

## REVIEW

Of the Doctoral Dissertation of Dmitry Antipov “Development of algorithms for special cases of genome assembly” submitted for defense of the degree of a Candidate of Sciences in mathematical biology, bioinformatics (Scientific specialty code 1.5.8).

### Relevance

The dissertation describes the development of new algorithms for biological sequence assembly from genomes and metagenomes. Recent advances in DNA sequencing technology have dramatically lowered the cost of whole-genome sequencing, making these technologies indispensable tools for research. Thus, the relevance of the dissertation topic is very high.

Sequence assembly is a foundational problem in genomics. Most DNA sequencing technologies can read only a short sequence of DNA at a time, and so computational methods are needed to reconstruct the original, longer fragments. The primary challenge of the sequence assembly problem is repeats, strings of DNA that occur in multiple regions of the genome. When attempting to reconstruct the original sequence from much shorter sequencing reads, these repeats confuse the process and can prohibit complete assembly. This ambiguity is typically represented as a graph of possible sequence reconstructions, and sequence assembly algorithms explore this graph to identify a correct reconstruction. However, with short sequencing reads, the resulting graph can be complex and the correct assembly of complete genomes and plasmids impossible.

The primary contribution of the dissertation is the development of new methods for the simplification and processing of sequence assembly graphs for microbial genomics. The relevance of this work is immediately evident from the large community impact and high number of citations achieved by the associated publications. The candidate's contributions are devoted to the widely used SPAdes assembly software and organized into three main areas: (1) development of hybridSPAdes, for the integration of both short and long sequencing reads for improved microbial genome assemblies, (2) development of plasmidSPAdes to detect and recover microbial plasmids during assembly, and (3) development of metaplasmidSPAdes for the recovery of microbial plasmids from metagenomes. Google Scholar reports >14,000 citations for the 2012 SPAdes paper, on which the candidate was a co-author. Within the genomics community, the SPAdes toolkit is widely used, and highly regarded for its expert engineering. The hybridSPAdes paper, for which the candidate was first author, has over 300 individual citations and describes the integration of multiple datatypes for improved assembly. Lastly, the more recent plasmidSPAdes and metaplasmidSPAdes papers have over 300 and 40 citations, respectively. These tools improve the recovery and assembly of microbial plasmids, which can be important factors in virulence, antibiotic resistance, and degradation, and are highly relevant to current research in microbial genomics and metagenomics.

### Validity

The presented work is valid. The dissertation is built upon the solid foundation of the SPAdes assembly toolkit. There have been multiple, independently published evaluations that show SPAdes is reliable, accurate, and routinely a top-performing assembler. With over 14,000 citations, the genomics community has clearly given this software its resounding approval. The dissertation extends SPAdes with important new functionality.

The new methods and results are presented in a clear and correct manner, and appropriate validation has been performed to demonstrate their accuracy. Benchmarking results are presented for hybridSPAdes, plasmidSPAdes, and metaplasmidSPAdes using appropriate datasets and competing tools. These results demonstrate that the tools are efficient, accurate, and extend the state of the art. I have found no technical errors in the presented work.

## **Novelty**

The primary contribution of the presented work is extensions to the SPAdes assembler to better resolve repeats and better recover plasmid sequences. The method developed for hybridSPAdes was one of the first implementations of an assembly method capable of using long sequencing reads to simplify short-read sequencing graphs. The problem of aligning noisy, long reads to a graph is difficult and the developed solution appears effective. This “hybrid” technique remains in use and is a common means for assembling high-quality, complete microbial genomes.

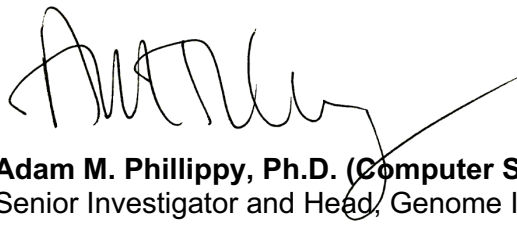
The methods developed for plasmid assembly are more straightforward applications of coverage filtering and graph simplification, but nonetheless address an important problem using novel approaches. Many assemblers will omit or otherwise have difficulty assembling microbial plasmids, and so dedicated tools that address this problem are appreciated.

Such well-designed and practical tools are incredibly powerful for advancing research. They enable discovery across the field of genomics and amplify the impact of the dissertation. The contributions of this dissertation do not simply answer a single hypothesis; they enable others to answer countless hypotheses, as is evident from the many studies that have made use of them for the discovery of new biology.

## **Conclusion**

The dissertation of Dmitry Antipov includes work from five papers published in highly regarded and indexed journals, including three as first author. 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, and novel. Thus, the dissertation of Dmitry Antipov meets the requirements necessary for the granting of the degree of Candidate of Sciences in mathematical biology, bioinformatics, and I support this action without hesitation.

17 January 2021, 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