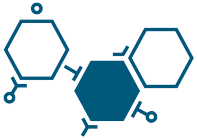
MS in Computer Science

Gemma Martini



FPLGED

FAST PACKING OF LARGE GENOMIC DATA

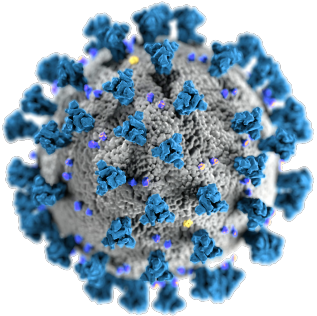
First supervisor: Filippo Geraci PhD

Second supervisor: Veronica Guerrini PhD

25 febbraio 2022

Data deluge

In January 2020 SARS-CoV-2 spreads in Wuhan (China)



Data deluge

Data deluge

30kB size of SARS-CoV-2 genome

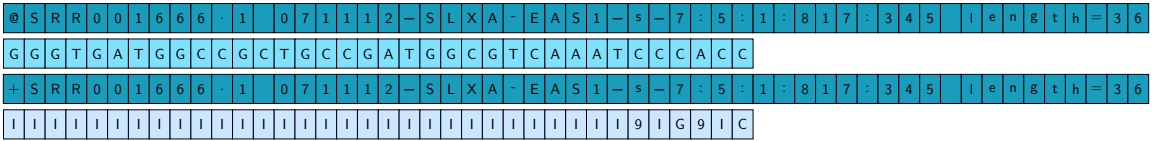
8'440'604 SARS-CoV-2 submissions (2020-2021)

241GB total size of SARS-CoV-2 samples GISAID[i]

[i] GISAID: the largest public repository for genomic research on Covid-19

File format

FASTQ



Quality score

The *quality score* is an output of the sequencing process: it has the same size of the nucleotide sequence, where each character (from ASCII 33 to 126) represents the *Phred* quality score.

Definition (Phred quality score)

We term *Phred quality score* and denote Q the logarithm of the error probability (P) of calling a certain nucleotide.

$$Q = -10 \cdot log_{10}(P)$$

FPLGeD structure

Header

@	EAS139:76:FC706VJ:6:1101:20273:23294 1:N:0:CGATGT

@	EAS139:89:FC706VJ:7:1101:7912:1314 2:N:0:GTGAAAGT

Sequence

DNA $\rightarrow$ CGATGTttttgcgcatcgATCGTAGAGAtTtTACGGCAGTGTATGA

RNA $\rightarrow$ CGAUGUuuuugcgcaucgAUCGUAGAGAuUuUACGGCAGUGUAUGA

Quality

!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCFCCCCCCC65

Theoretical results

Encoders	T. complexity	BC Compr. Ratio	WC Compr. Ratio
`Header`	O(n)	$8\times$	$1\times$
`Sequence`	O(n)	$49932\times$	$1.14\times$
`Quality_000`	O(n)	$8\times$	$1.1\times$
`Quality_001`	O(n)	$15\times$	$1.8\times$
`Quality_010`	O(n)	$8\times$	$2.7\times$
`Quality_011`	O(n)	$+\infty\times$	$2.5\times$
`Quality_100`	$O(L \cdot n)$[ii]	$1.25\times$	$1\times$

[ii] where L is the maximum possible number of characters that fit a 64-bits word

Competitors

SPRING: a next-generation compressor for FASTQ data

Shubham Chandak
Stanford University
ISMB/ECCB 2019

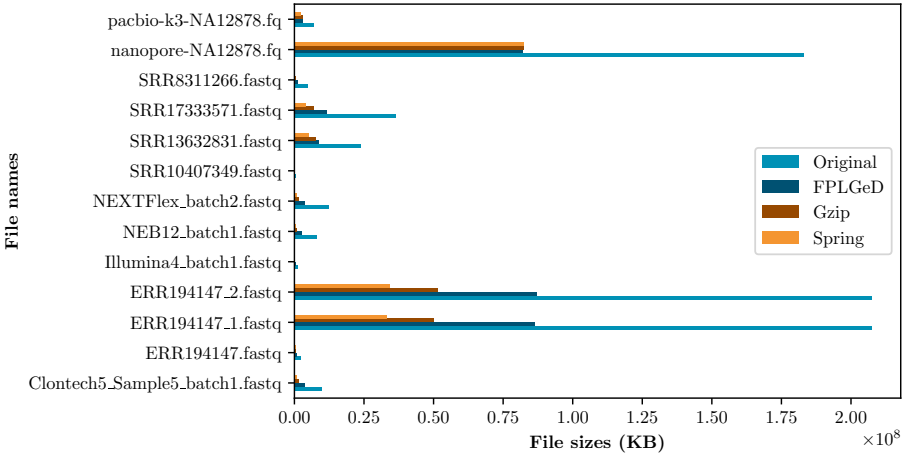
FPLGED

FAST PACKING OF LARGE GENOMIC DATA

Dataset

- 4 samples of human brain tissue, prepared with 4 different protocols and sequenced with an Illumina HiSeq-3000
- 3 sequencings of the genome of a specific individual belonging to the CEPH/UTAH PEDIGREE, obtained through 3 different technologies
- RNA sequencing from human, brook trout, and Chinese ginseng
- Reference genomes of homo sapiens, zebrafish, and house mouse

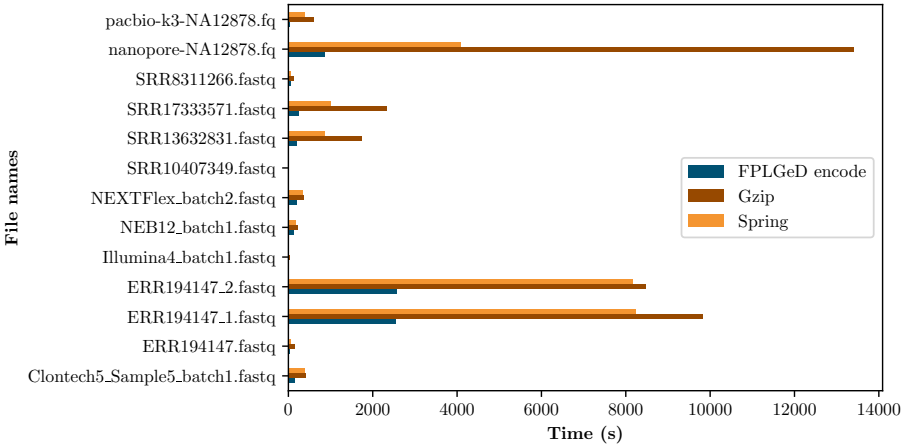
Compression ratio



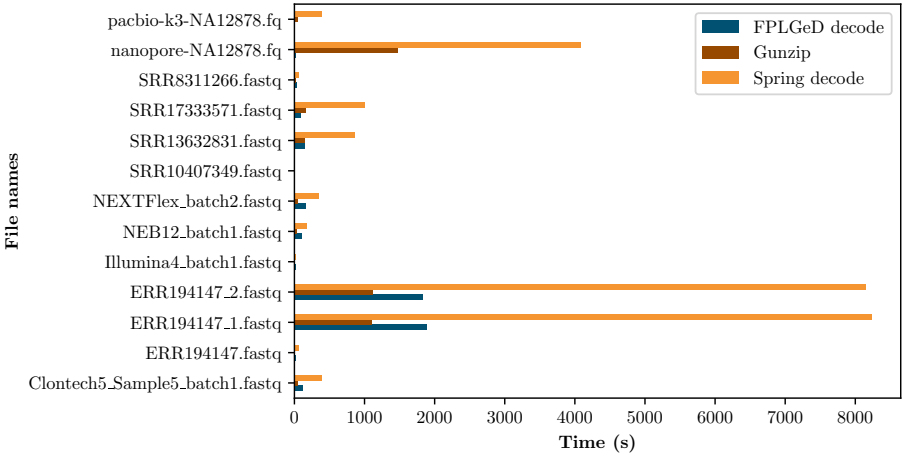
Compression ratio

O. Nanopore's MinION

Encode time


pacbio-k3-NA12878.fq
nanopore-NA12878.fq
SRR8311266.fastq
SRR17333571.fastq
SRR13632831.fastq
SRR10407349.fastq
NEXTFlex_batch2.fastq
NEB12_batch1.fastq
Illumina4_batch1.fastq
ERR194147_2.fastq
ERR194147_1.fastq
ERR194147.fastq
Clontech5_Sample5_batch1.fastq
File names
0
2000
4000
6000
8000
10000
12000
14000
Time (s)
FPLGeD encode
Gzip
Spring

Decode time


FPLGeD decode
Gunzip
Spring decode
pacbio-k3-NA12878.fq
nanopore-NA12878.fq
SRR8311266.fastq
SRR17333571.fastq
SRR13632831.fastq
SRR10407349.fastq
NEXTFlex_batch2.fastq
NEB12_batch1.fastq
Illumina4_batch1.fastq
ERR194147_2.fastq
ERR194147_1.fastq
ERR194147.fastq
Clontech5_Sample5_batch1.fastq
File names
0
1000
2000
3000
4000
5000
6000
7000
8000
Time (s)

Conclusions

We addressed the problem of reducing the size of files containing genomic information, specifically FASTA and FASTQ file types.

We compared FPLGeD with other algorithms frequently used to compress FASTA(Q) files, we found that the algorithm confirms in practice the characteristics highlighted by a theoretical analysis, offering a good reduction in terms of space and taking very little time to be executed.

We showed that FPLGeD outperforms its competitors on third generation sequencing data that is not only the most widely used nowadays, but it also represents the direction towards which all the modern sequencers are focusing on.

“Homo sum, humani nihil a me alienum puto”

P. Terenzio Afro

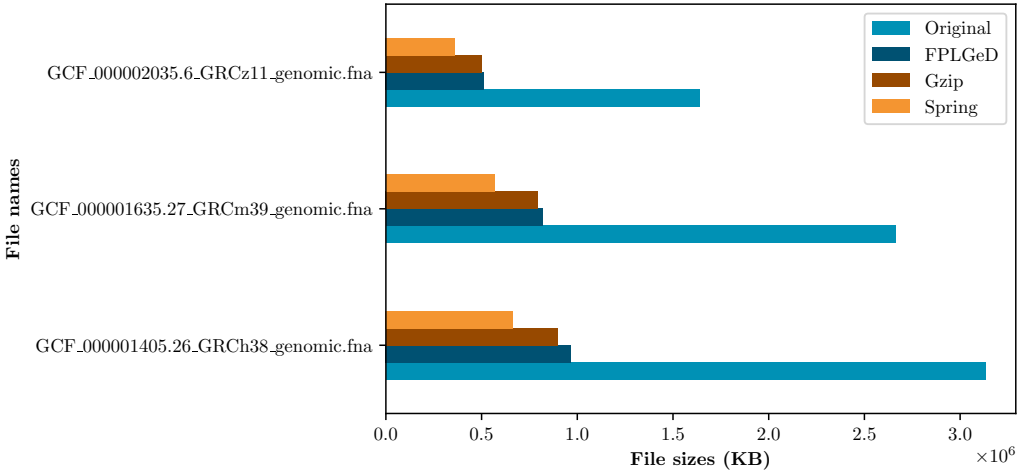
FASTQ Datasets

Sample	Name	Size(KB)	Type	Method	Instrument	N_{reads}	R_{length}
Human	np-NA12878	183 238 556	DNA	ONT	MinION	83 237 108	/
	pb-k3-NA12878	7 165 020	DNA	PacBio	SMRT	2 618 188	/
	ERR194147_1	207 494 172	DNA	Illumina	HiSeq2000	3 149 060 436	101
	ERR194147_2	207 494 172	DNA	Illumina	HiSeq2000	3 149 060 436	101
	ERR194147	2 173 144	DNA	Illumina	HiSeq2000	32 963 184	101
H. brain	Ct5_S5_b1	9 872 180	DNA	Clontech	HiSeq3000	238 508 524	51
	Ill4_batch1	1 225 876	DNA	Illumina	HiSeq3000	31 672 308	51
	NEB12_batch1	7 930 268	DNA	NEB	HiSeq3000	204 889 280	51
	NEXTF_batch2	12 349 932	DNA	NEXTFlex	HiSeq3000	319 073 916	51
Human	SRR10407349	480 116	RNA	Illumina	MiSeq	9 449 128	65
Human	SRR8311266	4 945 096	RNA	Illumina	HiSeq2500	81063812	51
B. trout	SRR13632831	23 965 044	RNA	Illumina	HiSeq2500	203 675 076	202
Ginseng	SRR17333571	36 457 676	RNA	Illumina	NovaSeq6000	216 678 820	302

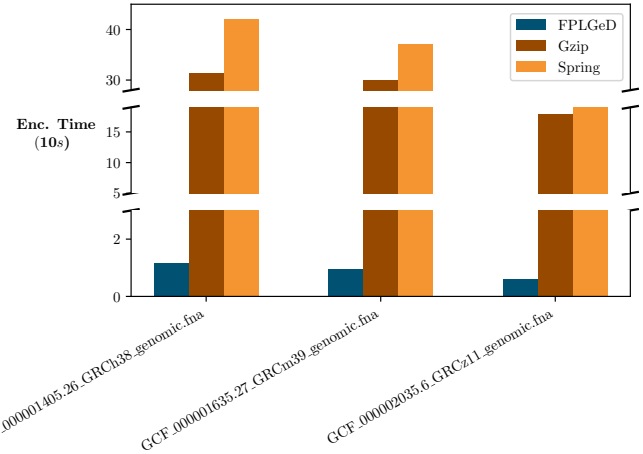
FASTA Datasets

Sample	Name	Size(KB)	Total seq. length	$N_{chromosomes}$
Human	`GCF_000001405.26_GRCh38_genomic`	3 134 124	3 099 734 149	24
Zebrafish	`GCF_000001635.27_GRCm39_genomic`	2 664 292	1 373 454 788	25
Mouse	`GCF_000002035.6_GRCz11_genomic`	1 640 076	2 728 222 451	22

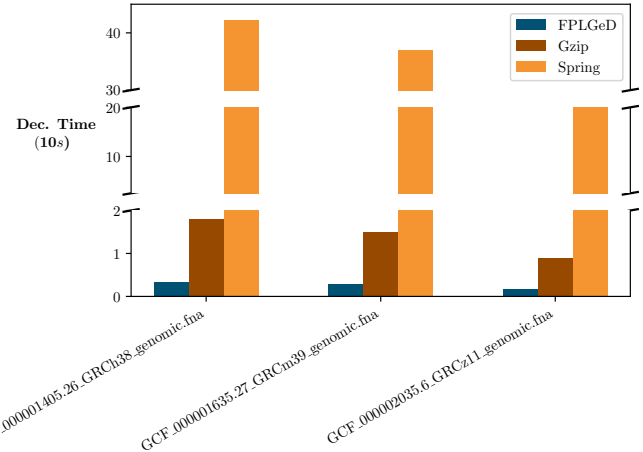
FASTA compression ratio



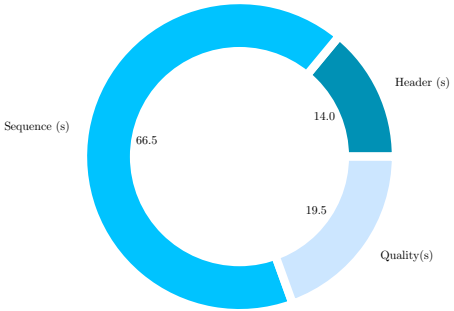
FASTA compression time



FASTA decompression time



Plots: time decompression



Plots: compression comparison