A Re-sequencing Tool for Next Generation Sequencing based on Burrows-wheeler Transform

Chen Hua Lu[1], Chun Yuan Lin[2,*], and Chuan Yi Tang[3]

[1]Department of Computer Science,
National Tsing Hua University,
Hsinchu 300, Taiwan, ROC
walterlu21@yahoo.com.tw

[2]Department of Computer Science and Information Engineering,
Chang Gung University,
Taoyuan 330, Taiwan, ROC
cyulin@mail.cgu.edu.tw

[3]Department of Computer Science and Information Engineering,
Providence University,
Taichung 433, Taiwan, ROC
cytang@pu.edu.tw

Received 5 October 2010; Revised 15 November 2010; Accepted 10 December 2010

Abstract. After the reference genomes of many organisms are sequenced in this post-genetic era, it has become an extremely important issue that how to do the re-sequencing and assembly for individual genomes from very large amount of reads. In this paper, we will present a re-sequencing tool designed for the Next Generation Sequencing (NGS) data. And these data are composed of a huge amount of short reads which will be aligned onto a reference genome. We modified and implemented the algorithm of Burrows-Wheeler Transform and FM-index to build the genome index of human, and proposed an idea to segment each short read into multiple non-overlapping seeds, which let us align short reads with large Hamming distance. Finally, we used 4 real data sets with different lengths form 1000 Genome Project to demonstrate the performance of our tool with a personal computer, and compared the results with a widely used tool, bowtie.

Keywords: NGS, BWT, FM-index, re-sequencing, short read, DNA

References

[1] P. Flicek and E. Birney, "Sense from Sequence Reads: Methods for Alignment and Assembly," *Nature Methods*, Vol. 6, pp. 6-12, 2009.

[2] H. Li, J. Ruan, R. Durbin, "Mapping Short DNA Sequencing Reads and Calling Variants Using Mapping Quality Scores," *Genome Research*, Vol. 18, No. 11, pp. 1851-1858, 2008.

[3] R. Li, Y. Li, K. Kristiansen, J. Wang, "SOAP: Short Oligonucleotide Alignment Program," *Bioinformatics*, Vol. 24, No. 5, pp. 713-714, 2008.

[4] D. Smith, Z. Xuan, M.Q. Zhang, "Using Quality Scores and Longer Reads Improves Accuracy of Solexa Read Mapping," *BMC Bioinformatics*, Vol. 9, No. 128, 2008.

[5] H. Jiang and W.H. Wong, "SeqMap: Mapping Massive Amount of Oligonucleotides to the Genome," *Bioinformatics*, Vol. 24, No. 20, pp. 2395-2396, 2008.

[6] B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, "Ultrafast and Memory-efficient Alignment of Short DNA Sequences to The Human Genome," *Genome Biology*, Vol. 10, No. 3, pp. R25.1-R25.10, 2009.

*Correspondence author

[7] R. Li, C. Yu, Y. Li, T.W. Lam, S.M. Yiu, K. Kristiansen, J. Wang, "SOAP2: An Improved Ultrafast Tool for Short Read Alignment," *Bioinformatics*, Vol. 25, No. 15, pp. 1966-1967, 2009.

[8] H. Li and R. Durbin, "Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform," *Bioinformatics*, Vol. 25, No. 14, pp. 1754-1760, 2009.

[9] M. Burrows and D. J. Wheeler, "A Block-Sorting Lossless Data Compression Algorithm," *Technical Report of Systems Research Center*, Palo Alto, CA, No. 124, 1994.

[10] P. Ferragina and G. Manzini, "Opportunistic Data Structures with Applications," in *Proceedings of the 41st Annual Symposium on Foundation of Computer Science*, pp. 390-398, 2000.

[11] P. Ferragina and G. Manzini, "An Experimental Study of an Opportunistic Index," in *Proceedings of the Twelfth Annual ACM-SIAM Symposium on Discrete algorithms*, pp. 269-278, 2001.

[12] NCBI Human Genome Resources, http://www.ncbi.nlm.nih.gov/projects/genome/guide/human/

[13] 1000 Genomes, http://www.1000genomes.org/