BarraCUDA - a Fast Sequence Mapping Software using Graphics Processing Units

Brian Y. H. Lam, Petr Klus, Simon Lam and Giles S. H. Yeo
Metabolic Research Laboratories, Institute of Metabolic Science, University of Cambridge, Addenbrooke’s Hospital, Cambridge, CB2 0QQ, United Kingdom

Email addresses: brian.lam@yoblog.com, Giles Yeo (ghy2@cam.ac.uk)

Introduction

High-throughput DNA sequencing (HTS) instruments today are capable of generating millions of sequencing reads in a short period of time, and this represents a serious challenge to current bioinformatics pipeline in processing such an enormous amount of data in a fast and economical fashion.

Modern graphics cards are powerful processing units that consist of hundreds of scalar processors in parallel in order to handle the rendering of high-definition graphics in real-time. It is this computational capability that we propose to harness in order to accelerate some of the time-consuming steps in analyzing data generated by the HTS instruments.

We have developed BarraCUDA, a novel sequence mapping software that utilizes the parallelism of Nvidia CUDA graphics cards to map sequencing reads to a particular location on a reference genome. While delivering a similar mapping fidelity as other mainstream programs, BarraCUDA is a magnitude faster in order to accelerate some of the time-consuming steps in analyzing data generated by the HTS instruments.

Inexact Matching Using FM-index Algorithm

The FM-index algorithm is a substring index based on Burrows-Wheeler transform (BWT) and it allows fast substring matching by prefix trie traversal. The algorithm is widely used in DNA sequence mapping programs such as BWA1 and Bowtie2, where the string is the reference genome and the query substrings are the sequencing reads generated by HTS instruments.

For inexact matching, a series of substitutions are introduced in the query string such that the string with a substituted character could lead to a match, e.g. ‘b’ in ‘abn’ is substituted with an ‘a’ to give an match ‘ana’ with one mismatch.

Inexact matching is performed by ‘backward search’ through a BWT prefix trie, e.g. ‘banana’.

Figures 1 and 2 illustrate the route to match ‘anb’.

CUDA Implementation

Substring matching is performed by traversal of the trie, e.g. ‘banan’.

Table 1 illustrates the mapping accuracy.

Mapping Accuracy

<table>
<thead>
<tr>
<th></th>
<th>BWA</th>
<th>BarraCUDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Mapped</td>
<td>96.50</td>
<td>96.64</td>
</tr>
<tr>
<td>% Error</td>
<td>0.04</td>
<td>0.06</td>
</tr>
</tbody>
</table>

Mapping Speed

The mapping throughput was examined by mapping a HTS sequencing library containing 14 million reads (ENA accession: SRRO03699) to the genome of Drosophila Melanogaster. The mapping throughput of BarraCUDA with one GPU was about 6X the speed of BWA with a single CPU core. With 8X GPUs, BarraCUDA outperforms BWA using all 12 CPU cores by 2.8 fold.

Conclusions

BarraCUDA is designed to take advantage of the parallelism of GPU to accelerate the mapping of millions of sequencing reads generated by HTS instruments. By doing this, we could, at least in part streamline the current bioinformatics pipeline such that the wider scientific community could benefit from the sequencing technology. BarraCUDA is currently available at http://seqbarracuda.sf.net

References


Acknowledgments

We would like to thank Graham Pullan, Tobias Brandvik, Ian McFarlane, Dag Lyberg, Simon Lam, Nicole Cheung, Thomas Bradley, and Giles S. H. Yeo for their support. This work is supported by the Medical Research Council Centre of Obesity and Related Metabolic Diseases, a funding to the Cambridge Biomedical Research Centre, NIHR Cambridge, UK.