

Error Correction in Nanopore Reads for *de novo* Genomic Assembly

Jacqueline Aldridge-Águila[1], Diego Álvarez-Saravia[2], Marcelo Navarrete[2], and Roberto Uribe-Paredes[1(✉)]

[1] Computer Engineering Department, University of Magallanes, Punta Arenas, Chile
{jacqueline.aldridge,roberto.uribe}@umag.cl
[2] School of Medicine, University of Magallanes, Punta Arenas, Chile
{diego.alvarez,marcelo.navarrete}@umag.cl

Abstract. The purpose of genome sequencing is to determine the DNA sequence of a given organism. Current sequencing technologies can be classified by the type of output data. Whereas Nanopore technology generates long reads with high error rates, short read technologies - such as Illumina sequencing - generate shorter reads but with low error rate. Since *de novo* genome assembly of sequencing reads is defined as a NP-hard problem, it remains as one of the major challenges for defining reference genomes of different species. This paper aims to improve the quality of reads obtained through Oxford Nanopore Technologies (ONT). We developed an algorithm to associate the reads obtained from Illumina with the ones obtained with Nanopore. Low accuracy ONT reads were corrected with the high quality Illumina reads to achieve an improved sequencing data. The inclusion of this algorithm as a preprocessing step resulted in improved coverage, contig length, and mismatch rate when performing *de novo* genome assembly of a bacterial genome with well known tools.

Keywords: Oxford nanopore · Illumina · De novo assembly · Correction reads

1 Introduction

Genomics has advanced notoriously in recent years, currently there are several technologies allowing the DNA sequencing of different species in a simpler process at a tiny fraction of the cost of 20 years ago [6]. Thanks to these advances many species have been sequenced, but the vast majority still have not gone through the process of defining their reference genome, which empirically implies that their genetic maps remains unknown. In order to obtain biologically meaningful information from sequencing data, unreferenced species must still undergo through a process known as *de novo* assembly.

This work has been partially supported by the MINEDUC under the project MAG1895, and by Conicyt under the project Fondecyt 1180882.

© Springer Nature Switzerland AG 2020
O. Gervasi et al. (Eds.): ICCSA 2020, LNCS 12253, pp. 754–762, 2020.
https://doi.org/10.1007/978-3-030-58814-4_63

De novo assembly aims to reconstruct the genome from the reads obtained in the sequencing process, without having any information aside of the reads.

There is a wide range of technologies to perform DNA sequencing. In this work we selected two technologies based on their cost and the read length they can generate. The first is Illumina, a short reads technology, with reads of maximum 300 nucleotides and very low error rates but a high processing cost when assembling an entire genome. The second, Oxford Nanopore Technologies (ONT), technology of long reads, that provides reads of variable size (commonly 10,000 nucleotides), but with a higher percentage of error and a low cost of processing when assembling in time and resources.

1.1 DNA Sequencing

All living being have a unique and unrepeatable deoxyribonucleic acid (DNA) which contains all the genetic information of the individual.

The DNA consists of four nucleotides, adenine (A), thymine (T), guanine (G) and cytosine (C), which are intertwined forming DNA sequences that give rise to genes. The order of the nucleotides will determine the genes that the organism can express, that is, the biological instructions.

DNA sequencing is the process that aims to determine the order of nucleotides of a DNA molecule [18].

1.2 Next Generation Sequencing

Next generation sequencing technologies (NGS) allow sequencing of a species using new strategies that generate a very high number of genome reads with less effort, lower cost, and much less time per sequenced nucleotide [15]

Today there are different NGS platforms (IonTorrent, Illumina, PacBio, Oxford Nanopore) that differ in the used sequencing chemistry. An important difference between the sequencers is the quantity and quality of the data, some tools deliver sequences of very small fixed sizes while others deliver very long reads but of variable size, also each technology has different percentages of error [1].

Sequencing technologies deliver the raw data, that is, a file with the genetic information divided into disordered fragments. Each fragment is called read.

The size and accuracy of the reads will depend on the used technology. To obtain the final sequence in an orderly manner, some assembly method must be used to join the reads and reconstruct the genome.

This project will use a technology of long reads, Oxford Nanopore and Illumina, as a technology of short reads.

Oxford Nanopore. This technology through the use of electrical signals determines the nucleotide to be sequenced. Through this method very long reads of variable length can be obtained.

It is a technology of high portability and low cost. Its ability to determine DNA fragments of greater length is one of its advantages [3], But its high percentage of error compared to short read technologies remains its bigger drawback.

Illumina. This technology uses the method of "synthesis sequencing", where nucleotides are labeled with fluorescent colors that generate light signals that will determine the nucleotide to be sequenced [12].

Illumina has the advantage of providing a low cost per base pair and low error rates. However, one of its biggest disadvantages is the short length of this reads. Therefore, the processing to rebuild the genome is computationally expensive, and often unachievable when there is no reference genome to map the reads.

Reads assembly remains major challenge in DNA sequencing, whereas longer reads may help to facilitate this process, the read length is inversely proportional to read accuracy. Therefore, we aimed to improve the accuracy of long reads by the development of a preprocessing algorithm to facilitate genome assembly. This paper presents a preprocessing strategy for sequences obtained with Oxford Nanopore, with the aim to improve reads accuracy. For this, we start with a correction algorithm, which uses Illumina sequences, to then make an assembly of improved reads. With this strategy we achieve improvements in genome coverage and lower error rates.

2 Genome Assembly

The sequencing process delivers DNA information divided into fragments. In order to extract the relevant information, sequencing reads must be assembled, i.e. aligned and merged in order to reconstruct the original sequence. Genome assembly and can be of two types, comparative assembly and *de novo* assembly.

2.1 *de novo* assembly

De novo assembly is a process used to reconstruct genomes, without having prior knowledge of its genomic organization. In this type of assembly a reference genome is not used, but only the information contained in the sequenced reads. This type of assembly is generally used when sequencing the genome of a species that has not been previously referenced.

Error Correction Algorithms in Long Reads. Currently there are different ways to perform *de novo* assembly. In this work the strategy will be to make an error correction before making the assembly to identify and correct errors.

The error correction algorithms for long reads can be divided into three approaches [19]:

– Algorithms based on alignment of short reads: Short reads are aligned with long reads and these alignments are used to correct errors in long reads.

– Algorithms based on assemblies of short reads: The short reads are assembled. Long reads align with the assembly and are used to correct long reads.
– Algorithms using only long reads: The reads are self-corrected through consensus sequences obtained by overlays or alignments of the same reads.

There are several tools for the correction, in this work we selected Lordec [17].

The Fig. 1 presents a scheme used in this work to make *de novo* assembly with the steps previously described. In the second step (preprocessing) we include the algorithm explained in this work.

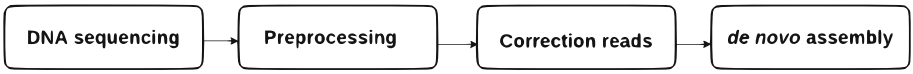

Fig. 1. Scheme of Pipeline used to *de novo* assembly.

3 Algorithm

In this work we developed an algorithm to associate the short reads obtained from Illumina with the ones obtained with Nanopore. We use a three step approach, starting with candidates selection, followed by filtering of short reads, and finally Nanopore read improvement through consensus.

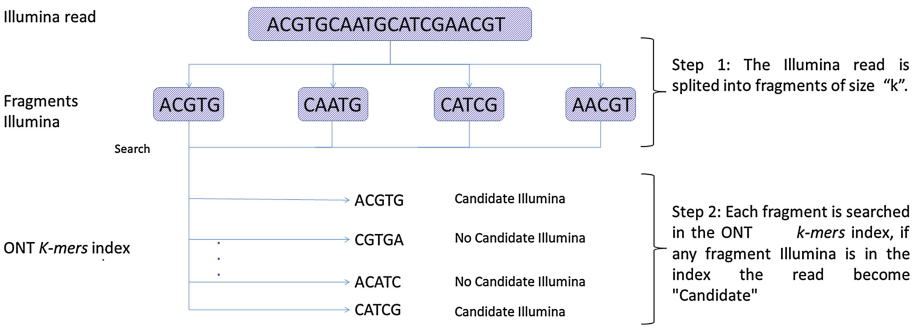

Fig. 2. First part of the algorithm: candidates selection

The first step is built upon the methodology presented by Kawulok J. [4]. This algorithm was adapted to perform a search for candidate reads within the Illumina library based on Levenshtein distance to Nanopore reads as presented in Algorithm 1. Given a Nanopore read, the Illumina reads that are similar are selected. First an index of *k-mers* is created, which stores all the sub-sequences of a defined size k along with their positions, then it is divide the Illumina read

into N fragments of a size k. Each fragment is searched in the index and those that are found become Illumina candidates. This is explained graphically in the Fig. 2.

Algorithm 1. Candidates selection

Input: *ont*: Nanopore read to improve
Input: *ill_reads*: Illuminas reads
Input: *k*: Size of Illumina fragments to use
Output: *new_ont*: New improved Nanopore read through Illuminas reads
Assumes: *create_index(ont, k)*: Create *k-mers* index for Nanopore read
Assumes: *get_slices(illumina, k)*: Split the Illumina read obtaining N fragments of a size k
Assumes: *search(fragment_illumina, ont)*: Search, align and calculate the distance between the Illumina read and the ONT read based on the Illumina fragment.

```
index_ont ← create_index(ont, k)
for illumina in illumina_reads do
    slices_ill ← get_slices(ill, k)
    for fragment_illumina in slices_ill do
        if fragment_ill in index_ont then
            pos, dist ← search(ont, fragment_illumina)
            Appends illumina to candidatos
cands_m, pos_m ← define_cand(cands, pos)
new_ont ← consensus(cands_m, pos_m, seq_ont)
return new_ont
```

In the second step reads were filtered on basis of the estimated error rate of Nanopore sequencing technology. After selecting the candidates, were analyzed with the function *define_cand* (Algorithm 2) where Illumina reads serve to improve the Nanopore reads.

Algorithm 2. Selection of Illumina candidates apt to improve Nanopore read

Input: *dist_max*: User defined percentage to determine accepted range of error between Illumina read and Nanopore read

```
function DEFINE_CAND(cands, pos)
    dist_max ← (size_ill ∗ (dist_max))
    for id_ill in cands do
        if cands[id_ill] ≤ dist_max then
            Appends id_ill to cands_mejora
            Appends pos[id_ill] to pos_mejora
    return cands_mejora, pos_mejora
```

Lastly, the Nanopore reads were improved with the information obtained in earlier steps (Algorithm 3).

Algorithm 3. Nanopore read improvement through consensus

Assumes: *get_max_nucleotide(cons, pos)*: Look at the position *pos* which nucleotide has greater distance weighting in the consensus array

```
1: function CONSENSUS(cands, pos, ont)
2:     for ill in cands do
3:         start ← pos
4:         end ← pos + size + 1
5:         p ← 0
6:         for x in start to end do
7:             for nucl in 'ACGT' do
8:                 if ill[p] == nucl then
9:                     cons[nucl][x] ← cons[nucl][x] + 1/dist
10:                p++
11:    new_ont ← ""
12:    for pos_ont in 0 to length(ont)-1 do
13:        max_nucl ← get_max_nucleotide(cons, pos_ont)
14:        Concatenates new_ont with max_nucl
15:
       return new_ont
```

4 Results

4.1 Experimental Environment

Different tools were used during the development of this project:

- Sequence alignment: BWA [7], Minimap2 [9]
- Analysis and manipulation of sequences: Samtools [10], Tablet [11], Quast [2], Qualimap [13].
- Error correction: Lordec [17].
- *De novo* assembly: Miniasm [8], Smartdenovo [16], Flye [5].

The alignment, analysis and manipulation of sequences were used to validate data and sintonize the algorithm.

The Python programming language was used to develop the algorithm along with different libraries, including Biopython for the management of biological data. This work was implemented a multicore version of the algorithm proposed.

Data Set. It was decided to use bacterial sequencing libraries of *Escherichia coli*. The sequencing libraries used were obtained from the NCBI database.

Below are the libraries used along with their most important features.

Illumina Library Metrics:

- Amount of reads: 893,383
- Mapped reads with the reference: 85.31%
- Duplicate Reads: 26.34%
- Error percentage with reference: 3.14%

Oxford Nanopore Library Metrics

- Amount of reads: 15,911
- Mapped reads with the reference: 95.87%
- Duplicate Reads: 4.13%
- Error percentage with reference: 22.67%

4.2 Experimental Results

After performing different tuning experiments, we tested the potential benefit of the preprocessing algorithm proposed here with three different *de novo* assemblers: Miniasm, Smartdenovo and Flye. We performed *de novo* assembly of a testing data-set detailed in the 4.1 section under three different conditions:

- Basic assembly: Assembly made with the original Nanopore reads (without any preprocessing or correction step).
- Lordec Assembly: Assembly performed only with the error correction of the original reads with the Lordec tool.
- Improved Assembly: Assembly performed after preprocessing with our tool (step two of Fig. 1) and correction with Lordec tool.

The Table 1 presents the results obtained by using the three assemblers mentioned above. Under the analyzed conditions the inclusion of a preprocessing step as proposed here results in a significant increase of total aligned fraction, the size of largest contig and genome fraction coverage.

Table 1. Statistics of Miniasm, Smartdenovo and Flye assemblers

		Genome fraction (%)	Total aligned length	Largest contig	Mismatches per 100 kbp	Indels per 100 kbp
Miniasm	Basic assembly	2.5%	206	259,582	2,444.65	14.81
	Lordec assembly	66.463%	3,453,423	308,407	3,541.52	97.95
	Improved assembly	69.497%	3,609,974	434,975	2,967.11	133.01
Smartdenovo	Basic assembly	2.918%	149,865	619,175	2,305.21	6.01
	Lordec assembly	63.571%	3,273,485	1,605,886	3,829.54	187.43
	Improved assembly	67.073%	3,454,852	1,609,702	3,224.38	193.48
Flye	Basic assembly	71.002%	3,675,484	880,453	3,306.01	1171.86
	Lordec assembly	76.772%	3,958,214	1,259,122	2,436.21	319.19
	Improved assembly	76.774%	3,963,040	1,661,394	2,361.44	293.19

The basic assembly covers a very small fraction of the genome when using Miniasm and Smartdenovo assemblers, this is due to the low coverage of the Nanopore data set used, the inclusion of the Lordec error correction tool improves this results. The integration of our preprocessing algorithm further improves the assembly. The integration of our tool no only resulted in increased coverage but also in a lower number of mismatches per 100 Kb. The number of indels may be increased or decreased depending on the analyzed assembly tool.

5 Conclusions

Currently, the availability and accessibility to NGS technology has increased dramatically, and consequently the amount of sequencing data generated. However, the assembly of new genomes remains challenging and is defined as a NP-hard problem [14]. This paper presents an algorithm to associate high accuracy short reads with error prone long reads, that could add potential benefits when included as a preprocessing step during *de novo* assembly.

Our approach allows additional benefit in terms of genome coverage, contig length, and mismatch rate when performing genome assembly with Miniasm, Smartdenovo, and Flye. This improvement is also sizable when including error correction steps with tools such as Lordec.

The benefit of the here proposed approach in terms of coverage ranges from 5 Kb to 181 Kb in a small sized bacterial genome, indicating its potential for helping the advancement in *de novo* assembly strategies. These results warrant further research in computational optimization of the algorithm, analysis of its performance with larger genomes, and its integration into more complex pipelines.

References

1. Cadena-Zamudio, J., Martínez-Peña, M., Guzmán-Rodríguez, L., Arteaga-Garibay, R., De Morelos, T.: Aplicación de secuenciación masiva para el estudio y exploración de diversidad microbiana y su aprovechamiento biotecnológico. Agroproductividad **9**(2), 70–83 (2016)
2. Gurevich, A., Saveliev, V., Vyahhi, N., Tesler, G.: Quast: quality assessment tool for genome assemblies. Bioinformatics **29**(8), 1072–1075 (2013)
3. Jain, M., Olsen, H., Paten, B., Akeson, M.: The oxford nanopore minion: delivery of nanopore sequencing to the genomics community. Genome Biol. **17** (2016). https://doi.org/10.1186/s13059-016-1103-0
4. Kawulok, J.: Approximate string matching for searching DNA sequences. Int. J. Biosci. Biochem. Bioinform. **3**, 145–148 (2013). https://doi.org/10.7763/IJBBB.2013.V3.183
5. Kolmogorov, M., Yuan, J., Lin, Y., Pevzner, P.A.: Assembly of long, error-prone reads using repeat graphs. Nat. Biotechnol. **37**(5), 540–546 (2019)
6. Land, M., et al.: Insights from 20 years of bacterial genome sequencing. Functi. Integr. Genomics **15**(2), 141–161 (2015)

7. Li, H.: Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997 (2013)
8. Li, H.: Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics **32**(14), 2103–2110 (2016). https://doi.org/10.1093/bioinformatics/btw152
9. Li, H.: Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics **34**(18), 3094–3100 (2018)
10. Li, H., et al.: The sequence alignment/map format and SAMtools. Bioinformatics **25**(16), 2078–2079 (2009)
11. Milne, I., et al.: Using tablet for visual exploration of second-generation sequencing data. Briefi. Bioinformatics **14**(2), 193–202 (2013)
12. Morozova, O., Marra, M.A.: Applications of next-generation sequencing technologies in functional genomics. Genomics **92**(5), 255–264 (2008)
13. Okonechnikov, K., Conesa, A., García-Alcalde, F.: Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics **32**(2), 292–294 (2016)
14. Phillippy, A.M.: New advances in sequence assembly (2017)
15. Rodríguez-Santiago, B., Armengol, L.: Tecnologías de secuenciación de nueva generación en diagnóstico genético pre-y postnatal. Diagnóstico Prenatal **23**(2), 56–66 (2012)
16. Ruan, J.: SMARTdenovo: Ultra-fast de novo assembler using long noisy reads (2015)
17. Salmela, L., Rivals, E.: LoRDEC: accurate and efficient long read error correction. Bioinformatics **30**(24), 3506–3514 (2014)
18. Tubbs, R., Stoler, M.: Cell and Tissue Based Molecular Pathology. ClinicalKey 2012, Churchill Livingstone/Elsevier (2009)
19. Zhang, H., Jain, C., Aluru, S.: A comprehensive evaluation of long read error correction methods. BioRxiv p. 519330 (2019)