REVIEW

Of the Doctoral Dissertation of Alla A. Mikheenko “DEVELOPMENT OF COMPUTATIONAL METHODS FOR ANALYSIS AND VISUALIZATION OF EUKARYOTIC GENOME ASSEMBLIES” submitted for defense of the degree of candidate of physico-mathematical sciences, specialization 03.01.09 – Mathematical biology, bioinformatics.

Relevance

This dissertation describes the development of new algorithms for the analysis and visualization of genome assemblies, with a particular focus on extended tandem repeat arrays. Recent efforts to finish the remaining gaps in the human reference genome have required assembly and validation approaches specifically designed to tackle the most complex regions of the genome. These regions often comprise multi-megabase arrays of satellite DNAs, which are arranged as head-to-tail tandem repeat arrays. Such arrays have remained unfinished and largely unexplored even twenty years after the first draft assemblies of the human genome were released. The advent of long-read sequencing technologies have finally opened these regions to analysis, but existing sequence analysis tools were unable to correctly resolve them and new methods were needed. The contributions of this thesis directly aided the development of such tools and enabled the validation of complex tandem repeats in the first complete assembly of a human genome. Thus, the relevance of the dissertation topic is very high.

As the co-chair of an international consortium that seeks to finish the remaining gaps in the human reference genome, I can personally attest to the impact of the work described in this thesis. The TandemMapper and TandemQUAST tools described in Chapter 2 were essential to our completion of the human centromeric repeat arrays, giving researchers a first glimpse at the sequence structure and evolution of these functionally important regions of the genome. The tools presented in this thesis were able to precisely identify errors in the initial draft assemblies of the human centromeric satellite arrays, which informed improved sequence assembly methods that corrected these errors. This improvement process continued for a number of iterations, where the validation tools identified errors that were then investigated in order to further improve the assembly methods. Without the help of such precise validation techniques, the quality of these assemblies would have remained inadequate. This also demonstrates the utility of visualization and validation tools in general, such as those described in Chapters 1 and 3, which have informed and improved the quality of many genome assemblies as well as the development of assembly tools themselves.

The relevance of this work is evident from the above noted contributions as well as the wide adoption of the developed tools. The candidate’s contributions are organized into three main areas: (1) development of the QUAST-LG tool for the quality assessment of large, eukaryotic genomes, (2) development of the TandemTools package for mapping and validating long tandem repeats, and (3) visualization of genome assemblies via the Icarus and AGB browsers. Papers describing this work have been published in journal *Bioinformatics*, which is a leading journal in the field, and the QUAST-LG paper has now been cited well over 100 times. The candidate is a primary contributor to the QUAST package, which is widely used throughout genomics and has been downloaded over 50,000 times. Lastly, as an active member of an international consortium, the candidate has directly contributed to the recent completion of the human reference genome. This work has not yet been published, but is a landmark achievement and demonstrates the clear impact of the candidate’s research.

Validity

Validity is, in some sense, the essence of this dissertation. All three chapters concern the validation of genome sequence assemblies. This is a noble goal, as accurate genome sequences are the foundation of all genomic research. The candidate's methods have unquestionably improved the validity of genomic sequences, by providing researchers with appropriate tools to identify and correct errors in their own genome assemblies.

The methods described are clear, well-reasoned, and effective, and appropriate validation has been performed to demonstrate their accuracy. Because the methods focus primarily on validation, simulations are relied upon to introduce known errors for testing. I also have personal experience using these tools for the validation of my own assemblies, and find them to be effective and practical on real assemblies as well.

The English portion of the defense is very well written and nearly without flaw. I am unable to assess the Russian portion of the dissertation, but based on the careful presentation of the English sections, I presume it is of equal quality.

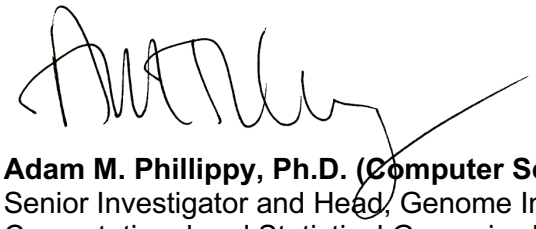
Novelty

The primary contribution of this thesis is a set of methods and tools for validating and visualizing genome sequence assemblies. I especially appreciated the cleverness of TandemQUAST, which applies a set of k-mer-based metrics that appear simple at first, but which are incredibly effective at identifying errors within complex tandem repeats. These powerful metrics are also accompanied by plots that are helpful for interpretation. Such a concise and effective solution to a difficult problem shows great insight, and it is clear that the candidate is an expert on the topic. Novel contributions include the TandemMapper algorithm, which makes use of adjacent pairs of solid k-mers, and a suite of new quality metrics associated with these k-mers. Such well-designed and practical methods are to be applauded.

Conclusion

The dissertation of Alla Mikheenko includes work from four first-author papers published in the indexed and field-standard journal *Bioinformatics*. Both the published papers and the dissertation itself demonstrate scholarship and cite the relevant prior works. As detailed above, the presented work is highly relevant, valid, novel, and accurately presented. Thus, the dissertation of Alla Mikheenko meets the requirements necessary for the granting of the degree of candidate of physico-mathematical sciences, specialty 03.01.09, and I support this action without hesitation.

27 December 2020, serving in my personal capacity,



Adam M. Phillippy, Ph.D. (Computer Science)
Senior Investigator and Head, Genome Informatics Section
Computational and Statistical Genomics Branch
National Human Genome Research Institute
National Institutes of Health
Bethesda, MD USA