Maciej DŁUGOSZ*[1]

Chapter 8. GENOME VARIANT CALLING IN CONTEXT OF SEQUENCING READS CORRECTION

8.1. Introduction

Illumina DNA sequencers produce a huge amount of data in lowering costs [16] Such a data has a form of short (about 100–150 bp) sequences called *reads*. A number of reads applications is long: cancer mutation discovery, genetic disorders analysis, genome *de novo* assembly, and many others. Data processing have to cope with many of technical problems, including huge data size, not uniform reads distribution within a genome, lack of reads covering (i.e. originating from) some of the genome regions, reads shorter than many of repeating fragments of a genome, and presence of sequencing errors.

The last problem is partially solved thanks to some built into algorithms strategies of errors tolerance. Besides, there was developed a group of specialized algorithms (correctors) aiming at the detection and elimination of the errors. The efficacy of those is often experimentally evaluated by simulating sequencing *in silico* or observing the impact of the correction on solving typical problems, like reads mapping. However, as the author knows, there is no thorough comparative analysis of such impact on real, full pipes of data processing. There were some tiny trials, but they were limited to two human chromosomes and generating simple statistics [1], or counting false-positive variants for simple model organism *C. elegans* [2]. More advanced experiments was presented in [3], however, the algorithm introduced therein, Coval, is a mixture of methods improving reads mapping, so it is not a correction algorithm *sensu stricto*. Moreover, the paper does not involve a comparison of other algorithms.

* Corresponding author: maciej.dlugosz@polsl.pl, Akademicka 2A, 44-100 Gliwice, PL.

[1] Department of Algorithmics and Software, Silesian University of Technology.

As variant calling (VC) is a crucial process in clinical genetic testing [4], it is necessary to scrupulously explore all of its aspects. In this paper, the author shows experimental results of the first thorough evaluation of different correction algorithms impact on variant calling utilizing Illumina sequencing reads.

The experiments were performed with reads of two full genomes. A number of correctors were chosen to test various ideas of a correction. The corrected reads were utilized to perform the entire short VC process. Resulting variants sets were compared to a ground truth sets to measure the VC quality and, indirectly, the correction quality. Such an approach is a novel method of error correction algorithms evaluation. The obtained results give an overview on existing correctors quality and verify reasonability of correction utilization in VC application.

8.2. Sequencing data processing

8.2.1. Error correction

Typically, algorithms deployed to correct reads errors exploit, i.a., redundancy of the sequencing data. The redundancy level is measured with *sequencing depth*, which is defined as a ratio of the sum of all reads lengths and the genome length.

Error correction efficacy and strategies are varied. In a paper [5] correctors were categorized into three types: (*i*) *k-spectrum-based*, (*ii*) *suffix tree/array-based*, (*iii) multiple sequence alignment-based*. Also a type (*iv*) *hidden Markov-model-based* was proposed [6]. Moreover, some of the algorithms fulfill traits of different groups (*hybrid* algorithms), and some of them are difficult to include to any of them.

8.2.2. Variant calling

The process of VC is a multi-stage task. Typically, it includes mapping of the reads to the reference genome, which is defined as aligning reads to the genome fragments, which sequences characterise high similarity to those reads, which suggests, that the reads originate from that fragments. Existing differences between reads and the genome fragment sequences can be caused by variants present in the sequenced genome or by other factors, like sequencing errors.

Reads mapping is a general task of different reads processing pipelines. In the context of VC it is followed by specialized processes responsible for proper variant determining, filtering, or annotating. In one of the simplest situations as a result two types of variants are obtained: single nucleotide variants (SNVs), which are differences of single genome bases, and short indels, which are leaks of short sequences in the proband genome or additional sequences appearing there. Evaluation of VC quality is possible with the utilization of available for some organisms *ground truth* variant sets, which represent our best knowledge about variants present in the sample.

8.3. Methods and data

8.3.1. Experimental data

The experiments aimed at determining the impact of Illumina reads correction on the quality of variant calling results. The experiments performed on a few sets of reads. It has to be emphasised, that availability of variant ground truth sets is limited. One of such sets for human is the oft-used Genome in a Bottle [7], however, the choice of the second organism posed a problem due to a leak of publicly available sets. Finally, *Arabidopsis thaliana* was selected, which is a model organism with ground truth sets introduced as a result of 1001 Genomes Project [8]. In the case of human, there were also available confident call regions file, which was utilized in the experiments.

The following reads sets were used: SRR1945754 of sequencing depth ca. 180× for *A. thaliana*, and ERR174324, ERR174325, ERR174326 for human, each of the depth ca. 15×. To parametrize the sequencing depth, the data was prepared as followes:

- for *A. thaliana* pairs of reads were randomly shuffled, and the subset of the pairs was extracted to achieve depths of 30×, 60×, 90×,
- for *H. sapiens* three sets of reads pairs were used: ERR174324, concatenation of ERR174324 and ERR174325 and concatenation of all of three ones; such an approach was possible due to performing sequencing on the same sample, with the same machine etc.; it allowed to obtain sets of depths ca. 15×, 30×, 45×.

8.3.2. Error correction algorithms

Due to a plenty of existing algorithms, the limited set of them was selected, treating as criteria of choice being in the group of newest, most popular, or characterising of best quality noticed by other authors. The highest popularity of (*i*) algorithm type legitimizes its overrepresentation, whereas (*iv*) is represented only by one algorithm (PREMIER [9]) and as it is not publicly available, it was not analysed in the experiments. Fiona [10], initially appointed to experiments, representing the hybrid solution of groups (*ii*) and (*iii*), was also omitted, as its resulting reads had quality indicators decreased to the level, that caused VC returning no results. Table 8.1 presents correction algorithms compared in the experiments.

Table 8.1

Correction algorithms

Algorithm	Type	Version	Describing paper
RECKONER	(*i*)	1.2	[11]
Musket	(*i*)	1.1	[12]
RACER	(*i*)	–	[13]
BLESS	(*i*)	1.02	[6]
Blue	(*i*)	1.1.3	[14]
Lighter	(*i*)	1.1.2	[15]
BFC	(*i*)	BFC-ht, version r181	[2]
Karect	(*ii*)	1.0	[16]
SAMDUDE	others	Uploaded 2 May, 2018	[1]

The majority of the algorithms need parametrization. The mostly required parameter is the oligomer (*k*-mer) length *k*. It was determined by performing a sequence of corrections with different odd *k* values and, if the correction succeeded, followed by VC. The value of *k* resulting in the best value in terms of F1-score was chosen. Actually, some of the algorithms have more sophisticated methods of the parameter determination, however, they are rather rules-of-thumb (e.g. BLESS needs to get *k* maximising the number of correcting changes in reads and such that the number of *k*-mers in data, speculated to represent correct sequences, be in some range, depending on a genome length). Choosing the best *k* gives insight into the potential of the algorithms. As *k*-mer length determining is an interesting problem itself, some of its results were observed.

The other parameters were genome length, sequencing depth (which ones were determined with information from a database [17]), number of reads *k*-mers (which was determined by a tool for *k*-mers counting KMC [18]), ploidy for Karect (haploid for *A. thaliana* and diploid for *H. sapiens*). For Lighter probability α, accordantly with authors guidelines, value 3.5/(sequencing depth) was chosen. Musket and Blue required so-called cutoff threshold and it was determined in a way, as RECKONER does.

All the input reads were paired. For correctors not supporting paired reads, the paired files were concatenated (by attaching a file of the latter reads to the end of a file of the first reads), then corrected, and finally split in a site of the concatenation.

8.3.3. Variant calling pipeline and its evaluation

To perform VC the reads were mapped with BWA[19] and variants were called with Strelka [20]. To evaluate the VC quality tool hap.py [21] was used. It returns statistics of variants set: TP, FP, FN, meaning as follows: number of detected true positive, false positive, and false negative (missing) variants. To compare results, it is convenient to define sensitivity (as TP/(TP+FN)), precision (as TP/(TP+FP)), and F1-score as their harmonic mean. All the values are independently defined for SNVs and indels.

8.4. Results

The experiments were performed on a server equipped with 256 GB of RAM and two Intel Xeon E5-2670 v3 processors, 12 cores (24 threads) each. The timeout limit of correction was set to 24 hours for every run.

8.4.1. Variant calling results

Fig. 8.1 shows the quality of VC for *A. thaliana*, separately for SNVs and short indels. To give a deeper view into the results, adequate sensitivity and precision graphs were shown. The “(raw)” designation denotes reads with no correction, acting as control cases. Figure 8.2 shows analogical results for *H. sapiens*.

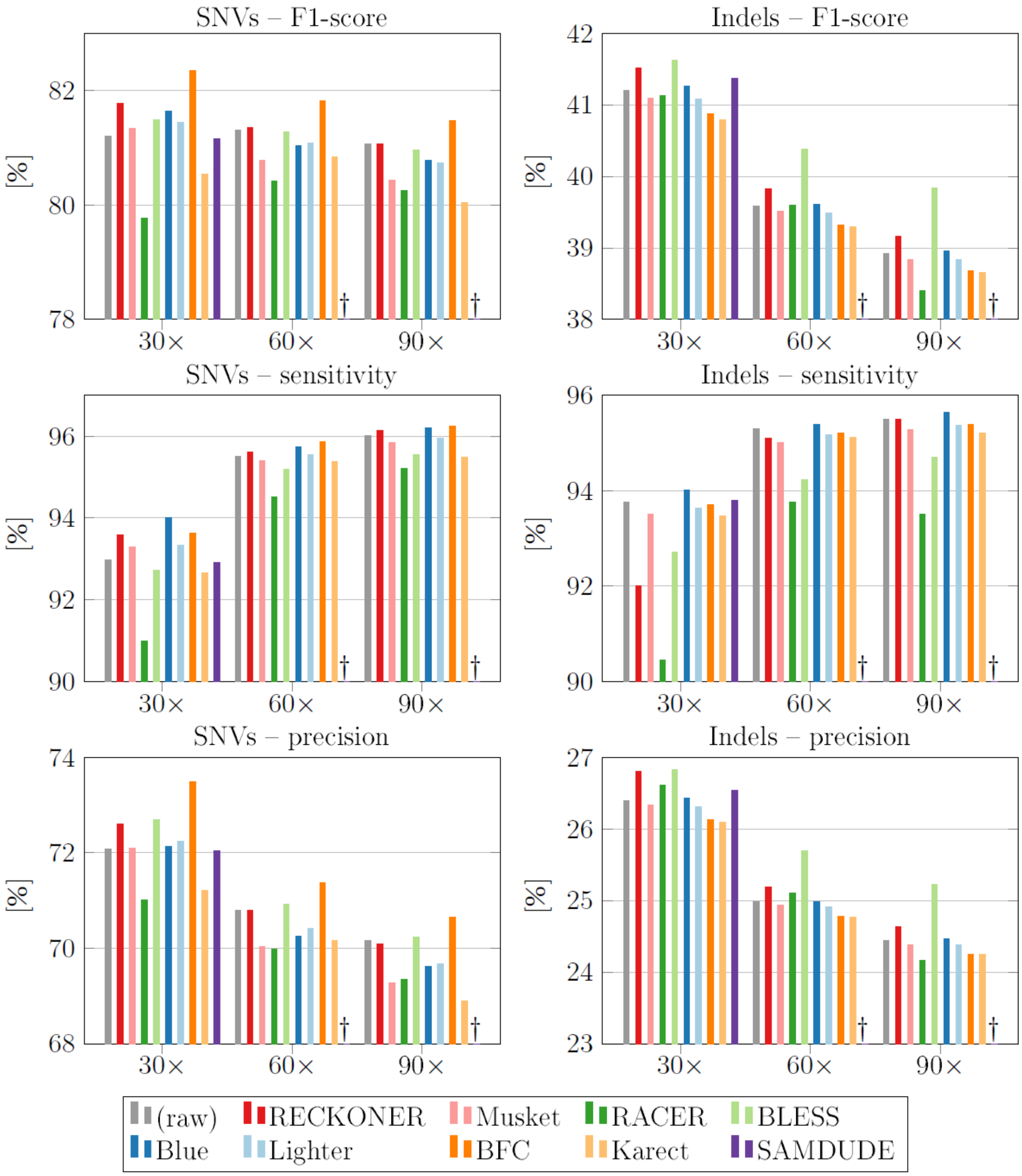

Fig. 8.1. VC results for *A. thaliana*; † – timeout.
Rys. 8.1. Wyniki detekcji wariantów dla *A. thaliana*.

Definitely better results of human VC are caused by the confident call regions utilization. It was observed, then after resignation of that, results were weaker, but the comparison of different correctors was analogical (results not shown).

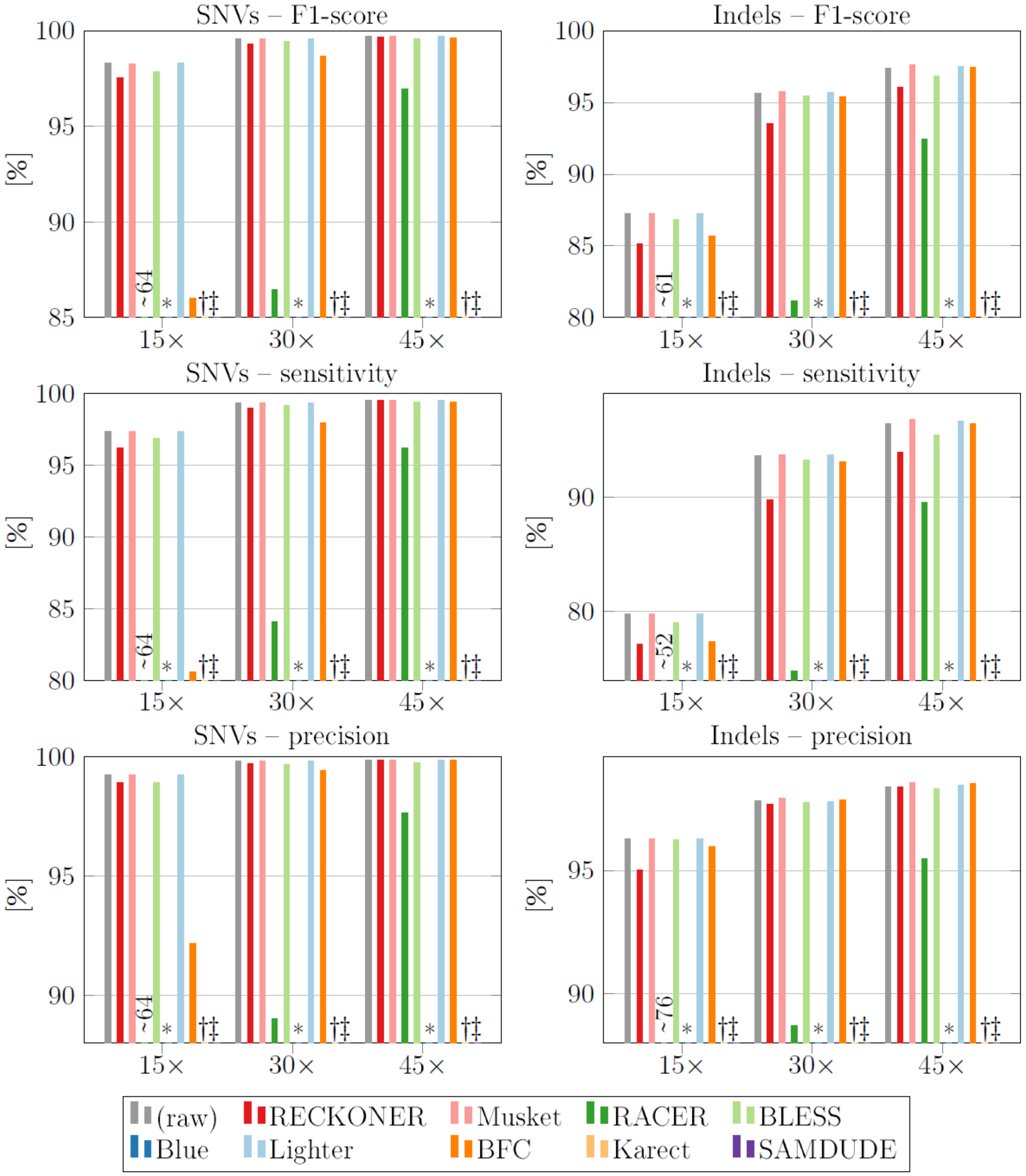

Fig. 8.2. VC results for *H. sapiens*; for selected RACER results numbers are given instead of the bars; ✻ – lack of memory; † – timeout; ‡ – algorithm crashes

Rys. 8.2. Wyniki detekcji wariantów dla *H. sapiens*; dla wybranych wyników algorytmu RACER zamieszczono liczby zamiast słupków

For all the *A. thaliana* experiments a low precision was observed. In many cases the correction caused increase of the quality, however, some of the algorithms caused its decline. The best algorithms was BFC (for SNVs) and Musket (for indels). Interesting is, that increasing the sequencing depth caused significant precision decline.

In the case of *H. sapiens*, Musket and Lighter return results of the same quality as the uncorrected data, what suggests, they actually do not correct the reads. They were run with a small value of k=13, as bigger values resulted in even weaker quality. Unfortunately, in most of the other cases the results were worse than before the correction. It means, that the correction destroyed the reads. Especially, the decrease of quality is visible for sensitivity, what means, that less number of variants is detected; as in most of the cases precision did not changed significantly, correction did not caused appearing a non-existent variants. For this organism increasing sequencing depth typically caused quality improvement in terms of all the measures.

It is noticeable, that RACER worked poorly for low sequencing depth reads. SAMDUDE finally was able to correct only the smallest set of the reads. As Blue crashed in all the human reads sets, its results are available only for *A. thaliana*.

8.4.2. Oligomer length impact

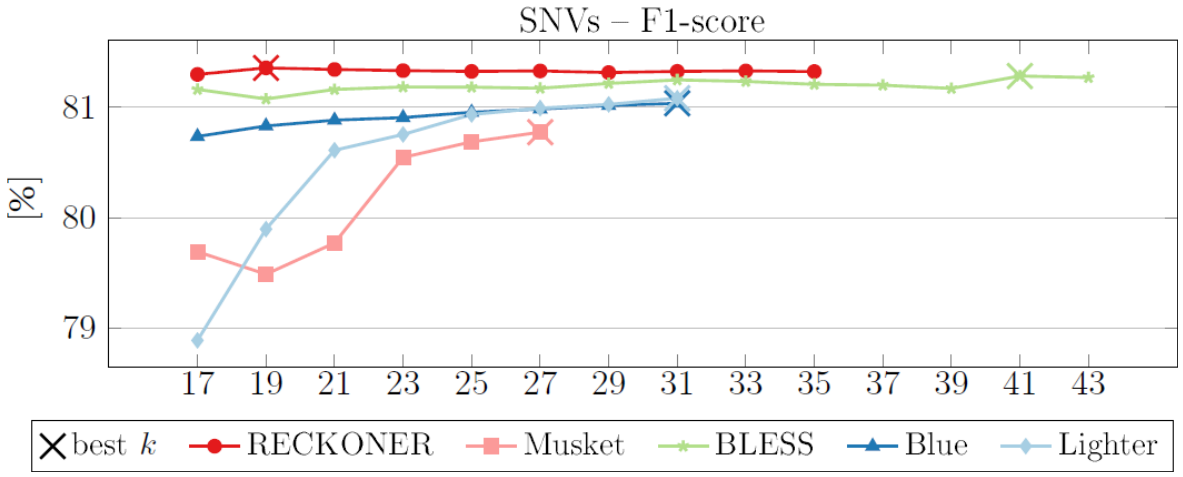

Fig. 8.3. Oligomer length impact on *A. thaliana* VC; sequencing depth 60×
Rys. 8.3. Wpływ długości oligomeru na detekcję wariantów *A. thaliana*; głębokość sekwencjonowania 60×

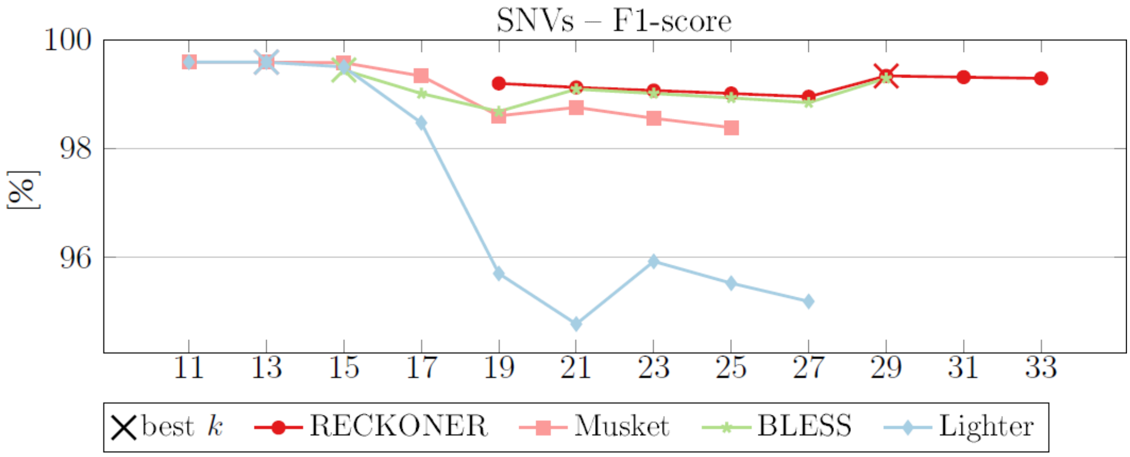

Fig. 8.4. Oligomer length impact on *H. sapiens* VC; sequencing depth 30×
Rys. 8.4. Wpływ długości oligomeru na detekcję wariantów *H. sapiens*; głębokość sekwencjonowania 30×

Fig. 8.3 show the *k*-mer length impact on F1-score for *A. thaliana* and *H. sapiens*, respectively. In some cases, the best value was the bound of the tested *k* values sequence. It means, that the following values caused some problems with its execution. Lighter undergoes the highest impact of *k*-mer length. In the case of Lighter and Musket the aforementioned observation, that they achieve the best results for small *k* values is visible (but they rather do not perform any correction).

8.5. Summary

The results show, that in the case of *A. thaliana* performing error correction is reasoned, as it allows slightly better results to be obtained. However, despite a significant number of correction algorithms and thorough analysis of the *k*-mer length, the differences are not huge, therefore utilizing them has not to be treated as a crucial part of variant calling. Unfortunately, in the case of long, mostly repeated human genome, the correction even causes the results quality decline.

In some cases correction *k*-mer length impact is not significant for variant calling, but in the others is crucial. It means, that it is necessary to put a big effort to determine this value properly.

It has to be in mind, that error correction is an additional stage of a data processing. As experiments prove, some of the algorithms are not able to correct reads in a day, which is rather a liberal limit. Requiring such a long time in practice disqualifies an algorithm.

Bibliography

1. Fischer-Hwang, I. Ochoa, T. Weissman, M. Hernaez, Denoising of Aligned Genomic Data, *Scientific reports* (2019) **9**(1):1-11.
2. H. Li, BFC: correcting Illumina sequencing errors, *Bioinformatics* (2015) **31**(17):2885-2887.
3. S. Kosugi, *et al.*, Coval: improving alignment quality and variant calling accuracy for next-generation sequencing data, *PloS one* (2013) **8**(10):e75402.
4. D.C. Koboldt, Best practices for variant calling in clinical sequencing, *Genome Medicine* (2020) **12**(1):1-13.

5. X. Yang, S.P. Chockalingam, S. Aluru, A survey of error-correction methods for next-generation sequencing, *Briefings in bioinformatics* (2012) **14**(1):56-66.
6. 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.
7. J. Zook *et al.*, Extensive sequencing of seven human genomes to characterize benchmark reference materials, *Scientific data* (2016) **3**(1):1-26.
8. C. Alonso-Blanco *et al.*, 1,135 genomes reveal the global pattern of polymorphism in Arabidopsis thaliana, *Cell* (2016) **166**(2):481-491.
9. X. Yin, Z. Song, K. Dorman, A. Ramamoorthy, *PREMIER–PRobabilistic error-correction using Markov inference in errored reads*, 2013 IEEE International Symposium on Information Theory, 7-12 July 2013, Istanbul (2013).
10. M. Schulz *et al.*, Fiona: a parallel and automatic strategy for read error correction, *Bioinformatics* (2014) **30**(17):i356-i363.
11. M. Długosz, S. Deorowicz, M. Kokot, *Improvements in DNA Reads Correction*, International Conference on Man–Machine Interactions, 3-6 October 2017, Kraków (2017).
12. Y. Liu, J. Schröder, B. Schmidt, Musket: a multistage k-mer spectrum-based error corrector for Illumina sequence data, *Bioinformatics* (2012) **29**(3):308-315.
13. L. Ilie, M. Molnar, RACER: Rapid and accurate correction of errors in reads, *Bioinformatics* (2013) **29**(19):2490-2493.
14. P. Greenfield, K. Duesing, A. Papanicolaou, D.C. Bauer, Blue: correcting sequencing errors using consensus and context, *Bioinformatics* (2014) **30**(19):2723-2732.
15. L. Song, L. Florea, B. Langmead, Lighter: fast and memory-efficient sequencing error correction without counting, *Genome biology* (2014) **15**(11):509.
16. 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.
17. National Center for Biotechnology Information: https://www.ncbi.nlm.nih.gov, accessed: 20 March, 2021.
18. M. Kokot. M. Długosz, S. Deorowicz, KMC 3: counting and manipulating k-mer statistics, *Bioinformatics* (2017) 33(17):2759-2761
19. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform, *Bioinformatics* (2009) **25**(14):1754-1760.
20. S. Kim *et al.*, Strelka2: fast and accurate calling of germline and somatic variants, *Nature methods* (2018) **15**(8):591-594.
21. GitHub - Illumina/hap.py: Haplotype VCF comparison tools: https://github.com/Illumina/ hap.py, accessed: 20 March, 2021.

GENOME VARIANT CALLING IN CONTEXT OF SEQUENCING READS CORRECTION

Abstract

One of the most widely used tool in genomics is DNA sequencing, especially of Illumina technology. Its multiple applications include variant calling, which aims at determining variants present in a proband genome. However, presence of sequencing errors could impact quality of calling results. There has not been performed thorough analysis of error correction efficacy in terms of variant calling. This paper addresses that problem, for different state-of-the-art correction algorithms and two different-sized genomes. Moreover, the impact of oligomer length – main parameter of many of correction algorithms – on results quality is shown.

Keywords: DNA sequencing, variant calling, read error correction.