Bioinformatics, 31(9), 2015, 1389–1395
doi: 10.1093/bioinformatics/btu844
Advance Access Publication Date: 22 December 2014
Original Paper

OXFORD

Sequence analysis

Disk-based compression of data from genome sequencing

Szymon Grabowski[1], Sebastian Deorowicz[2,*] and Łukasz Roguski[3,4]

[1]Institute of Applied Computer Science, Lodz University of Technology, Al. Politechniki 11, 90-924 Łódź, [2]Institute of Informatics, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, [3]Polish-Japanese Institute of Information Technology, Koszykowa 86, 02-008 Warszawa, Poland and [4]Centro Nacional de Análisis Genómico (CNAG), 08-028 Barcelona, Spain

*To whom correspondence should be addressed.
Associate Editor: John Hancock

Received on September 19, 2014; revised on November 27, 2014; accepted on December 17, 2014

Abstract

Motivation: High-coverage sequencing data have significant, yet hard to exploit, redundancy. Most FASTQ compressors cannot efficiently compress the DNA stream of large datasets, since the redundancy between overlapping reads cannot be easily captured in the (relatively small) main memory. More interesting solutions for this problem are disk based, where the better of these two, from Cox *et al.* (2012), is based on the Burrows–Wheeler transform (BWT) and achieves 0.518 bits per base for a 134.0 Gbp human genome sequencing collection with almost 45-fold coverage.
Results: We propose overlapping reads compression with minimizers, a compression algorithm dedicated to sequencing reads (DNA only). Our method makes use of a conceptually simple and easily parallelizable idea of minimizers, to obtain 0.317 bits per base as the compression ratio, allowing to fit the 134.0 Gbp dataset into only 5.31 GB of space.
Availability and implementation: http://sun.aei.polsl.pl/orcom under a free license.
Contact: sebastian.deorowicz@polsl.pl
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

It is well known that the growth of the amount of genome sequencing data produced in the last years outpaces the famous Moore's law predicting the developments in computer hardware (Deorowicz and Grabowski, 2013; Kahn, 2011). Confronted with this deluge of data, we can only hope for better algorithms protecting us from drowning. Speaking about big data management in general, there are two main algorithmic concerns: faster processing of the data (at preserved other aspects, like mapping quality in de novo or referential assemblers) and more succinct data representations (for compressed storage or indexes). In this article, we focus on the latter concern.

Raw sequencing data are usually kept in FASTQ format, with two main streams: the DNA symbols and their corresponding quality scores. Older specialized FASTQ compressors were lossless, squeezing the DNA stream down to about 1.5–1.8 bpb (bits per base) and the quality stream to 3–4 bpb, but more recently it was noticed that a reasonable solution for lossy compression of the qualities has negligible impact on further analyzes, for example, referential mapping or variant calling performance (Cánovas *et al.*, 2014; Illumina, 2012; Wan *et al.*, 2012). This scenario became thus immediately practical, with scores lossily compressed to about 1 bpb (Janin *et al.*, 2014) or less (Yu *et al.*, 2014). Note also that Illumina software for their HiSeq 2500 equipment contains an option to reduce the number of quality scores (even to a few), since it was shown that the fraction of discrepant single nucleotide polymorphisms grows slowly with diminishing number of quality scores in Illumina's CASAVA package (http://support.illumina.com/sequencing/sequencing_software/casava.ilmn). It is easy to notice that now the DNA stream becomes the main compression challenge. Even if

higher order modeling (Bonfield and Mahoney, 2013) or LZ77-style compression (Deorowicz and Grabowski, 2011) can lead to some improvement in DNA stream compression, we are aware of only two much more promising approaches. Both solutions are disk based. Yanovsky (2011) creates a *similarity graph* for the dataset, defined as a weighted undirected graph with vertices corresponding to the reads of the dataset. For any two reads s_1 and s_2 the edge weight between them is related to the 'profitability' of storing s_1 and the edit script for transforming it into s_2 versus storing both reads explicitly. For this graph, its minimum spanning tree (MST) is found. During the MST traversal, each node is encoded using the set of *maximum exact matches* between the node's read and the read of its parent in the MST. As a backend compressor, the popular 7zip is used. ReCoil compresses a dataset of 192 M Illumina 36 bp reads (http://www.ncbi.nlm.nih.gov/sra/SRX001540), with coverage below 3-fold, to 1.34 bpb. This is an interesting result, but ReCoil is hardly scalable; the test took about 14 h on a machine with 1.6 GHz Intel Celeron CPU and four hard disks.

More recently, Cox *et al.* (2012) took a different approach, based on the Burrows–Wheeler transform (BWT). Their result for the same dataset was 1.21 bpb in less than 65 min, on a Xeon X5450 (Quad-core) 3 GHz processor. The achievement is however more spectacular if the dataset coverage grows. For 44.5-fold coverage of real human genome sequence data, the compressed size improves to as little as 0.518 bpb (Actually in Cox et al. (2012) the authors report 0.484 bpb, but their dataset is seemingly no longer available and in our experiments we use a slightly different one.) allowing to represent the 134.0 Gbp of input data in 8.7 GB of space.

Note that if the reference sequence is available, either explicit or can be reconstructed ('presumed', in the terminology of Cánovas and Moffat 2013), then compressing DNA reads is much easier and high compression ratios are possible. Several FASTQ or SAM/BAM compressors make use of a reference sequence, to name Quip (Jones *et al.*, 2012), Fastqz and Fqzcomp (Bonfield and Mahoney, 2013) in one of their modes, SlimGene (Kozanitis *et al.*, 2011), CRAM (Fritz *et al.*, 2011), Goby (Campagne *et al.*, 2013), DeeZ (Hach *et al.*, 2014) and FQZip (Zhang *et al.*, 2015). Several techniques for compressing SAM files, including mapping reads to a presumed reference sequence, were also explored in Cánovas and Moffat (2013).

In this article, we present a new reference-free compressor for FASTQ data, Overlapping Reads COmpression with Minimizers (ORCOM), achieving compression ratios surpassing the best known solutions. For the two mentioned human datasets it obtains the compression ratios of 1.005 and 0.317 bpb, respectively. ORCOM is also fast, producing the archives in about 8 and 77 min, respectively, using eight threads on an AMD Opteron 6136 2.4 GHz machine.

2 Materials and methods

Let $s = s[0]s[1] \dots s[n-1]$ be a string of length n over a finite alphabet Σ of size σ. We use the following notation, assuming $0 \le i \le j < n$: $s[i]$ denotes the $(i+1)$th symbol of s, $s[i \dots j]$ the substring $s[i]s[i+1] \dots s[j]$ (called a factor of s), and $s \circ t$ the concatenation of strings s and t.

Our algorithm, ORCOM, follows the ancient paradigm of external algorithms: distribute the data into disk bins and then process (i.e. compress) each bin separately. Still, the major problem with this approach in reads compression concerns the bin criterion: how to detect similar (overlapping) reads, in order to pass them into the same bin? Our solution makes use of the idea of minimizers (Roberts *et al.*, 2004), a late bloomer in bioinformatics, cf. (Chikhi *et al.*, 2014; Deorowicz *et al.*, 2014; Li *et al.*, 2013; Movahedi *et al.*, 2012; Wood and Salzberg, 2014). Minimizers are a simple yet ingenious notion. The minimizer for a read s of length r is the lexicographically smallest of its all $(r-p+1)$ p-mers; usually it is assumed that $p \ll r$. This smallest p-mer may be the identifier of the bin into which the read is then dispatched. Two reads with a large overlap are likely to share the same minimizer. In the next paragraphs we present the details of our solution.

Assume the alphabet size $\sigma = 5$ (ACGTN). A reasonable value of p is about 10, but sending each bin to a file on disk would require $5^{10} = 9.77$ M files, which is way too much. Reducing this number to $4^{10} + 1$ (all minimizers containing at least one symbol N, which are rare, are mapped to a single bin, labeled N) is still not satisfactory. We solved this problem in a radical way, using essentially one (In fact, there is one extra file, with metadata, yet this one is of minor overall importance and we skip further description.) temporary file for all the bins, using large output buffers. To fetch the bin data in a further stage, the file has to be opened to read and the required reads extracted to memory from several locations of the file.

As DNA sequences can be read in two directions: forwards and backwards (with complements of each nucleotide), we also process each read twice, in its given and reverse-complemented form. Additionally, we introduce a 'skip zone', that is, do not look for minimizers in read suffixes of (default) length $z = 12$ symbols. This allows for some improvement in read ordering in the next step. The minimizers are thus sought over $2(r-z-p+1)$ resulting p-mers. We call them *canonical minimizers*.

However, a problem with strictly defined minimizers is uneven bin distribution. This not only increases the peak memory use, but also hampers parallel execution as the requirement for load balancing is harder to fulfill. To mitigate these problems we forbid some canonical minimizers, namely those that contain any triple AAA, CCC, GGG or TTT or at least one N (some of them, especially minimizers with runs of three or more As, are frequent). The allowed canonical minimizers are further called *signatures*, a term that we also used (with a slightly different definition) for the minimizers used in KMC 2, a k-mer counting algorithm (Deorowicz *et al.*, 2014).

In the next step, when the disk bins are built, we reordered the reads in bins to move overlapping reads possibly close to each other. From a few simple sort criteria tried out, the one that worked best was to sort the reads s_i, for all i, according to the lexicographical order of the string $s_i[j \dots r-1] \circ s_i[0 \dots j-1]$, where j is the beginning position of the signature for the read s_i. Such a reordering has a major positive impact on further compression. The reason is that overlaping reads are with high probability close to each other in the reordered array. The size of the skip zone should be chosen carefully. When too small, some signatures will be found close to the end of the read and the first factor of the sorting criterion, $s_i[j \dots r-1]$, will be too short to have a good chance of placing the read among those that overlap it in the genome. On the other hand, with the zone being too long many truly overlapping reads will be forbidden.

The last phase is the backend compression on bin-by-bin basis. We devised a specialized processing method, which produces several (interleaving) streams of data, finally compressed with either a well-known context-based compressor PPMd (Shkarin, 2002) or a variant of arithmetic coder (Salomon and Motta, 2010) (We use a popular and fast arithmetic coding variant by Schindler (http://www.compressconsult.com/rangecoder/), also known as a RC.). How we process the bins in detail, including careful mismatch handling, is presented in the next paragraphs.

We maintain a buffer (sliding window) of m previous reads, storing also the position of the signature in each read. For each read, we seek the read from the buffer which maximizes the overlap. The distance between a pair of considered reads depends on the number of elementary operations transforming one into another. For example, if the pair of reads is:

```
AACGTXXXXCGGCAT,
  CCTXXXXCGGCATCC,
```

where XXXX denotes a signature, we match them after a (conceptual) alignment:

```
AACGTXXXXCGGCAT,
  CCTXXXXCGGCATCC,
```

to find that they differ with one mismatch (`G` versus `C`) and 2 end symbols of the second read have to be inserted, hence the distance is $c_m \times 1 + c_i \times 2$, where c_m and c_i are the mismatch and the insert cost, respectively. The default values for the parameters are: $c_m = 2$ and $c_i = 1$, and they were chosen experimentally. In our example, the final distance is thus $2 \times 1 + 1 \times 2 = 4$. The read among the m previous ones that minimizes such a distance, and is not greater than max_dist, set by default to a half of read length, is considered a reference for the current read.

Next, the referential matching data are sent into a few streams.

Flags
Values from $\{f_{copy}, f_{diss}, f_{ex}, f_{mis}, f_{oth}\}$, with the following meaning:

- f_{copy}—the current read is identical to the previous one,
- f_{diss}—the read is not similar to any read from the buffer; more precisely, the similarity distance exceeds a specified threshold max_{dist},
- f_{ex}—the read overlaps with some read from the buffer without mismatches (only its trailing symbols are to be encoded),
- f_{mis}—the read overlaps with some read from the buffer with exactly one mismatch at the last position of the referenced read,
- f_{oth}—the read overlaps with some read from the buffer, but not in a way corresponding to flags f_{ex} or f_{mis}.

Lengths
Read lengths are stored here (1 byte per length in the current implementation, but a simple byte code can be used to handle the general case).

The five streams: lettersN, lettersA, lettersC, lettersG, lettersT
(These are used only if 'Flag' is f_{ex}, f_{mis} or f_{oth}.)

'LettersN' stores (*i*) all mismatching symbols from the current read where at the corresponding position of the referenced read there is symbol N, and (*ii*) all trailing symbols from the current read beyond the match (i.e. C and C in the example above).

'LettersX', for $X \in \{A, C, G, T\}$, stores all mismatching symbols from the current read where at the corresponding position of the referenced read there is symbol X (in our running example, the mismatching C would be encoded in the stream 'lettersG'. Note that the alphabet size for any 'LettersX' stream is 4, that is, $\{A, C, G, T, N\} \setminus \{X\}$.

Prev
(Used only if 'Flag' is f_{ex}, f_{mis} or f_{oth}.) Stores the location (id) of the referenced read from the buffer.

Shift
(Used only if 'Flag' is f_{ex}, f_{mis} or f_{oth}.) Stores the offsets of the current reads against their referenced read. The offset may be negative. For our running example, the offset is $+2$.

Matches
(Used only if 'Flag' is f_{oth}.) Stores information on mismatch positions. For our running example, the matching area has 13 symbols, but 4 of them belong to the signature and can thus be omitted (as the signature's position in the current read is known from the corresponding value in the stream 'Shift'). A form of RLE (run-length encoding) is used here. Namely, each run of matching positions (of length at least 1) is encoded with its length on 1 byte, and if the byte value is less than 255 and there are symbols left yet, we know that there must be a mismatch at the next position, so it is skipped over. 'Unpredicted' mismatches are encoded with 0. For our example, we obtain the sequence: 1, 7 (match of length 1; omitted mismatch; match of length 11, which is 7 plus 4 for the covered signature's area).

HReads
(Used only if 'Flag' is f_{diss}.) Here the 'hard' reads (not similar enough to any read from the buffer) are dispatched. They are stored almost verbatim: the only change in the representation is to replace the signature with an extra symbol (.). This helps a little for the compression ratio.

Rev
Contains binary flags telling if each read is processed directly or first reverse-complemented.

Some of the streams are compressed with a strong general-purpose compressor, PPMd (http://compression.ru/ds/ppmdj1.rar), using switches -o4 -m16m (order-4 context model with memory use up to 16 MB), others with our range coder (RC), also of order-4. Namely, the streams 'Flags' and 'Rev' are compressed with order-4 RC, the stream 'LettersX' with order-4 RC, where the context is formed of the four previous symbols, and all the other streams are compressed with PPMd.

The description presented above is somewhat simplified. We took some effort to achieve high processing performance. In particular, the input data (read from FASTQ files, possibly gzipped) are processed in 256 MB blocks (block size configurable as an input parameter) and added to a queue. Several worker threads find signatures in them, perform the necessary processing and add to an output queue, whose data are subsequently written to the temporary file. Also further bin processing is parallelized, to maximize the performance.

3 Results

We tested our algorithm versus several competitors on real and simulated datasets, detailed in Table 1. The test machine was a server equipped with four 8-core AMD Opteron 6136 2.4 GHz CPUs, 128 GB of RAM and a RAID-5 disk matrix containing 6 HDDs. We use decimal multiples for the units, that is, 'M' (mega) is 10^6, 'G' (giga) is 10^9, etc., for the file sizes and memory uses reported in this section.

3.1 Real datasets

The experimental results with real read data are presented in the upper parts of Tables 2 and 3. Apart from the proposed compressor

ORCOM, we tested DSRC 2 (Roguski and Deorowicz, 2014), Quip (Jones *et al.*, 2012), FQZComp (Bonfield and Mahoney, 2013), Scalce (Hach *et al.*, 2012), SRcomp (Selva and Chen, 2013) and BWT-SAP (Cox *et al.*, 2012). All these competitors, with the exception of BWT-SAP and SRcomp, are FASTQ compressors, and all of them present compression results of the separate streams. In Table 2 we also present the result of ReCoil (Yanovsky, 2011) on one dataset, copied from Cox *et al.* (2012). ReCoil is too slow to be run on all our data in reasonable time.

As we can see in Table 2, ORCOM wins on all datasets, in an extreme case (*Musa balbisiana*) with almost twice better compression ratio than the second best compressor, BWT-SAP. Figure 1 confirms the intuition that the performance of our algorithm improves with growing coverage. Table 3 presents the compression times and RAM consumptions. Our compressor is also usually the fastest, with rather moderate memory usage (up to 14 GB). We point out that the first phase, distributing the data into bins, is not very costly, for example, it takes less than 25% of the total time for the largest dataset, *Homo sapiens* 2. In the memory use the most frugal is BWT-SAP (which is another disk-based software), spending only 3 MB for each dataset. One should remember that compression times are related to the number of used threads: Quip, FQZComp, SRcomp and BWT-SAP are sequential (one thread), whereas DSRC 2, Scalce and ORCOM are parallel and use eight threads here. Moreover, ORCOM, BWT-SAP and SRcomp compress the DNA stream only, whereas the remaining compressors have full FASTQ files on the input (with fake remaining streams in case of simulated reads presented in Section 3.2), what hampers their performance in compression speed and memory use (the compression ratios are however given for the DNA stream only). For these reasons, the results from Table 3 should not be taken too seriously; they are given mostly to point out promising performance of our software.

ORCOM's compression performance depends on two parameters: the signature length and the skip zone length. How varying these parameters affects the compression ratio is shown in Figures 2 and 3, respectively, on the example of two datasets. It seems that choosing the signature length from {6, 7, 8} is almost irrelevant for the compression ratio, but with longer signature the ratio starts to deteriorate. The impact of the skip zone length depends somewhat on the chosen signature length, yet from our results we can say that any zone length between 8 and 16 is almost equally good.

We also show how replacing the straight minimizers with signatures affects the compression ratio and memory consumption (Table 4). The compression gain is slight, up to 2.3% on real data (*H.sapiens* 2) and up to 6.9% on simulated data (*H.sapiens* 3). Fortunately, the improvement is greater in memory reduction, sometimes exceeding factor 2. As we can see, in two cases using signatures required more memory than with minimizers, but this is for two relatively small datasets. Using signatures generally leads to more even data distribution across bins. The bin size distribution is still far from perfect though, as shown in Figure 4.

Finally, in Table 5 the sizes of individual streams *after compression*, for three datasets, are given. As one can see, the stream of hard reads is hardest to compress, which we think justifies its name.

3.2 Simulated datasets

In the experiments for simulated datasets the reads were obtained by randomly sampling *H.sapiens* reference genome (HG 37.3). The number of non-N-symbols in the reference is approximately 2859 M. The *H.sapiens* 3 dataset contains 1.25 G reads of length 100 bp reads (to have the genome coverage like for *H.sapiens* 2). Half of them were obtained directly from the reference and another half were reverse complemented.

The reads in *H.sapiens* 4 dataset were obtained from *H.sapiens* 3 dataset by modifying each base with a probability 1% (the probability is independent from the base position). It is important to stress that such a simulation of errors is far from what happens in real experiments (e.g. in most sequencers the quality of bases depends on the base position). Nevertheless, this simple error model allows us to compute the theoretically possible compression ratio and compare the

Table 1. Characteristics of the datasets used in the experiments

Dataset (Organism)	Genome length	Number of Gbases	Number of Mreads	Average read length	Accession number
G.gallus	1040	34.7	347	100	SRX043656
M.balbisiana	472	56.9	569	101	SRX339427
H.sapiens 1	3093	6.9	192	36	SRX001540
H.sapiens 2	3093	135.3	1340	101	ERA015743
H.sapiens 2-trim	3093	134.0	1340	100	ERA015743
H.sapiens 3	3093	125.0	1250	100	—
H.sapiens 4	3093	125.0	1250	100	—

Notes: Approximate genome lengths are in Mbases according to http://www.ncbi.nlm.nih.gov/genome/. *Homo sapiens* 3 and 4 datasets are artificial reads produced by sampling reference human genome. In *H.sapiens* 3 the sampled reads are exact and in *H.sapiens* 4 bases are modified with probability 1%.

Table 2. Compression ratios for various datasets

Dataset	DSRC 2	Quip	FQZComp	Scalce	SRcomp	ReCoil	BWT-SAP	ORCOM
G.gallus	1.820	1.715	1.419	0.824	1.581	—	0.630	**0.433**
M.balbisiana	1.838	1.196	0.754	0.342	0.522	—	0.208	**0.110**
H.sapiens 1	1.857	1.773	1.681	1.263	1.210	1.34	1.246	**1.005**
H.sapiens 2	1.821	1.665	1.460	1.117	NS	—	NS	**0.327**
H.sapiens 2-trim	1.839	1.682	1.474	0.781	Failed	—	0.518	**0.317**
H.sapiens 3	1.832	1.710	1.487	0.720	Failed	—	0.410	**0.174**
H.sapiens 4	1.902	1.754	1.568	1.022	Failed	—	0.810	**0.562**

Notes: Compressed ratios, in bits per base. The results of our approach are presented in the rightmost column. The best results are in bold. 'NS' means that the compressor was not examined as it does not support variable-length reads in a dataset. ReCoil was not examined in our experiments due to very long running times [the only result comes from Cox *et al.* (2012) paper].

Table 3. Compression times and memory usage of compressors

Dataset	DSRC 2		Quip		FQZComp		Scalce		SRcomp		BWT-SAP		ORCOM	
	Time	RAM	Time	RAM	Time	RAM	Time	RAM	Time	RAM	Time	RAM	Time	RAM
G.gallus	1.3	5.2	9.7	0.8	12.3	4.2	4.6	5.5	4.2	34.3	62.3	**0.003**	**1.1**	9.6
M.balbisiana	2.2	5.4	14.4	0.8	18.0	4.2	9.7	5.4	3.6	55.5	99.1	**0.003**	**2.1**	5.2
H.sapiens 1	**0.3**	6.1	1.6	0.8	2.3	4.2	1.2	5.5	**0.3**	6.9	6.0	**0.003**	0.5	9.7
H.sapiens 2	9.7	2.8	39.0	**0.8**	47.4	4.3	18.5	5.5	NS		NS		**5.6**	13.8
H.sapiens 2-trim	9.7	2.8	39.0	0.8	45.9	4.2	18.5	5.5	Failed		267.8	**0.003**	**4.6**	12.5
H.sapiens 3	2.2	5.2	35.3	0.8	45.0	4.2	12.8	5.4	Failed		246.1	0.005	2.8	11.7
H.sapiens 4	2.8	5.6	42.7	0.8	48.7	4.2	16.0	5.5	Failed		278.0	0.006	5.0	13.7

Notes: Times are in thousands of seconds. RAM consumptions are in GBs. The best results are in bold. 'NS' means that the compressor was not examined as it does not support variable-length reads in a dataset.

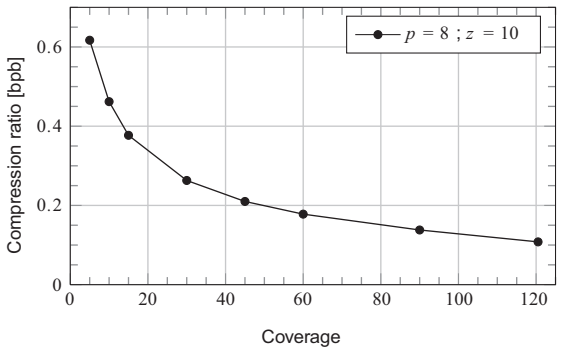

Fig. 1. Compression ratio for various coverage factors for *M.balbisiana* dataset

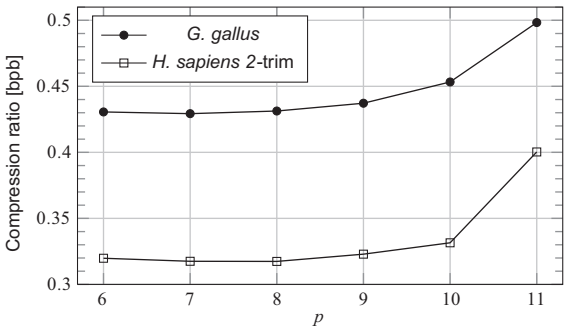

Fig. 2. Compression ratio for various signature lengths for *Gallus gallus* and *H.sapiens* 2-trim datasets ($z = 10$)

results of existing compressors with the estimated optimum and check what improvement is still possible in the reads compression area.

The obtained compression ratios are presented in two bottom rows of Table 2. We note that 'standard' FASTQ compressors achieve compression ratios similar to the ones on real reads, which is perhaps no surprise. BWT-SAP achieves a substantial improvement on *H.sapiens* 3, as the noise in the real data must have broken many long runs in the BWT-related sequence and have hampered the compression. Yet, even more improvement, close to 2-fold, is observed for ORCOM. This can be explained by the local search for similar reads in our solution: once an error affects a read's signature area, the read is moved to another bin. On the noisy *H.sapiens* 4 dataset all compression ratios are, as expected, inferior. It is also not surprising that the standard FASTQ compressors, unable to eliminate most of the redundancy of the DNA reads, lose less here than ORCOM and Scalce do. The compression of clean data (*H.sapiens* 3) is also faster (Table 3).

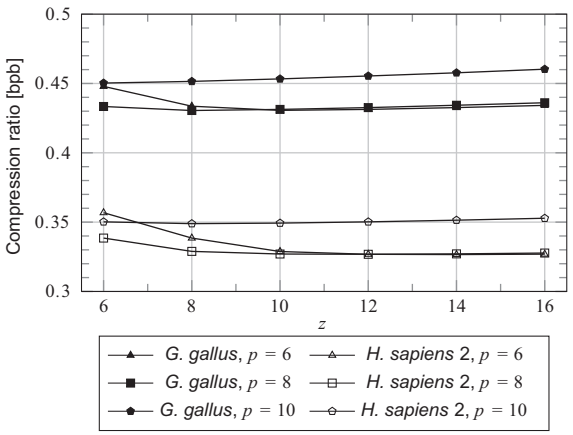

Fig. 3. Compression ratio for various skip zone lengths for *G.gallus* and *H.sapiens* 2 datasets

Table 4. Comparison of compression ratios and memory usage between minimizers and signatures in ORCOM

Dataset	Compression ratio		RAM	
	Minimizers	Signatures	Minimizers	Signatures
G.gallus	0.443	0.433	9.7	9.6
M.balbisiana	0.112	0.110	8.3	9.6
H.sapiens 1	1.007	1.005	7.1	9.7
H.sapiens 2	0.338	0.327	29.7	13.8
H.sapiens 2-trim	0.330	0.317	28.7	12.5
H.sapiens 3	0.188	0.175	26.5	11.7
H.sapiens 4	0.565	0.562	26.9	13.7

It is interesting to compare the ratios with estimations on how good compression ratio is possible. In theory, it is possible to perform a *de novo* assembly and to reproduce the genome from the reads.

The reads from *H.sapiens* 3 dataset can be reordered according to the position of the read in the assembled genome. Thus, to encode the dataset it is sufficient to encode the assembled genome and for each read also:

- the position of the read in the assembled genome,
- the length (in a case of variable-length reads); for our data it is unnecessary as all reads are of length 100 bp,
- the read orientation, that is, whether the read maps the genome directly or must be reverse complemented.

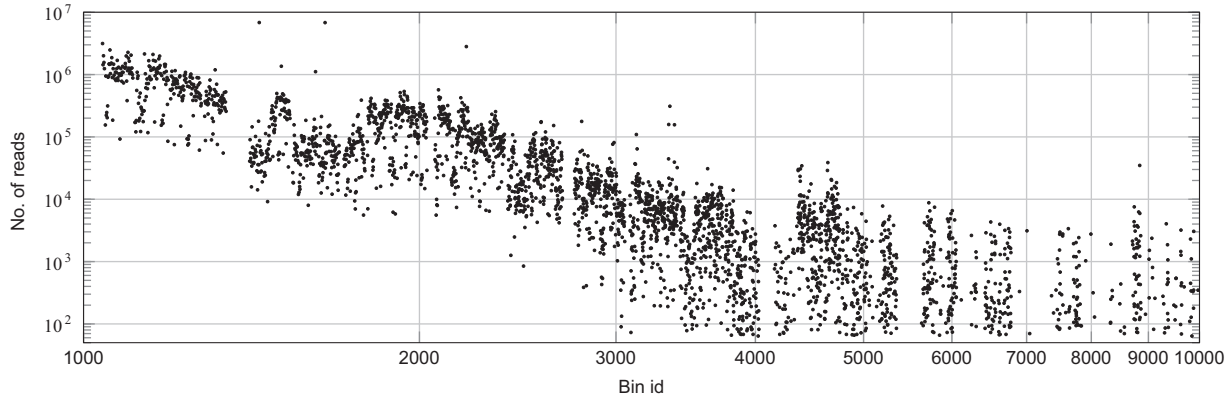

Fig. 4. Bin sizes for *G.gallus* dataset. Note that bin id must be at least 1024, since 1024 corresponds to AACAAAAA, and the prefix AAA is forbidden. Some samples of bin id to signature mappings: 2000—AACTTCAA, 3000—AAGTGTGA, 4000—AATTGGAA

Table 5. Stream sizes for selected datasets used in the experiments

Stream	*G.gallus*	*M.balbisiana*	*H.sapiens 2*
Flags	64	53	280
Lengths	0	0	157
LettersX	364	141	1,152
Prev	59	51	343
Shift	136	68	536
Matches	148	72	582
HReads	1072	326	2316
Rev	42	63	165

Notes: Sizes are given in Mbytes (1 Mbyte = 10^6 bytes).

The assembled genome can be encoded using 2 bits per base (there is no N symbols), so for 2859 Mb we need 5718 Mbit. Then, for each read it is necessary to use 1 bit for the orientation, that is, 1250 Mbit in total. The read positions are ordered but are not unique, so to encode the positions we need to store a multisubset of size 1250 M from a set of size 2859 M, which is equivalent to storing 1250 M *unique* and ordered integers from the range $\langle 0, 1250\text{M} + 2859\text{M}) = \langle 0, 4109\text{M})$.

The number of bits necessary to encode m unique and ordered integers from the range $\langle 0, n)$ is

$$\log_2 \binom{n}{m} \approx n\log_2 n - (n-m)\log_2(n-m) - m\log_2 m. \quad (1)$$

For $n = 4109 \times 10^6$ and $m = 1250 \times 10^6$ we obtain 3642.12 Mbit. Thus, in total we need 10 610.12 Mbit, so the compression ratio expressed in bits per base is 0.085.

In case of reads with 1% of wrong bases we need to encode also the bases in the reads and the positions of the differences between the reads and the assembled genome. There are 1250 Mbases to encode, but this time it is enough to use $\log_2 3$ bits per base as we are sure that the actual base differs to the base in the genome, so the total size of these data is 1981.20 Mbit. To encode the positions, we can conceptually concatenate the reads and encode the ordered and unique positions of wrong bases. We now apply Equation (1) with parameters $n = 125 \times 10^9$ and $m = 0.01n$ and obtain 10 099.14 Mbit. Thus, in total, to encode 1.25 G reads of length 100 bp with 1% of wrong bases, we need 22 690.46 Mbit, which translates into 0.182 bpb.

We can notice that the obtained results with simulated reads are much worse (roughly, by factor 2 for *H.sapiens* 3 and factor 3 for *H.sapiens* 4) than the estimated lower bounds. This is basically due to two reasons. One is that the proposed read grouping method belongs to crisp ones, that is, one read belongs to one and only one bin. In this way, reads with relatively small overlaps are likely to be scattered to different bins and their cross-correlation cannot be exploited. Moreover, even reads with a large overlap has some (albeit rather small) chance of landing in different bins. This harmful effect of separating similar reads is stronger for noisy data. The other reason is the simplicity of our modeling, in which read alignment is performed only in pairs of reads and thus some long matches may be prematurely truncated. Overcoming these limitations of our algorithm is an interesting topic for further research.

4 Conclusions and future work

We presented ORCOM, a lightweight solution for grouping and compressing overlapping reads in DNA sequencing data. We showed that the obtained compression ratio for large datasets is much better the one from the previously most successful, BWT-based, approach. Our algorithm is based on the recently popular idea of minimizers. For the human dataset comprising reads of 100 bp, with 44.5-fold coverage, we obtain 0.317 bpb compression ratio. This means that the 134.0 Gb dataset can be stored in as little as 5.31 GB. Also for the other tested datasets ORCOM attains, to our knowledge, compression ratios better than all previously reported.

ORCOM, as a tool, may be improved in a number of ways. Its performance depends, albeit mildly, on two parameters: signature length and skip zone length. Currently their default values are set *ad hoc*, while in the future we are going to work out quite a robust automated parameter selection procedure. Also, our plans include fine-tuning the backend modeling (e.g. the distance function between reads is rather crude now).

More importantly, perhaps, a memory-only mode can be added, convenient for powerful machines, but with more compact internal data representations affordable also for standard PCs, at least on small to moderate sized genomes. On the other hand, in the disk-based mode the memory use may be reduced, at least as a trade-off (less RAM, but also fewer threads, thus slower compression and/or somewhat worse compression ratio). From the algorithmic point, a more interesting challenge would be to come closer to the compression bounds estimated in Section 3.2. Finally, our ideas could be incorporated in a full-fledged FASTQ compressor, together with recent advances in lossy compression of the quality data, to obtain unprecedented compression ratios for FASTQ inputs in an industry-oriented massively parallel implementation.

Acknowledgement

The authors thank the reviewers for helpful comments.

Funding

The Polish National Science Centre under the project DEC-2012/05/B/ST6/03148 and also the European Union from the European Social Fund within the INTERKADRA project UDAPOKL-04.01.01-00-014/10-00 (partially). The infrastructure supported by POIG.02.03.01-24-099/13 grant: 'GeCONiI-Upper Silesian Center for Computational Science and Engineering'.

Conflict of Interest: none declared.

References

Bonfield,J.K. and Mahoney,M.V. (2013) Compression of FASTQ and SAM format sequencing data. *PLoS One*, **8**, e59190.

Campagne,F. *et al.* (2013) Compression of structured high-throughput sequencing data. *PLoS One*, **8**, e79871.

Cánovas,R. and Moffat,A. (2013) Practical compression for multi-alignment genomic files. In: B. Thomas (ed.) *Proceeding ACSC'13 Proceedings of the Thirty-Sixth Australasian Computer Science Conference*. Darlinghurst, Australia, Australian Computer Society, Inc., pp. 51–60.

Cánovas,R. *et al.* (2014) Lossy compression of quality scores in genomic data. *Bioinformatics*, **30**, 2130–2136.

Chikhi,R. *et al.* (2014) *On the representation of de Bruijn graphs*. arXiv preprint arXiv:1401.5383.

Cox,A.J. *et al.* (2012) Large-scale compression of genomic sequence databases with the Burrows–Wheeler transform. *Bioinformatics*, **28**, 1415–1419.

Deorowicz,S. and Grabowski,S. (2011) Compression of DNA sequence reads in FASTQ format. *Bioinformatics*, **27**, 860–862.

Deorowicz,S. and Grabowski,S. (2013) Data compression for sequencing data. *Algorithms Mol. Biol.*, **8**, 25.

Deorowicz,S. *et al.* (2014) *KMC 2: Fast and resource-frugal k-mer counting*. arXiv preprint arXiv:1407.1507.

Fritz,M.-Y. *et al.* (2011) Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Res.*, **21**, 734–740.

Hach,F. *et al.* (2012) SCALCE: boosting sequence compression algorithms using locally consistent encoding. *Bioinformatics*, **28**, 3051–3057.

Hach,F. *et al.* (2014) Deez: reference-based compression by local assembly. *Nat. Methods*, **11**, 1082–1084.

Illumina (2012) Reducing whole-genome data storage footprint. *Technical report*, Illumina.

Janin,L. *et al.* (2014) Adaptive reference-free compression of sequence quality scores. *Bioinformatics*, **30**, 24–30.

Jones,D. *et al.* (2012) Compression of next-generation sequencing reads aided by highly efficient de novo assembly. *Nucleic Acids Res.*, **40**, e171.

Kahn,S.D. (2011) On the future of genomic data. *Science(Washington)*, **331**, 728–729.

Kozanitis,C. *et al.* (2011) Compressing genomic sequence fragments using SlimGene. *J. Comput. Biol.*, **18**, 401–413.

Li,Y. *et al.* (2013) Memory efficient minimum substring partitioning. In: *Proceedings of the 39th International Conference on Very Large Data Bases*. VLDB Endowment. pp. 169–180.

Movahedi,N.S. *et al.* (2012) De novo co-assembly of bacterial genomes from multiple single cells. In: *BIBM*. IEEE Computer Society. pp. 1–5.

Roberts,M. *et al.* (2004) Reducing storage requirements for biological sequence comparison. *Bioinformatics*, **20**, 3363–3369.

Roguski,Ł. and Deorowicz,S. (2014) DSRC 2—industry-oriented compression of FASTQ files. *Bioinformatics*, **30**, 2213–2215.

Salomon,D. and Motta,G. (2010) *Handbook of Data Compression*. Springer, London.

Selva,J.J. and Chen,X. (2013) SRComp: Short read sequence compression using burstsort and elias omega coding. *PLoS One*, **8**, article no. e81414.

Shkarin,D. (2002) PPM: one step to practicality. In: *Data Compression Conference (DCC)*. Snowbird, UT, IEEE Computer Society Press. pp. 202–211.

Wan,R. *et al.* (2012) Transformations for the compression of FASTQ quality scores of next-generation sequencing data. *Bioinformatics*, **28**, 628–635.

Wood,D.E. and Salzberg,S.L. (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.*, **15**, R46.

Yanovsky,V. (2011) Recoil—an algorithm for compression of extremely large datasets of DNA data. *Algorithms Mol. Biol.*, **6**, 23.

Yu,Y. *et al.* (2014) Traversing the *k*-mer landscape of NGS read datasets for quality score sparsification. In: R. Sharan (ed.) Research in Computational Molecular Biology, *Vol. 8394* Lecture Notes in Computer Science. Springer International Publishing, Pittsburgh, PA, USA, pp. 385–399.

Zhang,Y. *et al.* (2015) FQZip: lossless reference-based compression of next generation sequencing data in FASTQ format. In: *18th Asia Pacific Symposium on Intelligent and Evolutionary Systems*, **Vol. 2**. Springer, Singapore, pp. 127–135.