GPU-Accelerated Pipeline For Next Generation Sequencing Data Simulation

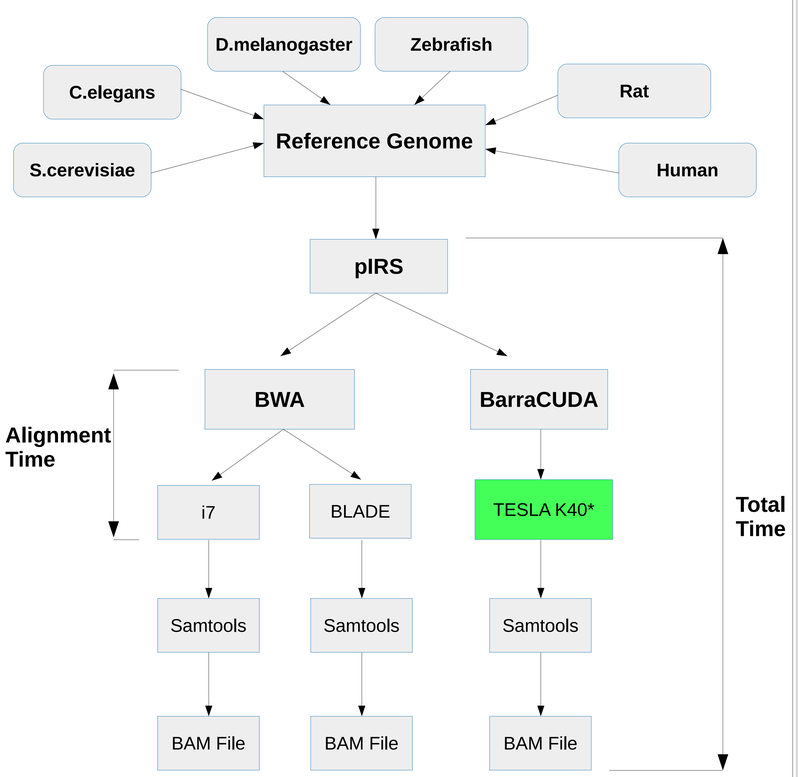
*Andrei **Rozanski**[1], Daniel T. Ohara[1], Pedro A. F. Galante[1]*

[1]- Bioinformatics group – Centro de Oncologia Molecular Hospital Sírio-Libanês; São Paulo, Brazil; ***Contact: arozanski@mochsl.org.br***

Why Simulate Next Generation Sequencing Data?

Cheap and reliable generation of biological information in large scale became possible through Next Generation Sequencing technology (NGS). An ever growing amount of data obtained from NGS is generated each year. Valuable information about development and evolution of several diseases have been learned from the analysis of such data. However, to convert data into useful information is a well known problem and a hard task to accomplish. To deal with that, bioinformaticians develop and improve several pipelines. Time and cost effective pipelines that are capable of dealing with a huge amount of data are desirable. Data simulation plays a key role in the development and optimization of pipelines. It helps to guarantee quality, better control and determination of pipeline reliability (i.e. determination of sensibility and specificity) and, at the end, it optimizes resources.

Figure - 1 Pipeline Steps



** "The Tesla K40 used for this research was donated by the NVIDIA Corporation."*

How GPU Can Help ?

Here we attempted to accelerate a NGS data simulation using GPU. Briefly, we obtained the reference genome for 6 different species - S.cerevisiae, C.elegans, D.melanogaster, Zebrafish, Rat and Human from UCSC GoldenPath - Figure – 1. Based on reference genome, we simulate paired-end reads with pIRS at 3X of coverage for each specie. After reads generation, we performed reads alignement. A CPU based strategy uses BWA (http://bio-bwa.sourceforge.net/) for the task and GPU based strategy were applied using GPU implementation of BWA – BarraCUDA. Three different computer configurations (Table – 1) were compared for pipeline run.

Results

All pipeline steps were performed sequentially and without interruptions. Genome and consequently simulated data sizes variates among species however perfect mapping for all species using different alignment softwares (BWA and BarraCUDA) were observed (Table - 2). A slight reduction on alignment speed were observed in i7 and Blade as genome size and genome/chromosome increases. Regarding alignment speed, i7 speed variated from ~60 to ~25% of GPU speed and Blade from ~60 to 40% of GPU speed. When total pipeline speed is considered, i7 variated from ~65 to ~40% and Blade from ~60 to 35% of GPU speed. This produces an average of ~44% and ~40% of GPU alignment speed for i7 and Blade respectively. Similarly, average of ~46% of GPU total pipeline speed for i7 and Blade were observed.

Table - 1 Computers Specifications

Alias	CPU Type	Processors	CPU Cores	RAM Memory	GPU
i7	Intel(R) Core(TM) i7-2600 CPU @ 3.40GHz	1	4	8Gb	None
Blade*	Intel(R) Xeon(R) CPU E5-2440 0 @ 2.40GHz	2	6	128Gb	None
GPU	Intel(R) Core(TM) i7-2600 CPU @ 3.40GHz	1	4	8Gb	Tesla K40

*Model Name: PowerEdge M1000e

Figure - 2 Alignment Speed

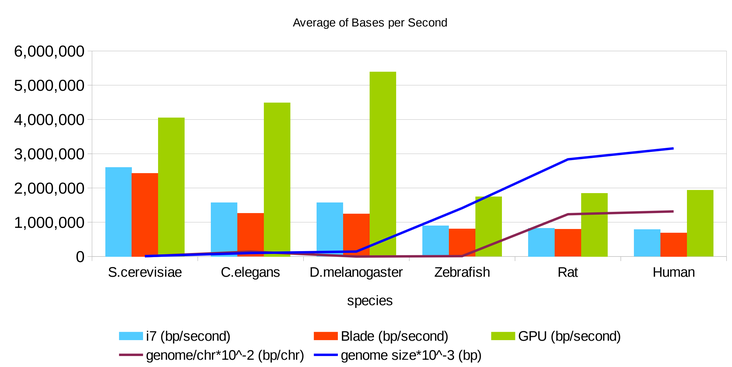
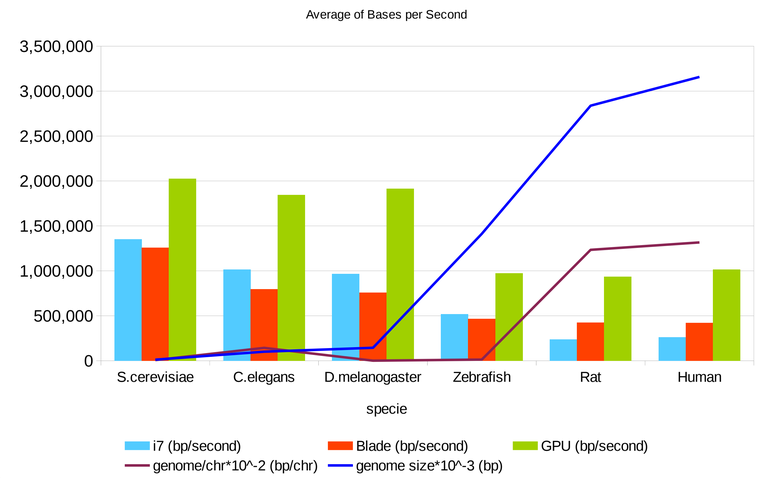


Table - 2 General Simulated Data Characteristics

	Genome Size *Total (bp)*	Simulated Read Pairs *Total (n)*	Simulated Bases *Total (bp)*
S.cerevisiae	12,157,105	182,348	36,469,600
C.elegans	100,286,070	1,504,287	300,857,400
D.melanogaster	143,726,002	2,154,943	430,988,600
Zebrafish	1,412,464,843	21,186,406	4,237,281,200
Rat	2,837,669,505	41,730,421	8,346,084,200
Human	3,157,590,972	46,349,211	9,269,842,200

Figure - 3 Pipeline Total Speed



Conclusions

To optimize pipelines that extract relevant information from public data is challenging. To accomplish this task, data simulation may be applied. Our findings support that GPU improves dramatically a CPU NGS Data Simulation performance. The accelerated pipeline showed marked improvement in alignment and total speed without jeopardizes in different genome sizes.

References:

1- **Klus**, P, Lam, S, Lyberg, D, Cheung, MS, Pullan, G, McFarlane, I, Yeo, GSh, Lam, BY (2012). BarraCUDA - a fast short read sequence aligner using graphics processing units. BMC Res Notes, 5:27.
2- **Hu**, X, Yuan, J, Shi, Y, Lu, J, Liu, B, Li, Z, Chen, Y, Mu, D, Zhang, H, Li, N, Yue, Z, Bai, F, Li, H, Fan, W (2012). pIRS: Profile-based Illumina pair-end reads simulator. Bioinformatics, 28, 11:1533-5.
3- **Li**, H, Durbin, R (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25, 14:1754-60.

Support:

FAPESP SÃO PAULO RESEARCH FOUNDATION
LUDWIG INSTITUTE FOR CANCER RESEARCH
nVIDIA.