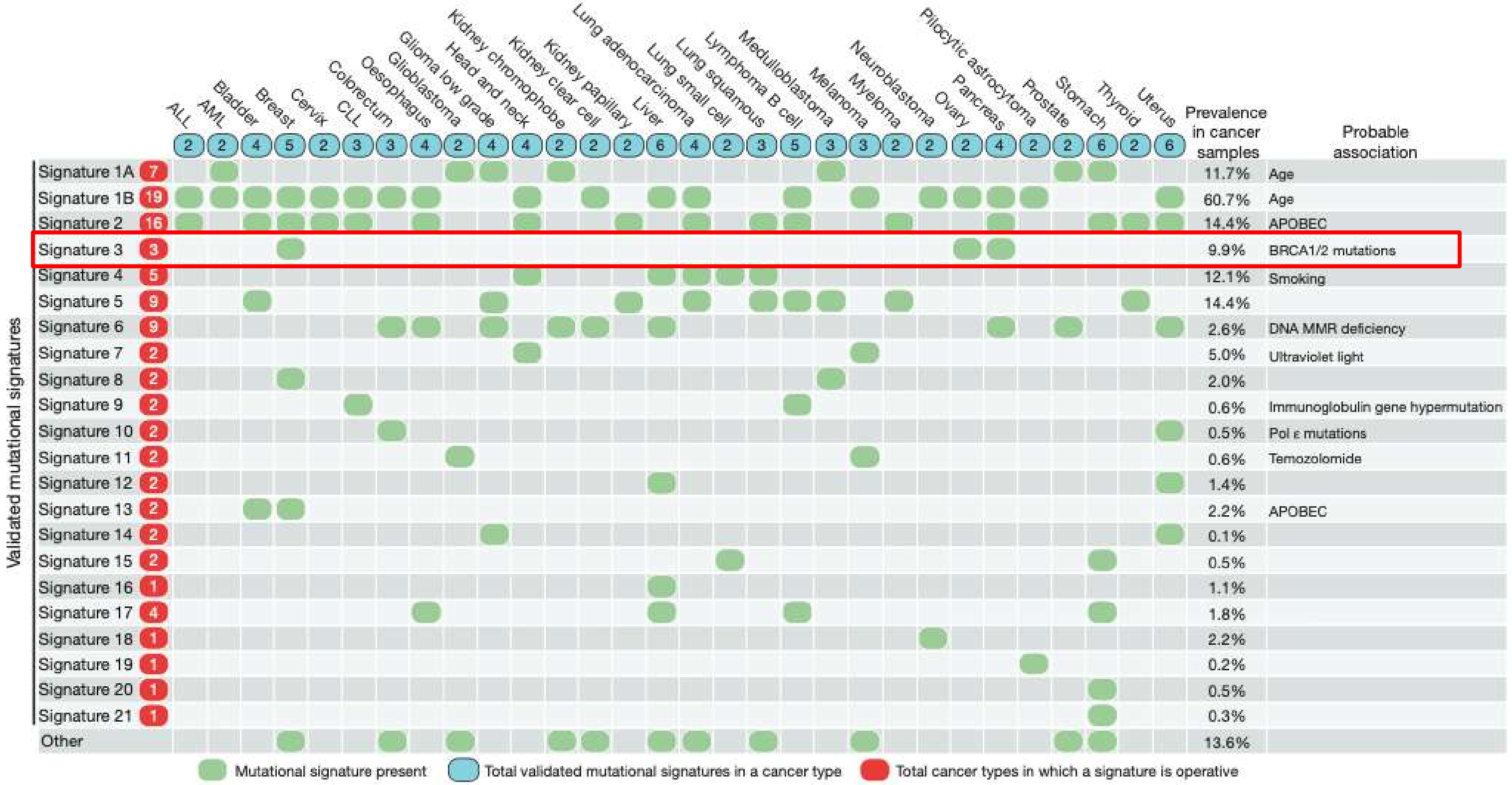
My internship at GIGA research center/ CHU Liège

Investigation of GPU methylation tools in the context of Nanopore sequencing

Olivier Renson

Context

- Genomic medicine initiative: cancer mutational signatures
- Homologous Recombination Deficiency (HRD): study of promotor methylation

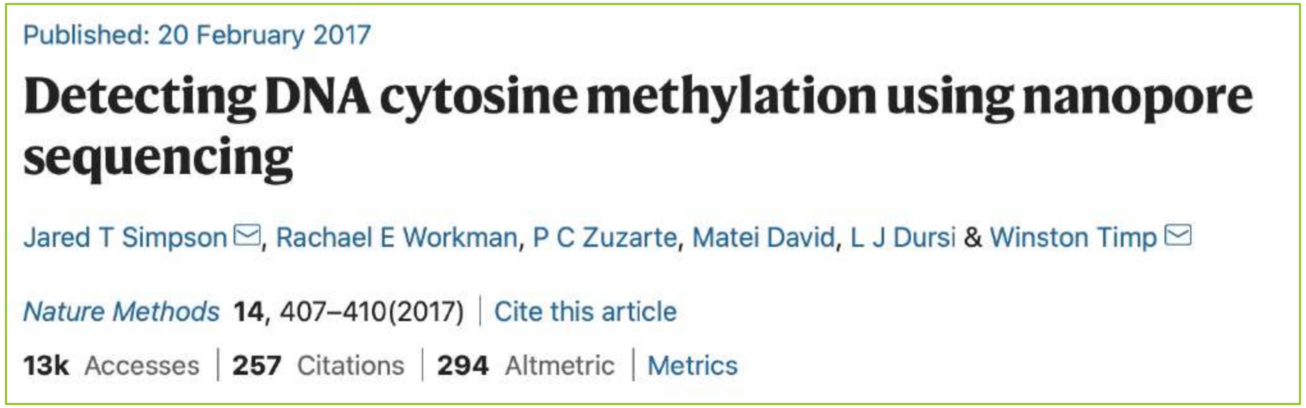


Validated mutational signatures		ALL	AML	Bladder	Breast	Cervix	CLL	Colorectum	Oesophagus	Glioblastoma	Glioma low grade	Head and neck	Kidney chromophobe	Kidney clear cell	Kidney papillary	Liver	Lung adenocarcinoma	Lung small cell	Lung squamous	Lymphoma B cell	Medulloblastoma	Melanoma	Myeloma	Neuroblastoma	Ovary	Pancreas	Pilocytic astrocytoma	Prostate	Stomach	Thyroid	Uterus	Prevalence in cancer samples	Probable association
		2	2	4	5	2	3	3	4	2	4	4	2	2	2	6	4	2	3	5	3	3	2	2	2	4	2	2	6	2	6		
Signature 1A	7		🟩							🟩	🟩		🟩								🟩							🟩	🟩			11.7%	Age
Signature 1B	19	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩			🟩		🟩		🟩	🟩			🟩		🟩		🟩	🟩	🟩	🟩				🟩	60.7%	Age
Signature 2	16	🟩		🟩	🟩	🟩	🟩		🟩			🟩			🟩		🟩		🟩	🟩			🟩			🟩			🟩	🟩	🟩	14.4%	APOBEC
Signature 3	3				🟩																				🟩	🟩						9.9%	BRCA1/2 mutations
Signature 4	5											🟩				🟩	🟩	🟩	🟩													12.1%	Smoking
Signature 5	9			🟩							🟩				🟩		🟩		🟩	🟩	🟩		🟩							🟩		14.4%	
Signature 6	9							🟩	🟩		🟩		🟩	🟩		🟩										🟩		🟩			🟩	2.6%	DNA MMR deficiency
Signature 7	2											🟩										🟩										5.0%	Ultraviolet light
Signature 8	2				🟩																🟩											2.0%	
Signature 9	2						🟩													🟩												0.6%	Immunoglobulin gene hypermutation
Signature 10	2							🟩																							🟩	0.5%	Pol ε mutations
Signature 11	2									🟩												🟩										0.6%	Temozolomide
Signature 12	2															🟩															🟩	1.4%	
Signature 13	2			🟩	🟩																											2.2%	APOBEC
Signature 14	2										🟩																				🟩	0.1%	
Signature 15	2																	🟩											🟩			0.5%	
Signature 16	1															🟩																1.1%	
Signature 17	4								🟩							🟩				🟩									🟩			1.8%	
Signature 18	1																							🟩								2.2%	
Signature 19	1																										🟩					0.2%	
Signature 20	1																												🟩			0.5%	
Signature 21	1																												🟩			0.3%	
Other					🟩			🟩		🟩			🟩	🟩		🟩	🟩		🟩			🟩						🟩	🟩			13.6%	

🟩 Mutational signature present ○ Total validated mutational signatures in a cancer type ● Total cancer types in which a signature is operative

Alexandrov, Ludmil B et al. "Signatures of mutational processes in human cancer." *Nature* vol. 500,7463 (2013)

Methylation detection on Nanopore data



Published: 20 February 2017

Detecting DNA cytosine methylation using nanopore sequencing

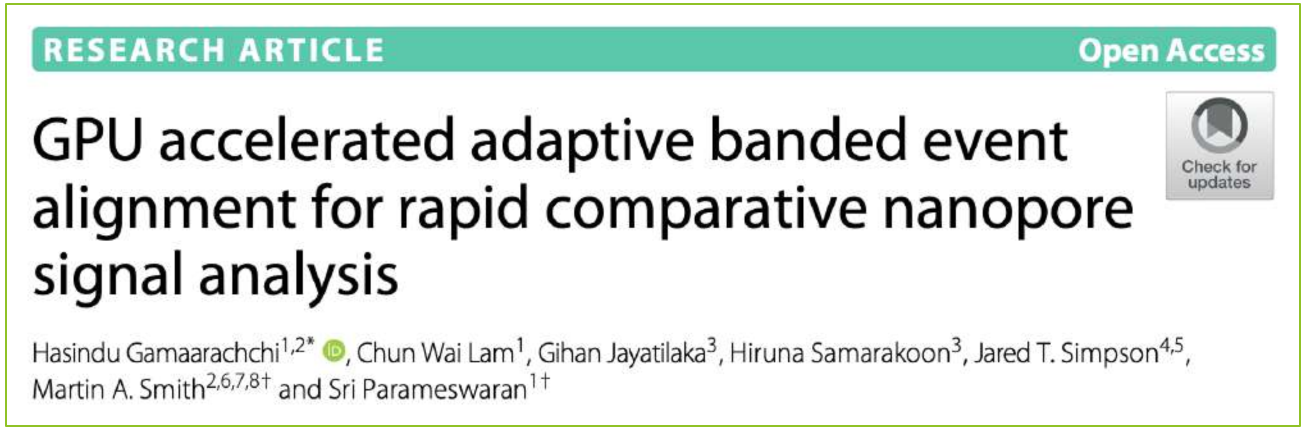
Jared T Simpson ✉, Rachael E Workman, P C Zuzarte, Matei David, L J Dursi & Winston Timp ✉

Nature Methods **14**, 407–410(2017) | Cite this article

13k Accesses | **257** Citations | **294** Altmetric | Metrics

Nanopolish (2017)

- Hidden Markov Model trained on synthetic DNA
- Multi-threading
- No GPU acceleration



RESEARCH ARTICLE **Open Access**

GPU accelerated adaptive banded event alignment for rapid comparative nanopore signal analysis

Hasindu Gamaarachchi[1,2*], Chun Wai Lam[1], Gihan Jayatilaka[3], Hiruna Samarakoon[3], Jared T. Simpson[4,5], Martin A. Smith[2,6,7,8†] and Sri Parameswaran[1†]

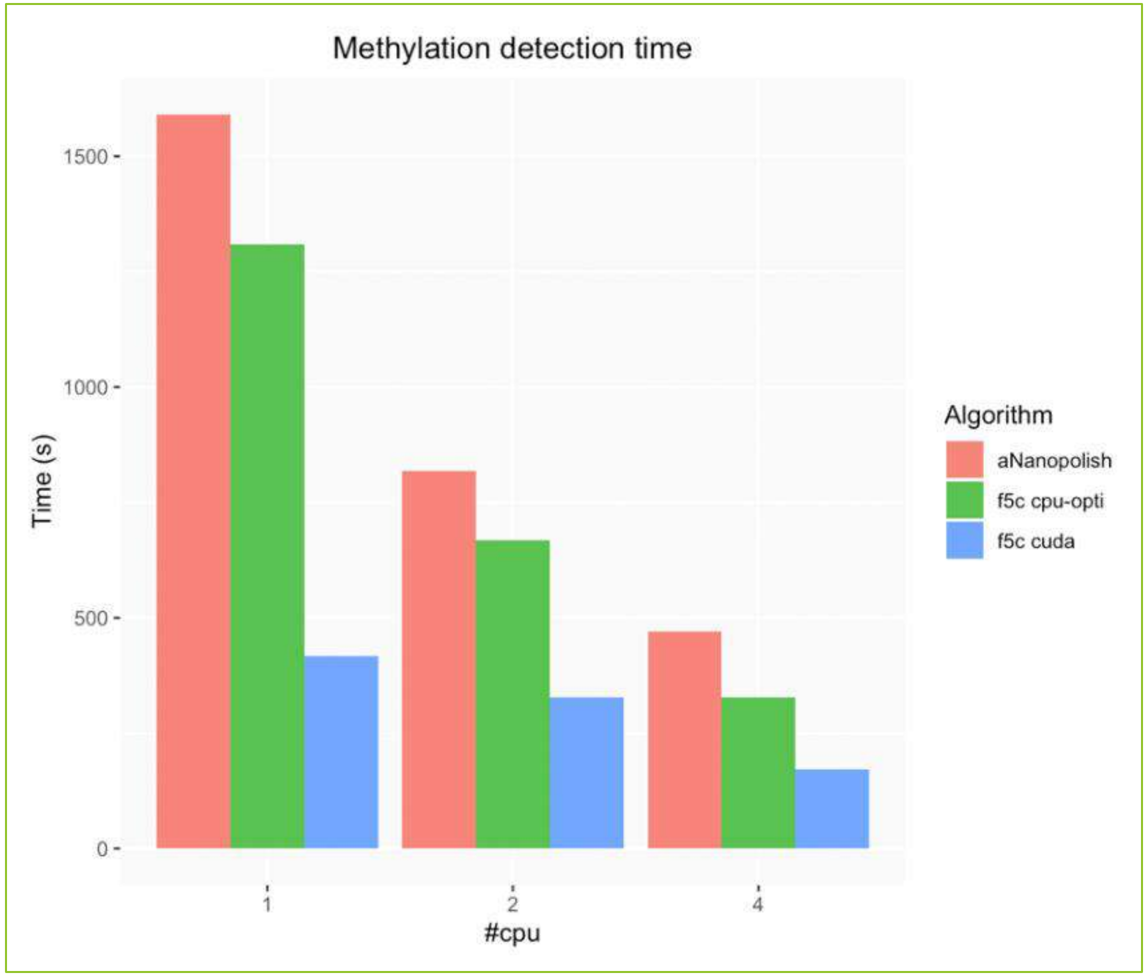
F5c (2020)

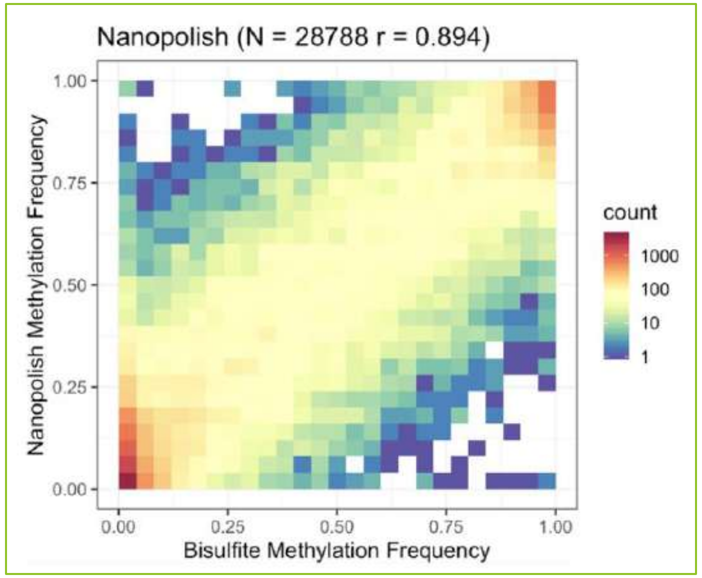
- Adaptative Banded Alignment
- Optimized multi-threading (cpu-opti)
- Support (1) GPU acceleration (cuda)

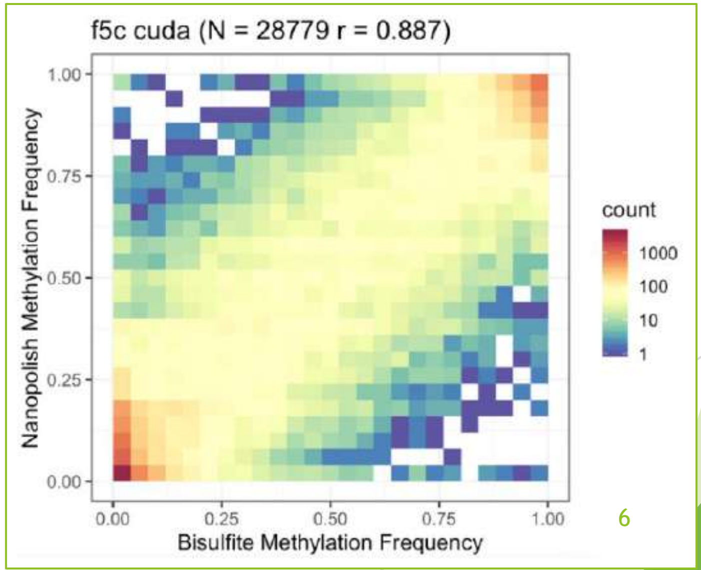
Setup

- Dataset
  - Human cell (NA12878)
  - Region: chr20:5,000,000-10,000,000 (fast5 + fastq + ref)
- Hardware
  - Dragon2 CPU nodes: 16-core Intel Xenon Gold
  - Dragon2 GPU nodes: 12-core Intel Xenon Gold + Tesla V100 (32Go)
- Jobs
  - Time evaluation with: echo Current time $(date+“%T“)
  - Run with 1, 2 and 4 CPUs

Nanopolish vs. f5c

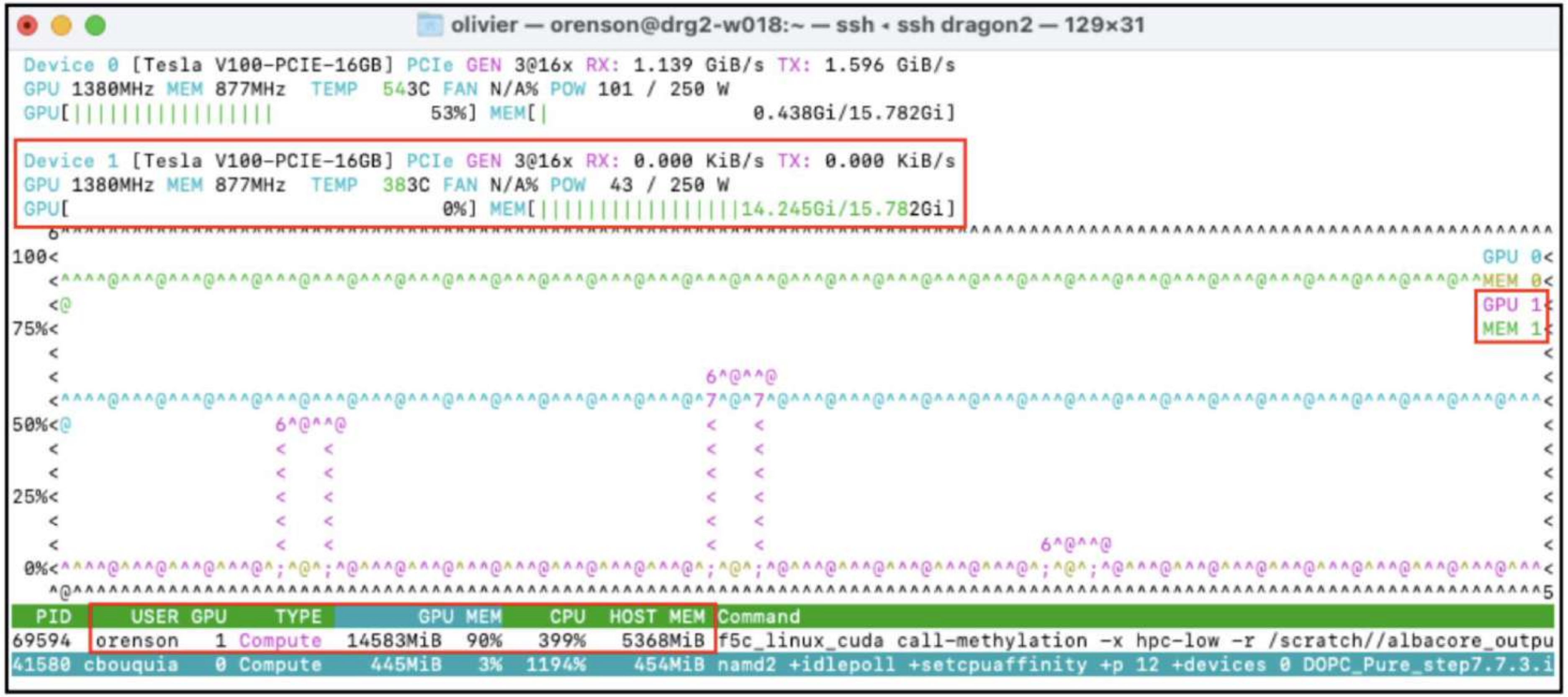






GPU monitoring

- nvtop

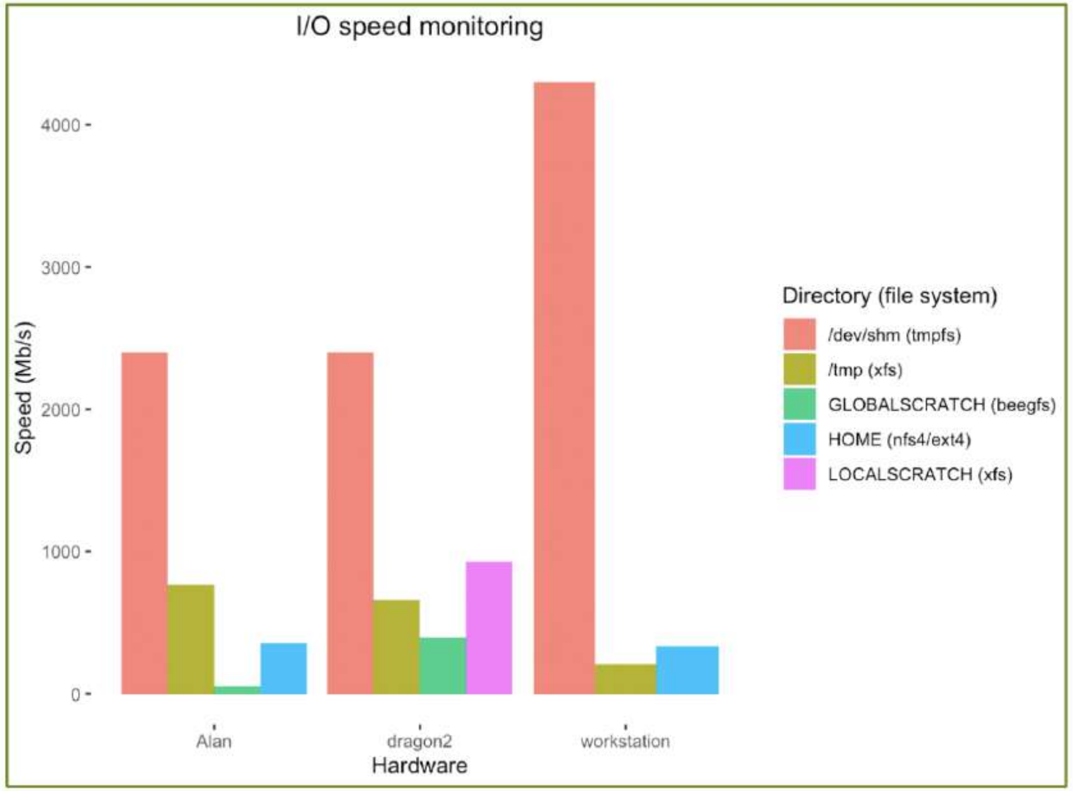


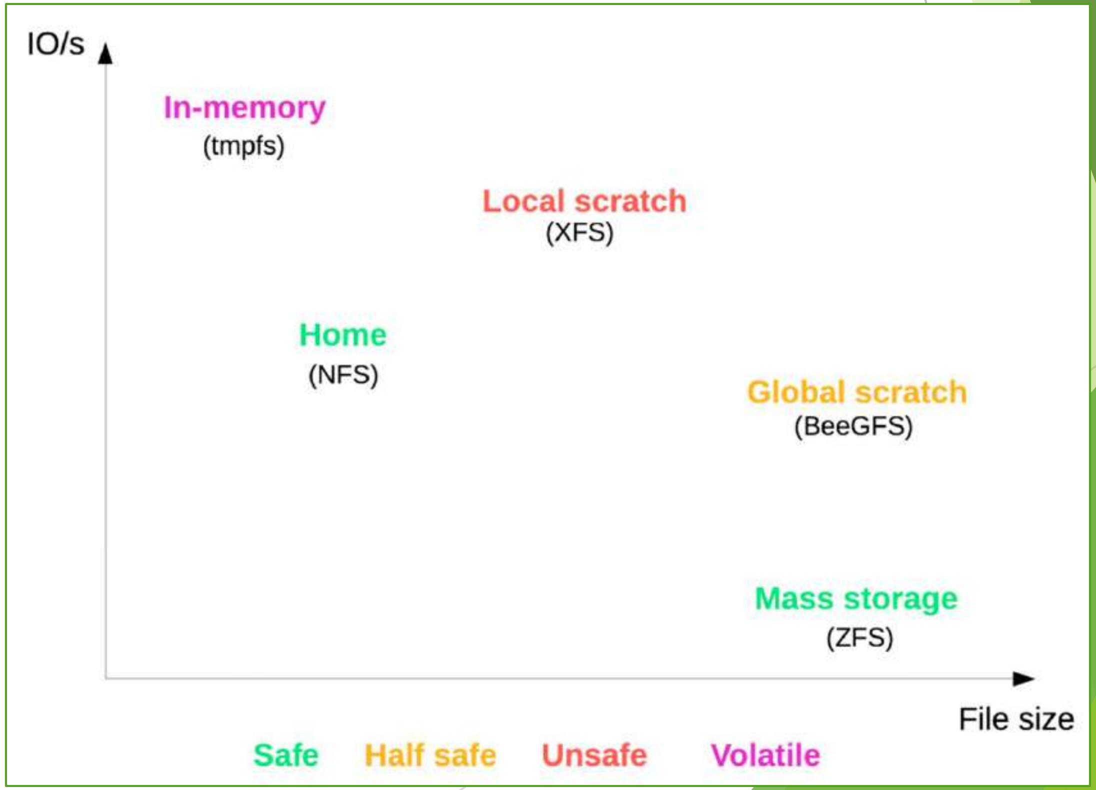
```
Device 0 [Tesla V100-PCIE-16GB] PCIe GEN 3@16x RX: 1.139 GiB/s TX: 1.596 GiB/s
GPU 1380MHz MEM 877MHz  TEMP  543C FAN N/A% POW 101 / 250 W
GPU[|||||||||||||||||                 53%] MEM[|                  0.438Gi/15.782Gi]

Device 1 [Tesla V100-PCIE-16GB] PCIe GEN 3@16x RX: 0.000 KiB/s TX: 0.000 KiB/s
GPU 1380MHz MEM 877MHz  TEMP  383C FAN N/A% POW  43 / 250 W
GPU[                                   0%] MEM[||||||||||||||||||14.245Gi/15.782Gi]
   6^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
100<                                                                                                                       GPU 0<
   <^^^^@^^^@^^^@^^^@^^^@^^^@^^^@^^^@^^^@^^^@^^^@^^^@^^^@^^^@^^^@^^^@^^^@^^^@^^^@^^^@^^^@^^^@^^^@^^^@^^^@^^^@^^^@^^^@^^^@^^MEM 0<
   <@                                                                                                                      GPU 1<
75%<                                                                                                                       MEM 1<
   <                                                                                                                            <
   <                                                           6^@^^@                                                           <
   <^^^^@^^^@^^^@^^^@^^^@^^^@^^^@^^^@^^^@^^^@^^^@^^^@^^^@^^^@^7^@^7^@^^^@^^^@^^^@^^^@^^^@^^^@^^^@^^^@^^^@^^^@^^^@^^^@^^^@^^^@^^^<
50%<@                          6^@^^@                          <   <                                                            <
   <                            <   <                          <   <                                                            <
   <                            <   <                          <   <                                                            <
25%<                            <   <                          <   <                                                            <
   <                            <   <                          <   <                                                            <
   <                            <   <                          <   <                         6^@^^@                             <
 0%<^^^^@^^^@^^^@^^^@^;^@^;^@^^^@^^^@^^^@^^^@^^^@^^^@^^^@^;^@^;^@^^^@^^^@^^^@^^^@^^^@^;^@^;^@^^^@^^^@^^^@^^^@^^^@^^^@^^^@^^^<
   ^@^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^5
  PID      USER GPU      TYPE        GPU MEM      CPU  HOST MEM Command
69594 orenson     1 Compute   14583MiB  90%    399%   5368MiB f5c_linux_cuda call-methylation -x hpc-low -r /scratch//albacore_outpu
41580 cbouquia    0 Compute     445MiB   3%   1194%    454MiB namd2 +idlepoll +setcpuaffinity +p 12 +devices 0 DOPC_Pure_step7.7.3.i
```

Monitoring file systems

- Monitoring new working directories I/O speed
  - df -Th $WORKINGDIR
  - dd if=/dev/zero of=$WORKINGDIR/test1 bs=1M count=2048 oflag=dsync




IO/s

In-memory (tmpfs)

Local scratch (XFS)

Home (NFS)

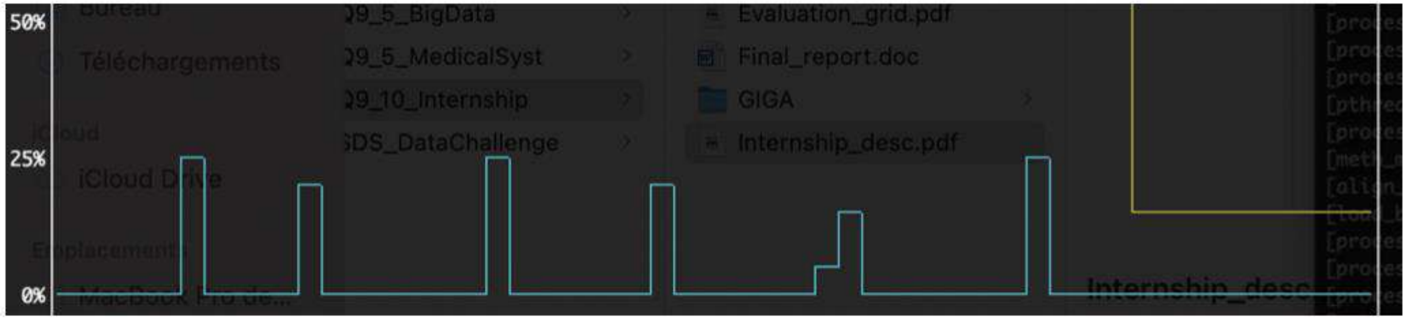
Global scratch (BeeGFS)

Mass storage (ZFS)

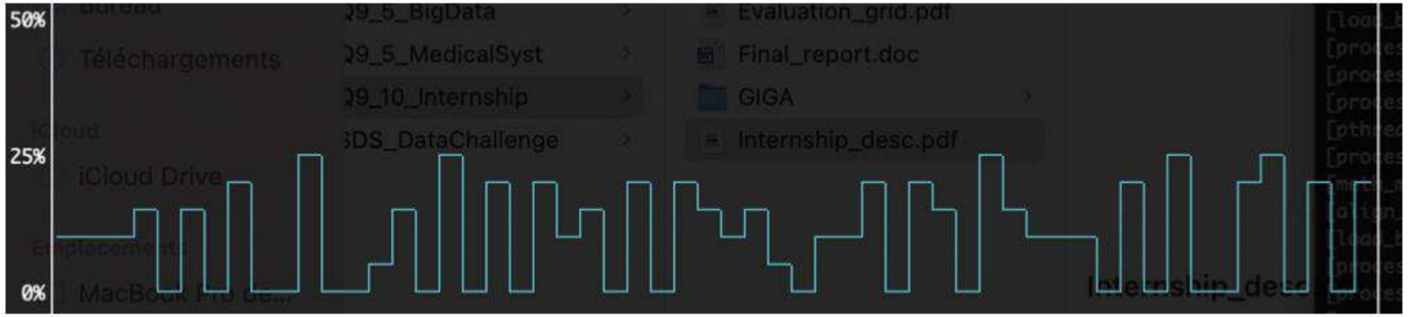
File size

Safe Half safe Unsafe Volatile

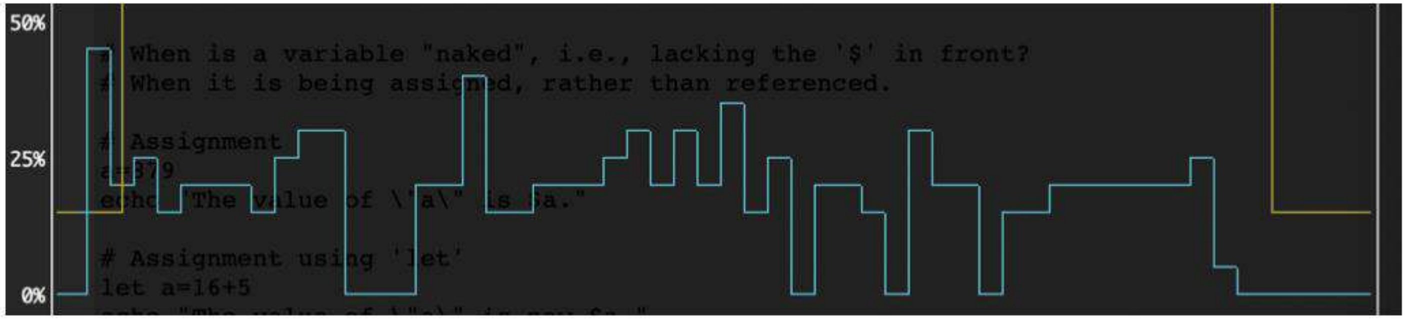
More CPU = more GPU activity



1 CPU



4 CPU



```
# When is a variable "naked", i.e., lacking the '$' in front?
# When it is being assigned, rather than referenced.

# Assignment
a=879
echo "The value of \"a\" is $a."

# Assignment using 'let'
let a=16+5
```

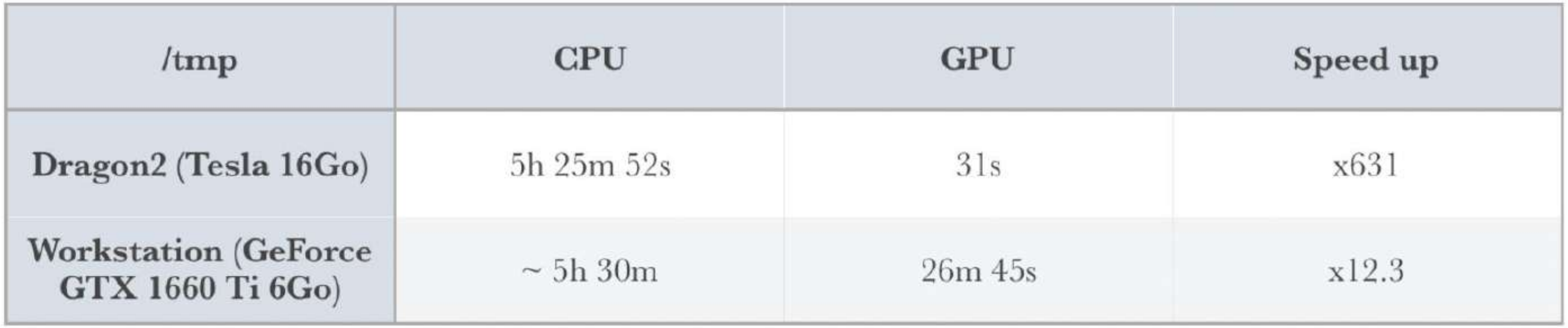
20 CPU

Conclusions on f5c

- f5c = **hybrid** software
- Balanced setup (workstation) >>> high end GPU (Alan)
- ! I/O bottleneck !
- CPU vs. GPU speed up : ~4.5x (Nanopolish 7m 50s → f5c, 1m 43s)
- f5c monitoring speed up : ~2x (4 CPUs, HOME, 1m 43s → 20 CPUs, /dev/shm, 52s)
- Multi-fast5 not worth it for a small dataset

Nanopore basecalling : Guppy

- Small dataset (512 Mo)



/tmp	CPU	GPU	Speed up
Dragon2 (Tesla 16Go)	5h 25m 52s	31s	x631
Workstation (GeForce GTX 1660 Ti 6Go)	~ 5h 30m	26m 45s	x12.3

Table 1 : execution time for each hardware setup

- F5c = hybrid CPU/GPU -> Workstation (balanced setup) wins
- Guppy = GPU only -> Dragon2 (big GPU) wins

MCF7 whole genome sequencing data

- Objectives
  - Test f5c on a new dataset
  - Compare methylation frequencies with bisulfite sequencing

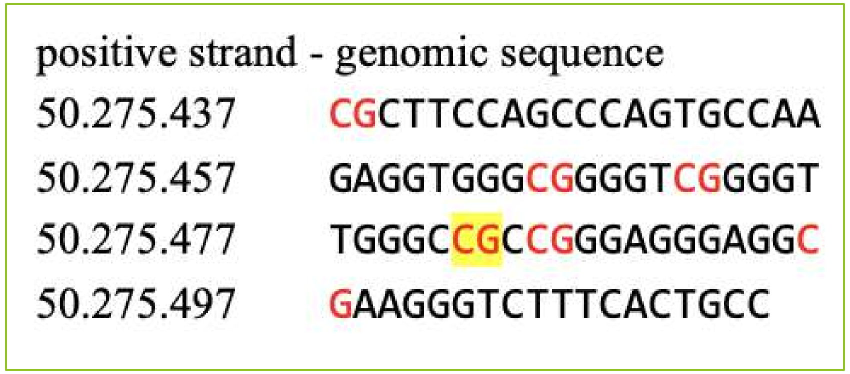
- Data (~ 1.5 To)
  - Fast5 directory : raw nanopore signal - 1023 Go - 9.26M reads
  - Fasta file : referance genome - 3 Go
  - Fastq file : basecalled reads - 136 Go – 8.59M reads – avg length 8369
  - Sam/bam files : alignment data- 158 Go/79Go – 11.46M alignments

Regions of interest

Genes	Chr	Start	End	Length	CpG	Function(s)
SEMA3B	3	50275437	50275514	77	6	Inhibits axonal extension, tumor suppressor by inducing apoptosis
RASSF1	3	50340943	50341036	93	6	Tumor suppressor, negatively regulate cell cycle at G1/S-phase
KLHL6	3	183555418	183555536	118	5	B-lymphocyte antigen receptor signalisation
INA	10	103276779	103276913	135	8	Type IV intermediate filament heteropolymers
PTPRCAP	11	67437695	67437765	70	2	T- and B-lymphocyte activation (transmembrane phosphoprotein)

- Split in single chromosomes for transfer on dragon2
  - CECI quotas : /CECI/home 100 Go, /CECI/trsf 1 To (10 days)
  - 1.5 To -> 125 Go, 72 Go, 55 Go

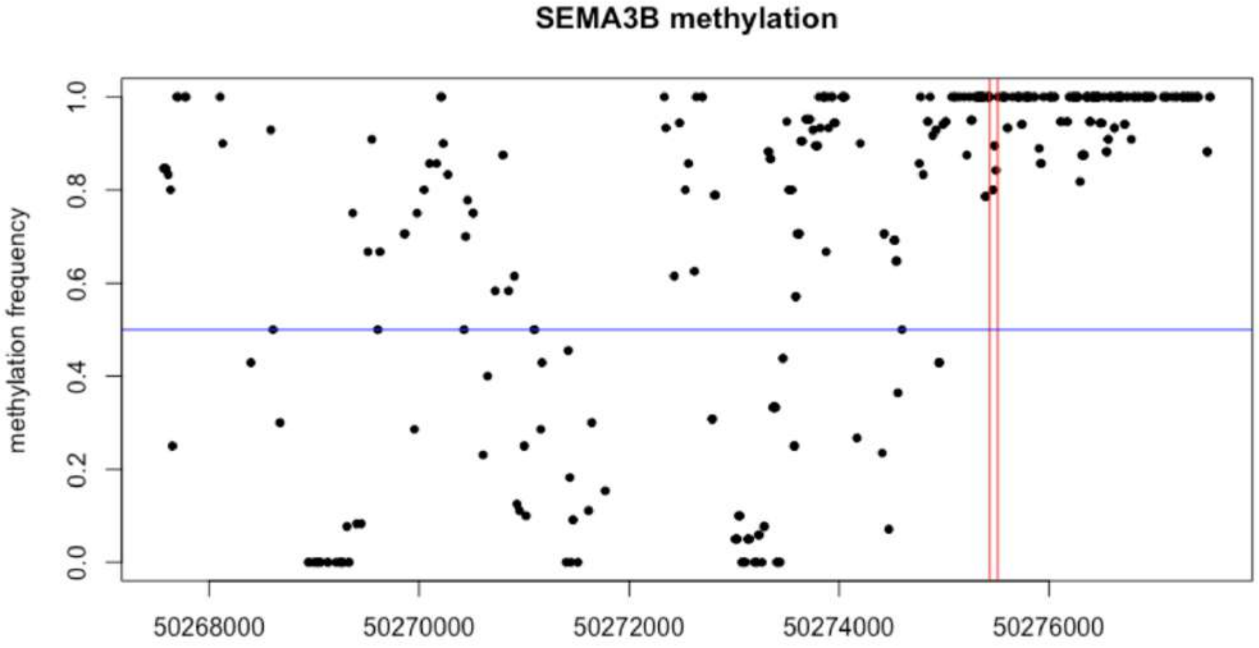
SEMA3B



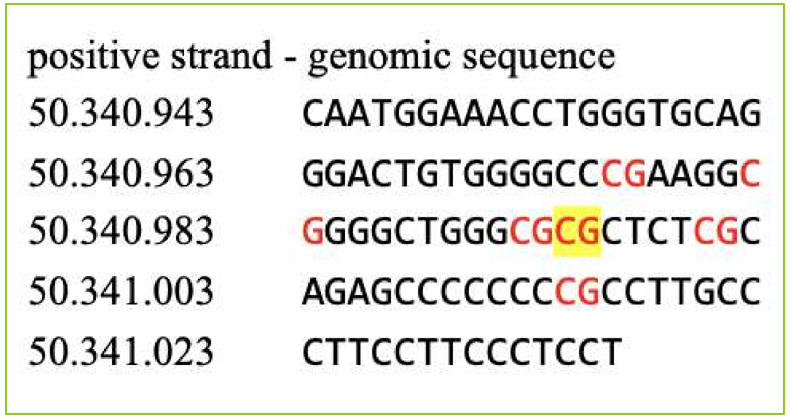
positive strand - genomic sequence

```
50.275.437      CGCTTCCAGCCCAGTGCCAA
50.275.457      GAGGTGGGCGGGGTCGGGGT
50.275.477      TGGGCCGCCGGGAGGGAGGC
50.275.497      GAAGGGTCTTTCACTGCC
```

CpG pos	bisulfites	nanopore
50,275,437	0.91	1.00
50,275,465	0.78	0.80
50,275,471	0.85	0.80
50,275,482	?	0.90
50,275,485	?	0.90
50,275,496	?	0.84



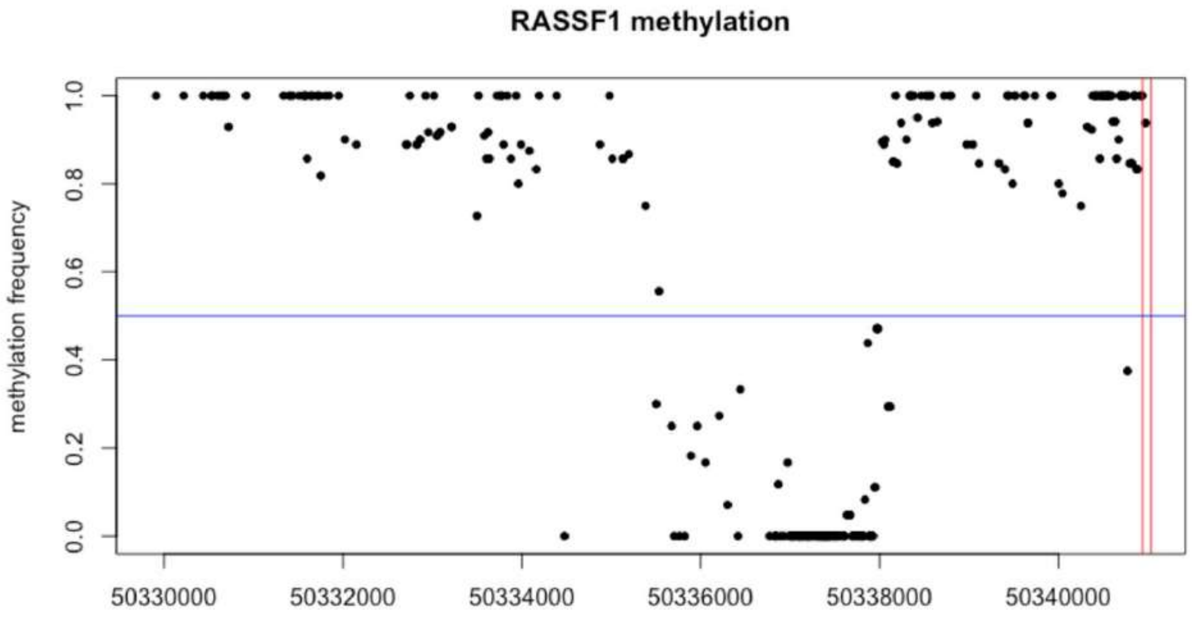
RASSF1



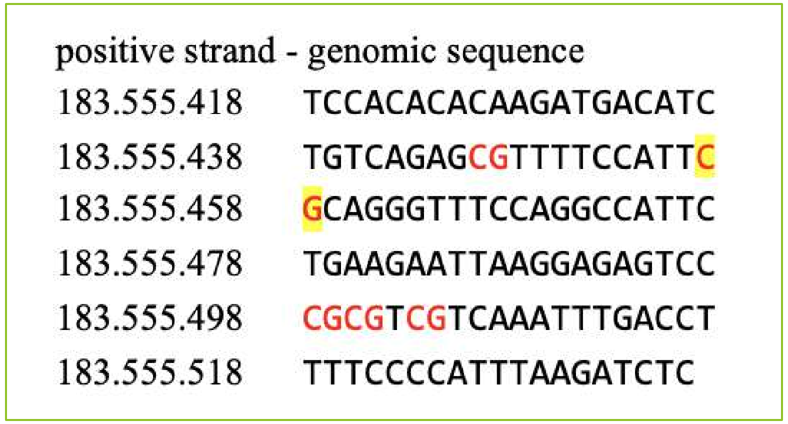
positive strand - genomic sequence

50.340.943	CAATGGAAACCTGGGTGCAG
50.340.963	GGACTGTGGGGCCCGAAGGC
50.340.983	GGGGCTGGGCGCGCTCTCGC
50.341.003	AGAGCCCCCCCGCCTTGCC
50.341.023	CTTCCTTCCCTCCT

CpG pos	bisulfites	nanopore
50,340,976	0.62	0.94
50,340,982	0.63	0.94
50,340,992	0.62	0.94
50,340,994	0.60	0.94
50,341,000	0.61	0.94
50,341,014	0.53	0.83



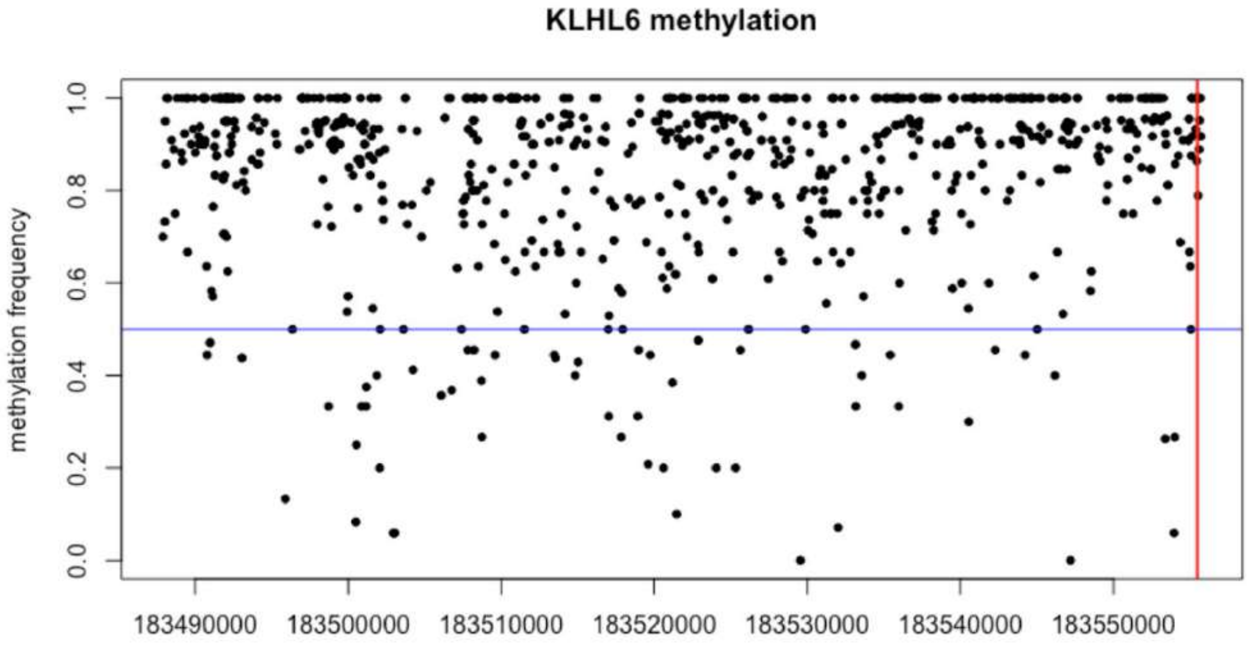
KLHL6



positive strand - genomic sequence

183.555.418	TCCACACACAAGATGACATC
183.555.438	TGTCAGAGCGTTTTCCATTC
183.555.458	GCAGGGTTTCCAGGCCATTC
183.555.478	TGAAGAATTAAGGAGAGTCC
183.555.498	CGCGTCGTCAAATTTGACCT
183.555.518	TTTCCCCATTTAAGATCTC

CpG pos	bisulfites	nanopore
183,555,446	0.85	0.86
183,555,456	0.92	1.00
183,555,498	0.94	1.00
183,555,600	0.95	1.00
183,555,603	0.93	1.00

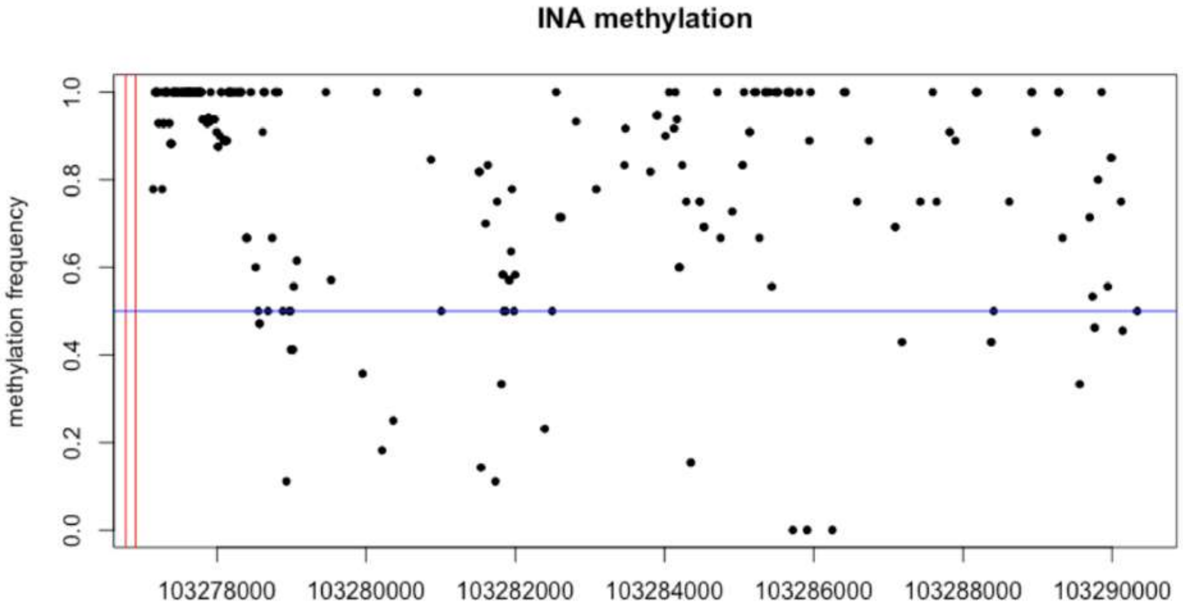


INA

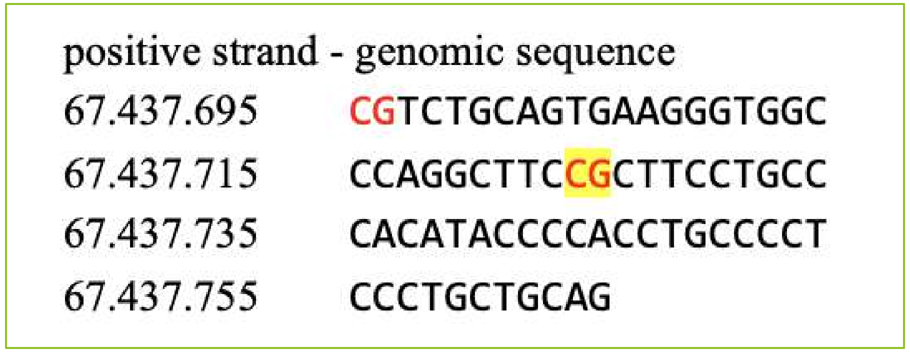
positive strand - genomic sequence

103.276.779	CAGAAACCCCTAACCTCCCA
103.276.799	GTCGGTTAAAGAAGAGGGGA
103.276.819	TAGGGTCAAGGGATGCGACA
103.276.839	GAGCTGTGTGGTTTCCGGAT
103.276.859	GGGAAACCTCAGTCGTTTAG
103.276.879	GCACCCCTCCGCTCGAGTCA
103.276.899	CTTCCGAAGCAGTCG

CpG pos	bisulfites	nanopore
103,276,801	0.89	0.82
103,276,834	0.94	0.89
103,276,854	0.93	1.00
103,276,872	0.79	1.00
103,276,888	0.96	1.00
103,276,892	?	1.00
103,276,903	?	1.00
103,276,912	?	1.00



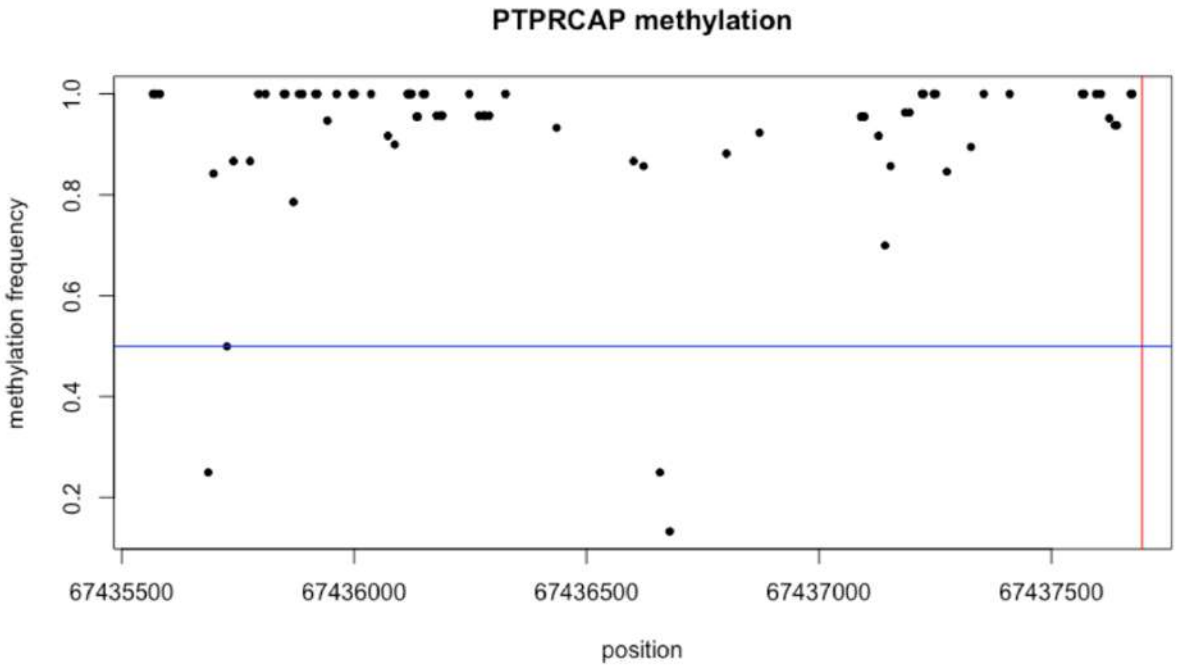
PTPRCAP



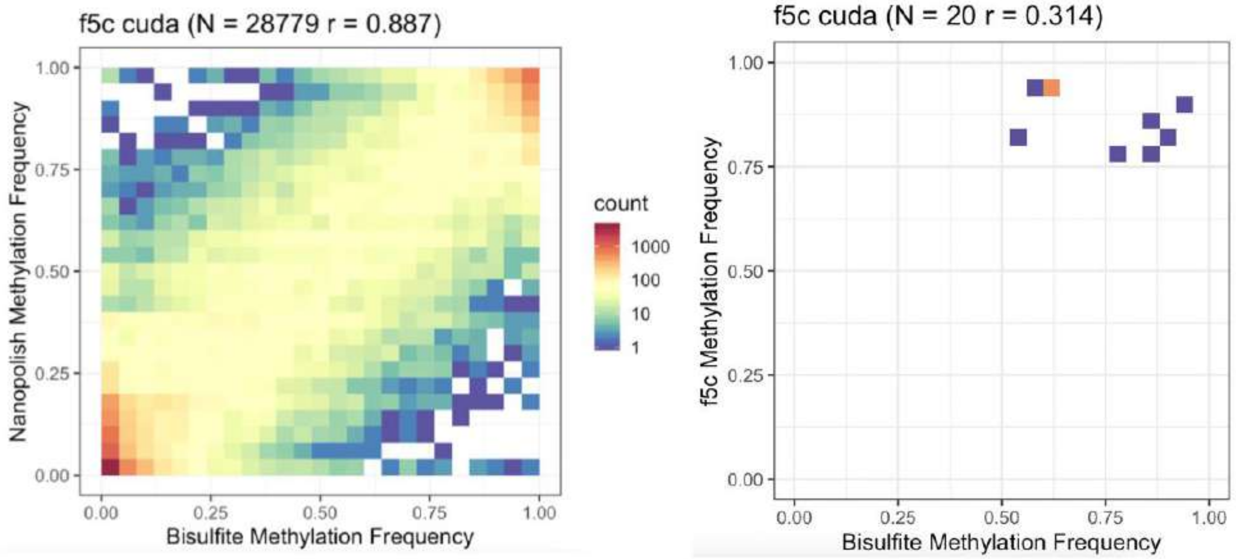
positive strand - genomic sequence

```
67.437.695      CGTCTGCAGTGAAGGGTGGC
67.437.715      CCAGGCTTCCGCTTCCTGCC
67.437.735      CACATACCCCACCTGCCCCT
67.437.755      CCCTGCTGCAG
```

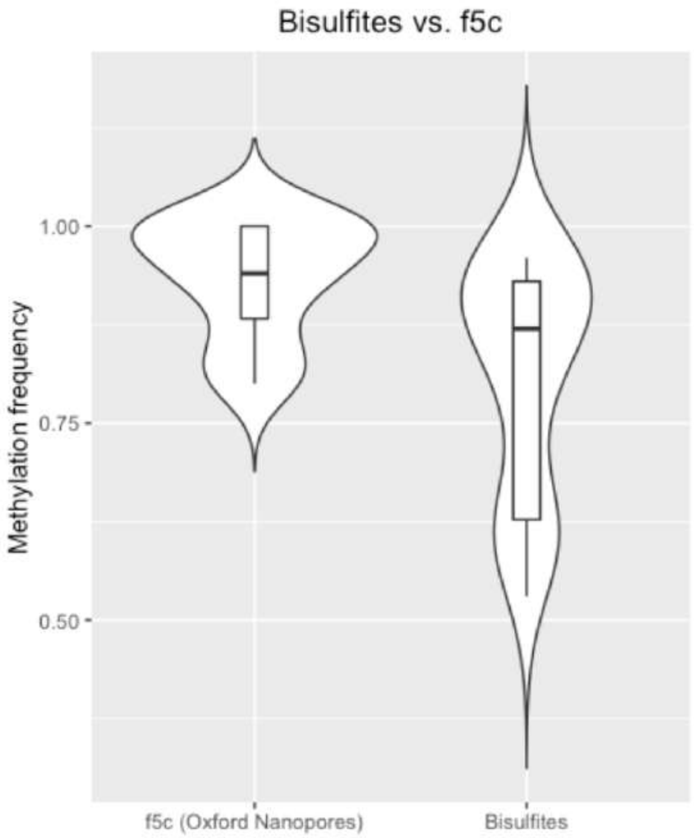
CpG pos	bisulfites	nanopore
67,437,695	0.89	1.00
67,437,724	?	0.92



Comparison with bisulfite sequencing



Img10: correlation between f5c output and bisulfite sequencing with the dataset provided by Oxford Nanopore (left) and with mcf7 data (right).



Thank you for your attention !