Send Orders for Reprints to reprints@benthamscience.net

The Open Bioinformatics Journal, 2014, *8*, 1-5

Open Access

Performances of Bioinformatics Pipelines for the Identification of Pathogens in Clinical Samples with the De Novo Assembly Approaches: Focus on 2009 Pandemic Influenza A (H1N1)

Tommaso Biagini[1], Barbara Bartolini[2], Emanuela Giombini[2], Maria R. Capobianchi[2], Fabrizio Ferrè[1], Giovanni Chillemi[3,*] and Alessandro Desideri[1,*]

[1]*Department of Biology, University of Rome "Tor Vergata", Via della Ricerca Scientifica, 00133, Rome, Italy and Molecular Digital Diagnostics (MDD), Via S. Camillo de Lellis 01100 Viterbo, Italy;* [2]*"L. Spallanzani" National Institute for Infectious Diseases, Via Portuense 292, 00149 Rome, Italy;* [3]*Inter-University Consortium for the Application of Super-Computing for Universities and Research-CASPUR, Via dei Tizii 6, Rome 00185, Italy*

Abstract: Diagnostic assays for pathogen detection are critical components of public-health monitoring efforts. In view of the limitations of methods that target specific agents, new approaches are required for the identification of novel, modified or 'unsuspected' pathogens in public-health monitoring schemes. Metagenomic approach is an attractive possibility for rapid identification of these pathogens. The analysis of metagenomic libraries requires fast computation and appropriate algorithms to characterize sequences. In this paper, we compared the computational efficiency of different bioinformatic pipelines ad hoc established, based on de novo assembly of pathogen genomes, using a data set generated with a 454 genome sequencer from respiratory samples of patients with diagnosis of 2009 pandemic influenza A (H1N1). The results indicate high computational efficiency of the different bioinformatic pipelines, reducing the number of alignments respect to the identification based on the alignment of individual reads. The resulting computational time, added to the processing/sequencing time, is well compatible with diagnostic needs. The pipelines here described are useful in the unbiased analysis of clinical samples from patients with infectious diseases that may be relevant not only for the rapid identification but also for the extensive genetic characterization of viral pathogens without the need of culture amplification.

Keywords: Bioinformatics pipeline, de novo assembly, genome reconstruction, metagenomics, pathogen detection, pyrosequencing.

INTRODUCTION

The trends in clinical diagnosis show gradual substitution of traditional methods with novel molecular biology technologies. In particular, advancement of sequencing technologies has disclosed unprecedented opportunities in many application areas. Metagenomics can be considered as an improved pathogen detection method, since it is based on a sequence-independent approach that does not rely on predefined target genome sequences; in addition, it can be applied to non cultivable organisms. The success of this approach, especially when applied to clinical samples, relies upon the advancement in bioinformatic methods [1]. Computational handling of the large amount of sequence data generated in a high-throughput sequencing run is an area of necessary research focus. The analysis of libraries requires fast computation and the right algorithms to characterize sequences; the majority of programs utilized so far to assemble metagenomic data have the limitation to rely on programs originally developed to assemble single genomes, which contain less complexity and generally have large sequence coverage [2]. On the contrary, microbial genomes are generally represented in a biological sample as a minute fraction of all the nucleic acid sequences, and may show genetic variability. We have addressed this question developing a bioinformatic pipeline, analyzing data obtained from a previous study and testing the pipeline in a defined biological context. We have already used ultra deep pyrosequencing (UDPS) to detect and characterize 2009 pandemic influenza A (H1N1) virus directly from nasopharyngeal swabs [3]. Identification of the H1N1 strain is of crucial importance because of its ability to escape drug treatment and initiate drug resistance [4-8]. The bioinformatic approach described in the previous study was based on the categorization of each read, obtained through an alignment search against a predefined series of influenza virus nucleotide and protein databases at NCBI, so it relied on a non- purely metagenomic approach.

In this paper, we have created and analyzed different bioinformatic pipelines based on the de novo assembly of pathogen genomes. We have compared their computational efficiency, underlining the improvements over the previous procedure [3] in terms of time per analysis and computational efficiency, using a data set previously generated with 454 genome sequencer, from patients with diagnosis of 2009 pandemic influenza A (H1N1).

*Address correspondence to these authors at the Department of Biology, University of Rome "Tor Vergata", Via della Ricerca Scientifica 1, 00133, Roma Italia; Tel: +39.06.72594376; Fax +39.06.2022798;
E-mails: desideri@uniroma2.it and giochillemi@gmail.com

1875-0362/14

MATERIAL AND METHODOLOGY

The key point of the bioinformatics procedure described in this paper is the de novo assembly of reads into contigs using the Overlap Layout Consensus (OLC) algorithm, as implemented in Newbler, a software specifically designed for assembling sequence data generated by the 454 GS-series of pyrosequencing, and CAP3 [9]. In the OLC algorithm, the relationships among the reads provided to an assembler can be represented as a graph, where the nodes represent each of the reads and an edge connects two nodes if the corresponding reads overlap. The assembly problem thus becomes the problem of identifying a path through the graph that contains all the nodes (Hamiltonian path). This formulation allows researchers to use techniques developed in the field of graph theory in order to solve the assembly problem.

After the quality trimming of the reads [11,12] an assembler following this paradigm starts with an overlap stage during which all overlaps between the reads are computed and the graph structure is computed. In a layout stage, the graph is simplified by removing redundant information. Graph algorithms are then used to determine a layout (relative placement) of the reads along the genome. In a final consensus stage, the assembler builds an alignment of all the reads covering the genome and infers, as a consensus of the aligned reads, the original sequence of the genome being assembled [10].

The data submitted to the OLC softwares had been generated by pyrosequencing (454/Roche GS FLX platform) of nasopharyngeal fluids from two patients positive for 2009 pandemic influenza A (H1N1), already analyzed by our group [3].

The flowchart describing the bioinformatic pipelines is shown in (Fig. **1**). The starting dataset consisted of the reads obtained after the filtering of the PCR primers. When the primer sequence position in the read was central, the read split into two parts and the sequence primer was eliminated [6, 7]. The new starting datasets were so made up of 186,011 reads for patient 1 (viral load 2x107 cp/ml) ranging in size from 50 to 491 (average length: 210) bp and 291,934 reads for patient 2 (viral load 2x105 cp/ml) ranging in size from 50 to 550 (average length: 221) bp.

After this step, there were two non-conditional branches, indicated with the letters A/B and C/D that have produced four different paths (AC; AD; BC; BD). In the first branch (path A), the reads were directly assembled with Newbler, while in the second branch (path B) the reads were first aligned against a database constructed with all mammalian sequences, downloaded from NCBI, and filtered using these cut-offs: identity >85%; overlap >70%; and E-value $<10^{-5}$. In this case, the new dataset, resulting from the removal of all the reads related to the human host, was then assembled with Newbler. The assembly of reads in each dataset was implemented using overlapping cut-off of 50 bp with 90% identity.

The contigs generated through paths A and B underwent path C or D. In path C, the output of Newbler, including assembled contigs and singletons (reads not assembled into contigs), was directly evaluated, without any additional elaboration. In path D, the output of Newbler was used as input for a second assembly with CAP3.

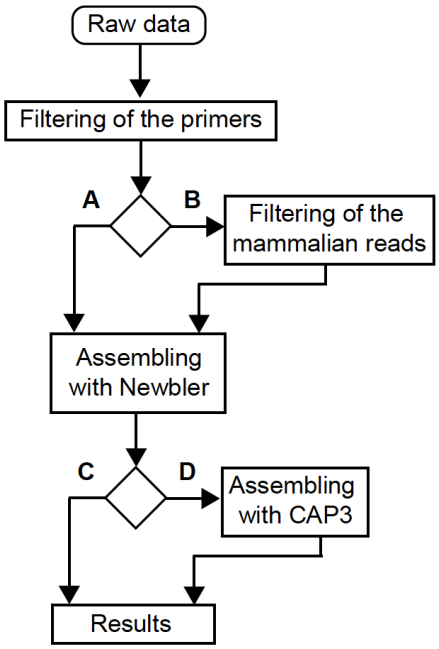

Fig. (1). Flowchart describing the various data analyses steps. In path A the reads were directly assembled with Newbler, while in path B the reads were aligned against a database with all mammalian sequences (NCBI), and filtered with the following cut-offs : identity >85%; overlap >70%; E-value <10-5). The contigs generated through paths A and B underwent path C or D. In path C the output of Newbler was directly evaluated; in path D the output of Newbler was used as input for a second assembly made with CAP3.

RESULTS

The assembly results for all the four bioinformatic pipelines are reported in (Table **1**). Comparing AD vs. AC and BD vs. BC, it can be inferred that the second assembly carried out with CAP3 on the contigs and singletons coming from Newbler analysis (path D) produces an increase in the number of contigs and a consequent significant decrease in the number of singletons for both patients. Comparing AC vs. BC and AD vs. BD, the number of contigs always decreases, as expected, due to the filtering on mammalian reads carried out before the assembly. No specific trend is observed for the average and maximum length of the contigs according to the different path combinations. The contigs obtained by the four pipelines have been aligned on the NCBI nucleotide database (BLAST analysis), to identify and characterize the microorganisms contained in the nasopharyngeal swab sample. The MEGAN taxon trees [13] relative to the bioinformatic pipeline AC and AD are shown in (Fig. **2**). The figure shows that the assembly of the original dataset, carried out with Newbler, is able to unambiguously identify the dominant pathogen species (influenza A virus), also in the case of patient 2, characterized by a lower viral load as compared to patient 1 ($7X10^5$ IU/ml vs. $2X10^7$IU/ml) and by the presence of reads coming from other pathogens (i.e. bacteria). In addition, from the BLAST analysis the best hit for all the contigs was 2009 pandemic influenza A (H1N1) (A/California/07/2009 strain) in both samples, providing univocal identification of the main pathogen present in these clinical samples.

Table 1. Assembly Results of Patient 1 (A) and Patient 2 (B) Using the Different Bioinformatic Pipelines (See Fig. 1).

A)

	Bioinformatic pipelines			
	AC	**AD**	**BC**	**BD**
Number of contigs	3342	9619	2447	4077
Average length	271 bp	282 bp	272 bp	282 bp
Longest contig	2138 bp	2138 bp	2056 bp	2056 bp
Number of singletons	58261	37383	21771	12882

B)

	Bioinformatic pipelines			
	AC	**AD**	**BC**	**BD**
Number of contigs	1230	4418	711	2849
Average length	250 bp	257 bp	252 bp	247 bp
Longest contig	2126 bp	2126 bp	2126 bp	2375 bp
Number of singletons	41753	31621	25436	17606

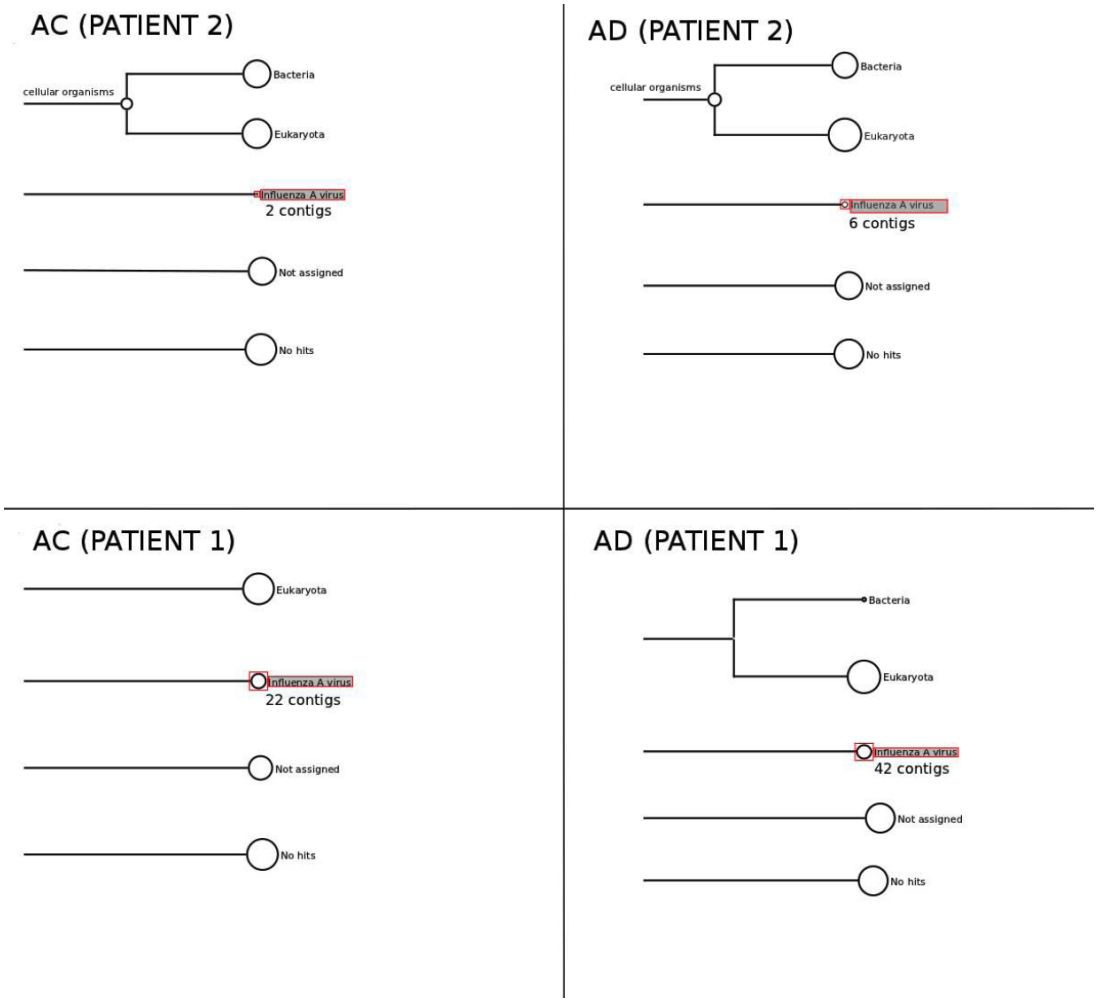

Fig. (2). Phylogenetic trees. The data coming from the AC and AD paths, described in figure 1, were analyzed by MEGAN to generate phylogenetic trees incorporating all taxa to which the contigs were assigned. The size of each circular node is proportional to the percentage of assignments at the taxonomic level.

In order to carry out a better comparison of the four assembly pipelines, only the contigs aligned against the H1N1 virus have been evaluated in terms of number of reads in contigs, percentage of identity and percentage of genome covered by all the contigs [14]. Moreover the number of reads present in these contigs but not aligning with influenza A virus (chimeric reads) has been considered. The results from patient 1 and patient 2 are shown in Table **2** (**A** and **B**, respectively).

As reported in Table **2B**, for patient 2 the paths AD and BD, where the second assembly with CAP3 has been carried out, are convenient when the aim of the analysis is the reconstruction of the pathogen genome after its identification. In fact, only after the application of the second assembly with CAP3 it is possible to identify four additional contigs, that bring the percentage of the 2009 pandemic influenza A (H1N1) virus genome covered up to 16%, maintaining the accuracy of the reconstruction, since no chimeric reads are contained in these additional contigs (Table **2B**). On the other hand, comparing paths BC and BD to paths AC and AD, respectively, indicates that the addition of the step of mammalian reads filtering does not produce any improvement in the assembly statistics.

For patient 1 (Table **2A**), the second assembly with CAP3 (path AD) produces an increase (from 22 to 42) in the number of H1N1 contigs as compared to path AC. However such an increase produces just a 2% increase of the genome coverage, since the new contigs have a lower percentage of identity as compared to path AC (95-100% versus 98-100%). Moreover, one of the new contigs is chimeric. A similar trend is also observed for the path BD where the number of contigs increases up to 52 when compared to the 21 observed in BC but the coverage only increases by 5%.

DISCUSSION AND CONCLUSION

The field of bioinformatics for metagenomics is very dynamic and new programs are continuously created to analyze the NGS-generated data. The challenge, mainly when the studies are aimed at the identification of a new viral species, is the setup of bioinformatic procedures able to produce fast and accurate identification of the pathogen(s).

In this study, we have used biological samples from patients with known diagnosis of infection with 2009 pandemic influenza A (H1N1), to validate a bioinformatic pipeline that should be used also for the investigation of diseases of unknown etiology. The results indicate that de-novo assembly is a key step for a unique identification of the pathogenic organism with a reasonable computation time.

In detail, our time of analysis ranges from a minimum of 24 hours for path AC to a maximum of 48 hours for path AD (Fig. **1**), greatly reducing the number of alignments (from hundreds of thousands to thousands) and the relative time necessary for an identification based on the alignment of individual reads (3). This computational time, added to the processing/sequencing time, is compatible with the diagnostic needs, allowing to obtain a complete and accurate picture of the pathogen(s) present in the sample within a few days overall. However, a limitation concerns the fact that, due to

Table 2. Characteristics of the Contigs Belonging to the 2009 Pandemic Influenza A H1N1 Virus Obtained from the Assembly of the Reads for Patient 1 (A) and Patient 2 (B), Assembled Using the Different Bioinformatic Pipelines (See Fig. 1)

A)

	Bioinformatic pipelines			
	AC	**AD**	**BC**	**BD**
Number of contigs	22	42	21	52
Number of reads in contigs	77296	79005	73161	75645
Identity (%)	98-100	95-100	98-100	95-100
Genome Covered (%)	87	89	85	90
Chimeric contigs (number of chimeric reads)	2 (52)	3 (70)	3 (111)	3 (124)

B)

	Bioinformatic pipelines			
	AC	**AD**	**BC**	**BD**
Number of contigs	2	6	2	6
Number of reads in contigs	153	173	153	173
Identity (%)	99-100	99-100	99-100	99-100
Genome Covered (%)	6	16	6	16
Chimeric contigs (number of chimeric reads)	0 (0)	0 (0)	0 (0)	0 (0)

not uniform coverage of the different regions of the genome, the approach cannot provide a uniform reliability of the reconstructed sequence.

As far as the reconstruction of the full genome of the pathogenic organism is concerned, the use of CAP3 on the output of Newbler (paths AD and BD) is strongly recommended since it improves the percentage of the reconstructed virus genome, maintaining the accuracy of the assembly. In detail, in both cases the number of chimeric reads is less than 1% of the total number of the reads inserted into the virus-specific contigs.

Application of the mammalian filter does not bring a significant improvement in the quality of reads to be assembled, while it greatly increases the time required to achieve the final assembly.

In conclusion, the pipelines here described may represent efficient tools for the unbiased analysis of clinical samples also from patients with unknown infectious diseases that may be relevant not only for the identification, but also for the extensive genetic characterization of viral pathogens without the need of culture amplification. This is particularly relevant for influenza viruses, where rapid characterization of new reassortants may have public health implications in respect of the treat represented by the appearance of new strains with pandemic potential. Finally, application of the pipeline to several patients affected by different diseases will also permit to gather the necessary statistics to propose possible optimizations of the procedure.

CONFLICT OF INTEREST

The author(s) confirm that this article content has no conflicts of interest.

ACKNOWLEDGEMENT

This work was partially supported by the Italian Ministry of Health (fondi Ricerca Corrente), by the Seventh Framework Program of European Union [FP7/2007-2013] under Grant Agreement n°278433-PREDEMICS, by Fondazione Cariplo [grant 2011] and by FILAS [project DIREI].

REFERENCES

[1] V. Kunin, A. Copeland, A. Lapidus, K. Mavromatis, and P. Hugenholtz, "A bioinformatician's guide to metagenomics," *Microbiol. Mol. Biol. Rev.*, vol. 72, no. 4, pp. 557-578, 2008.

[2] D. R. Mende, A. S. Waller, S. Sunagawa, A. I. Järvelin, M. M. Chan, M. Arumugam, J. Raes, and P. Bork, "Assessment of metagenomic assembly using simulated next generation sequencing data," *PLoS ONE*, vol. 7, no. 2, p. e31386, 2012.

[3] B. Bartolini, G. Chillemi, I. Abbate, A. Bruselles, G. Rozera, T. Castrignanò, D. Paoletti, E. Picardi, A. Desideri, G. Pesole, and M. R. Capobianchi, "Assembly and characterization of pandemic influenza A H1N1 genome in nasopharyngeal swabs using high-throughput pyrosequencing," *N. Microbiol.*, vol. 34, no. 4, pp. 3910397, 2011.

[4] Q.-S. Du, S.-Q. Wang, and K.-C. Chou, "Analogue inhibitors by modifying oseltamivir based on the crystal neuraminidase structure for treating drug-resistant H5N1 virus," *Biochem. Biophys. Res. Commun.*, vol. 362, no. 2, pp. 525-531, 2007.

[5] S.-Q. Wang, X.-C. Cheng, W.-L. Dong, R.-L. Wang, and K.-C. Chou, "Three new powerful oseltamivir derivatives for inhibiting the neuraminidase of influenza virus," *Biochem. Biophys. Res. Commun.*, vol. 401, no. 2, pp. 188-191, 2010.

[6] S.-Q. Wang, Q.-S. Du, and K.-C. Chou, "Study of drug resistance of chicken influenza A virus (H5N1) from homology-modeled 3D structures of neuraminidases," *Biochem. Biophys. Res. Commun.*, vol. 354, no. 3, pp. 634-640, 2007.

[7] R. M. Pielak, J. R. Schnell, and J. J. Chou, "Mechanism of drug inhibition and drug resistance of influenza A M2 channel," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 106, no. 18, pp. 7379-7384, 2009.

[8] J. R. Schnell and J. J. Chou, "Structure and mechanism of the M2 proton channel of influenza A virus," *Nature,* vol. 451, no. 7178, pp. 591-595, 2008.

[9] X. Huang and A. Madan, "CAP3: A DNA sequence assembly program," *Genome Res.*, vol. 9, no. 9, pp. 868-877, 1999.

[10] J. R. Miller, S. Koren, and G. Sutton, "Assembly algorithms for next-generation sequencing data," *Genomics*, vol. 95, no. 6, pp. 315-327, 2010.

[11] R. Schmieder and R. Edwards, "Quality control and preprocessing of metagenomic datasets," *Bioinformatics*, vol. 27, no. 6, pp. 863-864, 2011.

[12] H. H. Chou and M. H. Holmes, "DNA sequence quality trimming and vector removal," *Bioinformatics*, vol. 17, no. 12, pp. 1093-1104, 2001.

[13] D. H. Huson, A. F. Auch, J. Qi, and S. C. Schuster, "MEGAN analysis of metagenomic data," *Genome Res.*, vol. 17, no. 3, pp. 377-386, 2007.

[14] S. Mitra, B. Klar, and D. H. Huson, "Visual and statistical comparison of metagenomes," *Bioinformatics*, vol. 25, no. 15, pp. 1849-1855, 2009.

Received: October 25, 2013 Revised: December 14, 2013 Accepted: December 16, 2013