BreaKmer: detection of structural variation in targeted massively parallel sequencing data using kmers

Abo, MacConaill et al.
Genomic Structural Variations
Genomic Structural Variations

BCR-ABL fusion gene in Chronic Myeloid Leukemia
Genomic Structural Variations

• SV’s are one of the driving mechanisms of cancer

BCR-ABL fusion gene in Chronic Myeloid Leukemia
Genomic Structural Variations

• SV’s are one of the driving mechanisms of cancer
• InDels, Translocations, Rearrangements and genomic copy losses/gains

BCR-ABL fusion gene in Chronic Myeloid Leukemia
Genomic Structural Variations

• SV’s are one of the driving mechanisms of cancer
• InDels, Translocations, Rearrangements and genomic copy losses/gains
• Detecting known SV’s
Genomic Structural Variations

- SV’s are one of the driving mechanisms of cancer
- InDels, Translocations, Rearrangements and genomic copy losses/gains
- Detecting known SV’s
- Identifying novel SV’s

BCR-ABL fusion gene in Chronic Myeloid Leukemia
BreaKmer – A novel method for identifying SV’s
BreaKmer – A novel method for identifying SV’s

• Traditional clinical methods - slow, costly and challenging.
BreaKmer – A novel method for identifying SV’s

• Traditional clinical methods - slow, costly and challenging.
• WGS of tumors – optimal solution yet still very expensive and for now still unfeasible in a clinical setting.
BreaKmer – A novel method for identifying SV’s

• Traditional clinical methods - slow, costly and challenging.
• WGS of tumors – optimal solution yet still very expensive and for now still unfeasible in a clinical setting.
• BreaKmer
BreaKmer – A novel method for identifying SV’s

- Traditional clinical methods - slow, costly and challenging.
- WGS of tumors – optimal solution yet still very expensive and for now still unfeasible in a clinical setting.
- BreaKmer
  - Using WGS data but targeting specific regions – quicker
BreaKmer – A novel method for identifying SV’s

• Traditional clinical methods - slow, costly and challenging.
• WGS of tumors – optimal solution yet still very expensive and for now still unfeasible in a clinical setting.
• BreaKmer
  • Using WGS data but targeting specific regions – quicker
  • Using all alignment data available: unmatched pairs, mis-aligned reads and discordant reads.
BreaKmer – A novel method for identifying SV’s

• Traditional clinical methods - slow, costly and challenging.
• WGS of tumors – optimal solution yet still very expensive and for now still unfeasible in a clinical setting.
• BreaKmer
  • Using WGS data but targeting specific regions – quicker
  • Using all alignment data available: unmatched pairs, mis-aligned reads and discordant reads.
  • Sequence assembly from reads using k-mers is the core.
Discordant and misaligned reads
Discordant and misaligned reads

Discordant reads:
Discordant and misaligned reads

Discordant reads:
Discordant and misaligned reads

Discordant reads:
Discordant and misaligned reads

Discordant reads:

- Translocation
- Tandem duplication
- Inversion
Discordant and misaligned reads

Discordant reads:

Misaligned reads:
Discordant and misaligned reads

Discordant reads:

- Translocation
- Tandem duplication
- Inversion

Misaligned reads:

- Unmapped mates
- Soft-clipped reads
BreaKmer – General outline
BreaKmer – General outline

SV Calling
BreaKmer – General outline

SV Calling
• For each region, extract misaligned reads. (Save discordant reads for later)
BreaKmer – General outline

SV Calling

• For each region, extract misaligned reads. (Save discordant reads for later)
• Assemble contigs using kmers
BreaKmer – General outline

SV Calling

• For each region, extract misaligned reads. (Save discordant reads for later)
• Assemble contigs using kmers
• Align contigs to reference using BLAT
BreaKmer – General outline

SV Calling

• For each region, extract misaligned reads. (Save discordant reads for later)
• Assemble contigs using kmers
• Align contigs to reference using BLAT
• Report SV and BP
Contig assembly
Contig assembly

• Extract all misaligned reads for a region
Contig assembly

• Extract all misaligned reads for a region
• Enumerate all possible k-mers from these samples
Contig assembly

• Extract all misaligned reads for a region
• Enumerate all possible k-mers from these samples
• Enumerate all k-mers from the target reference sequence and keep only those that are also found in the sample.
Contig assembly

• Extract all misaligned reads for a region
• Enumerate all possible k-mers from these samples
• Enumerate all k-mers from the target reference sequence and keep only those that are also found in the sample.
• Start from a seed k-mer:
Contig assembly

- Extract all misaligned reads for a region
- Enumerate all possible k-mers from these samples
- Enumerate all k-mers from the target reference sequence and keep only those that are also found in the sample.
- Start from a seed k-mer:
  - Retrieve all reads containing the k-mer
Contig assembly

• Extract all misaligned reads for a region

• Enumerate all possible k-mers from these samples

• Enumerate all k-mers from the target reference sequence and keep only those that are also found in the sample.

• Start from a seed k-mer:
  • Retrieve all reads containing the k-mer
  • Assemble the reads into a contig
Contig assembly

• Extract all misaligned reads for a region
• Enumerate all possible k-mers from these samples
• Enumerate all k-mers from the target reference sequence and keep only those that are also found in the sample.
• Start from a seed k-mer:
  • Retrieve all reads containing the k-mer
  • Assemble the reads into a contig
  • Cache reads without an overlapping 90% homologous sequence for potential assembly later
Contig assembly

• Extract all misaligned reads for a region
• Enumerate all possible k-mers from these samples
• Enumerate all k-mers from the target reference sequence and keep only those that are also found in the sample.
• Start from a seed k-mer:
  • Retrieve all reads containing the k-mer
  • Assemble the reads into a contig
  • Cache reads without an overlapping 90% homologous sequence for potential assembly later
  • Expand the contig by repeating with other k-mers within the retrieved reads
SV Calling
SV Calling

For each contig:
SV Calling

For each contig:

• Align to the reference target area using BLAT
SV Calling

For each contig:

• Align to the reference target area using BLAT
• Use the BLAT to determine if there was an Indel
SV Calling

For each contig:

- Align to the reference target area using BLAT
- Use the BLAT to determine if there was an Indel
- Filter results (min size, read depth, etc..)
SV Calling

For each contig:

- Align to the reference target area using BLAT
- Use the BLAT to determine if there was an Indel
- Filter results (min size, read depth, etc.)
- Align again to the whole reference genome.
SV Calling

For each contig:

- Align to the reference target area using BLAT
- Use the BLAT to determine if there was an Indel
- Filter results (min size, read depth, etc..)
- Align again to the whole reference genome.
- If it’s aligned – is it aligned to a different region?
SV Calling

For each contig:

• Align to the reference target area using BLAT
• Use the BLAT to determine if there was an Indel
• Filter results (min size, read depth, etc.)
• Align again to the whole reference genome.
• If it’s aligned – is it aligned to a different region?
• Apply rearrangement (local) or translocation filters.
Results
Results

• 38 cancer samples were selected
Results

• 38 cancer samples were selected
  • 12 were replicated to assess reproducibility
Results

• 38 cancer samples were selected
  • 12 were replicated to assess reproducibility
  • 4 were replicated and diluted (to 50% and 20%) to assess sensitivity
Results

• 38 cancer samples were selected
  • 12 were replicated to assess reproducibility
  • 4 were replicated and diluted (to 50% and 20%) to assess sensitivity

• 80 normal samples were selected to use as positive controls
Results

• 38 cancer samples were selected
  • 12 were replicated to assess reproducibility
  • 4 were replicated and diluted (to 50% and 20%) to assess sensitivity
• 80 normal samples were selected to use as positive controls
• 2 Target region lists were compiled
Results

• 38 cancer samples were selected
  • 12 were replicated to assess reproducibility
  • 4 were replicated and diluted (to 50% and 20%) to assess sensitivity
• 80 normal samples were selected to use as positive controls
• 2 Target region lists were compiled
• Novel CV’s were validated using PCR
Results

• 38 cancer samples were selected
  • 12 were replicated to assess reproducibility
  • 4 were replicated and diluted (to 50% and 20%) to assess sensitivity

• 80 normal samples were selected to use as positive controls

• 2 Target region lists were compiled

• Novel CV’s were validated using PCR

• Comparison to 4 other methods – CREST, Meerkat, BreakDancer, Pindel
Results
Results
Results

• 28/29 translocation positive samples were called.
Results

- 28/29 translocation positive samples were called.
- 75/77 in translocations in non-diluted replicates were called
Results

• 28/29 translocation positive samples were called.
• 75/77 in translocations in non-diluted replicates were called
• 98.3% true positive calls amongst replicates
Results

- 28/29 translocation positive samples were called.
- 75/77 in translocations in non-diluted replicates were called.
- 98.3% true positive calls amongst replicates.
- 9/10 translocations in the 20% diluted replicates were identified.
Results

- 28/29 translocation positive samples were called.
- 75/77 in translocations in non-diluted replicates were called.
- 98.3% true positive calls amongst replicates.
- 9/10 translocations in the 20% diluted replicates were identified.
- Overall 97.4% sensitivity in detecting the 38 known events.
Results
Results

Known translocations
Results

• 21 unknown SV’s detected.
Results

• 21 unknown SV’s detected.
  • 9/11 translocations were validated
Results

• 21 unknown SV’s detected.
  • 9/11 translocations were validated
  • 8/9 indels were validated (1 sample didn’t have sufficient DNA)
Results

• 21 unknown SV’s detected.
  • 9/11 translocations were validated
  • 8/9 indels were validated (1 sample didn’t have sufficient DNA)

• 77.3% predictive value
Results

- 21 unknown SV’s detected.
  - 9/11 translocations were validated
  - 8/9 indels were validated (1 sample didn’t have sufficient DNA)

- 77.3% predictive value

- 5 SV’s detected in the 80 non-cancer samples – 3 of them later validated.
## Comparison to other methods

### Table 2. Counts for the number of true-positive results for all the replicates, listed by the known alterations and four SV detection methods

| Known alteration | Total replicates | BreakeR | CREST | Meerkat | BreakDancer |
|------------------|------------------|---------|-------|---------|-------------|
|                  | ND | D50 | D20 | ND | D50 | D20 | ND | D50 | D20 | ND | D50 | D20 |
| ABL1-BCR         | 24 | 3  | 3  | 24 | 3  | 3  | 24 | 3  | 3  | 24 | 3  | 3  |
| ALK- EML4        | 15 | 3  | 3  | 13 | 3  | 2  | 13 | 2  | 2  | 13 | 3  | 1  |
| EGFR-intergenic  | 9  | 3  | 3  | 9  | 3  | 3  | 7  | 2  | 0  | 8  | 3  | 3  |
| BCL2-IGH         | 11 | 0  | 0  | 11 | 0  | 0  | 1  | 0  | 0  | 10 | 0  | 0  |
| PML-RARA         | 5  | 3  | 3  | 5  | 3  | 3  | 5  | 3  | 3  | 5  | 3  | 3  |
| FLT3-ITD         | 8  | 0  | 0  | 8  | 0  | 0  | 2  | 0  | 0  | 0  | 0  | 0  |
| EWSR1-FLI1       | 2  | 0  | 0  | 2  | 0  | 0  | 2  | 0  | 0  | 2  | 0  | 0  |
| KMT2A-MLLT3      | 2  | 0  | 0  | 2  | 0  | 0  | 2  | 0  | 0  | 1  | 0  | 0  |
| KMT2A-MLLT10     | 1  | 0  | 0  | 1  | 0  | 0  | 1  | 0  | 0  | 1  | 0  | 0  |
| KMT2A-MLLT4      | 1  | 0  | 0  | 1  | 0  | 0  | 1  | 0  | 0  | 1  | 0  | 0  |
| KMT2A-MLLT6      | 1  | 0  | 0  | 1  | 0  | 0  | 1  | 0  | 0  | 1  | 0  | 0  |
| ERG-EWSR1        | 1  | 0  | 0  | 1  | 0  | 0  | 1  | 0  | 0  | 1  | 0  | 0  |
| EWSR1-WTI        | 1  | 0  | 0  | 1  | 0  | 0  | 1  | 0  | 0  | 1  | 0  | 0  |
| ANKRD13B-FGFR1   | 1  | 0  | 0  | 1  | 0  | 0  | 1  | 0  | 0  | 1  | 0  | 0  |
| FIPI1-PDGFR1     | 1  | 0  | 0  | 1  | 0  | 0  | 1  | 0  | 0  | 1  | 0  | 0  |
| ERG-FUS          | 1  | 0  | 0  | 1  | 0  | 0  | 1  | 0  | 0  | 1  | 0  | 0  |
| IGH-MYC          | 1  | 0  | 0  | 1  | 0  | 0  | 1  | 0  | 0  | 1  | 0  | 0  |
| KIT deletion     | 1  | 0  | 0  | 1  | 0  | 0  | 1  | 0  | 0  | 0  | 0  | 0  |

**Total replicates:** 86 12 12 84 12 11 66 10 8 70 12 10
**Total samples:** 38 4 4 37 4 4 30 4 3 27 4 4

ND: non-dilution replicates; D50: dilution replicates with 50% tumor purity; D20: dilution replicates with 20% tumor purity.
Comparison to other methods
Comparison to other methods

• Other methods identified a strikingly large number of previously unidentified SV’s compared to BreaKmer.
Comparison to other methods

• Other methods identified a strikingly large number of previously unidentified SV’s compared to BreaKmer.

• Very little overlap between the methods.

| Method    | Total Calls | BreakDancer | Meerkat | CREST |
|-----------|-------------|-------------|---------|-------|
| BreaKmer  | 494         | 17          | 9       | 11    |
| CREST     | 26246       | 451         | 2237    |       |
| Meerkat   | 15991       | 504         |         |       |
| BreakDancer | 15712      |             |         |       |
Comparison to other methods

• Other methods identified a strikingly large number of previously unidentified SV’s compared to BreaKmer.

• Very little overlap between the methods.

| Method    | Total Calls | BreakDancer | Meerkat | CREST |
|-----------|-------------|-------------|---------|-------|
| BreaKmer  | 494         | 17          | 9       | 11    |
| CREST     | 26246       | 451         |         | 2237  |
| Meerkat   | 15991       | 504         |         |       |
| BreakDancer | 15712     |             |         |       |

• 90% of additional calls were not identified in more than a single replicate.
Comparison to other methods

- Other methods identified a strikingly large number of previously unidentified SV’s compared to BreaKmer.

- Very little overlap between the methods.

- 90% of additional calls were not identified in more than a single replicate.

| Method    | Total Calls | BreakDancer | Meerkat | CREST |
|-----------|-------------|-------------|---------|-------|
| BreaKmer  | 494         | 17          | 9       | 11    |
| CREST     | 26246       | 451         | 2237    |       |
| Meerkat   | 15991       | 504         |         |       |
| BreakDancer | 15712     |             |         |       |
Results
Results
Results

• Dilution expectedly affects the SV evidence.
Results

- Dilution expectedly affects the SV evidence.
- Read support lowers as the tumor content in the sample grows smaller.
Results

• Dilution expectedly affects the SV evidence.
• Read support lowers as the tumor content in the sample grows smaller.
• (EGFR went through a big somatic amplification which also affected the read depth).
conclusion
conclusion

• Targeting specific areas
conclusion

- Targeting specific areas
- Using all read mapping data (discordant, unmatched and soft-clipped)
conclusion

- Targeting specific areas
- Using all read mapping data (discordant, unmatched and soft-clipped)
- Using k-mers for assembly
conclusion

• Targeting specific areas
• Using all read mapping data (discordant, unmatched and soft-clipped)
• Using k-mers for assembly
• Very high sensitivity, reproducibility and predictive results.
conclusion

• Targeting specific areas
• Using all read mapping data (discordant, unmatched and soft-clipped)
• Using k-mers for assembly
• Very high sensitivity, reproducibility and predictive results.
• Maybe too good?
conclusion

• Targeting specific areas
• Using all read mapping data (discordant, unmatched and soft-clipped)
• Using k-mers for assembly
• Very high sensitivity, reproducibility and predictive results.
• Maybe too good?
• Designed with detecting known SV’s quickly and cheaply as the primary goal.