

SPRING: FASTQ compression - Supplementary Data

Shubham Chandak, Kedar Tatwawadi, Idoia Ochoa, Mikel Hernaez, Tsachy Weissman

May 21, 2018

This document contains details about the datasets, installing and running the compression tools and some additional results. Source code for SPRING and related instructions can be found at <https://github.com/shubhamchandak94/SPRING>.

1 Datasets

We list below the links for the datasets used for evaluation. After downloading, the files were unzipped using gunzip command. In some cases FASTQ files were concatenated to get higher coverage datasets.

P. aeruginosa - SRR554369

ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR554/SRR554369/SRR554369_1.fastq.gz
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR554/SRR554369/SRR554369_2.fastq.gz

S. cerevisiae - SRR327342_1

ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR327/SRR327342/SRR327342_1.fastq.gz
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR327/SRR327342/SRR327342_2.fastq.gz

Metagenomic - ERR532393

ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR532/ERR532393/ERR532393_1.fastq.gz
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR532/ERR532393/ERR532393_2.fastq.gz

T. cacao - SRR870667_2

ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR870/SRR870667/SRR870667_2.fastq.gz

H. sapiens - ERP001775

ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR174/ERR174324/ERR174324_1.fastq.gz
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR174/ERR174325/ERR174325_1.fastq.gz
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR174/ERR174324/ERR174324_2.fastq.gz
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR174/ERR174325/ERR174325_2.fastq.gz

The first two files were concatenated and the last two files were concatenated to obtain a 28x coverage paired-end dataset.

H. sapiens - NA12878 Rep 1, Lane 1

This dataset was downloaded from Illumina's BaseSpace public data (<https://basespace.illumina.com/datacentral>) from the project *NovaSeq S2: Nextera DNA Flex (8 replicates of NA12878)*. The following FASTQ files comprise this 25x paired-end dataset.

NA12878-Rep-1_S1_L001_R1_001.fastq
NA12878-Rep-1_S1_L001_R2_001.fastq

These datasets were variable length and were trimmed to 147bp for evaluation with FaStore. Trimming was done using util/trimmer.cpp, available in the SPRING source code.

H. sapiens - NA12878 Rep 1 & 2

This dataset was downloaded from Illumina’s BaseSpace public data (<https://basespace.illumina.com/datacentral>) from the project *NovaSeq S2: Nextera DNA Flex (8 replicates of NA12878)*. The following FASTQ files were downloaded.

NA12878-Rep-1_S1_L001_R1_001.fastq
NA12878-Rep-1_S1_L002_R1_001.fastq
NA12878-Rep-2_S2_L001_R1_001.fastq
NA12878-Rep-2_S2_L002_R1_001.fastq
NA12878-Rep-1_S1_L001_R2_001.fastq
NA12878-Rep-1_S1_L002_R2_001.fastq
NA12878-Rep-2_S2_L001_R2_001.fastq
NA12878-Rep-2_S2_L002_R2_001.fastq

The first four files were concatenated together and the last four files were concatenated together to obtain 100x paired-end data. These datasets were variable length and were trimmed to 147bp for evaluation with FaStore. Trimming was done using util/trimmer.cpp, available in the SPRING source.

For the 50x human dataset used for read compression experiments (Table 8), we concatenated the first and second files above to get file 1, and the fifth and sixth files above to get the corresponding paired-end reads of file 1.

Accession no.	Species	Genome length (Mb)	Read length	#reads (M)	Coverage	PE/SE	Technology
SRR554369	<i>P. aeruginosa</i>	6	100	3.3	50	PE	GAIIx
SRR327342	<i>S. cerevisiae</i>	12.1	63, 75	30	175	PE	GAII
ERR532393	Metagenomic	-	100	72	-	PE	HiSeq 2000
SRR870667_2	<i>T. cacao</i>	350	74	69	15	SE	GAIIx
ERP001775	<i>H. sapiens</i>	3137	101	879	28	PE	HiSeq 2000
NA12878 Rep 1, Lane 1	<i>H. sapiens</i>	3137	147	540	25	PE	Novaseq
NA12878 Rep 1 & 2	<i>H. sapiens</i>	3137	147	2173	100	PE	Novaseq

Table 1: Datasets used for evaluation. PE denotes paired-end, SE denotes single-end. For SRR327342, the read length of the first read in each pair is 63, and that of the second read is 75.

2 Installing and running tools

2.1 Installation

SPRING

7-zip should be already installed. On Linux, run `sudo apt-get install p7zip-full`.

```
git clone https://github.com/shubhamchandak94/SPRING.git
cd SPRING
./install.sh
```

FaStore

```
git clone https://github.com/refresh-bio/FaStore.git
cd FaStore
make
```

DSRC 2

Boost should already be installed.

```
git clone https://github.com/refresh-bio/DSRC.git
cd DSRC
make
```

pigz

```
wget https://zlib.net/pigz/pigz-2.4.tar.gz
tar -xzf pigz-2.4.tar.gz
cd pigz-2.4
make
```

2.2 Running compression algorithms

2.2.1 SPRING

General usage:

Compression

```
./spring -c -1 Fastq_file_1 [-2 Fastq_file_2] [-p] [-t num_threads] \
[-q mode] [-r qvz_ratio] [-i] -o outputfile

-2 second file for paired end reads
-p Preserve order of reads. If not specified, read pairing information is still preserved
  for paired-end datasets.
-t num_threads - Default 8
-q Retain quality values. Possible modes:
    qvz - qvz specify bits/quality ratio using -r flag (default 8.0 lossless)
    bsc - use bsc compressor.
    illumina_binning_bsc - bin quality using Illumina 8 level binning and compress with bsc
    illumina_binning_qvz - bin quality using Illumina 8 level binning and compress with qvz lossless
-r bits/quality ratio if -q qvz used [default 8.0 lossless]
-i Retain read IDs, if not specified fake ids will be generated during decompression
-o Output file name
```

Decompression

```
./spring -d compressed_file -o outputfile [-t num_threads]
-o outputfile name. For paired-end datasets, two files created: outputfile.1 and outputfile.2.
-t num_threads - Default 8
```

For compression in the perfectly lossless mode (lossless quality compression, read identifiers retained and read order preserved), run

```
./SPRING/spring -c -1 in_1.fastq -2 in_2.fastq -i -p -q qvz -t 8 -o compressed_file
```

For the NovaSeq datasets, qvz was replaced by bsc.

For compression in the information preserving mode (Illumina binning of quality, read identifiers no retained and read order not preserved), run

```
./SPRING/spring -c -1 in_1.fastq -2 in_2.fastq -q illumina_binning_qvz -t 8 -o compressed_file
```

For the NovaSeq datasets, illumina_binning_qvz was replaced by bsc.

For decompression (in either mode), run,

```
./SPRING/spring -d compressed_file -o out.fastq -t 8
```

More examples for usage of SPRING with various options are available in the Github README (<https://github.com/shubhamchandak94/SPRING/blob/master/README.md>).

2.2.2 Other algorithms

FaStore

To compress in_1.fastq and in_2.fastq in the perfectly lossless mode (lossless quality compression and read identifiers retained), run

```
./FaStore/bin/fastore_bin e -z -H -q0 -t8 -i"in_1.fastq in_2.fastq" -otemp.bin
./FaStore/bin/fastore_rebin e -z -t8 -itemp.bin -otemp.rebin
./FaStore/bin/fastore_pack e -z -v -t8 -itemp.rebin -ocompressed.fastore
```

The compressed files are compressed.fastore.cdata and compressed.fastore.cmeta.

In the information preserving mode (Illumina binning for quality and read identifiers not retained), run

```
./FaStore/bin/fastore_bin e -z -q2 -t8 -i"in_1.fastq in_2.fastq" -otemp.bin
./FaStore/bin/fastore_rebin e -z -t8 -itemp.bin -otemp.rebin
./FaStore/bin/fastore_pack e -z -v -t8 -itemp.rebin -ocompressed.fastore
```

For NovaSeq datasets, Illumina binning is not applied, so the -q2 in the first line should be replaced by -q0.

To decompress compressed.fastore to out_1.fastq and out_2.fastq, run

```
./FaStore/bin/fastore_pack d -z -t8 -icompressed.fastore -o"out_1.fastq out_2.fastq"
```

For single-end datasets, the -z flag should be removed.

In the perfectly lossless mode, FaStore retains the order of the reads through the read identifiers. However, on decompression, the reads are not outputted in their original order. While it is possible to sort the FASTQ file using the identifiers to get back the original order, we did not include this step when measuring the decompression time/memory. On using the -v flag, FaStore outputs the sizes of various streams (reads, quality etc.), which is used to calculate the size occupied by reads for Table 8.

pigz

Compression:

```
./pigz-2.4/pigz -k -p 8 in_1.fastq in_2.fastq
```

Decompression:

```
./pigz-2.4/unpigz -k -p 8 in_1.fastq.gz in_2.fastq.gz
```

DSRC 2

Compression:

```
./DSRC/bin/dsrc -c -t8 -v in_1.fastq in_1.fastq.dsrc
./DSRC/bin/dsrc -c -t8 -v in_2.fastq in_2.fastq.dsrc
```

Decompression:

```
./DSRC/bin/dsrc -d -t8 -v in_1.fastq.dsrc out_1.fastq
./DSRC/bin/dsrc -d -t8 -v in_2.fastq.dsrc out_2.fastq
```

3 Results

All the experiments were run on a server with a 40-core Intel(R) Xeon(R) 2.20 GHz processor, 258 GB of RAM, 7.3 TB disk space and Ubuntu 16.04. All tools were run with 8 threads. In all the results, 1 MB denotes 10^6 bytes and 1 GB denotes 10^9 bytes.

Main results

Tables 2-6 contain the compression results along with the time and memory requirements.

Sample	SRR554369	SRR327342	ERR532393	SRR870667_2	ERP001775	NA12878 Rep 1, Lane 1	NA12878 Rep 1 & 2
Organism	<i>P. aeruginosa</i>	<i>S. cerevisiae</i>	Metagenomic	<i>T. cacao</i>	<i>H. sapiens</i>	<i>H. sapiens</i>	<i>H. sapiens</i>
Technology	GAIx	GAII	HiSeq 2000	GAIx	HiSeq 2000	NovaSeq	NovaSeq
Coverage	50x	175x	Unknown	15x	28x	25x	100x
Read length	100	63 & 75	100	74	101	147	147
Uncompressed Size	768	5,986	19,284	13,847	227,246	195,748	787,616
Perfectly Lossless							
pigz	279	2,062	6,911	4,926	74,250	36,132	144,927
DSRC 2	198	1,507	5,155	3,540	52,049	26,520	106,665
FaStore	145	-	3,625	2,736	35,753	11,189	34,169
SPRING	134	940	3,131	2,449	28,033	6,859	25,538
Information Preserving							
FaStore	87	-	1,947	1,321	17,547	10,039	29,374
SPRING	82	373	1,745	1,240	13,203	5,575	20,127

Table 2: Compression sizes in MB. FaStore wasn’t run on SRR327342 since it does not support variable length reads. All datasets except SRR870667_2 are paired-end.

Sample	SRR554369	SRR327342	ERR532393	SRR870667_2	ERP001775	NA12878 Rep 1, Lane 1	NA12878 Rep 1 & 2
Perfectly Lossless							
pigz	11s	1m19s	4m14s	3m	49m	33m	2h18m
DSRC 2	4s	15s	1m12s	1m9s	12m	11m	35m
FaStore	1m	-	7m	5m30s	2h13m	1h45m	7h30m
SPRING	3m	6m	15m	11m	3h17m	2h38m	10h25m
Information Preserving							
FaStore	40s	-	7m	4m15s	1h48m	1h40m	7h10m
SPRING	3m	5m	14m	9m	2h54m	2h26m	9h53m

Table 3: Compression times. All tools were run with 8 threads.

Sample	SRR554369	SRR327342	ERR532393	SRR870667_2	ERP001775	NA12878 Rep 1, Lane 1	NA12878 Rep 1 & 2
Perfectly Lossless							
pigz	4s	28s	1m46s	1m12s	27m	12m	1h7m
DSRC 2	3s	13s	40s	36s	13m	9m	55m
FaStore	12s	-	2m45s	2m	28m	17m	1h6m
SPRING	35s	2m18s	8m11s	6m15s	1h40m	48m	4h11m
Information Preserving							
FaStore	7s	-	1m34s	1m	19m	13m	58m
SPRING	16s	1m32s	5m34s	3m27s	1h8m	38m	3h8m

Table 4: Decompression times. All tools were run with 8 threads.

Sample	SRR554369	SRR327342	ERR532393	SRR870667_2	ERP001775	NA12878 Rep 1, Lane 1	NA12878 Rep 1 & 2
Perfectly Lossless							
pigz	0.01	0.01	0.01	0.01	0.01	0.01	0.01
DSRC 2	0.23	0.21	0.21	0.21	0.4	0.14	0.22
FaStore	2.2	-	18	5.1	37	42	160
SPRING	0.75	2.2	4.7	6.9	45	30	119
Information Preserving							
FaStore	2.1	-	18	4.4	26	39	147
SPRING	0.69	1.7	4.3	5.5	45	39	158

Table 5: Compression memory (RAM) in GB.

Sample	SRR554369	SRR327342	ERR532393	SRR870667_2	ERP001775	NA12878 Rep 1, Lane 1	NA12878 Rep 1 & 2
Perfectly Lossless							
pigz	0.002	0.002	0.002	0.002	0.002	0.002	0.002
DSRC 2	0.25	0.29	0.31	0.43	0.42	0.35	0.36
FaStore	0.9	-	2	2.3	30	44	167
SPRING	0.5	0.5	2.7	2.5	12	12	12
Information Preserving							
FaStore	0.6	-	1.5	1.6	23	34	133
SPRING	0.4	0.5	2.7	0.7	12	12	12

Table 6: Decompression memory (RAM) in GB.

Analysis

Table 7 contains relative sizes of reads, quality values and read identifiers for two datasets in two modes. Table 8 contains read compression results for different compression modes and coverages.

Sample	Mode	Reads	Quality	Identifiers	Total
ERP001775	Perfectly Lossless	4,114	23,020	899	28,033
ERP001775	Information Preserving	2,503	10,700	0	13,203
NA12878 Rep 1, Lane 1	Perfectly Lossless	2,982	3,586	291	6,859
NA12878 Rep 1, Lane 1	Information Preserving	1,995	3,580	0	5,575

Table 7: Sizes (in MB) of reads, quality values and read identifiers after compression by SPRING for two human datasets. In the compressed archive, the streams with id in their name correspond to the identifiers, those with quality in their name correspond to quality values, and the rest of the streams correspond to reads. To view the sizes of the streams, run `tar -tvf archive_name`.

	SPRING-SE		SPRING-PE		FaStore-PE
Coverage	Perfectly lossless	Information preserving	Perfectly lossless	Information preserving	Information preserving
25x	3,589	1,631	2,982	1,994	6,061
50x	6,641	2,552	5,354	3,320	9,227
100x	12,520	4,095	10,000	5,752	14,203

Table 8: Read compression sizes in MB for different modes of read compression: (i) SE perfectly lossless: Concatenate the two files and compress with order preserved, (ii) SE information preserving: Concatenate the two files and compress without order preserved, (iii) PE perfectly lossless: Compress the two files with order preserved and utilizing the paired end structure, (iv) PE information preserving: Compress the two files with only pairing information preserved (i.e., the read pairs are arbitrarily reordered in the decompressed file).

Additional Results

Table 9 contains compression results for NovaSeq variable length reads.

Sample	NA12878 Rep 1, Lane 1	NA12878 Rep 1 & 2
Organism	<i>H. sapiens</i>	<i>H. sapiens</i>
Technology	NovaSeq	NovaSeq
Coverage	26x	105x
Maximum Read length	151	151
Uncompressed Size	205,386	826,117
Perfectly Lossless		
pigz	38,007	152,430
DSRC 2	28,448	114,393
SPRING	7,755	27,992
Information Preserving		
SPRING	6,094	22,399

Table 9: Compression sizes in MB for the variable-length NovaSeq datasets. Only tools supporting variable length reads were tested.