BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 11 2014, pages 1627–1629
doi:10.1093/bioinformatics/btu078

Sequence analysis

Advance Access publication February 12, 2014

BIMA V3: an aligner customized for mate pair library sequencing

Travis M. Drucker[1], Sarah H. Johnson[2], Stephen J. Murphy[2], Kendall W. Cradic[3], Terry M. Therneau[4] and George Vasmatzis[2,*]

[1]Department of Information Technology, MN 55905, [2]Department of Molecular Medicine, [3]Department of Laboratory Medicine and Pathology and [4]Department of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN 55905, USA

Associate Editor: Micheal Brudno

ABSTRACT

Summary: Mate pair library sequencing is an effective and economical method for detecting genomic structural variants and chromosomal abnormalities. Unfortunately, the mapping and alignment of mate-pair read pairs to a reference genome is a challenging and time-consuming process for most next-generation sequencing alignment programs. Large insert sizes, introduction of library preparation protocol artifacts (biotin junction reads, paired-end read contamination, chimeras, etc.) and presence of structural variant breakpoints within reads increase mapping and alignment complexity. We describe an algorithm that is up to 20 times faster and 25% more accurate than popular next-generation sequencing alignment programs when processing mate pair sequencing.

Availability: http://bioinformaticstools.mayo.edu/research/bima/

Contact: vasm@mayo.edu or vasmatzis.george@mayo.edu

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on August 19, 2013; revised on January 8, 2014; accepted on February 3, 2014

1 INTRODUCTION

Large genomic chromosomal rearrangements play an important role in the development and progression of many cancers. Amplifications, inversions, deletions and translocations can modify the expression and/or functionality of many tumor suppressors and oncogenes. Thus, identification of large genomic rearrangements in cancer cells can greatly aid in our understanding of the mechanisms underlying cancer progression.

Mate pair (MP) library sequencing offers an efficient, a cost-effective and a comprehensive method to detect the presence of and to determine the fine grain structure of structural variants throughout the entire genome. MPs' large 2–5-kb fragment size increases the likelihood that a read pair will span a structural variant and minimizes the number of read pairs required to detect and characterize an event. Multiple samples can be processed on a single lane of a flow cell, reducing costs without degrading accuracy. Coverage analysis can accurately determine aneuploidy, amplification and deletion on multiplexed samples. For a detailed description of the MP sequencing protocol, please refer to http://www.illumina.com/technology/mate_pair_sequencing_assay.ilmn.

*To whom correspondence should be addressed.

The MP library preparation protocol introduces several sequencing artifacts that complicate MP read pair mapping and alignment. After shearing, biotin end-labeling and size selection, 2–5-kb DNA fragments are circularized by joining two ends of a single fragment. In the circularization step, chimeras, or false MPs, can be generated when two separate fragments ligate. Before sequencing, the circularized DNA is randomly sheared into 500-bp fragments, and the biotin-containing fragments are isolated. Because shearing is random, the biotin junction can be within the sequencers' read length of either end of the fragment, producing reads containing sequence from two different genomic locations. Finally, the biotin isolation step is not 100% efficient, 10–15% of the sequenced read pairs consist of paired-end reads.

The use of large insert sizes and the presence of paired-end reads, biotin junctions, chimeras and breakpoint junctions in MP sequencing make alignment more difficult and time-consuming. Large insert sizes and chimeric reads increase the expected paired read mapping search space. Sequencing through biotin junctions (20–25% of 100-bp reads) and structural variant breakpoint junctions produces reads containing sequence from multiple genomic locations. The presence of pair-end reads, chimeras and discordant read pairs spanning structural variants increases the complexity of correctly scoring paired genomic locations.

To address the difficulties of MP sequencing alignment, we have created Binary Indexing Mapping Algorithm (BIMA) V3, the latest version of a mapping and alignment tool designed to handle MP sequencing artifacts. Building on BIMA's proven track record (Feldman *et al.*, 2011; Kovtun *et al.*, 2013; Vasmatzis *et al.*, 2007, 2012), V3's enhancements allow it to accurately align up to 25% more reads in up to 1/20 the time of currently popular next-generation sequencing (NGS) alignment programs.

2 METHODS

BIMA V3 is a hash-based algorithm (Li and Homer, 2010) that indexes the entire reference genome with three unique hash tables. Each hash table uses a slightly different hashing algorithm that converts individual bases to a single binary bit. Each encoding (A/G = 1 & T/C = 0, C/G = 1 & A/T = 0 and A/C = 1 & T/G = 0) generates keys that are unaffected by a specific pair of transition (G ⇔ A, C ⇔ T) or transversion (C ⇔ G, A ⇔ T, A ⇔ C, G ⇔ T) mutations or sequencing errors. Using three separate encodings ensures that at least one key correctly maps to the reference genome when a single (potentially multiple) mismatching base is present.

Keys are 32 bits in length (32 encoded consecutive base pairs) and index into an array of 2^{32} (4 294 967 296) slots, requiring ~30 GB of RAM per hash table. During hash table creation, keys are generated every base pair (31-bp overlap between consecutive keys), with the associated slot being populated with the reference genome position. A single slot may contain up to 2^{15} (32 768) reference genome positions. During read mapping, keys are typically generated every 8 bp (24-bp overlap between consecutive keys) over the entire read in both orientations (forward and reverse compliment). A 100-bp read will have 60 keys generated (10 keys per orientation × two orientations × three encodings) during mapping. Overlapping keys generated from the entire read ensure that a subset of keys do not bridge biotin or breakpoint junctions.

A list of candidate positions for each read is generated from the reference positions associated to the keys created from the read. To reduce runtime, keys associated to >750 reference genome positions are excluded from processing, unless too few keys generate candidate positions (reference 0 or >750 positions). If two candidate positions for a single read are within 50 bp (e.g. keys from separate parts of the read associate to reference genome positions close in locality), the read is evaluated for an InDel at the leftmost candidate position.

All candidate positions are scored using fast binary operations (XOR and POPCOUNT) to count the base pairs from the read that do not match the reference genome. An optimal concordant alignment is generated by evaluating the combined score of every pair of candidate positions from each read that is within the expected insert size. An optimal discordant alignment is generated by selecting the overall best combined score paired alignment. If a junction is detected in either the concordant and/or discordant alignments, their score/s are adjusted. A 5% scoring penalty is applied to the discordant alignment (favoring concordant alignments), and the highest scoring paired alignment is reported.

BIMA V3 is implemented in C, leverages OpenMP for index and alignment parallelization, supports 64-bit Linux operating systems and uses standard file formats (fastq and SAM). BIMA supports read pairs with individual read lengths between 64 and 512 bp and efficiently scales up to 16 concurrent threads. BIMA does not currently support single-end reads.

3 RESULTS

The accuracy and execution time of BIMA V3 were compared with two popular NGS alignment programs, Burrows-Wheeler Aligner (BWA; Li and Durbin, 2009) and Novoalign (http://www.novocraft.com). Two data sets, one synthetic and one real, were used to measure alignment accuracy and single threaded execution time. Each data set consisted of 1 000 000 MP read pairs (2 000 000 total reads), each 100 bp in length. The whole-genome simulator for Next-Generation Sequencing (dwgsim) (https://github.com/nh13/DWGSIM/wiki) were used to generate and evaluate the synthetic data set (Table 1). The real data set was sequencing from the lymph node prostate cancer cell-line (LNCAP) (prostate adenocarcinoma) and was evaluated by comparing the reference genome position predicted by the NGS alignment programs with reference genome positions predicted by BLAST-like alignment tool (BLAT; Kent, 2002; Table 2).

The synthetic data set (Table 1) portrays all evaluated aligners being roughly equivalent in accuracy (BIMA having a slight advantage) but vary significantly in alignment time. Unfortunately, synthetic MP sequencing does not accurately model all artifacts present in real MP sequencing (biotin/breakpoint junctions, paired-end read contamination, chimeras, etc). Both Novoalign and BWA require a considerable amount of additional processing time (~50% to ~250%) to deliver significantly fewer accurately mapped reads (~10% to ~25%) on real MP sequencing (Table 2).

Table 1. Evaluation of synthetic mate-pair read pairs (dwgsim)

	BIMA	BWA	Novoalign
Time to align(s)	489	870	7378
Correctly mapped (%)	92.27	90.82	91.04
Incorrectly mapped (%)	2.27	3.51	3.50
Unmapped that should be mapped (%)	0.00	0.21	0.00
Unmapped that should not be mapped (%)	5.46	5.46	5.46

One million synthetic mate-pair read pairs were generated with dwgsim from the hg19 human reference genome (100-bp reads, 4 000-bp insert, 0.5–1.5% error rate, 0.1% mutation rate, etc). Alignment time, in seconds, was captured on a single core of an Intel Xeon 2.9-GHz E5-2690. See Supplemental materials for an expanded comparison (including Burrows-Wheeler Aligner Maximal Exact Matches; BWA-MEM), algorithm invocation details and the full dwgsim_eval reports.

Table 2. Evaluation of real mate-pair read pairs (LNCAP)

	BIMA	BWA	Novoalign
Time to align(s)	606	3260	11 091
Within 50 bp of a BLAT result (%)	85.17	60.67	74.35
Not mapped by algorithm (%)	7.29	34.74	21.50
Not mapped by BLAT (%)	1.03	0.49	0.69
Exactly matches a BLAT result (%)	68.76	56.96	68.44
1–50 bp distant from a BLAT result (%)	16.42	3.71	5.91
More than 50 bp distant from a BLAT result (%)	6.51	4.11	3.46

The LNCAP cell line was sequenced using an optimized Illumina mate pair protocol (Murphy *et al.*, 2012), 100-bp reads, ~2400-bp insert, etc. One million read pairs were aligned on a single core of an Intel Xeon 2.9-GHz E5-2690. Reported reference genome positions were compared with individual read alignments predicted by BLAT. See Supplemental materials for an expanded comparison (including BWA-MEM), algorithm invocation details and a detailed explanation of the analysis performed.

BIMA V3's utilization of three hash tables with 32-bit keys requires a significant amount of process memory, 128 GB during hash table creation and 100 GB during alignment, compared with Novoalign and BWA (8 GB and 5 GB, respectively). It is faster for BIMA to generate indexes during every invocation (~800 seconds using three threads), than read in a compressed pre-computed index from a shared file system. Because Novoalign and BWA pre-compute their indexes, the alignment times reported in Tables 1 and 2 do not include index generation.

BIMA V3 accurately aligns more MP reads in less time than popular NGS alignment programs, enabling researchers to discover more structural variants with less coverage. Researchers are already using it to develop new genetic testing methods (K.Cradic *et al.*, manuscript submitted) and reprocess 100s of existing MP data sets to identify previously undetected structural variants.

Funding: Center for Individualized Medicine, Mayo Clinic

Conflict of Interest: none declared.

REFERENCES

Feldman,A.L. *et al*. (2011) Discovery of recurrent t(6;7)(p25.3;q32.3) translocations in ALK-negative anaplastic large cell lymphomas by massively parallel genomic sequencing. *Blood*, **117**, 915–919.

Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res*., **12**, 656–664.

Kovtun,I.V. *et al*. (2013) Lineage relationship of Gleason patterns in Gleason score 7 prostate cancer. *Cancer Res*., **73**, 3275–3284.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li,H. and Homer,N. (2010) A survey of sequence alignment algorithms for next-generation sequencing. *Brief. Bioinform*., **11**, 473–483.

Murphy,S.J. *et al*. (2012) Mate pair sequencing of whole-genome-amplified DNA following laser capture microdissection of prostate cancer. *DNA Res*., **19**, 395–406.

Vasmatzis,G. *et al*. (2007) Quantitating tissue specificity of human genes to facilitate biomarker discovery. *Bioinformatics*, **23**, 1348–1355.

Vasmatzis,G. *et al*. (2012) Genome-wide analysis reveals recurrent structural abnormalities of TP63 and other p53-related genes in peripheral T-cell lymphomas. *Blood*, **120**, 2280–2289.