BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 12 2009, pages 1554–1555
doi:10.1093/bioinformatics/btp255

Sequence analysis

MapView: visualization of short reads alignment on a desktop computer

Hua Bao[1,†,*], Hui Guo[1,†], Jinwei Wang[2], Renchao Zhou[1], Xuemei Lu[1] and Suhua Shi[1,*]

[1]State Key Laboratory of Biocontrol and Key Laboratory of Gene Engineering of Ministry of Education, School of Life Sciences, Sun Yat-Sen University, Guangzhou, 510275 and [2]School of Manufacturing Science and Engineering, Sichuan University, Chengdu 610064, China

Received on March 10, 2009; revised and accepted on April 8, 2009

Advance Access publication April 15, 2009

Associate Editor: Joaquin Dopazo

ABSTRACT

Summary: We introduce a new visual analytics tool named MapView to facilitate the representation of large-scale short reads alignment data and genetic variation analysis. MapView can handle hundreds of millions of short reads on a desktop computer with limited memory. It supports a compact alignment view for both single-end and paired end short reads, multiple navigation and zoom modes and multi-thread processing. Moreover, MapView offers automated genetic variation detection. MapView has been used in our lab and by over 10 research labs worldwide.

Availability: http://evolution.sysu.edu.cn/mapview/.

Contact: baohua100@hotmail.com; lssssh@mail.sysu.edu.cn

Supplementary information: Supplementary data are available at http://evolution.sysu.edu.cn/mapview/MVF.pdf

1 INTRODUCTION

Next-generation sequencing technologies generate huge amounts of DNA sequence reads at a fraction of the cost of Sanger sequencing. The advent of these technologies opens opportunities to a variety of biological applications including genome re-sequencing, whole-transcriptome sequencing, ChIP-seq and miRNA discovery (Shendure and Ji, 2008). While the promise of next-generation sequencing technologies has become a reality, they also present substantial informatics challenges. One of the main challenges is data visualization on desktop computers. Existing viewers such as Consed (Gordon *et al.*, 1998) and Hawkeye (Schatz *et al.*, 2007) were designed for genome assemblies of Sanger capillary sequence reads and do not yet have effective support for next-generation sequence reads. Moreover, loading large alignment data into Consed and Hawkeye requires huge amount of memory not typically available to desktop computer users. EagleView (Huang and Marth, 2008) is the only visualization tool specifically designed for next-generation sequencing technologies. But it has no support for variation detection and visualization of paired end reads. Moreover, EagleView also has memory limitations. Existing viewers only read assembly file in the traditional ACE format. But short sequence alignments can be generated by a number of software programs such as MAQ (Li,H. *et al.*, 2008), SOAP (Li,R. *et al.*, 2008) and SeqMap (Jiang and Wong, 2008). It will be more convenient for users to take these programs' alignment result file as input file.

Here, we present the first freely available software called MapView for visualization of large-scale alignment data, data validation and genetic variation analysis on desktop computers. MapView supports a compact alignment view for both single-end and paired end short reads, multiple navigation and zoom modes, and different formats of input file. It can handle large-scale data with super high computational efficiency. Moreover, MapView offers automated variant detection.

2 METHODS

2.1 MapView format

MapView format (MVF) is a novel file format designed for fast and memory efficiency visualization of huge amount of short reads alignment data. The MVF (Supplementary Material) consists of four sections: file header, data, index and statistics information. The MVF binary file, combined with effective compressing and indexing of the alignments, will enable reduction in disk usage and fast retrieval of alignments in a specified region.

2.2 Loading and navigation algorithm

MapView is a disk-based viewer and it only loads a tiny portion of the MVF file into memory. Specifically, MapView loads current displayed page and six neighbor pages. This fractional loading and neighbor caching algorithm leads to a small burden on memory resources while not compromising the speed. The MVF file offset of alignment data is indexed by reference position. To quickly find alignments associated with a specified region, MapView uses the index to locate the offset address of short reads mapped on the specified region and then retrieve the alignments data. This indexed navigation algorithm enables the users to quickly jump to different regions.

3 RESULTS

3.1 Computational efficiency

A typical alignment dataset of next-generation sequencing technologies may contains hundreds of millions of reads, reaching alignment file size of many gigabytes. The great majority of alignment viewers were designed for loading and processing big assembly file in the ACE format. This memory-based design requires huge amount of memory not typically available to

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First authors.

Table 1. Computational efficiency comparison

Tool	Version	Memory usage (GB)	CPU time (s)
Consed	18.0	12.06	208
Hawkeye	2.0.8	14.14	297
EagleView	2.2	3.91	207
MapView	3.1	0.04	2

The alignment data for the assessment are of reference length 43 596 771 bases and 6 615 627 Illumina 44-bp reads. We converted the reference and alignment file (MAQ output) to ACE format for Consed and EagleView, bank format for Hawkeye and MVF format for MapView as input file. The memory usage and CPU time for each tool were measured after it was loaded and displayed alignment view.

desktop computer users. We compared MapView's computational requirements to Consed, Hawkeye and EagleView on an alignment dataset that contains a reference chromosome of length 43 596 771 bases and 6 615 627 Illumina 44-bp reads. We found that MapView required <0.5% of the memory used by Consed and Hawkeye, and required 1.5% of the memory used by EagleView (Table 1). The CPU time used by MapView during loading the alignment was significantly lower than the three other programs. We also tested the four tools on a larger alignment dataset of *Oryza sativa indica*'s chromosome I consisting of 26 531 161 Illumina 44-bp reads. Consed, Hawkeye and EagleView were unable to load these data on our server with 16 GB memory. MapView opened these data with 0.04 GB memory in 2 s.

3.2 MapView features

MapView is an information-rich viewer with a single-window GUI. Its feature set is specifically designed for visualization of large alignment file of next-generation sequence reads (Table 2). MapView converts the FASTA reference sequence file and alignment result file which output by MAQ, SOAP or SeqMap program to a single binary file in the MVF. MapView also allows users to input new alignment result file format. In order to utilize screen space effectively, reads are optimally placed in multiple lines. MapView can display the orientation of the read (forward/reverse), quality score of a given nucleotide on the read and coverage information of a given position on the reference. It allows navigation by reference and variant location. It also supports zooming, customizable brightness and highlight of particular item. Moreover, MapView can display coverage and quality score distribution. Instead of manually checking all the conflicts of a contig to discover significant variants, MapView offers automated single-nucleotide variation detection (Altschuler *et al.*, 2000; Brochman *et al.*, 2008). Based on your specifications on what you consider a valid variant, the variant detection will scan through the entire alignment and report all the variants that meet the requirements. MapView also provides preliminary capability of structure variation detection. MapView is also well established for its multitasking and multithreading. It can process multiple tasks (i.e. visualization of alignment, variant detection and coverage computation) in parallel. MapView is written in object-oriented style in the C# computer language with Framwork.NET 2.0. At present, it runs under Windows with Framework.NET 2.0 and has been tested on the Windows XP and Vista operating systems.

Table 2. MapView feature list

Feature categories	Features
View	Compact view of alignment with zooming capability
	Pinpoint view of base quality
	Pinpoint view of coverage
	Highlight and list view of variants
	List view of quality score and coverage distribution
Navigation	Navigation by reference position
	Navigation by variant position
	Navigation by page
Efficiency	Handling hundreds of millions of reads on a desktop computer
	Multi-thread processing
Alignment format	MAQ, SOAP, SeqMap and user-defined format
Genetic variation detection	Single-nucleotide variation detection
	Preliminary capability of structure variation detection

ACKNOWLEDGEMENTS

We thank Ziwen He for providing Illumina sequence data for our software testing. We also thank all MapView beta testers for their helpful feedback.

Funding: National Basic Research Program of China (2007CB815701); National Natural Science Foundation of China (30730008, 30800060, 30600064 and 40876075).

Conflict of Interest: none declared.

REFERENCES

Altshuler,D. *et al.* (2000) An SNP map of the huaman genome generated by reduced representation shotgun sequencing. *Nature*, **407,** 513–516.

Brochman,W. *et al.* (2008) Quality scores and SNP detection in sequencing-by-synthesis system. *Genome Res.*, **18,** 763–770.

Gordon,D. *et al.* (1998) Consed: a graphical tool for sequence finishing. *Genome Res.*, **8,** 195–202.

Huang,W. and Marth,G. (2008) EagleView: a genome assembly viewer for next-generation sequencing technologies. *Genome Res.*, **18,** 1538–1543.

Jiang,H. and Wong,W.H. (2008) SeqMap: mapping massive amount of oligonucleotides to the genome. *Bioinformatics*, **24,** 2395–2396.

Li,H. *et al.* (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, **18,** 1851–1858.

Li,R. *et al.* (2008) SOAP: short oligonucleotide alignment program. *Bioinformatics*, **24,** 713–714.

Schatz,M.C. *et al.* (2007) Hawkeye: an interactive visual analytics tool for genome assemblies. *Genome Biol.*, **8,** R34.

Shendure,J. and Ji,H. (2008) Next-generation DNA sequencing. *Nat. Biotechnol.*, **26**, 1135–1145.