BIOINFODIAG

Réseau Français de Bioinformatique pour le Diagnostic

API HOUR

GPU accelerated Bioinformatics

Yannis Duffourd
CHU Dijon – INSERM U1231

11 janvier 2023

Computation times on GS analysis

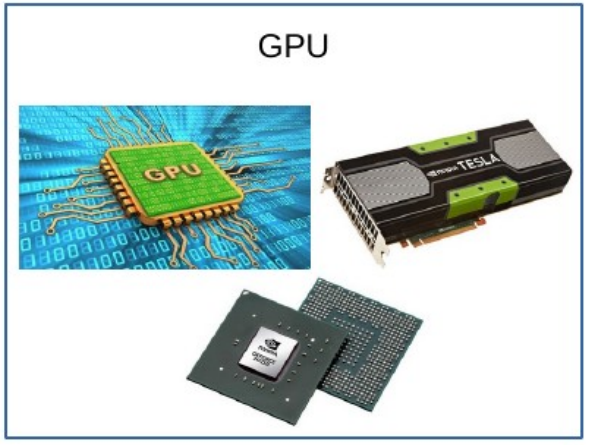
1 human GS ~ 30x : CPU driven analysis

	computation time	waiting time
Classical CPU analysis	255,4 h	~72h
Fast analysis	100 h	28,8 h
Improved fast analysis	100 h	12 h

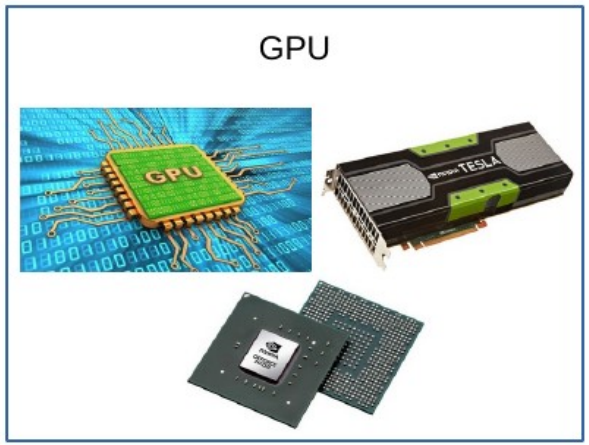
Computation times on GS analysis

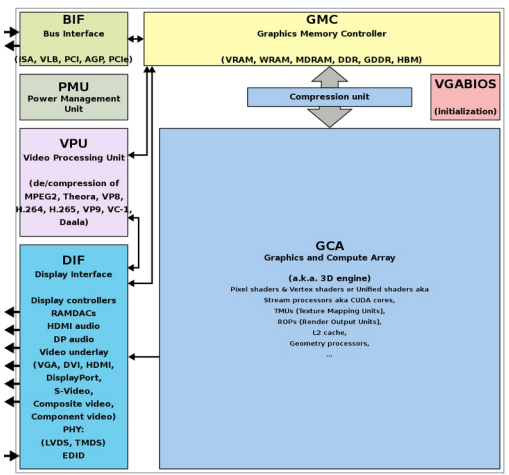
1 human GS ~ 30x : CPU driven analysis

	computation time	waiting time
Classical CPU analysis	255,4 h	~72h
Fast analysis	100 h	28,8 h
Improved fast analysis	100 h	12 h



Computation times on GS analysis





From Wikipedia

Graphical Processing Unit

A processor dedicated and specialized in graphical operations

Video games, 3D modeling, image rendering, etc …

Matrix computation

Specialized cores in a large number

2560 CUDA cores + 640 Tensor cores

Tesla V100 GPU : 8.2 TFLOPS (double precision)

vs i9 16c 5.2 GHz : 1.4 TFLOPS

Dedicated softwares needed …

GPU Softwares

BarraCUDA - a fast short read sequence aligner using graphics processing units

Petr Klus[1†], Simon Lam[2†], Dag Lyberg[3], Ming Sin Cheung[4], Graham Pullan[5], Ian McFarlane[2], Giles SH Yeo[1] and Brian YH Lam[1*]

2012 - BMC Research Notes

Basé sur une réimplémentation de BWT en CUDA (Compute Unified Device Architecture)

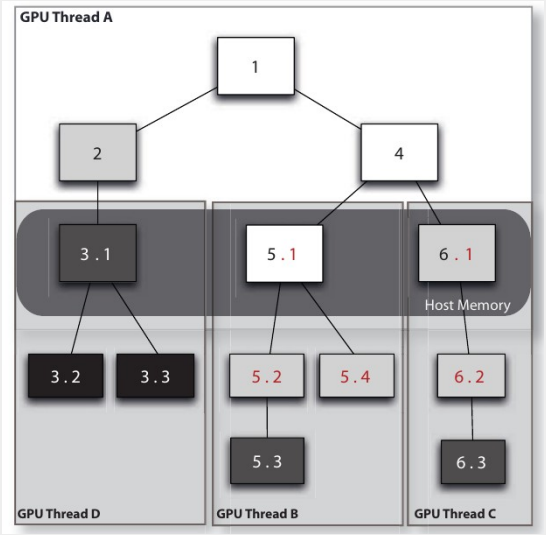
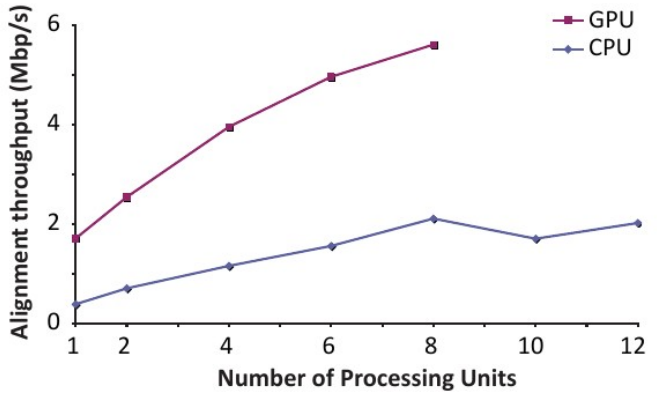


Table 1 Alignment accuracy compared to BWA using simulated reads

	BWA		BarraCUDA	
	Gap enabled	Gap disabled	Gap enabled	Gap disabled
Single-end				
% Mapped	89.95	91.01	91.42	91.14
% Error	0.06	0.04	0.05	0.05
Paired-end				
% Mapped	96.53	96.50	96.61	96.64
% Error	0.04	0.04	0.06	0.06



GPU Softwares

Arioc: high-throughput read alignment with GPU-accelerated exploration of the seed-and-extend search space

Richard Wilton,[1] Tamas Budavari,[2] Ben Langmead,[3,6] Sarah J. Wheelan,[4,7] Steven L. Salzberg,[3,5,6] and Alexander S. Szalay[1,3]

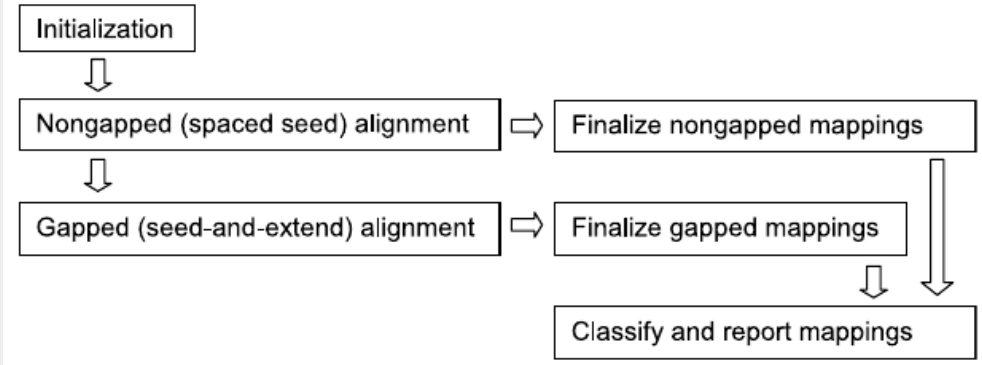
2015 - PeerJ

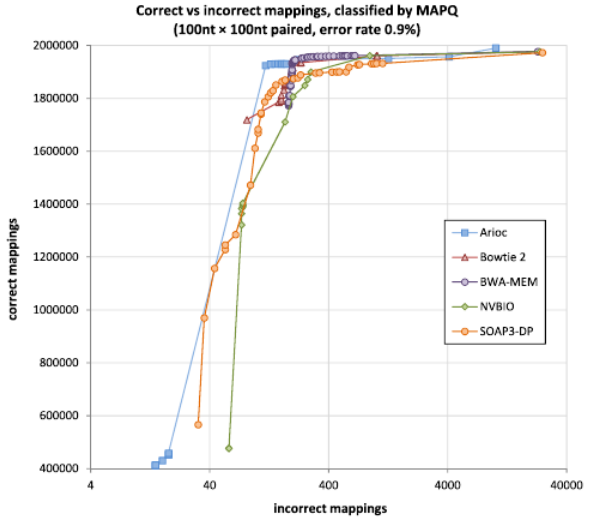
Arioc: GPU-accelerated alignment of short bisulfite-treated reads

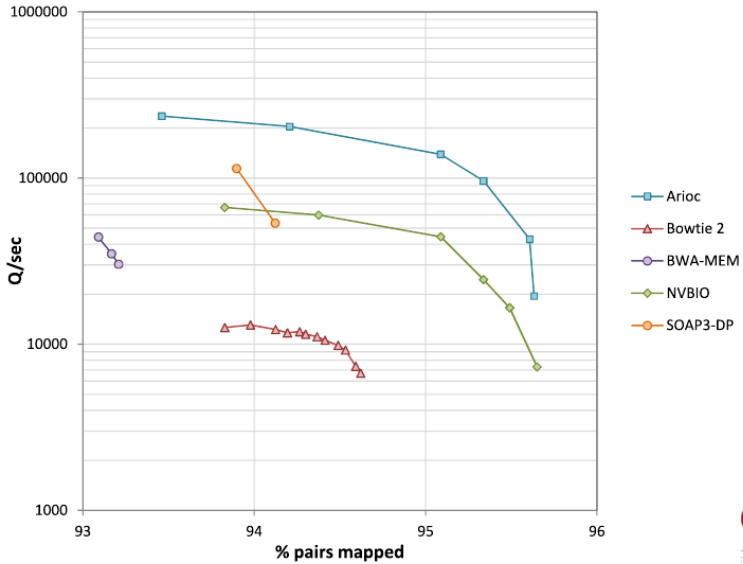
FREE

Richard Wilton ✉, Xin Li, Andrew P Feinberg, Alexander S Szalay

Bioinformatics, Volume 34, Issue 15, 1 August 2018, Pages 2673–2675, https://doi.org/10.1093/bioinformatics/bty167







GPU Softwares

Turki and Roshan *BMC Genomics* 2014, **15**:969
http://www.biomedcentral.com/1471-2164/15/969

BMC Genomics

SOFTWARE Open Access

MaxSSmap: a GPU program for mapping divergent short reads to genomes with the maximum scoring subsequence

Turki Turki[1,2*] and Usman Roshan[2*]

Smith-Waterman sur GPU

CUDA / OpenCL

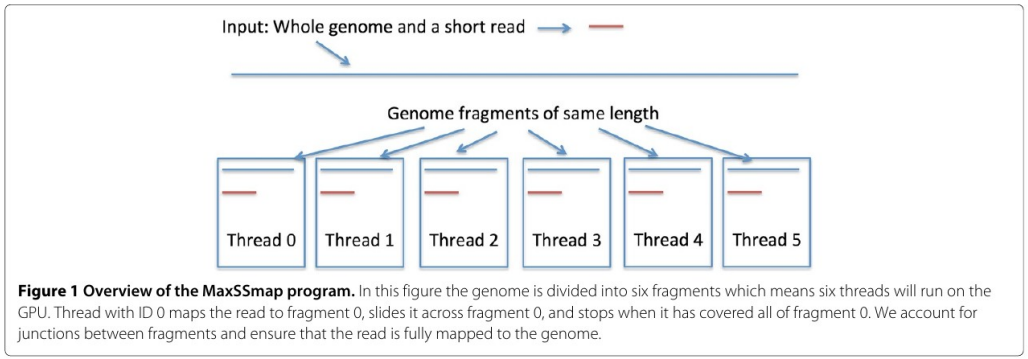


Figure 1 Overview of the MaxSSmap program. In this figure the genome is divided into six fragments which means six threads will run on the GPU. Thread with ID 0 maps the read to fragment 0, slides it across fragment 0, and stops when it has covered all of fragment 0. We account for junctions between fragments and ensure that the read is fully mapped to the genome.

Table 8 Percent of 100,000 ancient horse DNA reads (SRR111892) of length 76 bp mapped to the horse genome Equus_caballus EquCab2 (GCA_000002305.1) and human reference genome

Horse genome				
BWA	**NextGenMap**	**NextGenMap+ MaxSSmap_fast**	**NextGenMap+ MaxSSmap**	**NextGenMap+ CUDASW++**
2.2	16	20.5	23.1	26 (estimated)
Time in minutes to map reads				
0.6	2.4	1609.6	2836	14820 (estimated)
Human genome				
BWA	**NextGenMap**	**NextGenMap+ MaxSSmap_fast**	**NextGenMap+ MaxSSmap**	**NextGenMap+ CUDASW++**
0.16	14	18.6	21	NA
Time in minutes to map reads				
0.82	2.9	2375.5	4108	NA

We ran NextGenMap+CUDA-SW++ for a maximum of 168 hours and estimated the time to align all rejected reads. Also shown is time in minutes.

- Pas une vraie concurrence à bwa en terme de performance de calcul

- Utilisé pour aligner les lectures non alignées par bwa...

Sur l'humain : 2375 minutes / 100 000 reads

1 WGS 30x = ~ 1 Md de reads

23 750 000 minutes …

soit 23 ans … ??

GAD

GPU Softwares

BALSA: integrated secondary analysis for whole-genome and whole-exome sequencing, accelerated by GPU

Ruibang Luo[1,3], Yiu-Lun Wong[1,3], Wai-Chun Law[1,3], Lap-Kei Lee[1,2], Jeanno Cheung[1], Chi-Man Liu[1] and Tak-Wah Lam[1]

2014 - PeerJ

CUSHAW: a CUDA compatible short read aligner to large genomes based on the Burrows–Wheeler transform

FREE

Yongchao Liu ✉, Bertil Schmidt ✉, Douglas L. Maskell Author Notes

Bioinformatics, Volume 28, Issue 14, 15 July 2012, Pages 1830–1837,
https://doi.org/10.1093/bioinformatics/bts276

Optimizing Data Intensive GPGPU Computations for DNA Sequence Alignment

Cole Trapnell† and Michael C. Schatz†

2009 – Parallel Comput.

PLoS One. 2018; 13(1): e0190279. PMCID: PMC5749749
Published online 2018 Jan 2. doi: 10.1371/journal.pone.0190279 PMID: 29293576

pyPaSWAS: Python-based multi-core CPU and GPU sequence alignment

Sven Warris, Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Resources, Software, Supervision, Validation, Writing – original draft, Writing – review & editing,[1,2,*] N. Roshan N. Timal, Methodology, Software, Writing – review & editing,[3] Marcel Kempenaar, Investigation, Methodology, Software, Writing – review & editing,[1] Arne M. Poortinga, Software,[1] Henri van de Geest, Software,[2] Ana L. Varbanescu, Conceptualization, Investigation, Methodology, Resources, Software, Writing – review & editing,[3] and Jan-Peter Nap, Conceptualization, Funding acquisition, Resources, Supervision, Writing – original draft, Writing – review & editing[1,2]

2.5 reads / s
Soit 13 ans pour un WGS 30x

BitMapper2: a GPU-accelerated all-mapper based on the sparse *q*-gram index

Haoyu Cheng, Yong Zhang and Yun Xu*

Samples

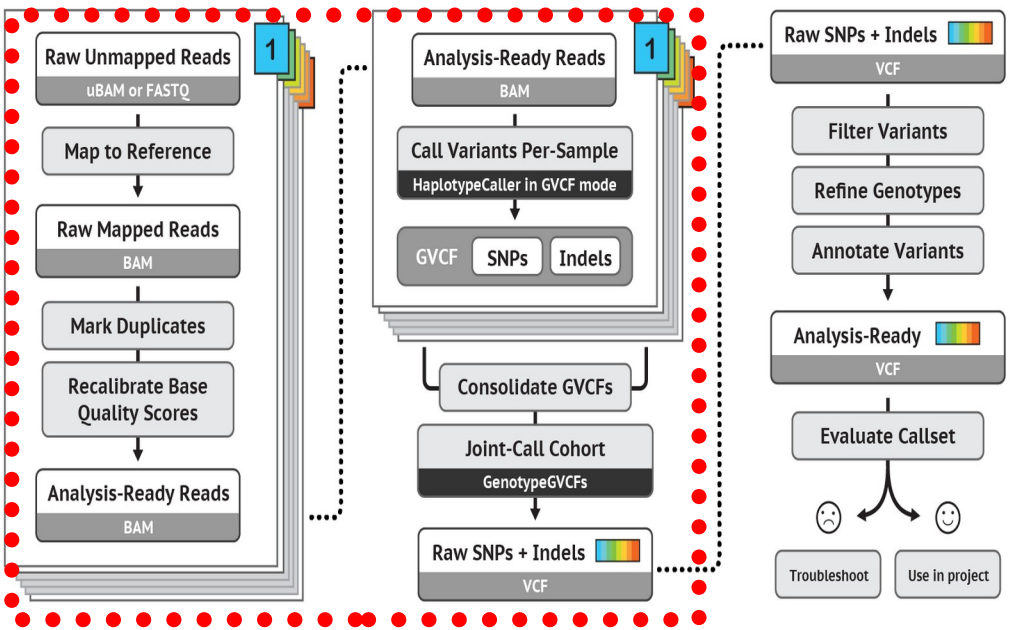
10 WGS samples

dijen203
dijen006
dijen088
dijen103
dijen104
dijen105
dijen112
dijen192
dijen193
dijen194

~ 30-40 x

HiSeq X Ten

2014-18



Sample	total exec time (s)	*On CPU*
dijen203	62834	
dijen006	69477	
dijen192	67705	
dijen193	71039	
dijen194	67812	

1h = 3600 s

Samples

10 WGS samples

dijen203
dijen006
dijen088
dijen103
dijen104
dijen105
dijen112
dijen192
dijen193
dijen194

~ 30-40 x

HiSeq X Ten

2014-18

dijen203 Ran 3 times for repeatability

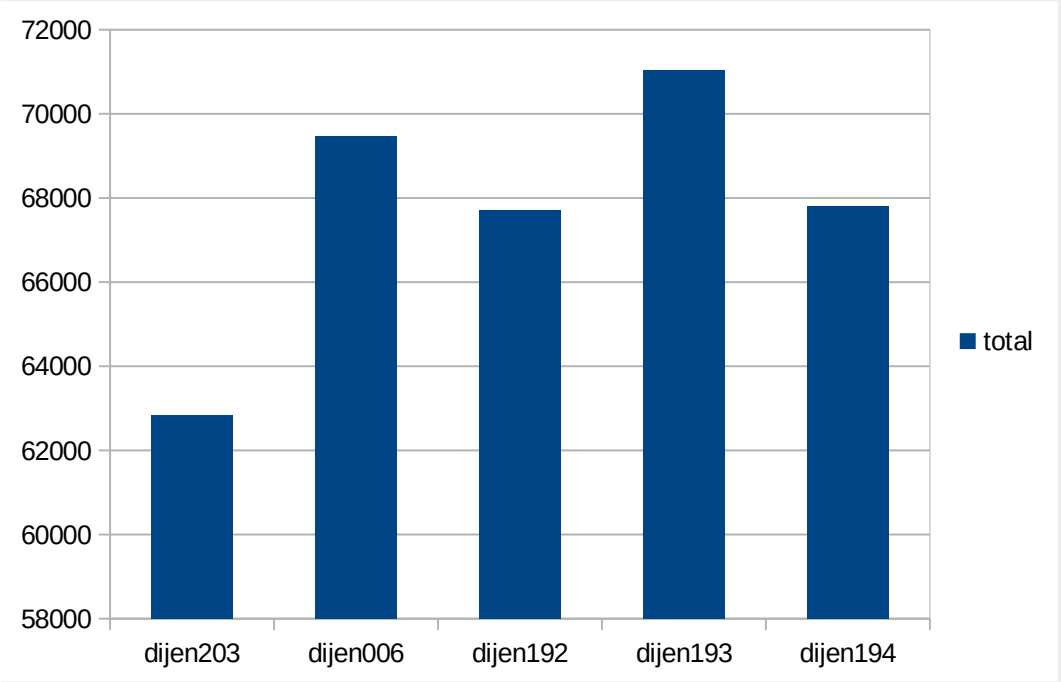
Parabricks accelerated Genomics Pipeline

Version v2.1.5

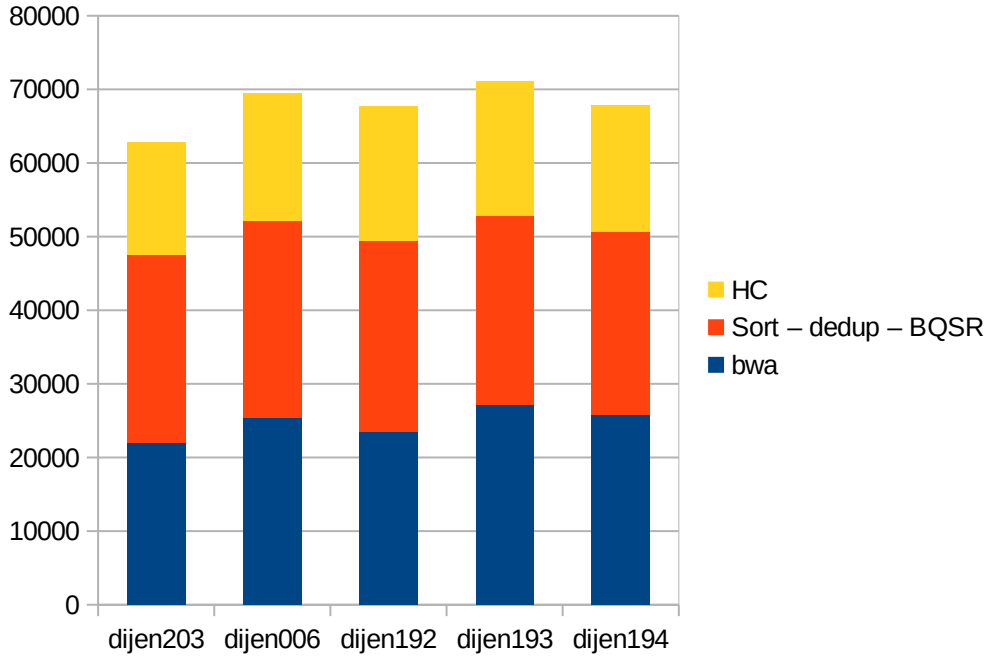
4 Nvidia GPUs on a DGX Server
(Tesla v100 – 32 Go dedicated RAM)

HDD : 3 x 2 To SSD (Raid)

Performances



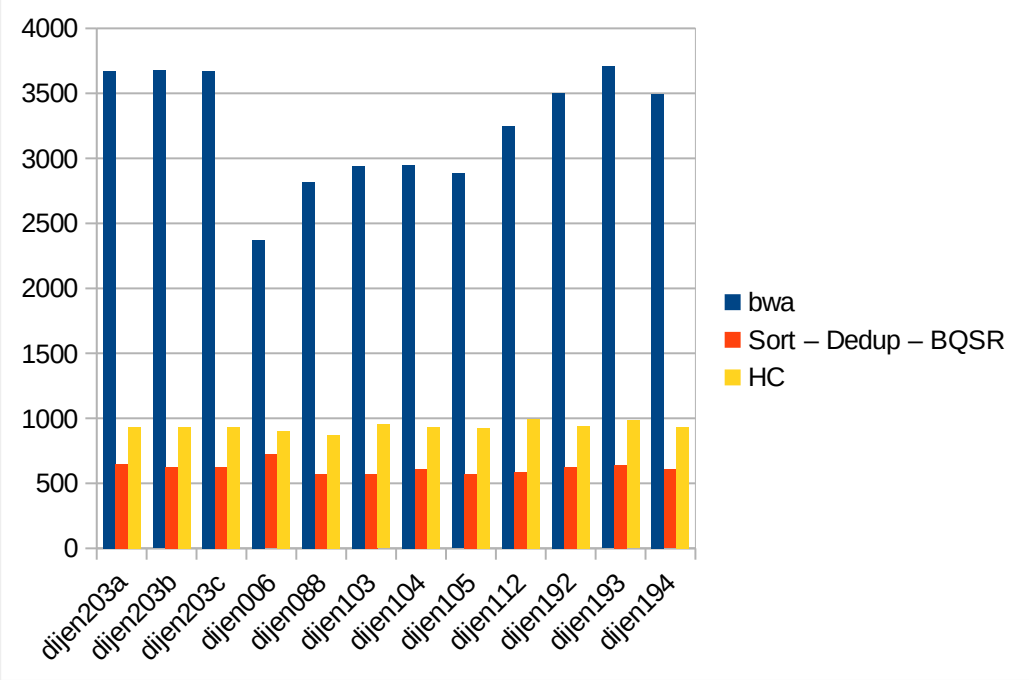
CPU Driven analysis on 16 cores

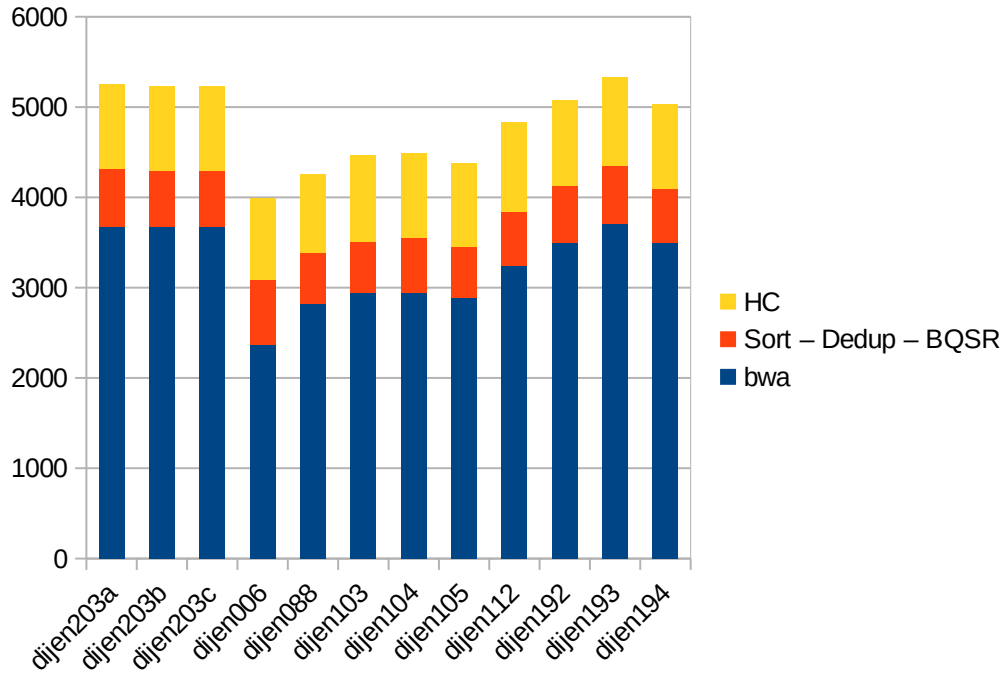


All times are expressed in seconds (s)

Performances

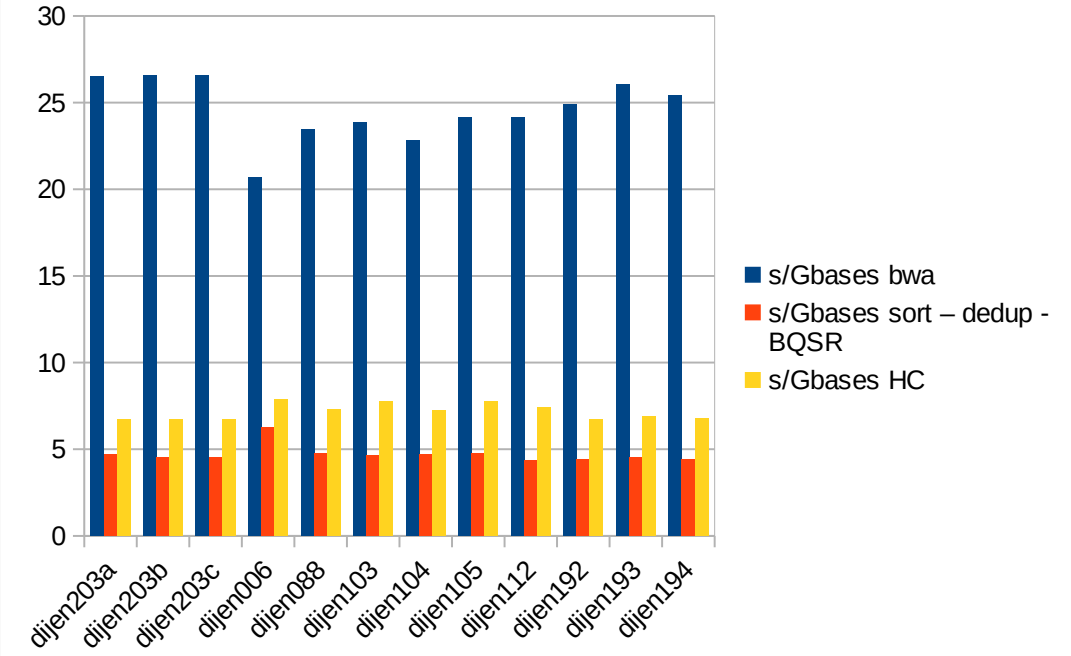
Steps running time (s)

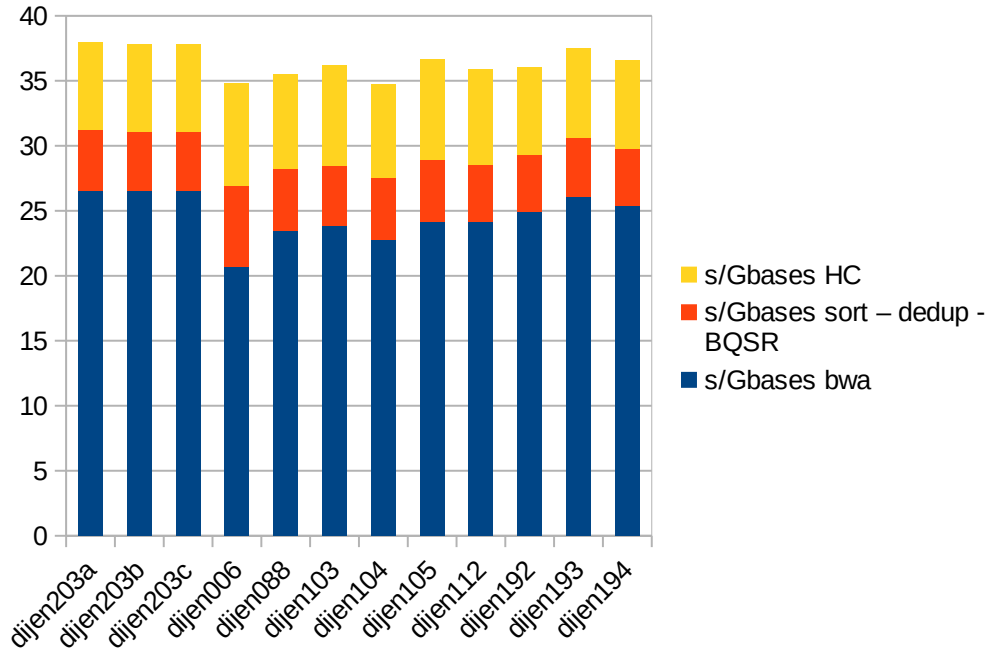




Performances

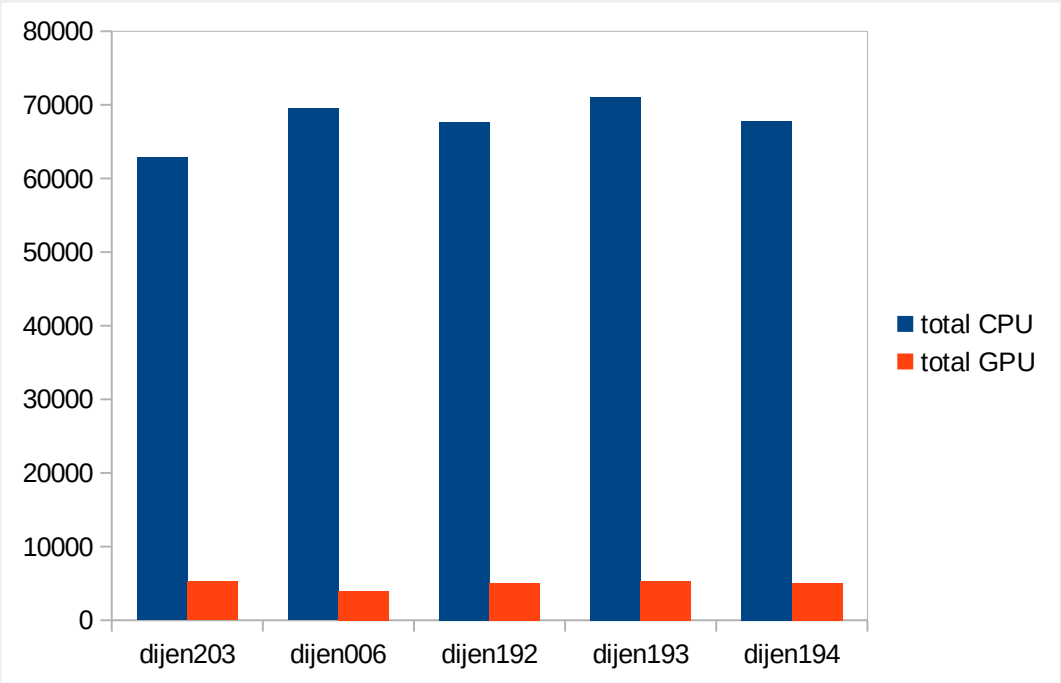
Steps running time (s/Gbase)

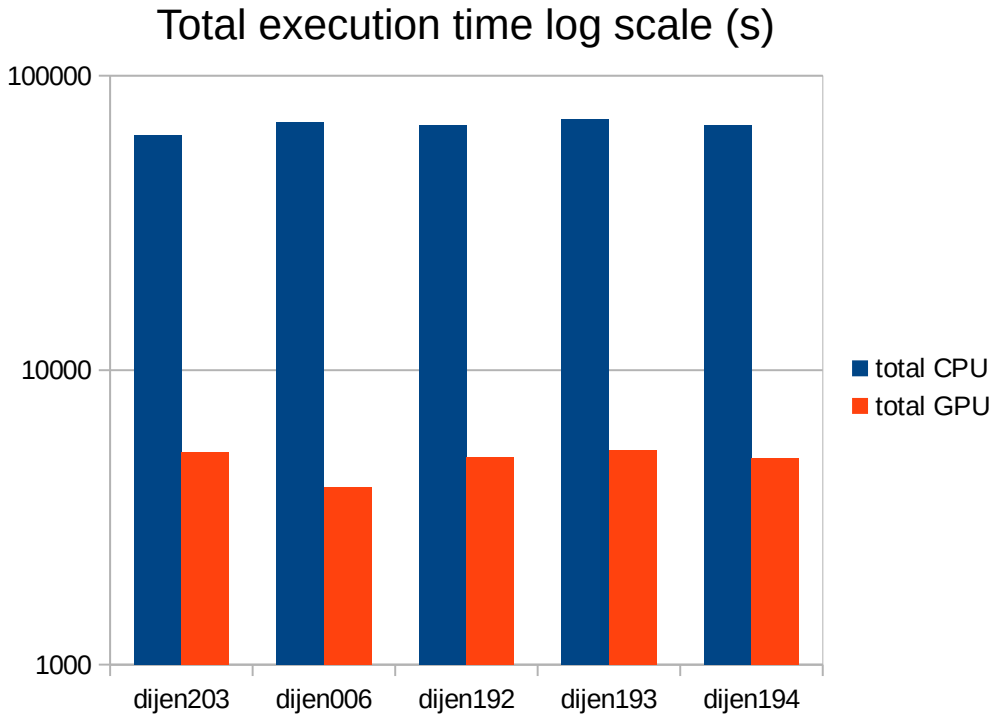




Performances

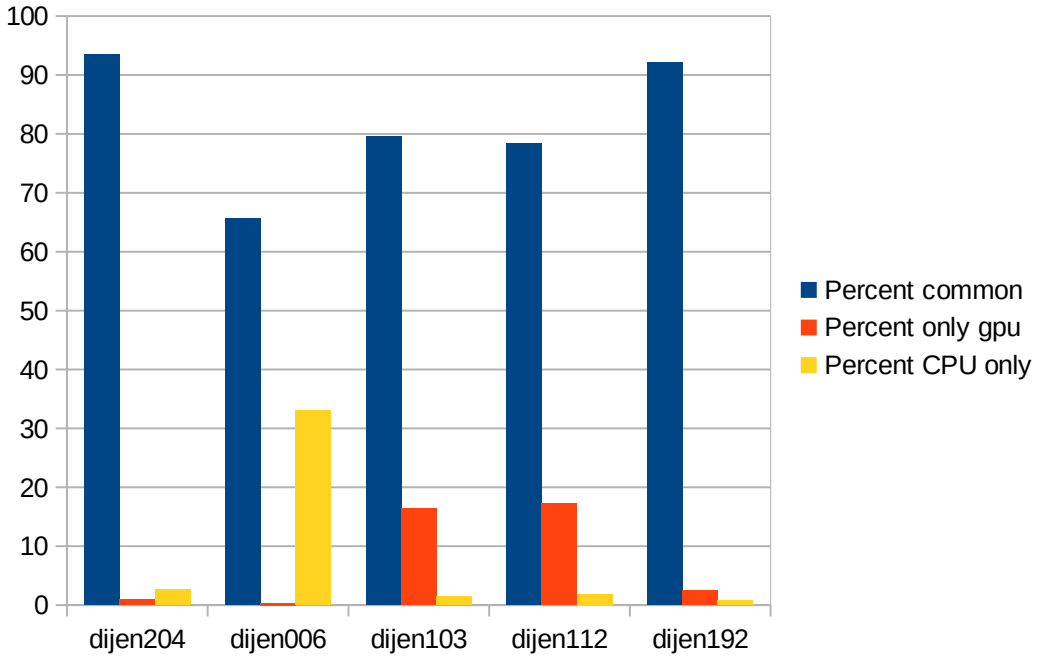
Comparison
4 GPU vs CPU on 16 cores

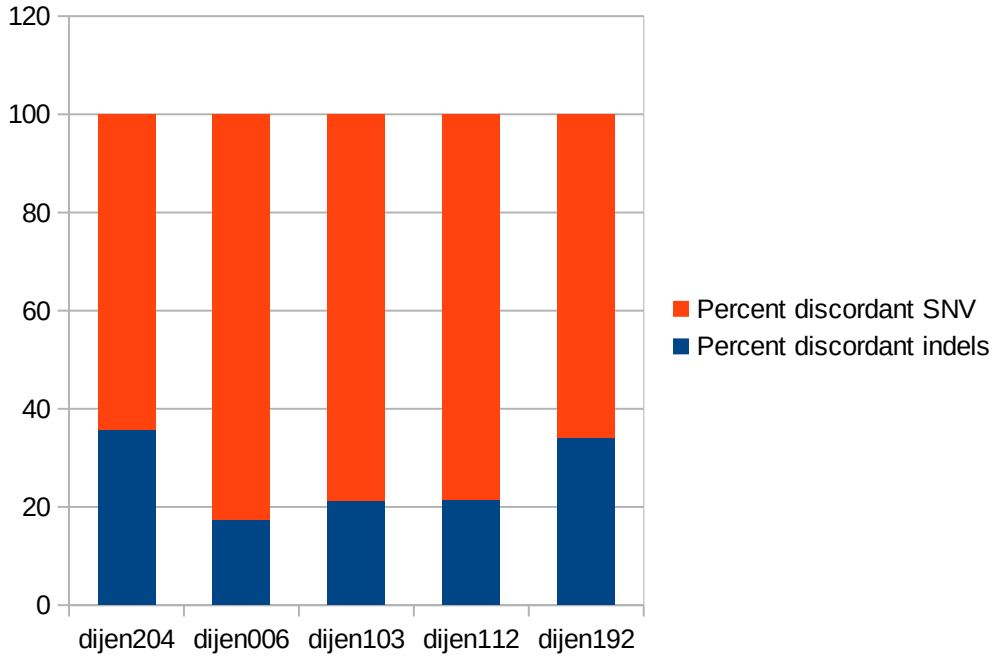




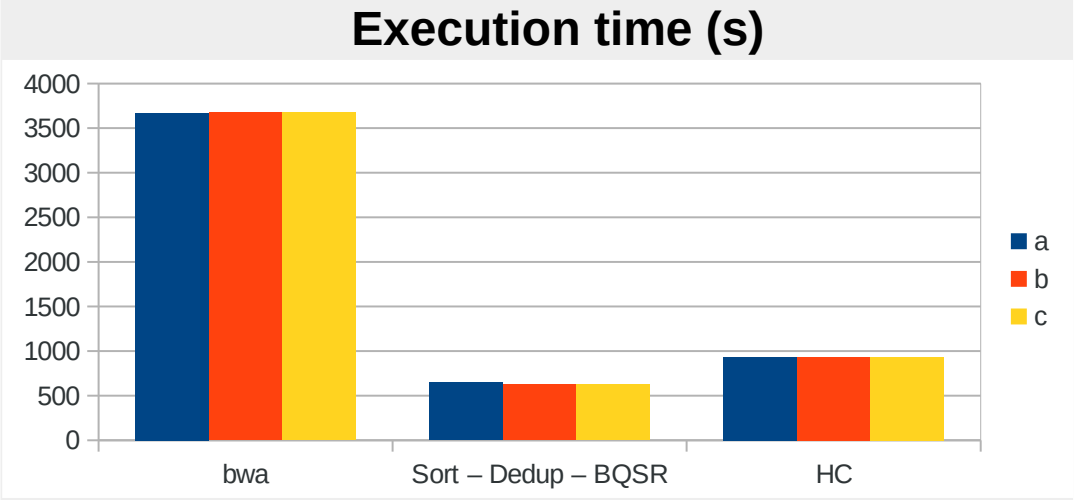
Faster with GPU with a factor of : 13,9 average

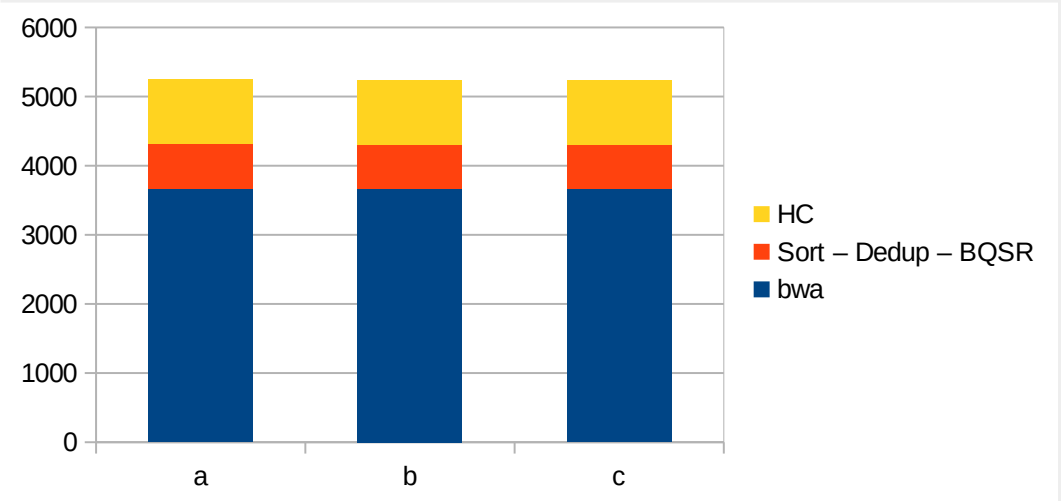
Variation detection Results





Repeatability





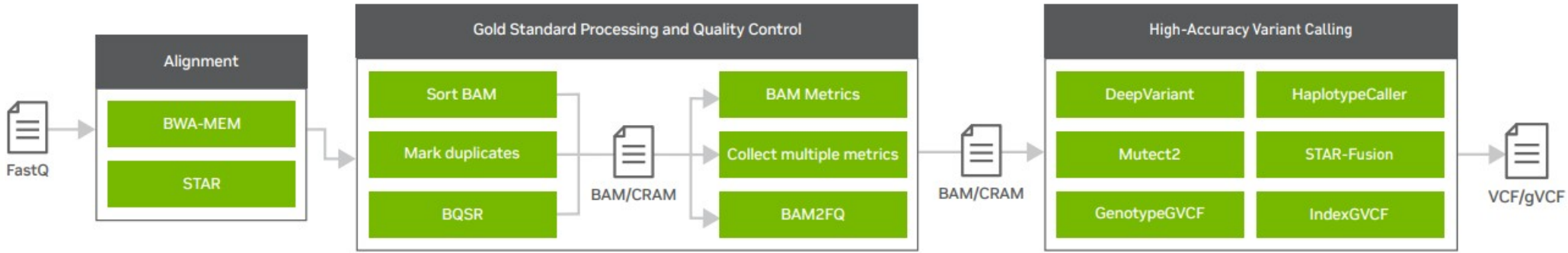
dijen203

Strictly the same results(variations)

Exact same variation list

Exact same bioinformatics parameters

Meanwhile in parabricks (2022)



Performance Comparison

Germline End-To-End Secondary Analysis

Data was generated using publicly available data (https://precision.fda.gov/challenges/truth) for NA12878, deprecating the data to 30X coverage. For the 22-minute runtime, DGX A100 with 320G memory was used. The native GATK4.1 numbers were generated using 32 vCPU (3.1 GHz Intel Xeon® Platinum 8175M) using 28Gb RAM.

Take Home Messages

GPU driven bioinformatics is largely possible

up to 30 times faster (~ 15 times in 2019)
results are very close from cpu driven analysis

Free – open source softwares are mostly deprecated

proprietary code is available under licence

Development is needed

but not really interesting for most of bioinformatics teams …
nor really doable

FPGA exist & is competitive ?

Dragen, supported by the Broad Institute & Illumina

Energetic / Ecologic performances ?