TIDE: Inter-Chromosomal Translocation and Insertion Detection using Embeddings

Rosanne Vetro, Roshanak Farhoodi, Rohith Kotla, Nurit Haspel, David Weisman, Jennifer Rosen[1], Dan Simovici

University of Massachusetts Boston, Department of Computer Science

[1]MedSTAR Washington Hospital Center

- Identification of genomic mutations have implications for disease nature understanding, diagnostic and therapeutic approaches
- Among other types of mutations, large scale Structural Variations (SVs) are often described as one of the primary causes of cancer

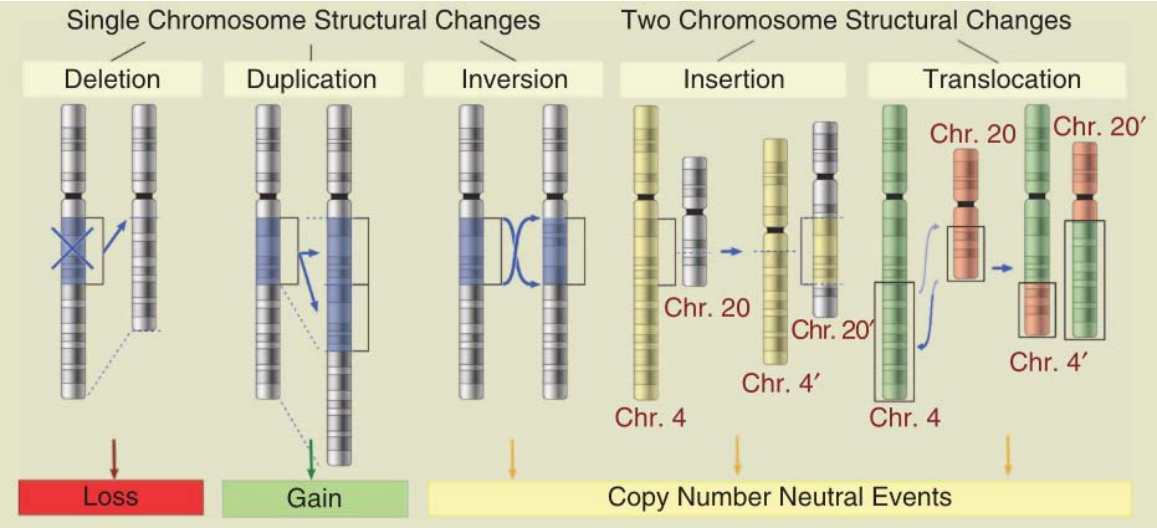
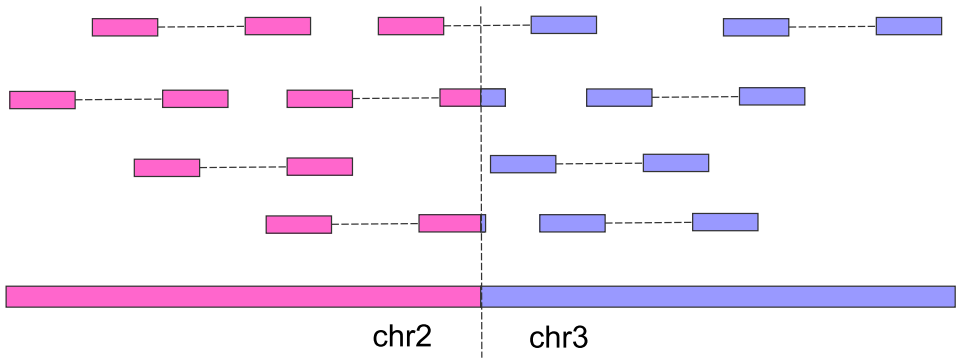


Figure: Abdullah K. Alqallaf, Fuad M. Alkoot and Mash'el S. Aldabbous (2013). Discovering the Genetics of Autism, Recent Advances in Autism Spectrum Disorders - Volume I, Prof. Michael Fitzgerald (Ed.), ISBN: 978-953-51-1021-7, InTech, DOI: 10.5772/53797.

- Identify reads from whole-genome NGS paired-end data that contain breakpoints generated by inter-chromosomal translocations and insertions

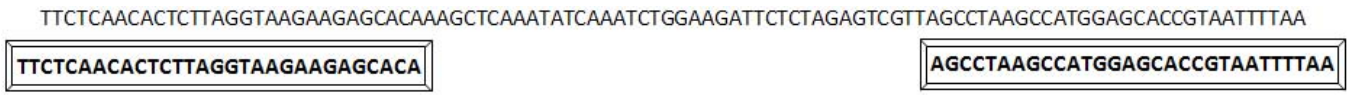


- Due to the very large volume of data, we use an approximate method to detect breakpoints

- Predicting the precise location of breakpoints is hard
- Exact similarity search and mapping methods are inefficient having very large volume of query data and such a vast search space
- Breakpoint detection using Next Generation Sequencing (NGS) data remains an open problem, as recent studies show 30% overlap in predictions across multiple calling methods (pilot work by International Cancer Genome Consortium, The Cancer Genome Atlas and Sage Bionetworks)

- Input: Next-generation sequencing paired-end reads
- Read Alignment: Reads are Locally aligned to reference genome
- Identify Discordant paired-end reads: Find reads where both mates have unique alignments, but the mates aren't in the expected relative orientation, or aren't within the expected distance range, or both
- Candidate reads are selected according to the information in the BAM file such as mapping quality, flag field and surrounding discordant pairs

- Fingerprint is a list of integers; elements of this list are the indices of the array containing all the k-mers over the symbols A,C,T,G and N
- Two windows with n nucleotides each are extracted from the beginning and end of each read. Fingerprints representing information about the k-mers of the selected windows are generated and stored



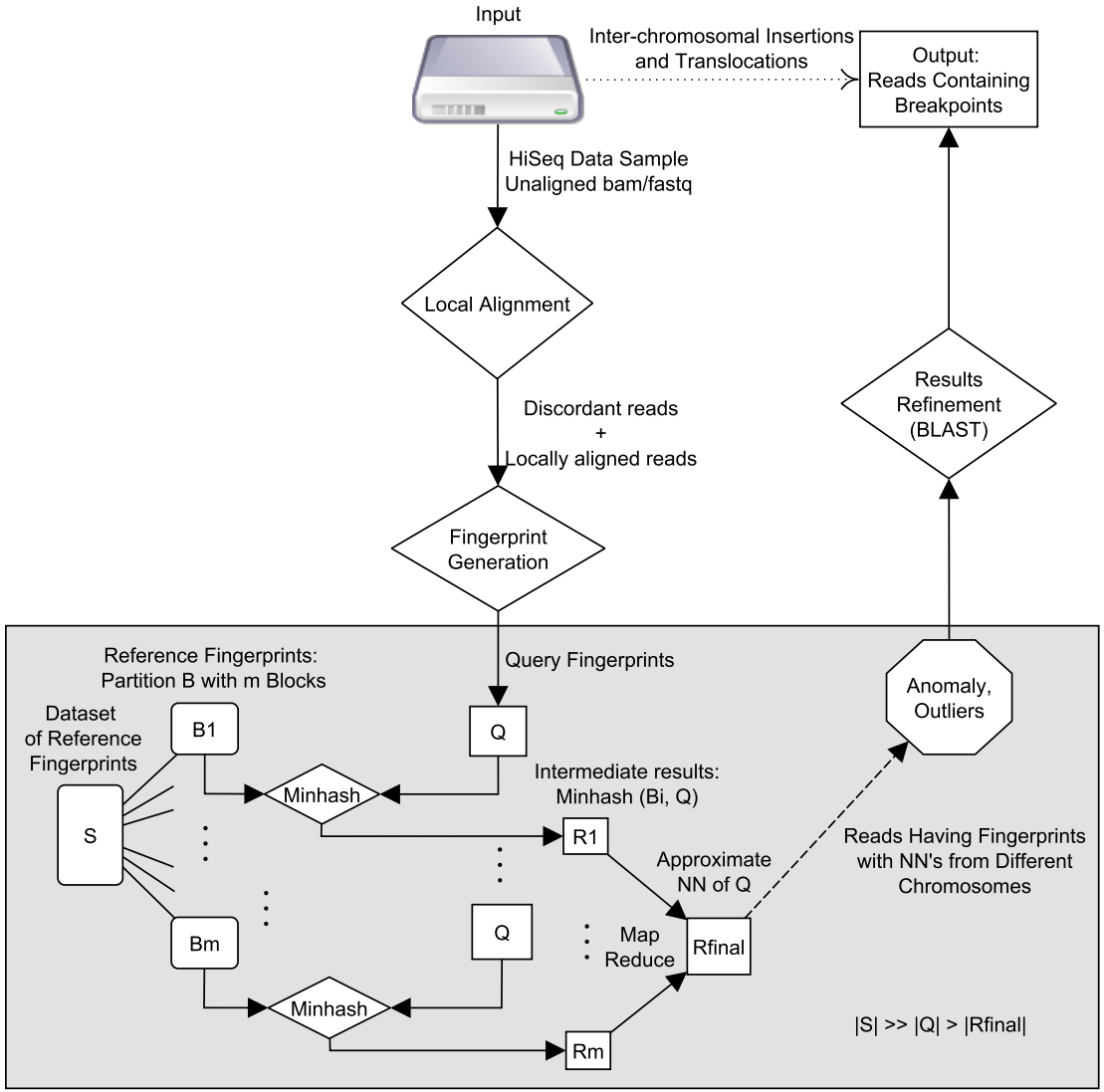
- The reference genome is also converted into fingerprints using 1-nucleotide sliding windows

- Let (X, D) be a metric space and let P be a subset of X.
- Given a query point $q \in X$ and a radius r,
  find $\{p \in P | D(p, q) \leq r\}$
- We need a data structure to return p if it exists.

- LSH is used to find the approximate nearest neighbor of each sample fingerprint among those fingerprints obtained from the reference genome
- Hash the input items so that similar items are mapped to the same buckets with high probability. The goal is to maximize probability of collision of similar items rather than avoid collisions
- A LSH family for a variant of the Jaccard index is used
- Near neighbors are fingerprints from different classes (sample or reference genome) having high similarity

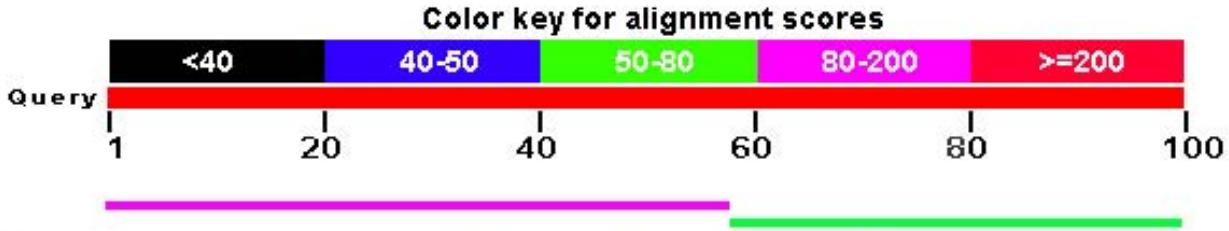
- TIDE uses the Apache Hadoop software library and the Hadoop MapReduce module to combine, sort and process the collection of results generated by several instances of the approximate nearest neighbor search
- All near(est) neighbors (obtained from different parts of the reference genome) of a given query fingerprint are gathered and the most similar one is selected
- Reads represented by nearest neighbors that do not belong to the same reference chromosome in the genome are considered strong candidates for spanning a breakpoint

- The result of the previous stages may still contain a number of false positive matches due to sequencing errors and the approximate nature of the nearest-neighbor search
- In the refinement stage, the resulting reads are mapped to the human genome using the BLAST sequence alignment program
- All matches that align partially to two different chromosome are reported


Input
Inter-chromosomal Insertions and Translocations
Output: Reads Containing Breakpoints
HiSeq Data Sample Unaligned bam/fastq
Local Alignment
Discordant reads + Locally aligned reads
Fingerprint Generation
Results Refinement (BLAST)
Reference Fingerprints: Partition B with m Blocks
Query Fingerprints
Anomaly, Outliers
Dataset of Reference Fingerprints
B1
Q
Minhash
Intermediate results: Minhash (Bi, Q)
S
R1
Approximate NN of Q
Reads Having Fingerprints with NN's from Different Chromosomes
Bm
Q
Map Reduce
Rfinal
Minhash
Rm
|S| >> |Q| > |Rfinal|

- Raw data: NGS paired-end reads was produced by Illumina High-Seq plataform
- Data description: Three DNA samples taken from tissues of patients having thyroid cancer and known to have PAX8-PPARγ rearrangements, datasets were composed of approximately 195 million paired-end reads of length 100bp and insert size 500bp
- BWA is used to align the reads to the UCSC GRCh37/hg19 human reference
- Input data: Set of aligned reads in a compressed sequence map format (BAM)
- The ground truth is composed of three reads containing breakpoints, one for each sample

- TIDE accurately detected the expected reads for two samples
- TIDE was not able to detect the expected read for the third sample due to the high number of mutated bases around its start and end positions. Nevertheless, TIDE was able to detect additional reads containing PAX8-PPARγ rearrangement for the third sample
- The clinical samples were also tested using two recently published methods to detect SV's: Delly and Bellerophon
- Delly detected thousands of translocations between chromosomes 2 and 3 but did not identify the precise location of the target breakpoints
- Bellerophon found a set of breakpoints between other chromosomes, but no rearrangements between chromosomes 2 and 3



Homo sapiens chromosome 3, GRCh37.p13 Primary Assembly

Sequence ID: ref|NC_000003.11| Length: 198022430 Number of Matches: 1

Range 1: 12412324 to 12412381 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
108 bits(58)	6e-22	58/58(100%)	0/58(0%)	Plus/Plus

Features: peroxisome proliferator-activated receptor gamma isoform 2

```
Query  1         AAATGCCAGATAAGACATAGACTTTCCTGGCTGCTAAAGAGCAATTAAGCATTGGAAT  58
                 ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  12412324  AAATGCCAGATAAGACATAGACTTTCCTGGCTGCTAAAGAGCAATTAAGCATTGGAAT  12412381
```

Homo sapiens chromosome 2, GRCh37.p13 Primary Assembly

Sequence ID: ref|NC_000002.11| Length: 243199373 Number of Matches: 1

Range 1: 113986454 to 113986495 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
78.7 bits(42)	5e-13	42/42(100%)	0/42(0%)	Plus/Minus

Features: paired box protein Pax-8 isoform PAX8A
paired box protein Pax-8 isoform PAX8E

```
Query  59         AGGAGGAGGTAGTCACTGCAGTATTGGCAACTTACAGGCTTT  100
                  ||||||||||||||||||||||||||||||||||||||||||
Sbjct  113986495  AGGAGGAGGTAGTCACTGCAGTATTGGCAACTTACAGGCTTT  113986454
```

Average running time of TIDE on the clinical datasets

Stage	Time
Local alignment	$\sim$ 10 h
LSH (SketchSort with default parameters)	$\sim$ 1 h
LSH (SketchSort with modified parameters for the max Jaccard distance)	$\sim$ 5 h
Map-Reduce	$\sim$ 20 min
Refinement	$\sim$ 4 h

Environment: Ubuntu 12.4, 64Gb of memory, 8 cores

- TIDE provides an efficient model (distance preserving projections of fingerprints combined with a parallel framework and result refinement) to perform similarity search and anomaly detection in sample genomes
- Probabilities can be adjusted in order to tolerate matches with higher rates of mutation
- TIDE successfully located expected breakpoints in the data sets provided

- Test TIDE with a larger number of tumor/normal whole-genome sequencing data
- Compare TIDE with other state of the art method
- Improve the filtering of candidate reads by utilizing other statistics such as the genome mappability score (GMS) to avoid ambiguously mapped reads
- Detect and classify all types of SVs including duplications and deletions

Thank you!

This research is funded in part by a seed grant from the Massachusetts Green High Performance Computing Center (MGHPCC).