G-SNPM - A GPU-based SNP mapping tool

Alessandro Orro[1]✉, Andrea Manconi[1], Emanuele Manca[2], Giuliano Armano[2], Luciano Milanesi[1]

[1]Institute for Biomedical Technologies, National Research Council, Milano, Italy
[2]Department of Electrical and Electronic Engineering, University of Cagliari, Cagliari, Italy

Motivation and Objectives

In genotyping analysis often researchers need to merge together genetic datasets coming from different genotyping platforms that use different sets of Single Nucleotide Polymorphisms (SNPs) to represent genetic polymorphisms. In order to do this, it is necessary to know the exact position of a SNP in a chromosome and update this information when new builds of the reference genome are available.

In this work, we present G-SNPM (GPU SNP Mapping) a GPU-based tool to map SNPs on a genome.

Methods

G-SNPM is a tool that maps a short sequence (read) representative of a SNP against a reference DNA sequence in order to find the absolute position of the SNP in that sequence.

Several tools have been devised to perform short-read mapping. Without aiming to be exhaustive, we can cite some solutions: MAQ (Li and Durbin, 2008), RMAP (Smith et al., 2008; Smith et al., 2009), Bowtie (Langmead et al., 2009), BWA (Li and Durbin, 2009), CloudBurst (Schatz, 2009), and SHRiMP (Rumble et al., 2009). A comparative study aimed at assessing the accuracy and the runtime performance of six state-of-the-art next-generation sequencing read alignment tools (Ruffalo et al., 2011) highlighted that among all SOAPv2 (Li et al., 2009) is the one that shows the higher accuracy.

Recently, it has been proposed SOAPv3 (Liu et al., 2012) the GPU-based evolution of the SOAPv2 aligner. Experimental results shown that SOAPv3 outperforms notably both BWA and Bowtie. When tested to align millions of 100-bp read pairs to the human genome it resulted at least 7.5 times faster than BWA, and 20 times faster than Bowtie. Moreover, SOAPv3 that not exploits heuristics is able to align correctly slightly more reads than BWA and Bowtie. The current release of SOAPv3 supports alignments with up to four mismatches while it does not support indels.

In G-SNPM each SNP is mapped on its related chromosome by means an automatic three stage pipeline. In the first stage, G-SNPM uses SOAPv3 to parallel align on a reference chromosome its related reads representative of a SNP. Due to the fact that SOAPv3 does not support indels, it might not be able to align some reads. Then, in the second stage G-SNPM uses another short-read mapping tool to align the unmapped reads. In particular, in this stage it is used SHRiMP which exploits specialized vector computing hardware to speed-up the dynamic programming algorithm of Smith-Waterman. Finally, in the third stage, G-SNPM analyses the alignments of the reads mapped by SOAPv3 and SHRiMP to calculate the absolute position of each SNP. An output file is generated which for each SNP reports its name, the related chromosome, the original SNP position, and the mapped SNP position. Moreover, information about the alignment as the strand, number of mismatches, and indels are also provided (see Figure 1).

In G-SNPM reference DNA sequences are accepted in standard FASTA format, whereas SNPs must be represented through two files: a FASTA file with the representative reads of the SNPs, and

Name	CHR	SNP	Map	S	M	I	D
rs13305024	Y	17038316	17038318	-	1	0	0
rs9786448	Y	17115298	17115299	+	1	0	0
rs9785704	Y	17175506	17175507	+	0	0	0
MitoA9073G	MT	9073	9073	+	1	0	0
MitoA9094G	MT	9094	9094	+	1	0	0

Figure 1: Screenshot of the generated output file.

another flat file with information about the SNP, in particular the original absolute SNP position and its alleles. Currently, automatic generation of these files is provided for SNP probes of the Illumina Chip. G-SNPM analyses Illumina files to automatically generate the previous described files for each chromosome.

Results and Discussion

The tool has been tested in the problem of re-mapping all the SNP probes of the Illumina Chip HumanOmni 1S (version 1), in order to find the map positions of each SNP in the build 37.3 of the refseq.

To assess the performance of G-SNPM we compared its performance with those obtained by mapping the same SNPs with the state-of-the-art short-read mapping tool BWA. Experimental results shown that in the task of mapping around 1.2 million of SNPs BWA has been unable to map 55 SNPs (maximum edit distance 4% and up to two gap opening), whereas G-SNPM mapped correctly all SNPs. In particular, 178 SNPs has been mapped with SHRiMP in the second stage of the pipeline.

Results shown that BWA has been able to map more reads than SOAPv3. Since SOAPv3 does not support indels, it might be unable to align some reads. However, it should be pointed out that differently that SOAPv3, BWA is designed not to miss any potential alignment resulting in many incorrect mapped reads (Ruffalo et al., 2011).

Currently, G-SNPM runs on linux and it is freely available as a standalone application at the address http://www.itb.cnr.it/web/bioinformatics/g-snpm .

To use G-SNPM is required a computer equipped with a CUDA enabled GPU card based on the Fermi architecture. We assessed G-SNPM with a NVIDIA GeForce GTX 480 card.

Acknowledgements

This work has been supported by the Italian Ministry Education and Research (MIUR) through the Flagship "InterOmics", ITALBIONET (RBPR05ZK2Z), HIRMA (RBAP11YS7K) and the European "MIMOMICS" projects.

References

1. Langmead B, Trapnell C, et al. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25.
2. Li H, Ruan J, Durbin R (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. , 18(11):1851–8.
3. Li H and Durbin R (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25(14):1754-1760
4. Li R, Yu C, Li Y, et al. (2009). SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics, 25(15).
5. Liu CM, Wong T, et al. (2012). SOAP3: ultra-fast GPU-based parallel alignment tool for short reads. Bioinformatics, 28(6):878-9.
6. Ruffalo M, LaFramboise T, Koyutürk M (2011). Comparative analysis of algorithms for next-generation sequencing read alignment. Bioinformatics 15;27(20):2790-6.
7. Rumble SM, Lacroute P, et al. (2009). SHRiMP: Accurate Mapping of Short Color-space Reads. PLoS Comput Biol 5(5):e1000386. doi:10.1371/journal.pcbi.1000386.
8. Schatz MC (2009). CloudBurst: highly sensitive read mapping with MapReduce. Bioinformatics, 25(11):1363–9.
9. Smith AD, Chung WY, et al. (2009). Updates to the RMAP short-read mapping software. Bioinformatics, 25(21):2841–2842.
10. Smith AD, Xuan Z, Zhang MQ (2008). Using quality scores and longer reads improves accuracy of solexa read mapping. BMC Bioinformatics, 9:128.