DOI:10.34918/87331

Maciej DŁUGOSZ[1,*]

Chapter 5. DNA SEQUENCING READS CORRECTION EVALUATION: REAL VS SIMULATED READS

5.1. Introduction

Techniques of a DNA sequencing give an opportunity to achieve a variety of scientific goals. The process typically is performed with automated machines called sequencers, resulting in a set of short DNA sequences called reads. The reads undergo a further processing, including matching them mutually in a task called *de novo* assembly to obtain long sequences called contigs. The aim is to generate the contigs possibly well-resembling sequences of the sequenced chromosomes. Another task is reads mapping, i.e. aligning them to a previously known sequence of the genome (reference genome). Mapping may be a part of workflows aiming at genetic disorders analysis, genome variants calling, identifying functional elements in genomes, and many others.

This work focuses on reads generated with Illumina sequencers, as it is one of the most popular technique. The Illumina reads are characterized by short length and relatively low amount of errors appearing in a sequencing process: some part (about a few percent, the fraction is called an error rate) of sequences symbols are incorrectly altered in reads by another ones. Such changes are called substitution errors, and are typically tens times more frequent than errors of another type, called insertions and deletions (indels). They include situations, when (typically single) symbols additionally appear in the read sequences or, respectively, they are missing. These errors tend to degrade the quality of downstream analysis. Consequently, a number of algorithms

[1] Department of Algorithmics and Software, Faculty of Automatic Control, Silesian University of Technology, Gliwice, Poland.
[*] Corresponding author: maciej.dlugosz@polsl.pl.

(correctors) aiming at detection and correction of the errors were devised. Detection can be facilitated with Phred quality scores, associated with the read symbols, however, in a step of the correction, all the algorithms utilize a data redundancy in the read. That follows, that the total length of the reads typically is significantly bigger than the length of the genome, hence one can expect, that a specified fragment of the genome is represented in multiple reads. The ratio of the lengths is called sequencing depth.

The introduction of a new corrector or performing a comparative analysis of the existing ones require choosing a correction evaluation method. One of the utilized strategies is simulating reads *in silico* and performing the correction of artificial reads. The other one is correcting reads obtained in real sequencing experiments. An emerging question concerns a concordance of these methods conclusions. In this work we focus on comparing correctors quality results obtained with three methods: two ones based on reads simulation and the one based on real reads. Our strategy is similar to the one adapted in a work [1], where it was used to evaluate correctors, however, now we observe another set of correctors, some correctors and an auxiliary assembler are of newer versions, we utilize another tool for reads simulation and obtain results for a wider range of sequencing depths. Moreover, here we consider strictly the comparison of the evaluation methods, rather than correctors evaluation.

5.2. Reads correction evaluation methods

In the literature, there are two main approaches of reads correction efficacy considered:

- direct, utilizing reads simulated with a computer by random probing a given reference genome and introducing errors; the method allows for simply comparing of corrected and erroneous reads with the original sequence and observing changes introduced by the corrector,
- indirect, utilizing reads obtained from real sequencing experiments, which are corrected and then passed as an input for another algorithms, aiming at acheiving some specific goals, as *de novo* assemblers, reads mappers or genome variant calling pipelines, and observing the results quality of such downstream processes.

5.2.1. Reads simulation methods and rationale

Utilizing simulated reads has several advantages over real ones:

- the position in the genome, where the real read originates from is not known [2],
- the period between the sequencing experiment and publishing the reads in a database can be long [2],
- a simulation allows reads with the specified characteristics to be obtained [2],
- real reads can be “contaminated” with another organisms sequences [3],
- a cost of a reads simulation is lower [3].

A simple method, based on Phred quality scores allows to simulate reads as follows. Given the reference genome and a real reads set (profile set), the sequences of the quality scores from the profile set are extracted. Then the probabilities of bases (reads symbols) substitution errors are computed by decoding the scores. Next, subsequences from the random positions of the reference genome are extracted. The subsequences are of the length equal to the profile reads, and they constitute the output reads set. Finally, the read sequences are disturbed by generating the errors with the obtained probabilities. As a result, a reads file is generated, containing the subsequences with introduced errors and accompanying quality score sequences from the profile set. The method is popular as an *ad hoc* strategy, as in [4]. In [5] it is enhanced by including different probabilities of various subtitution errors, e.g. the probability of substituting a symbol A with C is higher than with G or T.

Another approach is to engage specialized simulation tools. A group of them was presented in an overview work [6], but with no experimental veryfication. One of the tools is ART [7]. It was designed to simulate reads of a few types, including the ones from Illumina sequencers. Authors included a set of generating read models (error profiles). The profiles correspond to rather outdated versions of sequencers, nevertheless, ART allows to simply generate own error profiles based on given profile sets, hence its utilization is still adequate. Another tool, SInC [8], has a built-in model to simulate only Illumina reads of length 100 bp (base pairs). It allows to generate sequencing reads, but the tool is rather desinged to simulate genome variants, which manifest itself in the reads similarly to errors, however, they definitely are not the same. Both of them are a differences with the reference genome, but the errors differs also with the genome being sequenced. CuReSim [9] was developed to simplify a reads mappers evaluation. It has a few predefined error probability distributions, can be parametrized with real error scores and a read length, and is designed to generate reads of any type. FASTQsim [10], as the authors stated, also would be capable to simulate reads of any techniques, as it is equipped with a tool for real reads analysis and may be parametrized with a read length, substitution probability distribution, indel length distribution, etc.

Regardless of the simulation method, the evaluation is in general similar: simulated (erroneous) reads with the errors are corrected and the resulting (corrected) reads are compared with the ones devoid of the errors (proper ones, also returned by the simulator), which allows to simply measure number of errors before and after the correction, including number of errors introduced by the corrector. The comparison may be performed in two granulation modes: in terms of single symbols (a single base error correction is treated as a success), or of entire reads (the entire read must contain just proper symbols to be successfull).

The correction measure typically is defined as

$$\mathrm{gain} = \frac{\mathrm{TP} - \mathrm{FP}}{\mathrm{TP} + \mathrm{FN}} \tag{1}$$

where:
TP – number of symbols/reads perfectly corrected
FP – number of symbols/reads originally correct, but disrupted by the corrector
FN – number of uncorrected or miscorrected symbols/reads

The value of gain is negative, when a number of disruptions is higher than a number of effective corrections.

5.2.2. Comparison with real reads

Correctors evaluation can be performed by correcting a set of real reads and processing the output by another algorithm. In such a situation, a quality of that algorithm is measured. One of the most useful, however not straighforward method, is apdapting a *de novo* assembler. The output quality, a set of obtained contigs, typically is rated with a set of measures, i.a.: N50, NG50, NGA50, genome coverage (cov), number of missassemblies (misasm), or LGA50 values. N50 is such a contig length, that the total length of the contigs of length N50 or more constitutes half of all the contigs length. NG50 is such a contig length, that the total length of the contigs of length NG50 or more constitute a half of the entire genome. NGA50 is similar to NG50, but before rating the contigs are splitted in places of assembly errors. Genome coverage (or genome fraction) is the percentage of contig bases aligned to the reference genome, whereas number of missassemblies refers to wrong alignments of the contigs positions.

LGA50 is defined similarly to NGA50, however, it is a number of contigs rather than the length. Number of misassemblies and LGA50 indicate better results when are lower, the other ones – when higher.

5.3. Results and discussion

Table 1 contains reference genomes identifiers, whereas table 2 contains accession numbers of the real read sets used in the experiments. To simulate reads and evaluate *de novo* assembly results, we chose three model organisms: *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Musa acuminata*. We simulated reads of various characteristics: read lengths, error rates, and sequencing depths, with two methods: Phred and ART, both based on predefined profile sets, accordant to the intended read length and error rate. The choice of the ART is motivated by possibility to generate a profile with a given profile reads set. It allows the output read characteristics to be flexibly configured with an expected length and error rate. In a context of the abovementioned arguments, the potentially competing tool for ART would be not discussed Mason 2 [2], however, due to technical problems, described in [11], we will not take it into account. Moreover, in the overview work [6] ART was stated as the best simulator.

The results were compared with a granulation of the entire reads, due to a higher usefullness of the "completely" corrected reads [12] and the reason stated in a supplementary material of [13]: a small part of Illumina reads have a huge number of errors, hence such reads can strongly affect the whole rate. Moreover, expecting the entire reads to be error-free is a stricter requirement.

The real reads evaluation was performed by assembling them with Minia [14] and achieving the aforementioned measures with Quast [15]. In experiments we have included the correctors presented in Table 3.

Table 1

Reference genomes

Organism	Genome length [Mbp]	Accession number
S. cerevisiae	11.63	GCF_000313865.2
C. elegans	102.8	GCF_000002985.6
M. acuminata	461.5	GCF_000313855.2

Table 2

Read sets accession numbers

Accession no.	Read length [bp]	Depth	Application
ERR422544	100	41	Real reads for *S. cerevisiae*
SRR543736	101	57	Real reads for *C. elegans*
ERR204808	108	16	Real reads for *M. acuminata*
DRR031158	100	–	Profile set for $p \approx 2\%$
SRR1802178	150	–	Profile set for $p \approx 2\%$
SRR065390	100	–	Profile set for $p \approx 4 – 5\%$
SRR650760	151	–	Profile set for $p \approx 4 – 5\%$

Table 3

Correction algorithms

Algorithm	Version	Paper	Algorithm	Version	Paper
RECKONER	2.0	[16]	Lighter	1.1.2	[21]
Musket	1.1	[4]	BFC	BFC-ht, version r181	[13]
RACER	–	[17]	Karect	1.0	[22]
BLESS	1.02	[18]	SAMDUDE	Uploaded 2 May, 2018	[23]
Fiona	2.4.0	[19]	CARE	2.1	[24]
Blue	1.1.3	[20]			

The reads sets were generated for two lengths (100 bp and about 150 bp), two error rates *p* (about 2% and 4 up to 5%) and three sequencing depths (20×, 30×, 60×), for three different-sized genomes, representing a wide range of possible reads characteristics. The real reads sets were downloaded from a public database ENA, and represent the mentioned three genomes. In a case, when a corrector has to be aparametrized with a *k*-mer length, we run it multiple times and chose the value resulting the best output (for simulated reads we chose the length for a sets of 20× depth).

Figures 1, 2 show, respectively, result for Phred simulation and simulation with ART. Notation L*x*D*y* means a read length *x* [bp] and a sequencing depth *y*. Gain measure values were multiplied by 100. Figures 3, 4 show results for *de novo* assembly of the real reads. As a reference, we added results for not corrected real reads, marked as (raw). With a dagger (†) we marked cases, when a corrector failed due to a timeout (more than 24 hours), with a section mark (§), when it crashed, and with an asterisk (*), when it returned a message about wrong quality scores (BLESS). For *M. acuminata* measures referring to a genome size (NG50, NGA50, LGA50), due to a low total contigs lengths, are unavailable.

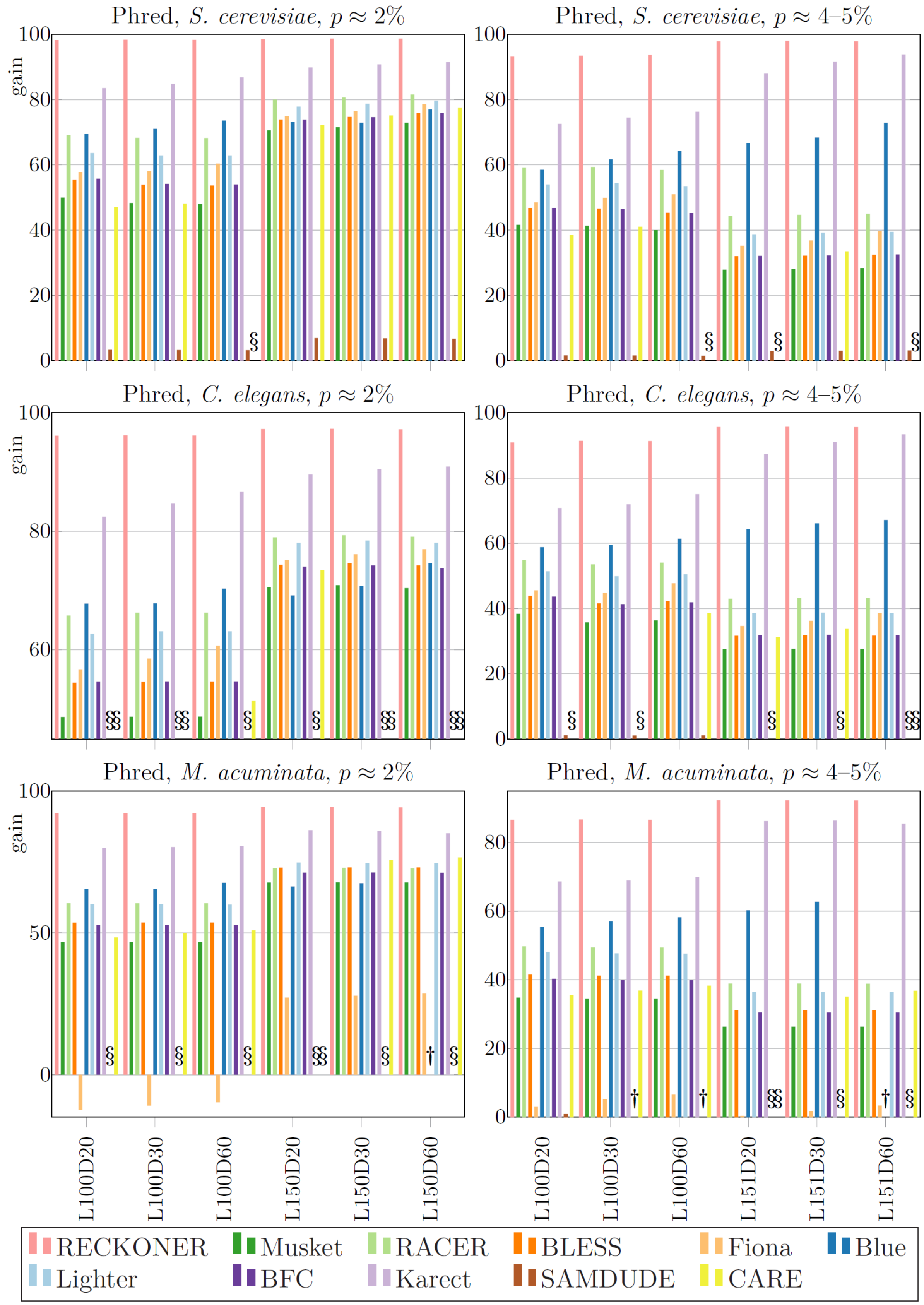

Fig. 1. Results for reads simulated with a Phred method
Rys. 1. Wyniki dla odczytów symulowanych metodą Phred

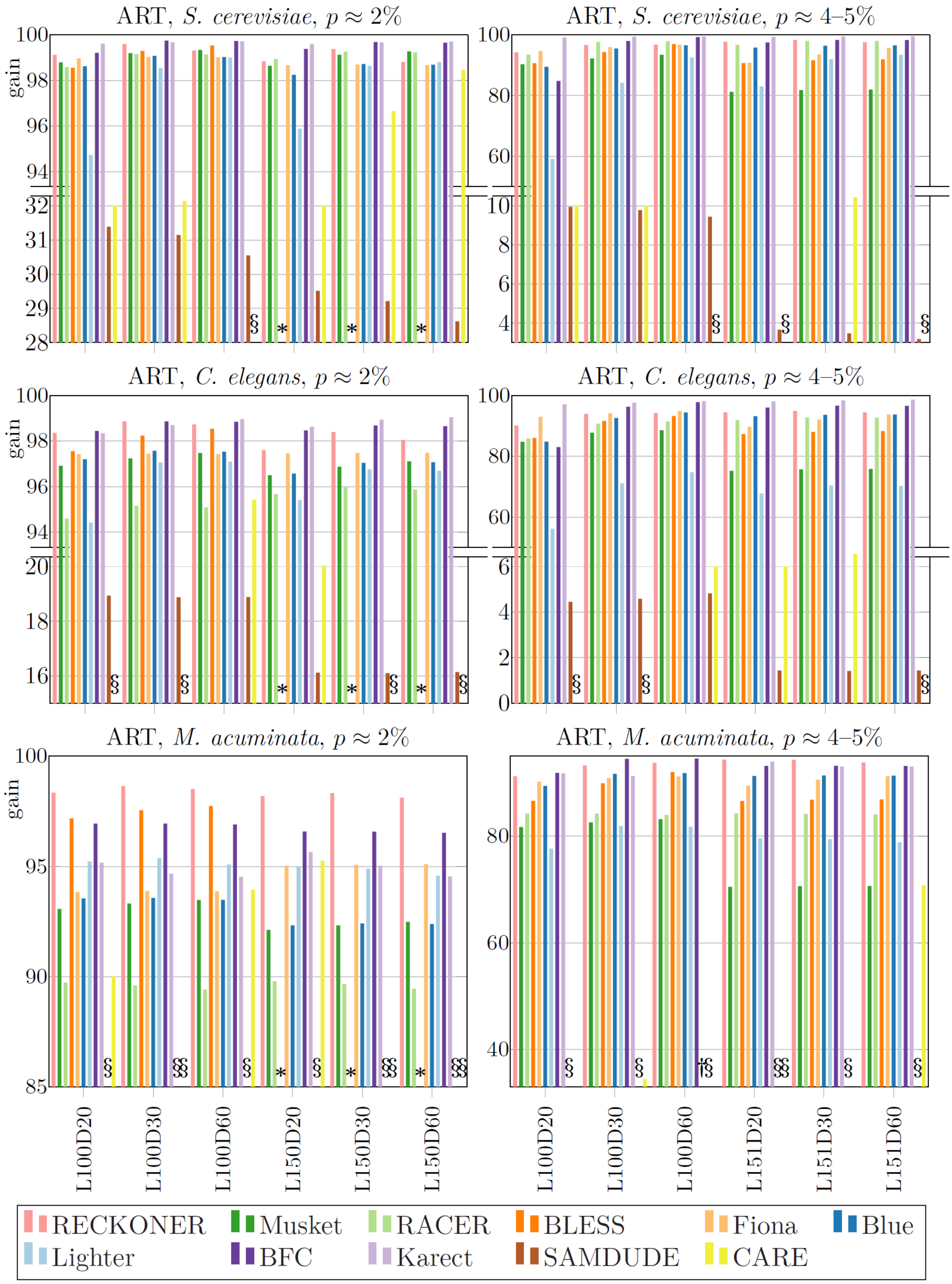

Fig. 2. Results for reads simulated with ART
Rys. 2. Wyniki dla odczytów symulowanych narzędziem ART

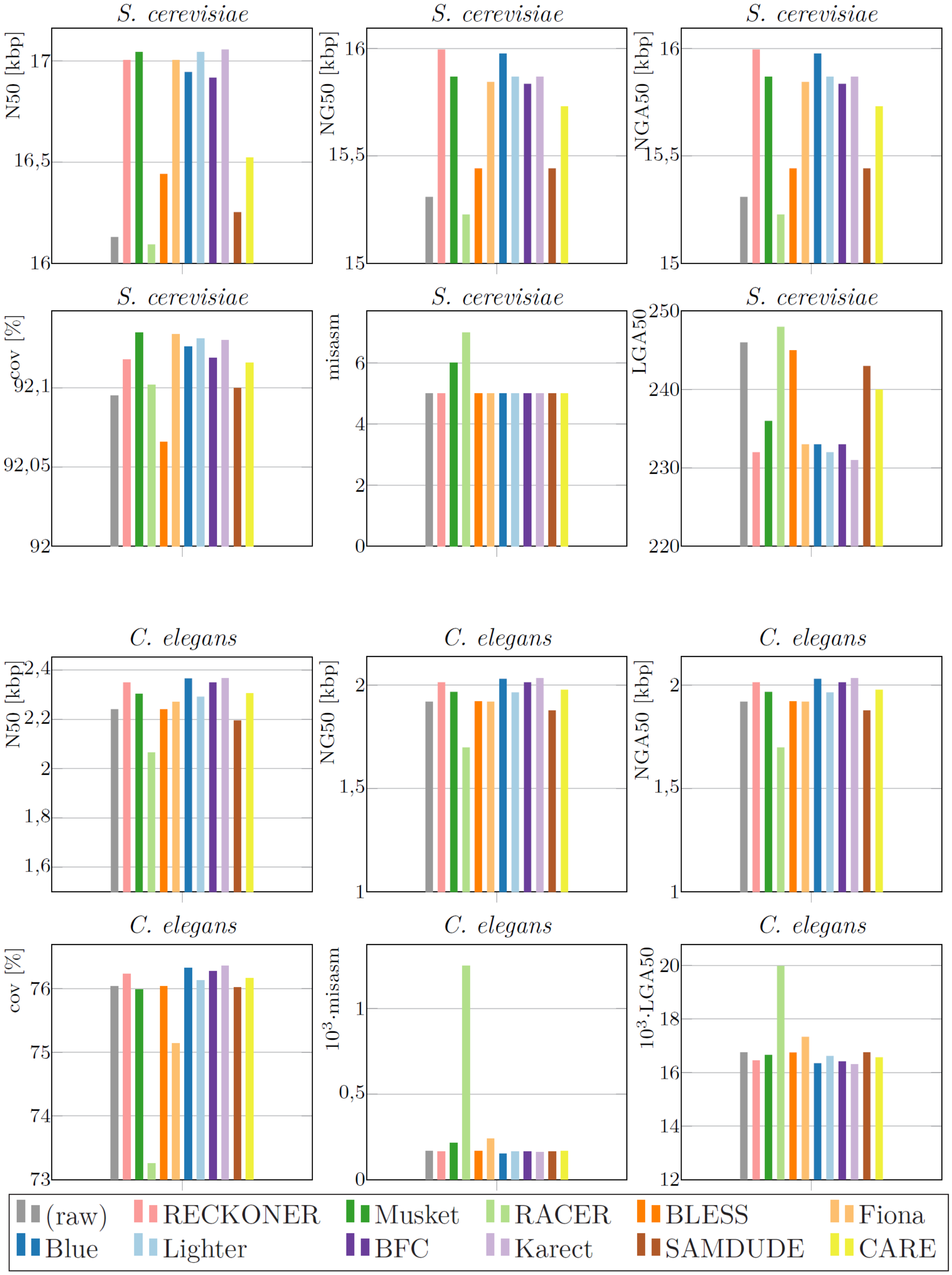

Fig. 3. Results for *de novo* assembly of the real reads
Rys. 3. Wyniki asemblacji *de novo* odczytów rzeczywistych

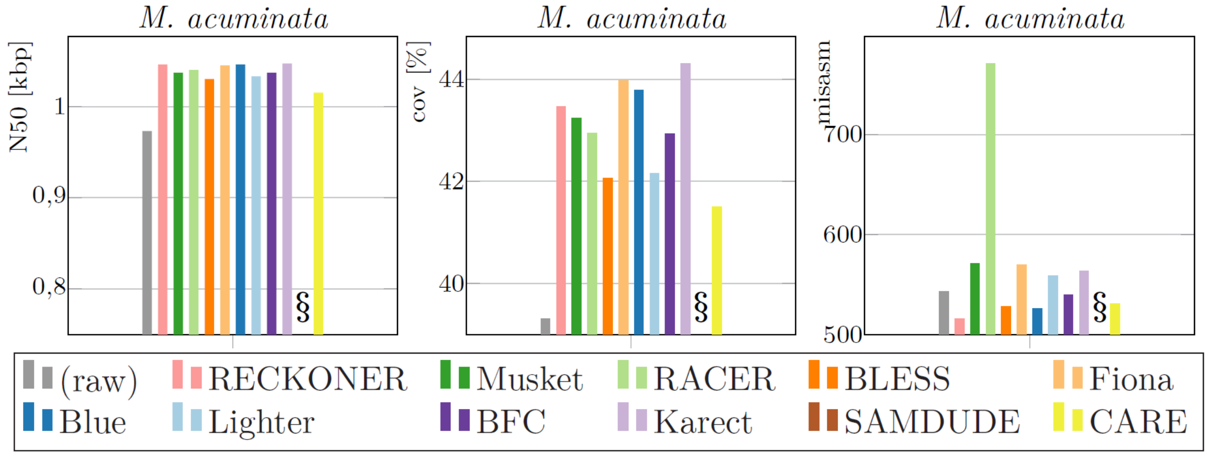

Fig. 4. Results for *de novo* assembly of the real reads (cont'd)
Rys. 4. Wyniki asemblacji *de novo* odczytów rzeczywistych (cd.)

The results for Phred simulation show, that RECKONER is extremely efficient and in most of the cases it allows to correct over 90% of erroneous reads, in some cases achieving almost 100%. Another good algorithm is Karect, and as a third appears Blue. RACER and Lighter achieve similar, average results. BLESS and BFC should be rated rather poorly, but clearly negatively outstanding algorithms are Musket, CARE and, especially, SAMDUDE. For the third genome very weak results was obtained by Fiona, which in three cases caused a general data degradation. An interesting observation is a read length impact. In the case of good-quality reads ($p \approx 2\%$) for longer reads better results were obtained. In a case of the other reads ($p \approx 4$–5%) the trend is inverted.

In the case of ART reads the general observation is that the results are more even, both between different genomes and algorithms. The correctors ranking is more varied, depending on the cases. Typically, the best is one of BFC or Karect, however, for *M. acuminata* reads Karect results are out of the top and usually the best one is RECKONER. Moderated results are obtained for BLESS and Blue. There are no cases, when Musket or Fiona are severely outstanding. The weakest algorithms are Lighter, sometimes RACER, SAMDUDE and CARE. The observed read length impact is different and more subtle. For the reads of both error rates the longer ones cause lower quality, but in some cases such phenomenon is not visible.

The relative values of *de novo* assembly evaluation, performed for the real reads corrected with different algorithms vary slightly, which is an effect of some assembler tolerance for the sequencing errors and not a straightforward transfer of the errors to the contigs quality. It seems, that the best algorithm is Karect, followed by RECKONER and Blue, however the top varies, depending on the genome and a measure. For example, for *S. cerevisiae* the top algorithms change in terms, respectively, N50 and

NG50. In general, Musket, Fiona, Lighter, BFC characterise by moderated results. The poorest algorithms are RACER, BLESS, CARE, and SAMDUDE. All of them have weak values of contigs length measures, but additionaly RACER has huge numbers of missassemblies.

To summarize, in many cases a relative and an absolute correctors qualities differ for various evaluation strategies. We will list the most visible cases. RACER achieves an average results in a case of Phred reads, rather weak for ART reads, however in a case of *de novo* assembly, the results are definitely bad. Results for BLESS are weak for the Phred and average for the ART simulated reads, but for the real ones are very poor. Fiona achieves very weak results for *M. acuminata* Phred simulated reads, but for the simulated with ART and the real ones the quality does not seem to be negative. For the Phred simulated and the real reads BFC is classified as an average algorithm, in contrast to the ones generated with ART, where it is one of the best. On the other hand, in some cases all the comparisons yielded similar conclusions. RECKONER is one of the best one in all the cases (except for the Phred reads, where it advantageously outstands), similarly to Karect. SAMDUDE and CARE in all the cases achieve weak results.

5.4. Conclusions

The experiments showed, that the correction algorithms evaluation results differ depending on the adopted strategy. A simple Phred method results are dispersed depending on the chosen corrector, probably privileging the ones accepting idealized conditions of the method. Results for the reads generated with ART are more levelled, but still does not clearly correspond to the real reads. On the other hand, the outcome for real reads undergo an impact of the addtional step of *de novo* assembly and inconsistent results for a single reads set may be obtained. Moreover, the results of the various algorithms differ just slightly.

In contrast, the comparison of methods showed, that in some cases the results are consistent. In all the strategies RECKONER and Karect are in a group of the best algorithms, while SAMDUDE and CARE almost always are the weakest, but these situations are not a clear rule. The observations give no basis to conclude, which method of simulation is more concordant with an indirect method of *de novo* assembly utilization.

Acknowledgements

The work was performed using the infrastructure supported by POIG.02.03.01-24-099/13 grant: “GeCONiI – Upper Silesian Center for Computational Science and Engineering”.

Bibliography

1. M. Długosz, S. Deorowicz, RECKONER: read error corrector based on KMC, *Bioinformatics* (2017) **33(7)**:1086–1089.
2. M. Holtgrewe, Mason: a read simulator for second generation sequencing data, Technical Report FU Berlin (2010).
3. A. Shcherbina, FASTQSim: platform-independent data characterization and in silico read generation for NGS datasets, *BMC research notes* (2014) **7(1)**:1–12.
4. L. Yongchao, J. Schröder, B. Schmidt, Musket: a multistage k-mer spectrum-based error corrector for Illumina sequence data, *Bioinformatics* (2013) **29(3)**:308–315.
5. D.R. Kelley, M.C. Schatz, S.L. Salzberg, Quake: quality-aware detection and correction of sequencing errors, *Genome biology* (2010) **11(11)**:1–13.
6. M. Escalona, S. Richa, D. Posada, A comparison of tools for the simulation of genomic next-generation sequencing data, *Nature Reviews Genetics* (2016) **17(8)**:459.
7. W. Huang, L. Li, J.R. Myers, G.T. Marth, ART: a next-generation sequencing read simulator, *Bioinformatics* (2012) **28(4)**:593–594.
8. S. Pattnaik, S. Gupta, A.A. Rao, B. Panda, SInC: an accurate and fast error-model based simulator for SNPs, Indels and CNVs coupled with a read generator for short-read sequence data, *BMC bioinformatics* (2014) **15(1)**:1–9.
9. S. Caboche, C. Audebert, Y. Lemoine, D. Hot, Comparison of mapping algorithms used in high-throughput sequencing: application to Ion Torrent data, *BMC genomics* (2014) **15(1)**:1–16.
10. A. Shcherbina, FASTQSim: platform-independent data characterization and in silico read generation for NGS datasets, *BMC research notes* (2014) **7(1)**:1–12.
11. M. Długosz, S. Deorowicz, M. Kokot, Improvements in DNA Reads Correction, *International Conference on Man-Machine Interactions, October 3–6, 2017, Kraków* (2017):115–124.
12. M. Molnar, L. Ilie, Correcting illumina data, *Briefings in bioinformatics* (2015) **16(4)**:588–599.

13. H. Li, BFC: correcting Illumina sequencing errors, *Bioinformatics* (2015) **31(17)**:2885–2887.
14. R. Chikhi, G. Rizk, Space-efficient and exact de Bruijn graph representation based on a Bloom filter, *Algorithms for Molecular Biology* (2013) **8**: 1–9.
15. A. Gurevich, V. Saveliev, N. Vyahhi, G. Tesler, QUAST: quality assessment tool for genome assemblies, *Bioinformatics* (2013) **29(8)**:1072–1075.
16. M. Długosz, S. Deorowicz, Illumina reads correction – evaluation and improvements, *In Review* preprint (2023), doi:10.21203/rs.3.rs-2715541/v1.
17. L. Ilie, M. Molnar, RACER: Rapid and accurate correction of errors in reads, *Bioinformatics* (2013) **29(19)**:2490–2493.
18. Y. Heo, A. Ramachandran, W.M. Hwu, J. Ma, D. Chen, BLESS 2: accurate, memory-efficient and fast error correction method, *Bioinformatics* (2016) **32(15)**: 2369–2371.
19. M.H. Schulz, *et al.*, Fiona: a parallel and automatic strategy for read error correction, *Bioinformatics* (2014) **30(17)**:i356–i363.
20. P. Greenfield, K. Duesing, A. Papanicolaou, D.C. Bauer, Blue: correcting sequencing errors using consensus and context, *Bioinformatics* (2014) **30(19)**:2723–2732.
21. L. Song, L. Florea, B. Langmead, Lighter: fast and memory-efficient sequencing error correction without counting, *Genome biology* (2014) **15**:1–13.
22. A. Allam, P. Kalnis, V. Solovyev, Karect: accurate correction of substitution, insertion and deletion errors for next-generation sequencing data, *Bioinformatics* (2015) **31(21)**:3421–3428.
23. I. Fischer-Hwang, I. Ochoa, T. Weissman, M. Hernaez, Denoising of aligned genomic data, *Scientific reports* (2019) **9(1)**:1–11.
24. F. Kallenborn, J. Cascitti, B. Schmidt, CARE 2.0: reducing false-positive sequencing error corrections using machine learning, *BMC bioinformatics* (2022) **23(1)**:1–17.

DNA SEQUENCING READS CORRECTION EVALUATION: REAL VS SIMULATED READS

Abstract

The problem of Illumina whole genome sequencing reads correction is widely considered in the literature. Nevertheless, new correction algorithms are still introduced, requiring reliable methods of the process efficiency evaluation to be available. The methods include utilization of both reads simulated *in silico* and real ones, obtained from DNA sequencing process. The paper focuses on comparing these approaches, trying to give the answer, if a convenient reads simulation allows the correction quality to be reliably verified. We evaluated a set of correction algorithms by processing reads generated with a simple method based on a Phred scores utilization and simulated with a specialized tool. We also *de novo* assembled sets of reads obtained from real sequencing experiments, observing an impact of the correction to the contigs characteristics. Finally, a concordance of the three methods results was analyzed.

Keywords: DNA sequencing, Illumina, read error correction, reads simulation, genome assembly