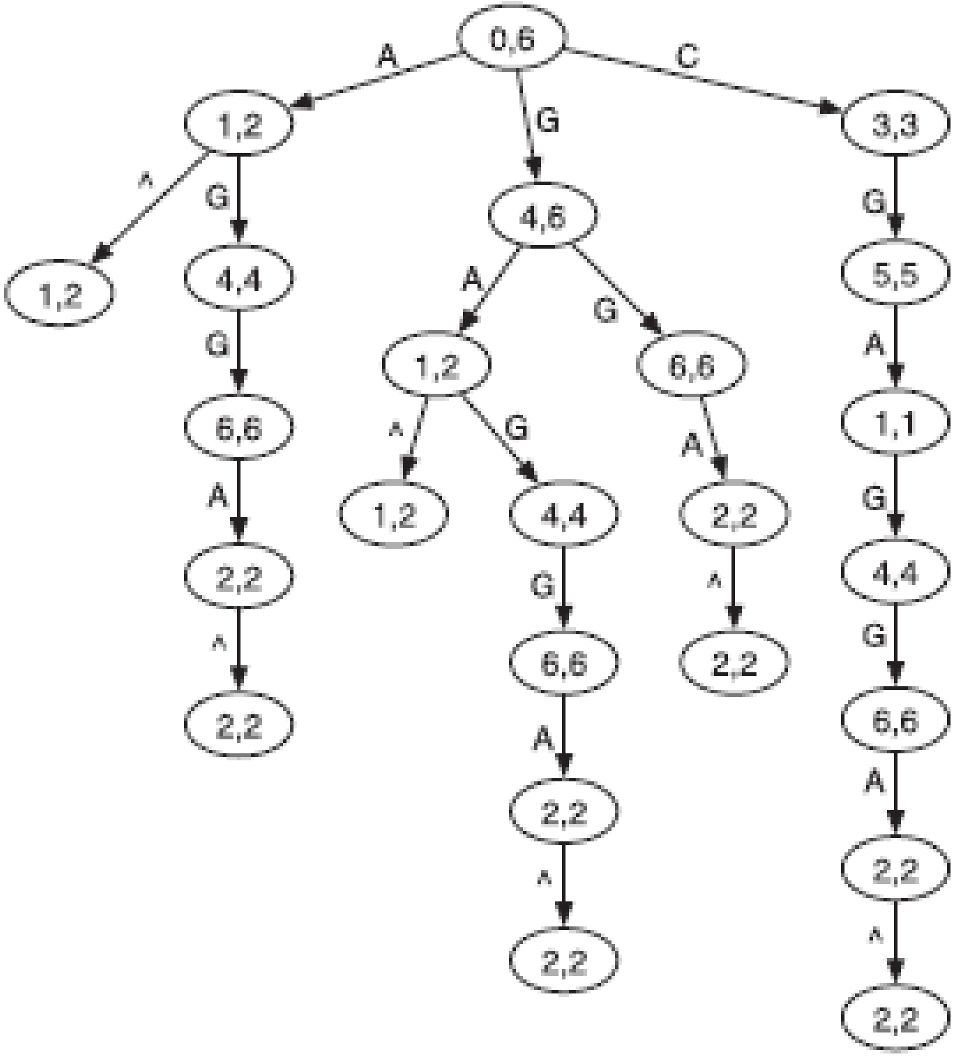
  SHRiMP

  RazerS

  SSAHA2, BLAT (long-read alignment)

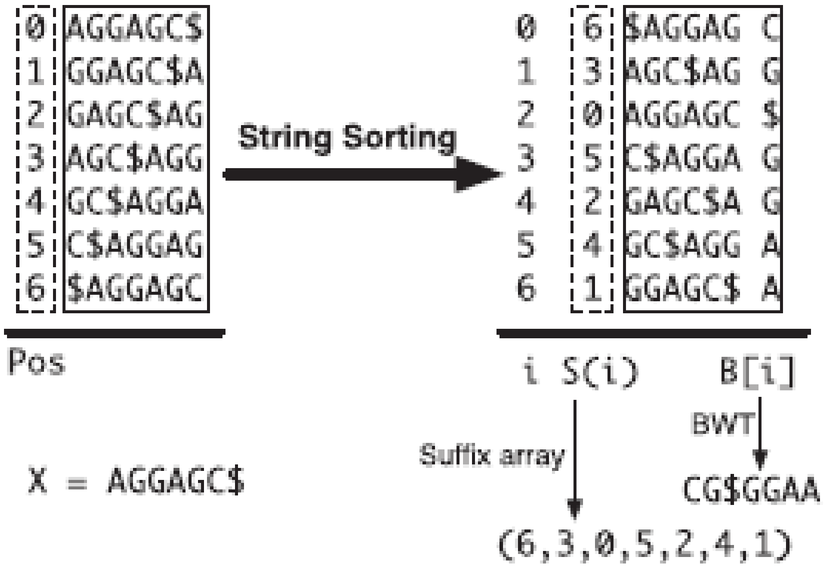
Algorithms based on prefix/suffix tries

- Exact matches => inexact matches
- Burrows-Wheeler transform of the reference sequence



REF = AGGAGC$

Find „AG”!



```
0 AGGAGC$
1 GGAGC$A
2 GAGC$AG
3 AGC$AGG
4 GC$AGGA
5 C$AGGAG
6 $AGGAGC
```

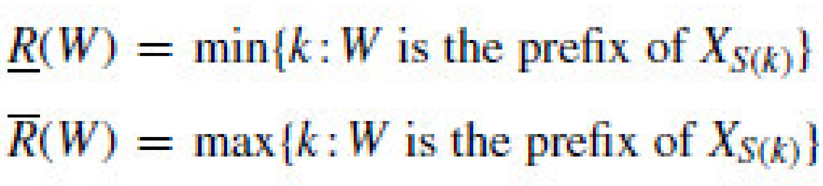
```
0  6  $AGGAG C
1  3  AGC$AG G
2  0  AGGAGC $
3  5  C$AGGA G
4  2  GAGC$A G
5  4  GC$AGG A
6  1  GGAGC$ A
```

X = AGGAGC$

CG$GGAA

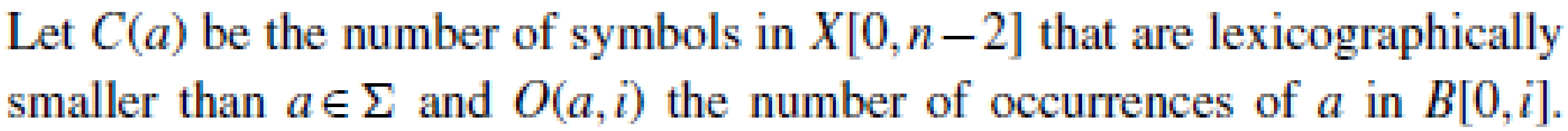
(6,3,0,5,2,4,1)

Suffix array interval, exact matching: backward search

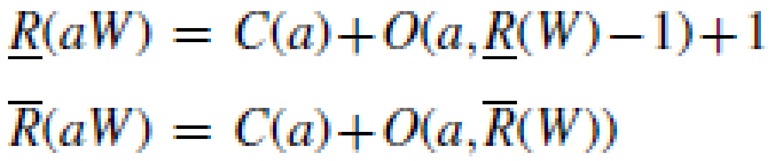


$$\underline{R}(W) = \min\{k : W \text{ is the prefix of } X_{S(k)}\}$$

$$\overline{R}(W) = \max\{k : W \text{ is the prefix of } X_{S(k)}\}$$



Let $C(a)$ be the number of symbols in $X[0,n-2]$ that are lexicographically smaller than $a \in \Sigma$ and $O(a,i)$ the number of occurrences of a in $B[0,i]$.

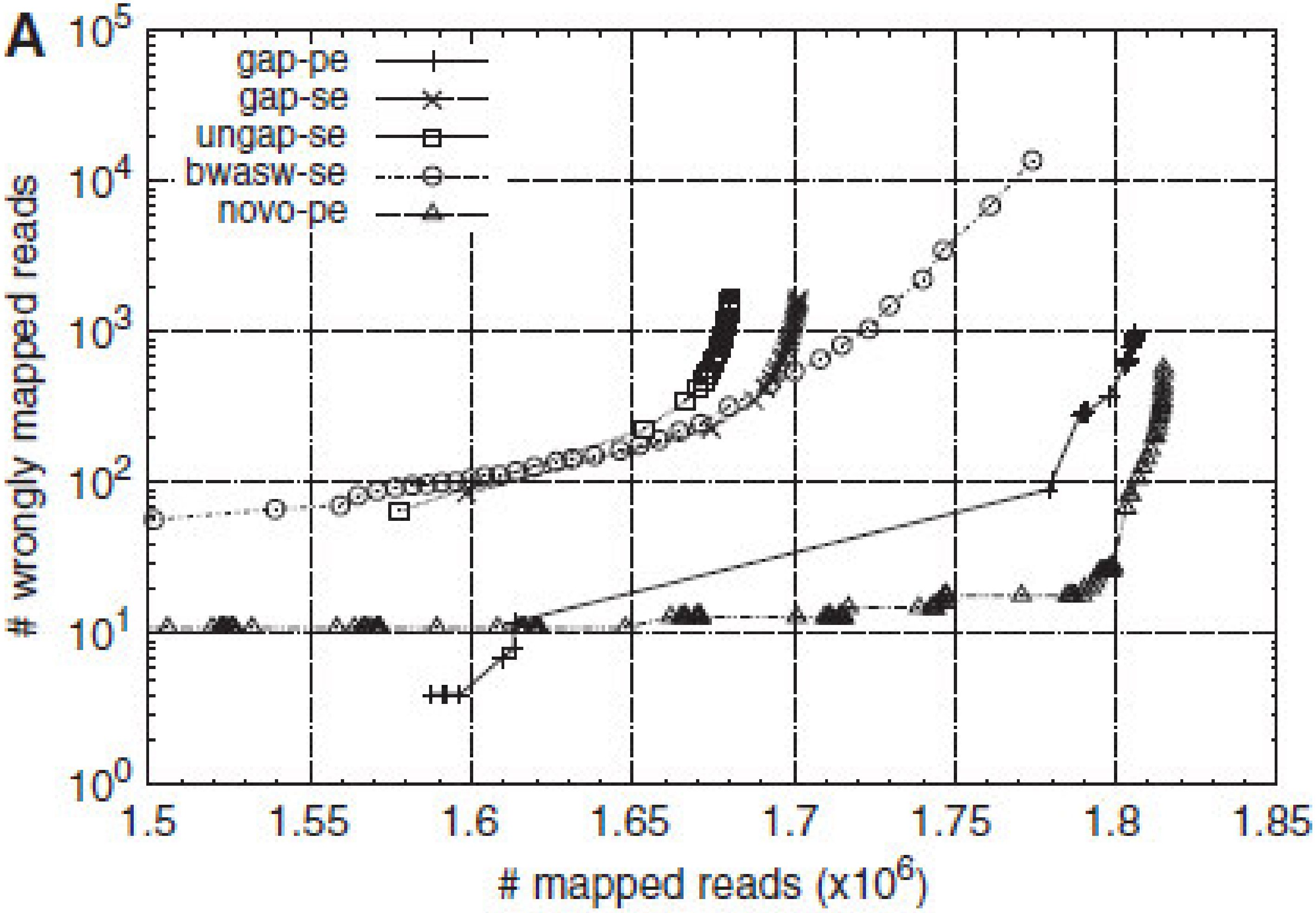


$$\underline{R}(aW) = C(a) + O(a, \underline{R}(W) - 1) + 1$$

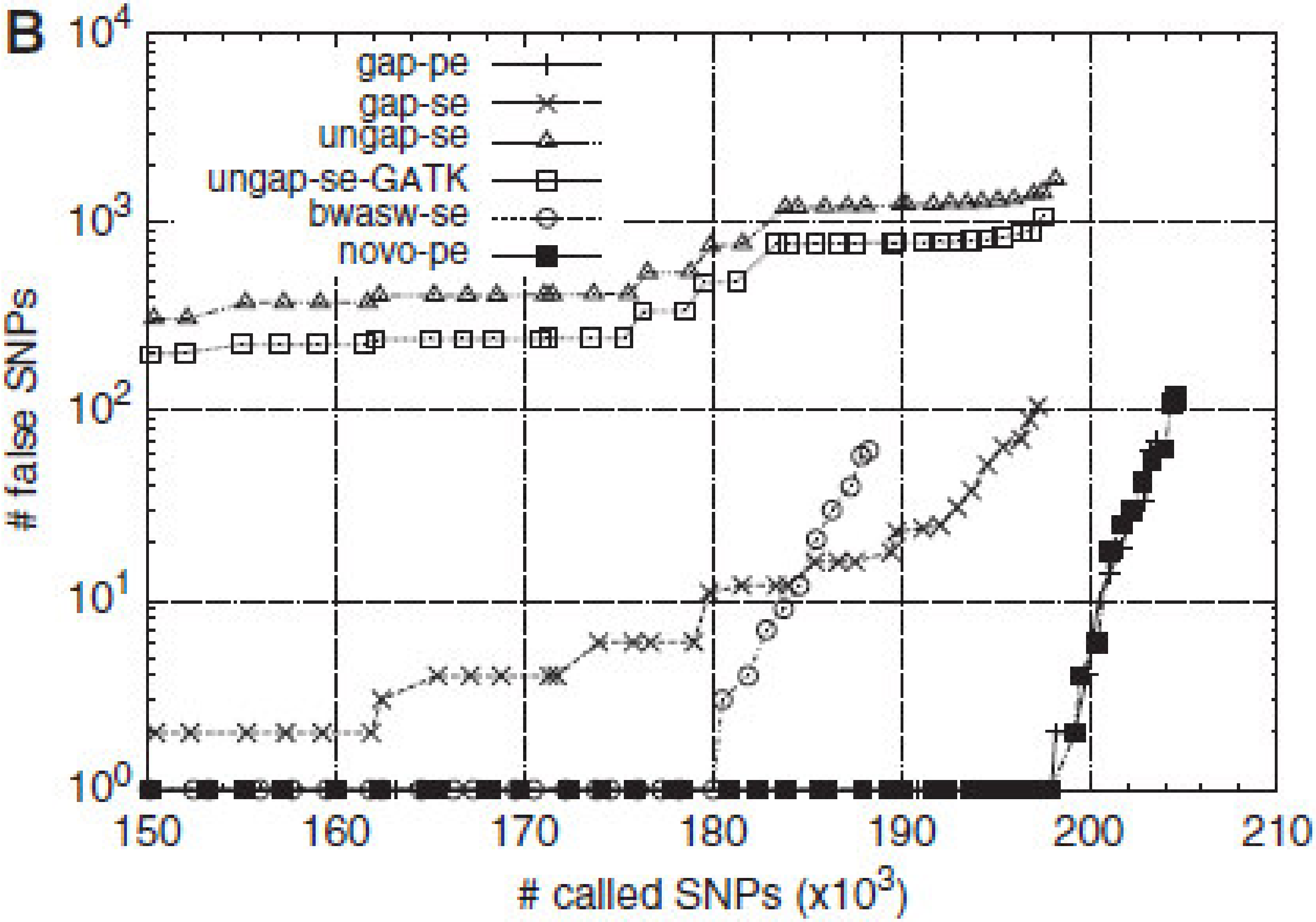
$$\overline{R}(aW) = C(a) + O(a, \overline{R}(W))$$

- BWA, BWA-SW
- Bowtie
- SOAPv2

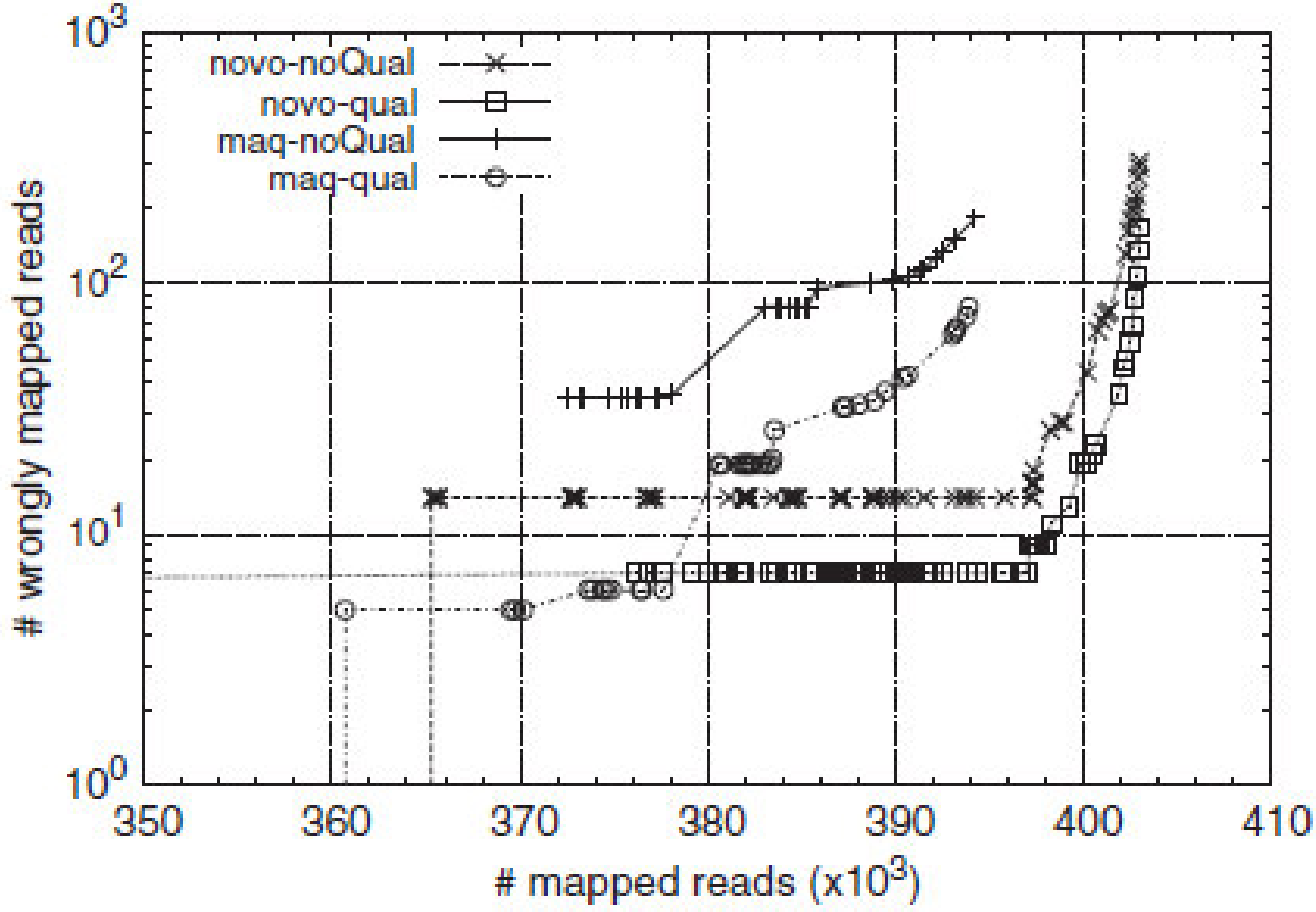
Effect of gapped alignment



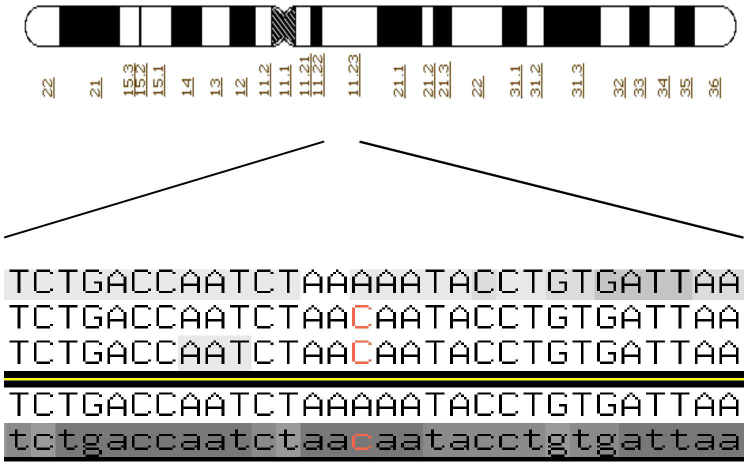
Effect of gapped alignment



Using base quality in alignment



Variant calling - SNPs

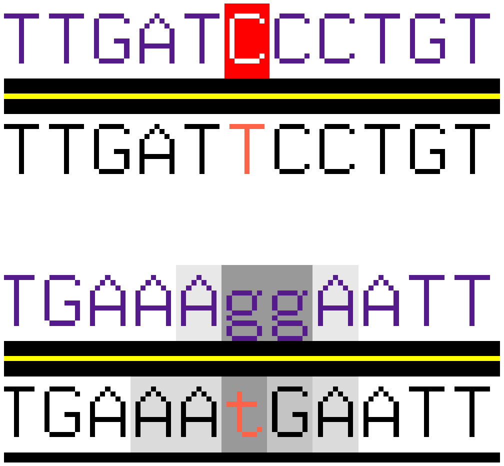


```
TCTGACCAATCTAAAAATACCTGTGATTAA
TCTGACCAATCTAACAATACCTGTGATTAA
TCTGACCAATCTAACAATACCTGTGATTAA

TCTGACCAATCTAAAAATACCTGTGATTAA
tctgaccaatctaacaatacctgtgattaa
```

- look at multiple sequences from the same genome region

- use base quality values to decide if mismatches are true polymorphisms or sequencing errors



```
TTGATCCCTGT

TTGATTCCTGT

TGAAAggAATT

TGAAAtGAATT
```

SNP variant callers

- SAMTools
- GATK
- VarScan
- BreakDancer

Structural variants

- Polymorphisms that **change the structure of the genome**, including all insertions, deletions and inversions.
- Categories:
  - Copy-number variants: insertions, deletions
  - Copy-count invariant: inversions, translocations
- Quality of discovery methods:
  - *Sensitivity, specificity*
  - *Location of breakpoints*
  - *Variant's size*
  - *Change in copy count*

Structural variant discovery with NGS

1. Detect *signatures*

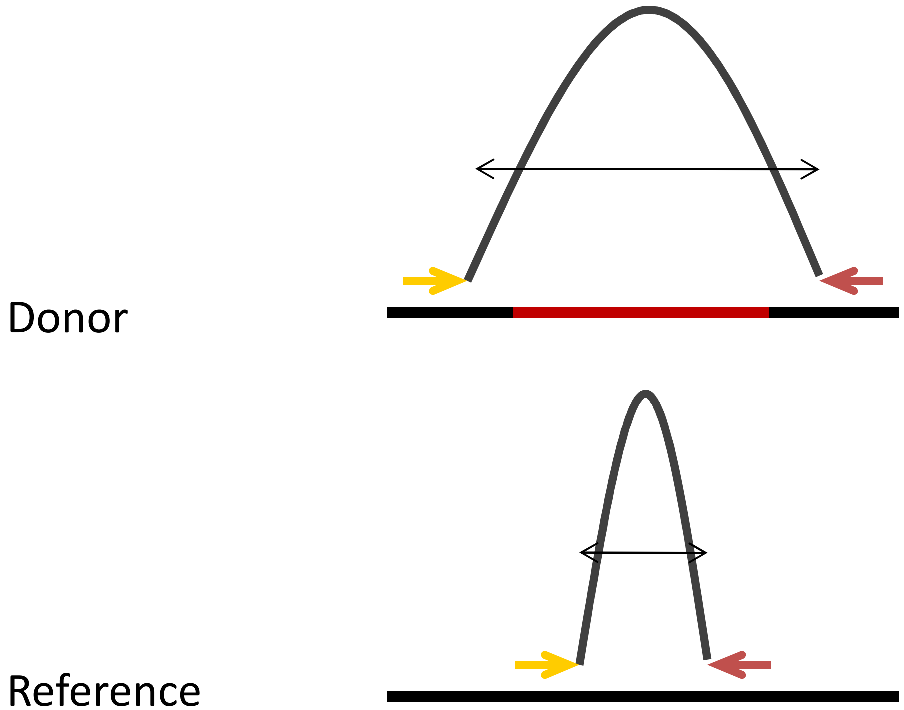
    Patterns of paired-end mappings (PEMs) and/or read depth (RD) that are created by structural variations.

2. Call the underlying variant

    Different events => different signatures

    Noise (sequencing errors, insert size distribution, ambiguous mapping, sequencing bias, etc.) => many false positives => **statistical significance**

Signatures based on PEM
Basic insertion

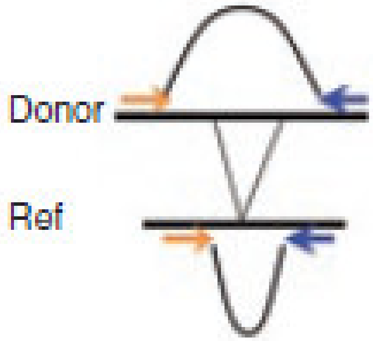


Donor

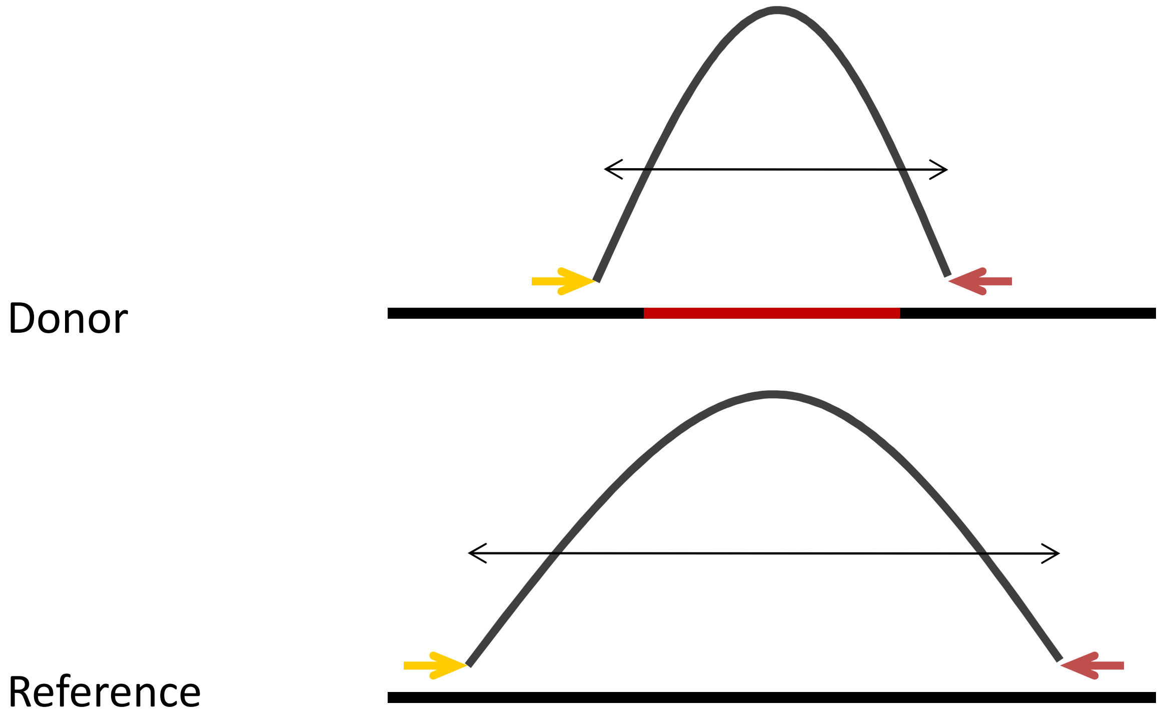
Reference

Signatures based on PEM
Basic insertion

- A mate pair that spans an isolated insertion event maps to the corresponding regions of the reference genome, but the mapped distance is smaller than the insert size.
- The size of insertion must be smaller than the insert size.
- Does not indicate the inserted segment.

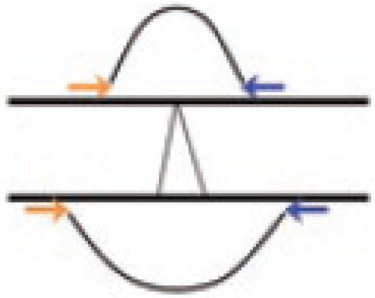


Signatures based on PEM
Basic deletion

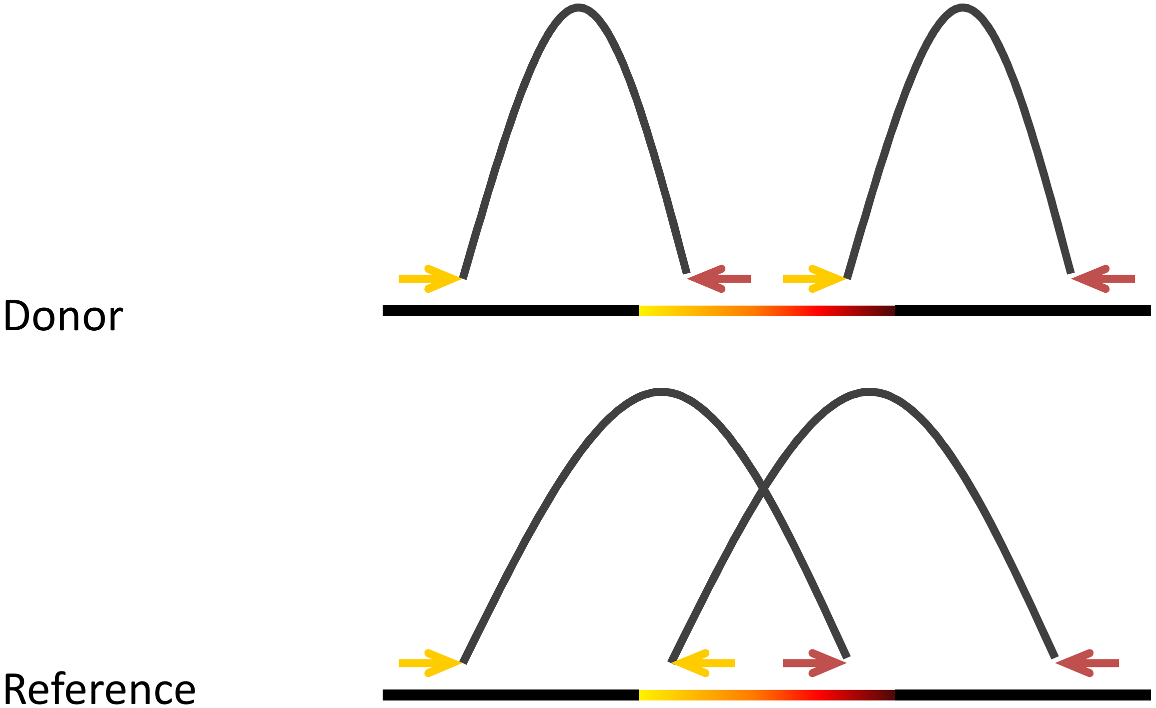


Signatures based on PEM
Basic deletion

- A mate pair that spans an isolated deletion event maps to the corresponding regions of the reference genome, but the mapped distance is greater than the insert size.



Signatures based on PEM
Basic inversion

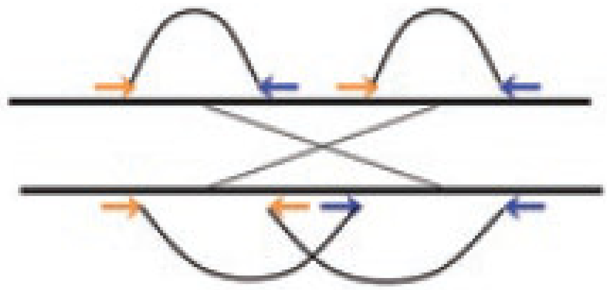


Donor

Reference

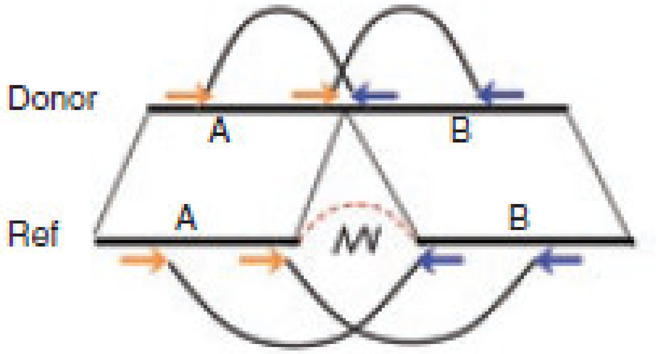
Signatures based on PEM
Basic inversion

- A mate pair that spans either (but not both) of its breakpoints maps to the reference genome with the orientation of the read, lying within the inversion, flipped.
- Two mate pairs form this *basic inversion signature*.

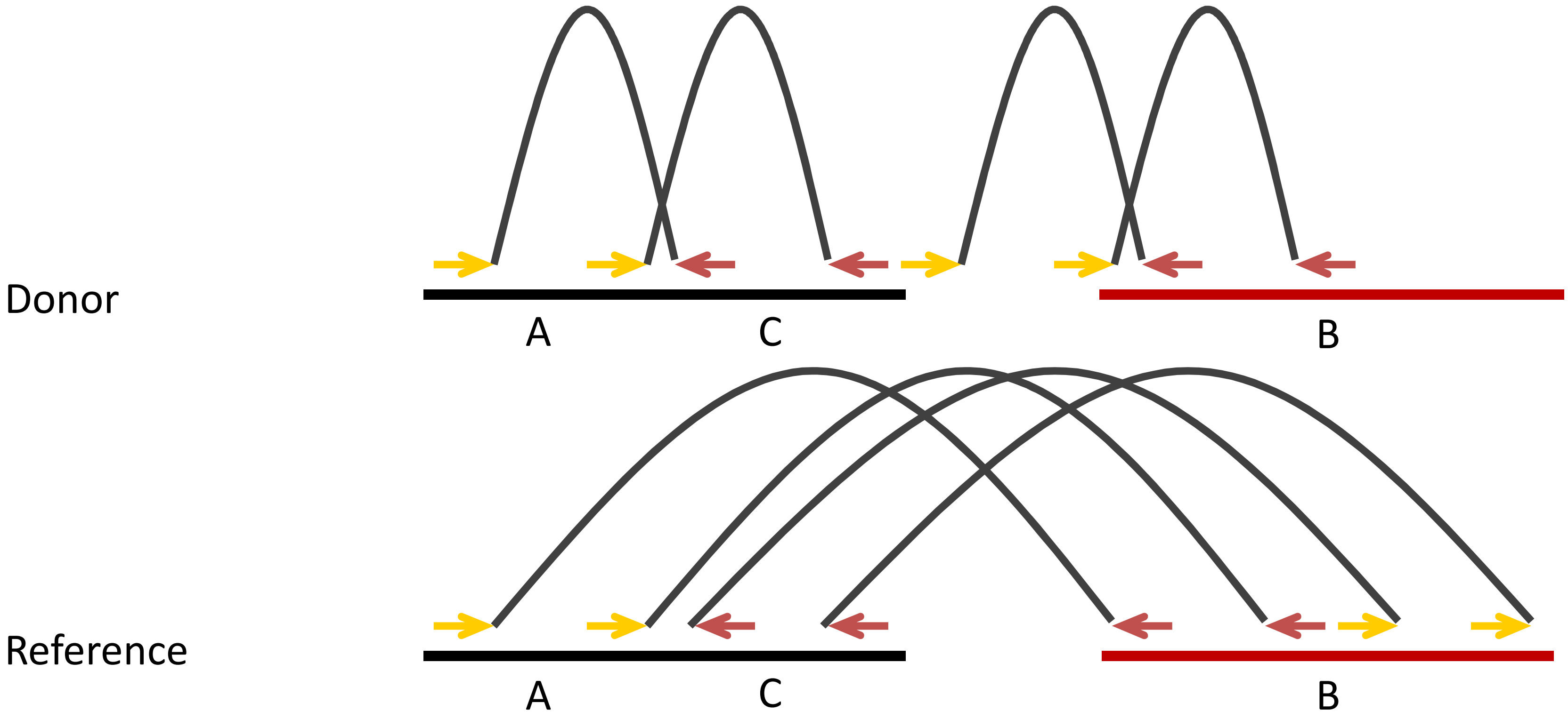


Signatures based on PEM
Linking signature

- Consider: two distant regions of the reference genome that are adjacent in the donor.
- A mate pair spanning the donor's breakpoint maps with a distance much greater than the insert size.
- The two spanning mate pairs that are closest to the breakpoint form a *linking signature*.

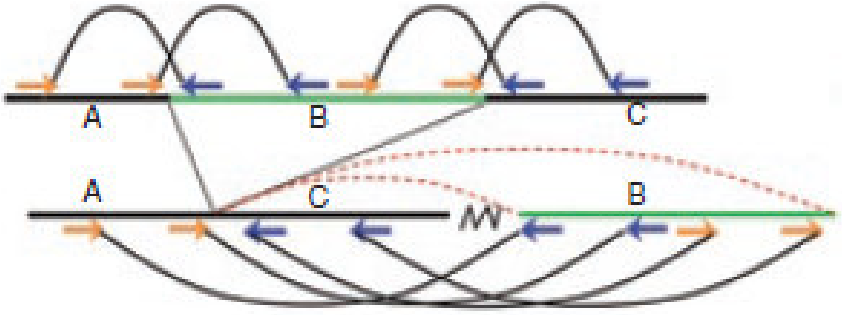


Signatures based on PEM
Linked insertion

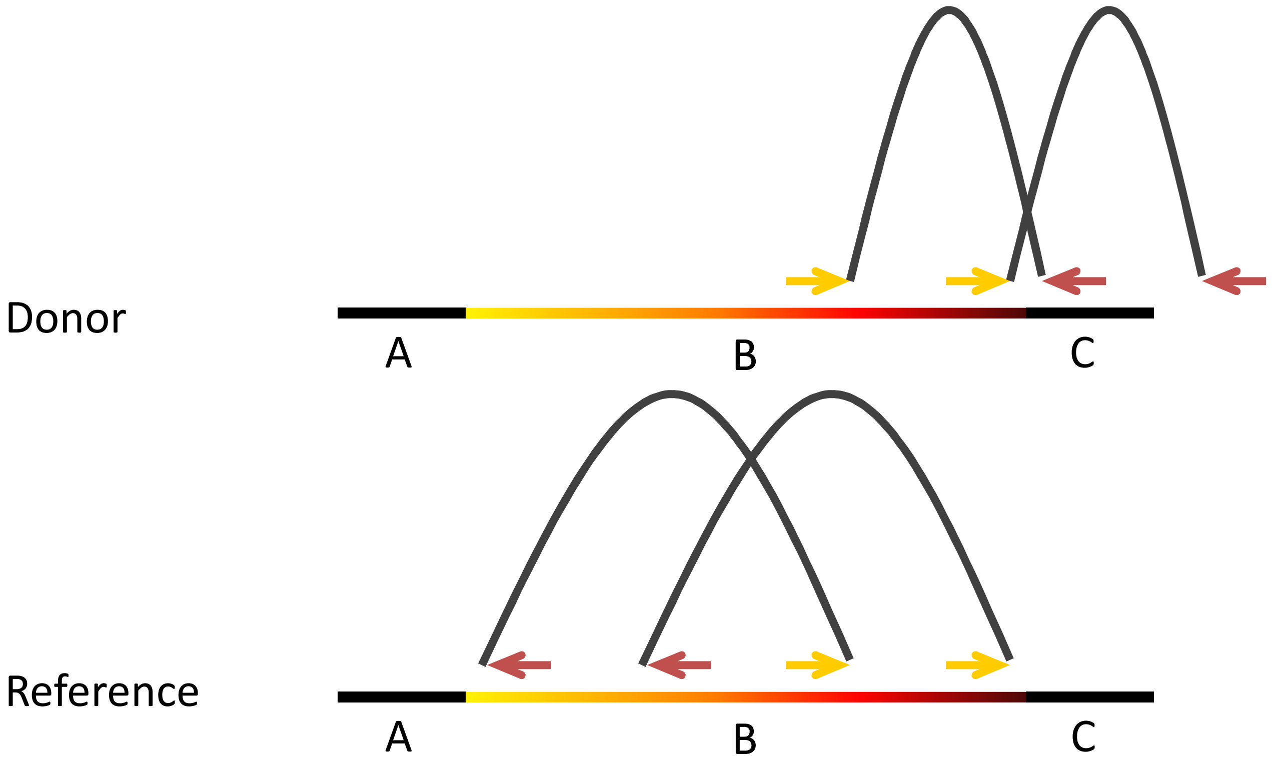


Signatures based on PEM
Linked insertion

- The inserted sequence is present elsewhere in the reference genome.
- Two linking signatures where the linked regions are close to each other.
- The inserted region can be identified.
  (if not very large)

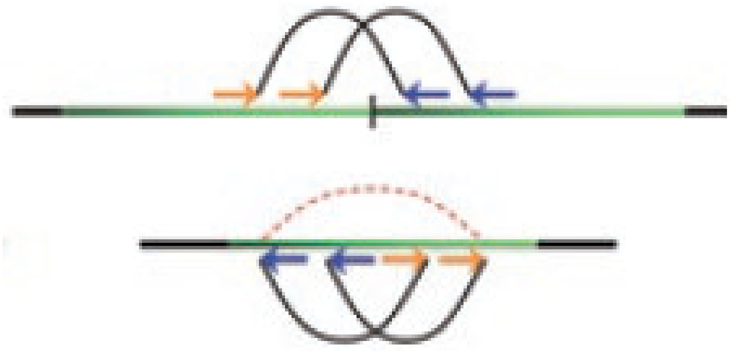


Signatures based on PEM
Everted duplication

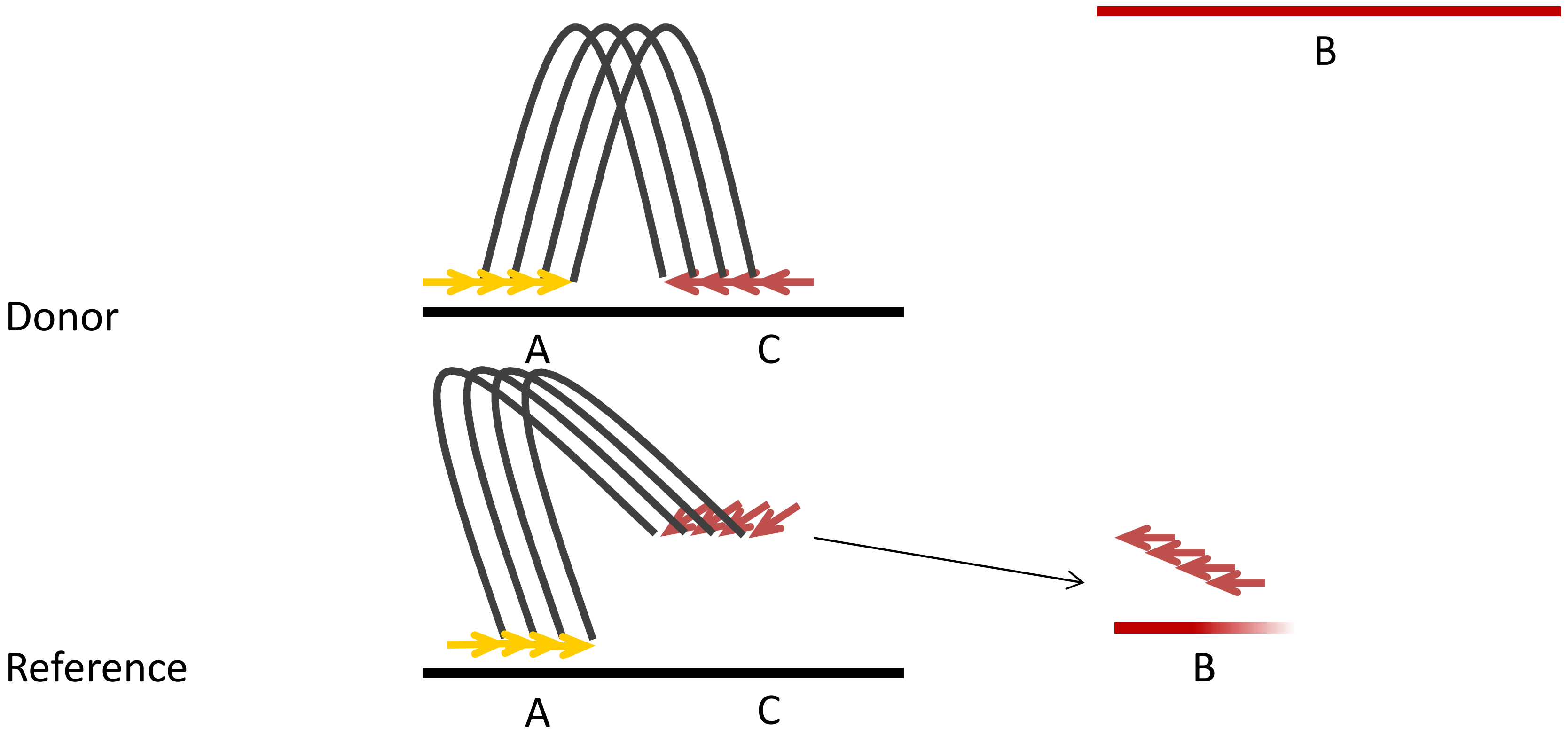


Signatures based on PEM
Everted duplication

- A region of the reference genome has been tandemly duplicated in the donor.
- The order of the mates is reversed while the orientation stays the same.
- Only novel tandem duplications are detectable.

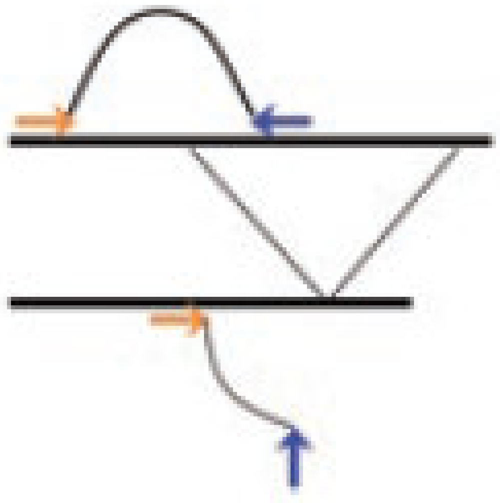


Signatures based on PEM
Hanging insertions

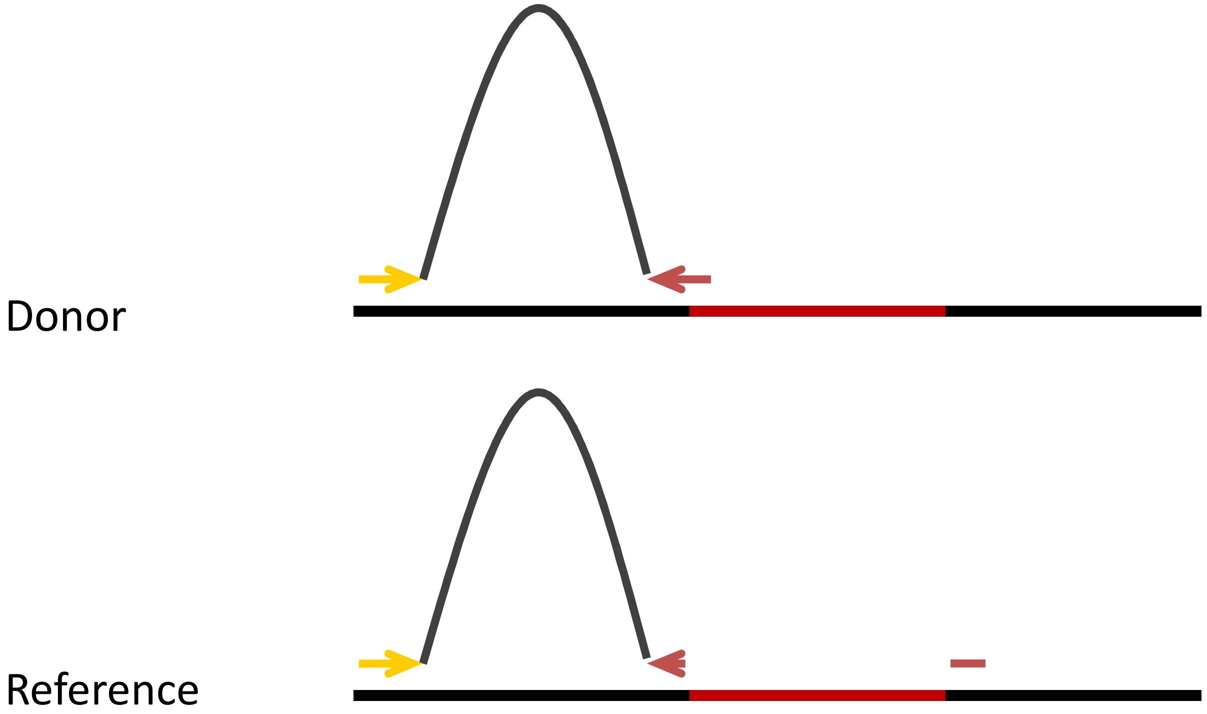


Signatures based on PEM
Hanging insertions

- Insertion of novel genomic segment: One read of a mate pair maps to the reference genome while the other read does not.
- With *de novo* assembly of the hanging reads one can reconstruct a small inserted segment (if it is substantially larger than the insert size, it will not cover the entire insertion).

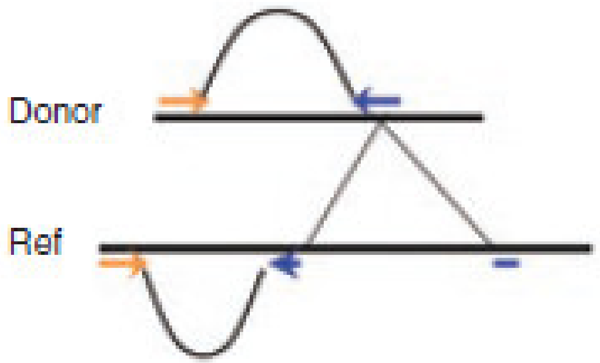


Signatures based on PEM
Anchored split mapping

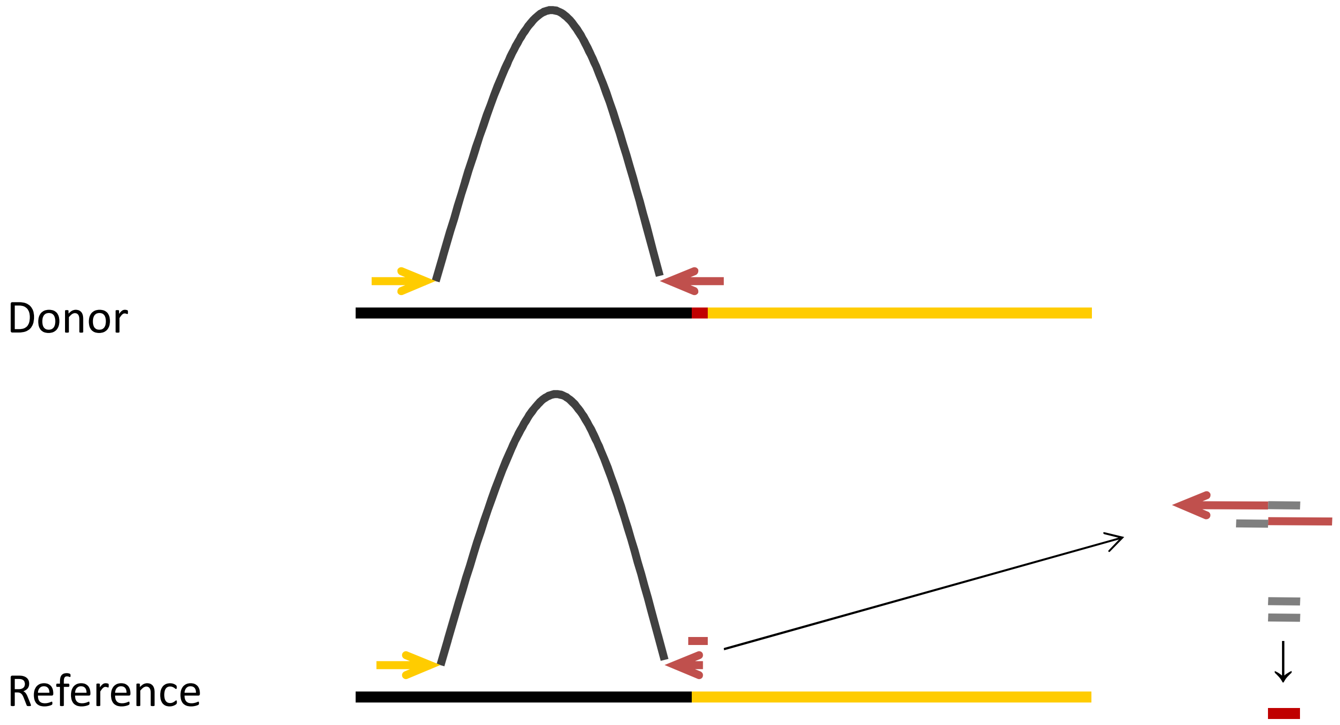


Breakpoint identification
Anchored split mapping

- Deletion breakpoint: a prefix and suffix of a read map to different locations.
- Detectable with longer reads.
- One of the mates maps to the reference and the other has a split mapping with one of its parts about 1 insert size away.
- Base pair precision.
- Deleted region can not be too large.

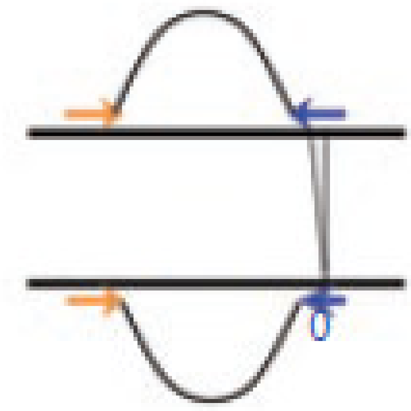


Signatures based on PEM
Anchored split mapping



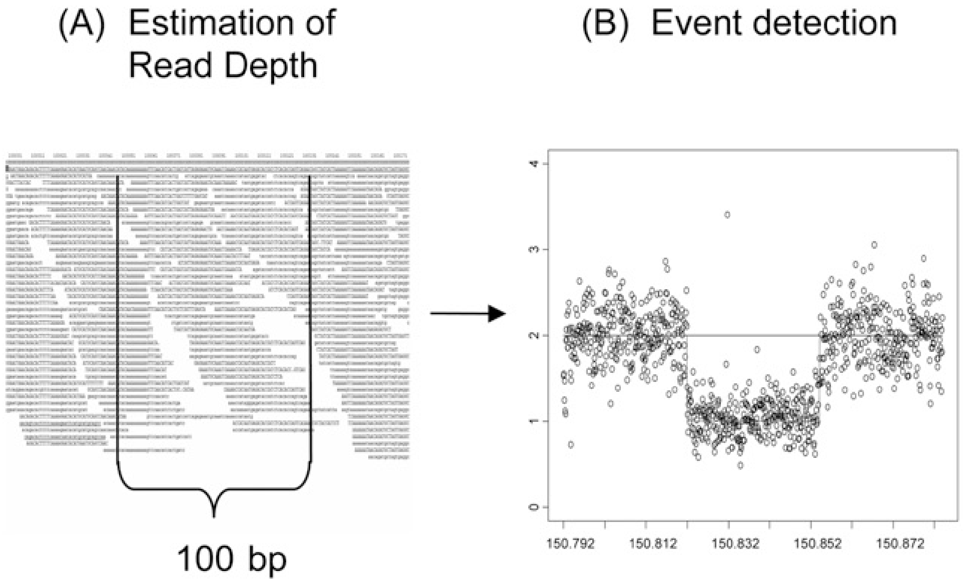
Breakpoint identification
Anchored split mapping

- Insertion of a few base pairs: the split read will have a prefix and suffix mapping to adjacent locations, and there will be a middle part of the read (the inserted bases) that will not be part of either the prefix or suffix mapping.
- Detects insertions of only a few base pairs.



Signatures based on depth of coverage

- Depth of coverage
  - Assuming uniform sequencing process => the number of reads mapping to a region follows Poisson distribution and is expected to be proportional to the number of times the region appears in the donor.
  - A region that has been deleted (duplicated) will have less (more) reads mapping to it.
  - ‘Gain/loss’ signature.

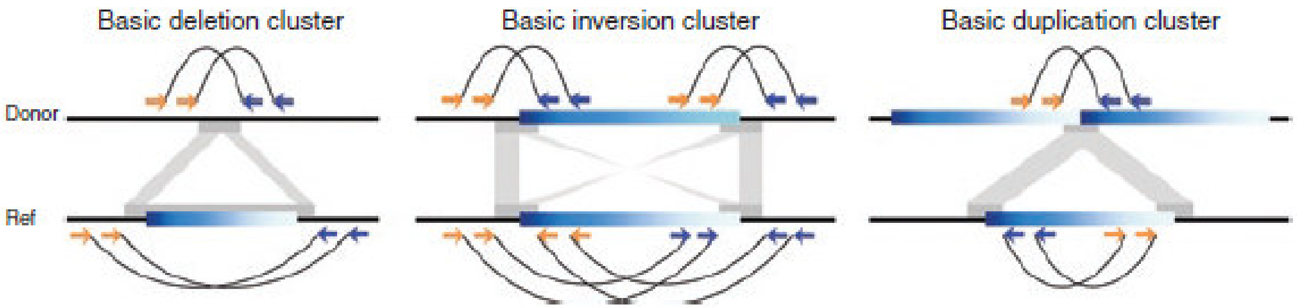


Signatures based on depth of coverage

- Pros/cons:
  - The DOC signature does not indicate where an insertion occured, but rather what duplicate sequence has been inserted; it is *not able to detect insertions of novel sequence.*
  - Directly *related to the coverage and to the size of the CNV.*
  - It can *detect very large CNVs* – the larger the event, the stronger the signature.
  - *Not able to identify smaller events*, that PEM signatures, even with low coverage, are able to detect.
  - *Much poorer at localising breakpoints.*

Variant calling

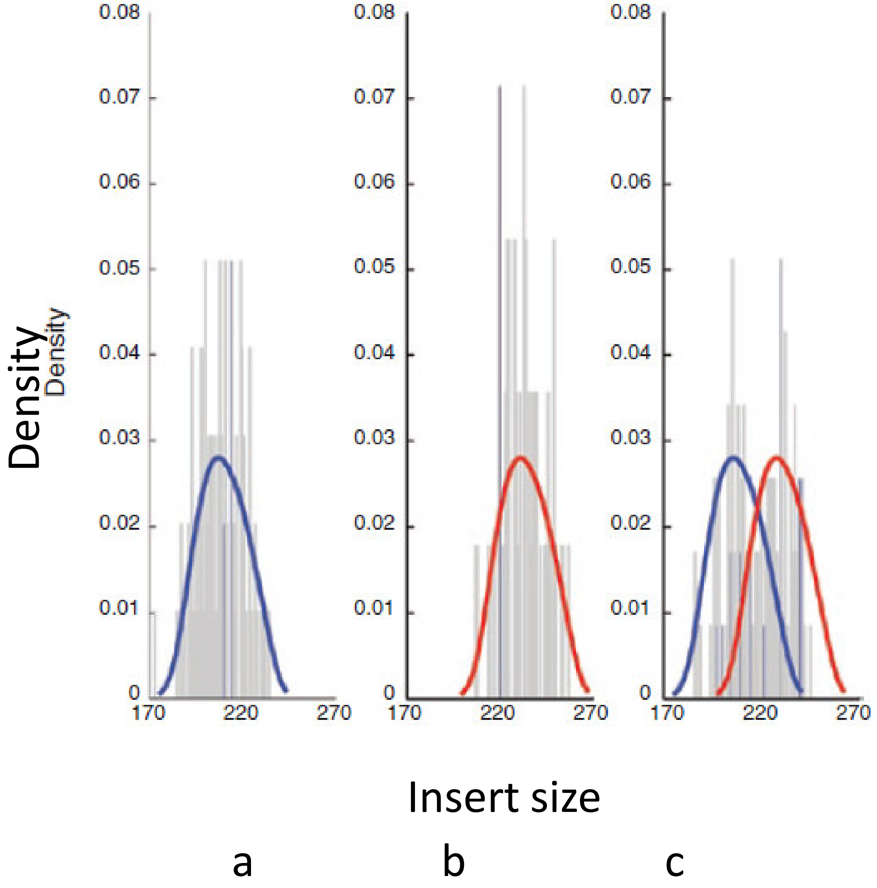
- Grouping signatures
  - Improve confidence
  - Increase the precision of the predicted breakpoints and event size
- Paired-end mapping clusters
- Depth-of-coverage windows



Paired-end mapping clusters

- Standard clustering strategy (widely used)
  - Considers mate pairs that do not have a ‘concordant’ mapping (one with correct orientation and with a mapped distance within 2-4 s.d. of the mean insert size).
  - Ignores any mate pair that has more than one good mapping.
  - Cluster: minimum (usually two) signatures of the same type and with similar size and location.
- Weaknesses:
  - does *not allow the detection of signatures within repetitive regions* of the genome. => *‘soft’ approaches*: optimization procedures to assign each mate pair to a cluster where it will have the most support from other mate pairs
  - Can not detect smaller indels than the fixed cutoff for the ‘discordant’ definition => distribution based clustering

Distribution based clustering



Distribution of all the mappings spanning a given location on the genome.

a)No variation. The mean of the distribution is at about the insert size (208 bp).

b)Homozygous deletion of size 24 bp. The mean is at ~232 bp.

c)Hemizygous deletion of size 22 bp, which results in a mixture of two separate distributions with means of 208 bp and 230 bp.

Depth-of-coverage windows

- Standard method:
  - Partition the reference into windows so that the coverage depth is consistent within a window but has a sharp difference between adjacent windows.
  - Each window correspond to a gain/loss/no event.

Tools	Downloadable	Basic deletion	Basic insertion	Basic inversion	Linking	Linked insertion	Hanging insertion	Anchored split mapping	Everted duplication	Gain/loss	Clustering and/or windowing strategies
PEMer	+	*	*	*	*	*					Standard
SeqSeq	+									*	Local change-point analysis
VariationHunter	+	*	*	*					*		Soft
MoDIL	+	*	*								Soft, distribution-based
Pindel	+							*			Standard
BreakDancer	+	*	*	*			*				Standard, distribution-based
ABI Tools	+	*	*	*						*	Standard, distribution-based
Roche RM	-	*	*					*			Standard
Illumina	-										?

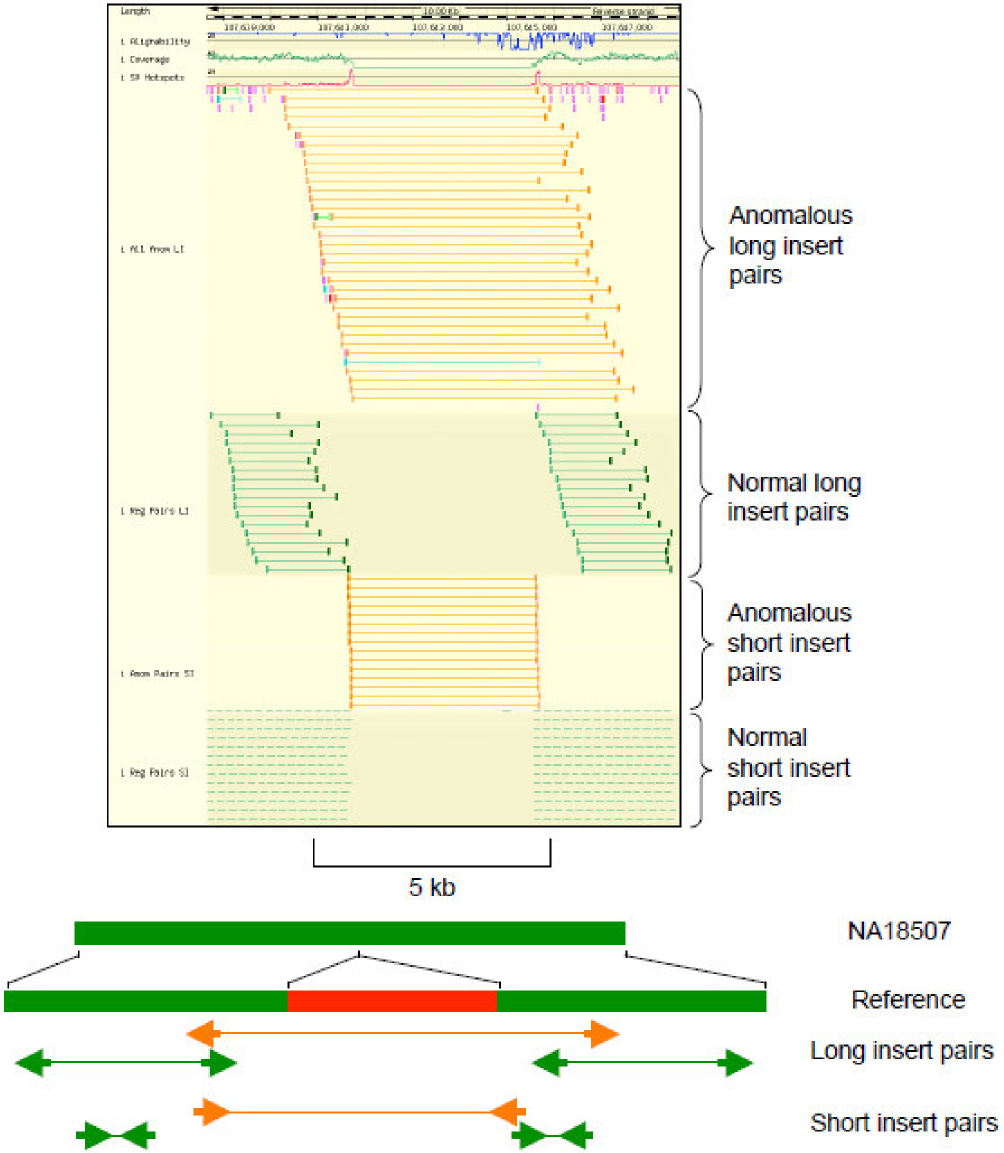
Limitations

- Repeating regions
- Low resolution of Depth of Coverage methods
- Insert size of library
  - Long insert size => detection of larger events (e.g. basic insertion)
  - Short insert size => detection of smaller events

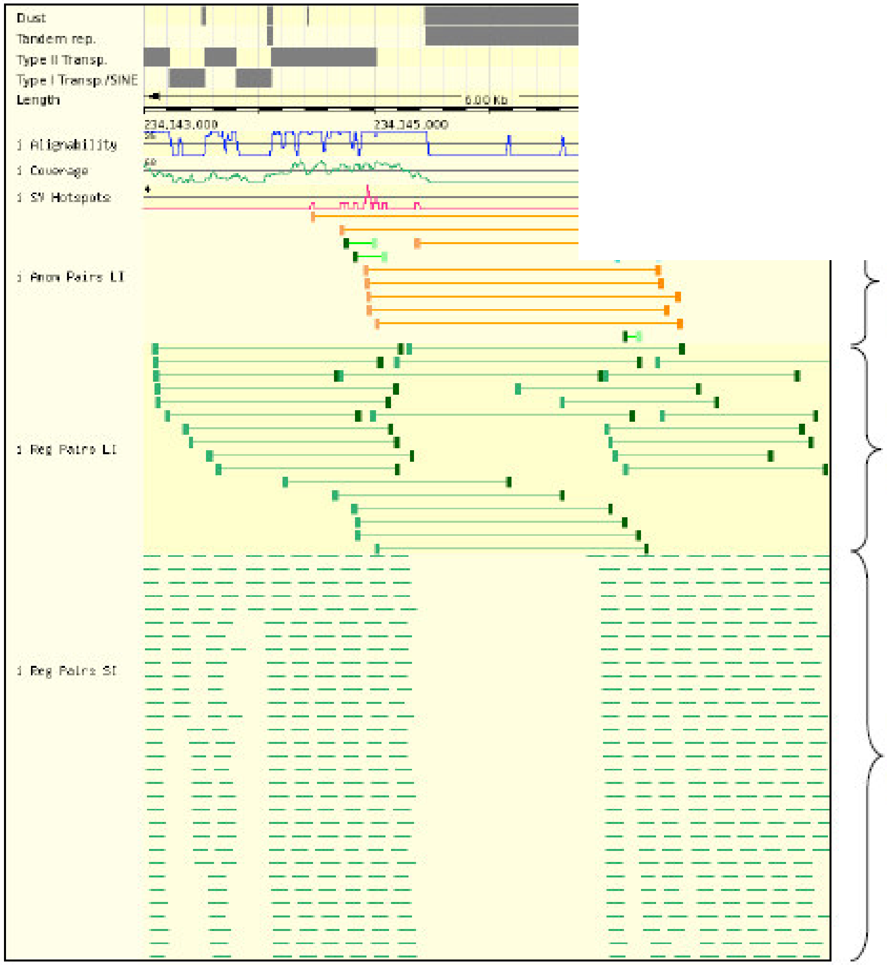
  => Use multiple libraries with varying insert sizes

Bentley et al.
Accurate whole genome sequencing using reversible terminator chemistry.
***Nature* 456, 53-59 (2008)**

A. Homozygous 3.6 kb deletion in NA18507 relative to the reference, supported by anomalously spaced long and short insert read pairs (orange) compared to regularly spaced read pairs (green).


Anomalous long insert pairs
Normal long insert pairs
Anomalous short insert pairs
Normal short insert pairs
5 kb
NA18507
Reference
Long insert pairs
Short insert pairs

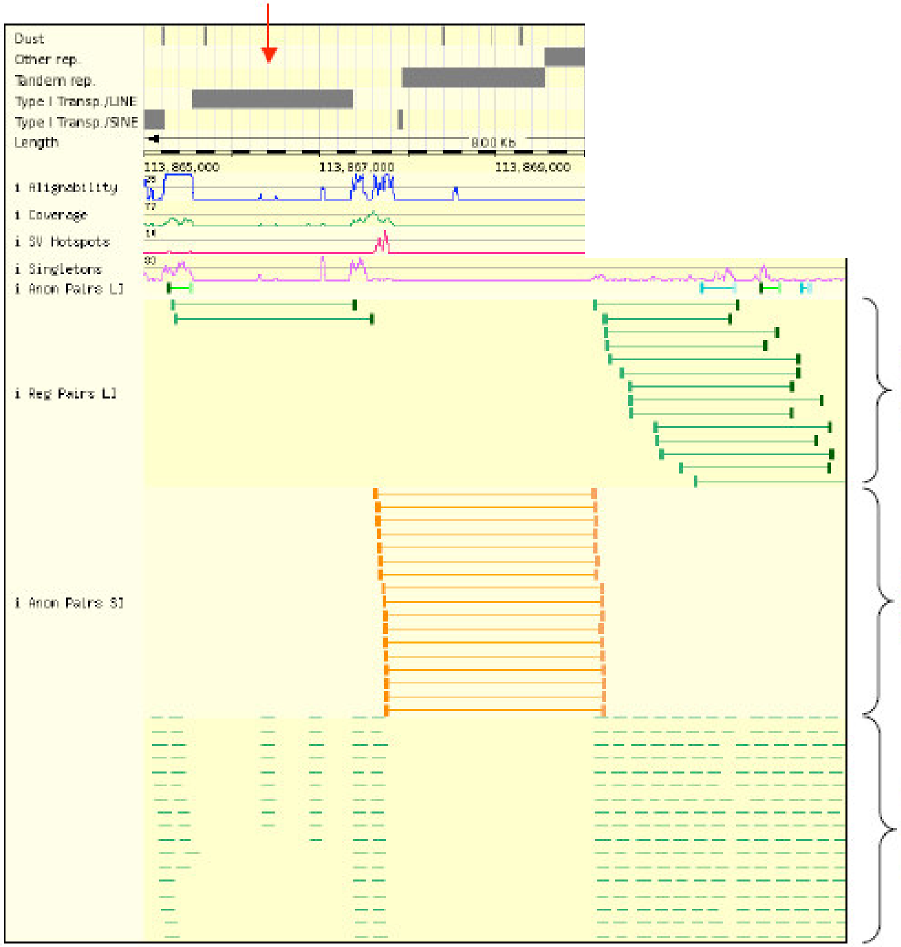
B. Deletion in NA18507 relative to the reference, seen with anomalous long insert pairs but not short insert data.



long insert pairs

Normal long insert pairs

Normal short insert pairs



Normal long insert pairs

Anomalous short insert pairs

Normal short insert pairs

C. Deletion in NA18507 relative to the reference, seen with anomalous short insert pairs but not long insert data.