LLC-GNUMAP: Scalable, Precise, and High-Coverage Genomics Mapping*

Natacha P. Leitão[1], João Leitão[2], and Francisco M. Couto[1]

[1]LaSIGE, Department of Informatics, Faculty of Sciences, University of Lisbon
[2]CITI, Department of Informatics, Faculty of Sciences and Technology, NOVA University of Lisbon

Next generation sequencing (NGS) technologies revolutionized biomedical research by bringing the promise of enabling the production of large amounts of sequence data at relatively low cost in a short time. NGS technologies are now being applied to a growing number of biological applications, most of which rely on accurate and fast read mapping mechanisms (i.e., aligning reads to a reference genome) in which a fundamental step is the ability to distinguish between technical sequencing errors and natural genetic variation. Many mapping tools have been developed, such as Bowtie[2], BWA[3], MAQ[4], and SHRiMP[5]. Each of these solutions rely on different approaches that have inherent limitations, either in terms of scalability (the time required to execute the mapping), coverage (the percentage of reads that are effectively mapped), and precision (mapping reads to the correct location in the reference genome). In particular, coverage is an extremely challenging aspect, as when a read occurs at multiple locations of the genome, finding the exact location from which it came is nearly impossible, and many existing solutions simply discard those reads. A key recent improvement in NGS technologies which might enable addressing the limited coverage of existing solutions is the quality information which is now included in every run; each position in a read correspond to one of the four bases, the quality score assigned to each position indicates the probability of the reported base in that position being incorrect. In fact, this information so far is being used by some solutions [2, 3, 4] through different approaches that focus in the probability of the called base being incorrect. A recent solution, named GNUMAP[1], has relied on this information to devise a rigorous probabilistic approach that uses the probabilities of all four bases being called. Therefore, this solution enables mapping of reads that occur in several regions, with a high coverage of approximately 70% albeit with no reported precision. In this manuscript, we propose a new mapping GNUMAP-base algorithm, dubbed *LLC-GNUMAP*, witch offers the potential not only to improve the coverage of GNUMAP, but also to improve its precision. At the core of LLC-GNUMAP lies a novel technique which increases the search space over the reference genome for each individual read, by using a hash-based approach that leverages quality information and biological constraints. As we increase the search space for each read, we designed our algorithm to be highly parallelized, allowing executions to span across an arbitrary number of machines - for instance in a cloud computing platform - with near linear scalability. Additionally, we propose a methodology to measure the precision of mapping algorithms based on artificial read sets, which are constructed to emulate common errors found in NGS datasets.

Acknowledgements

The authors would like to thank Alysson Bessani and Vinícius Cogo for the hardware provided, their availability and technical support. This work was financially supported by FCT through funding of LaSIGE Strategic Project, ref. PEst-OE/EEI/UI0408/2014.

References

[1] N. L. Clement, Q. Snell, M. J. Clement, P. C. Hollenhorst, J. Purwar, B. J. Graves, B. R. Cairns, and W. E. Johnson. The GNUMAP algorithm: unbiased probabilistic mapping of oligonucleotides from next-generation sequencing. *Bioinformatics*, 26(1):38–45, 2010.

[2] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, 10:R25, 2009.

[3] H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[4] H. Li, J. Ruan, and R. Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, 18:1851–1858, 2008.

[5] S. M. Rumble, P. Lacroute, A. V. Dalca, M. Fiume, A. Sidow, and M. Brudno. SHRiMP: Accurate Mapping of Short Color-space Reads. *PLoS Comput. Biol.*, 5(5):e1000386, 2009.

*Corresponding Author: Natacha P. Leitão. Address: LaSIGE, Faculdade de Ciências da Universidade de Lisboa, Departamento de Informática, Edifício C6 Piso 3, Campo Grande 1749 - 016 Lisboa. E-mail: nleitao@lasige.di.fc.ul.pt.