New Frontiers of Genome Assembly with SPAdes 3.1

Andrey D. Prjibelski[1,5], Dmitry Antipov[1], Anton Bankevich[1], Alexey Gurevich[1], Sergey Nurk[1], Yana Safonova[1], Irina Vasilinetc[1], Anton Korobeynikov[1,2], Alla Lapidus[1,3] and Pavel Pevzner[1,4]

[1] Algorithmic Biology Laboratory, St. Petersburg Academic University, St. Petersburg, Russia
[2] Department of Mathematics and Mechanics, St. Petersburg State University, St. Petersburg, Russia
[3] Theodosius Dobzhansky Center for Genome Bioinformatics, St. Petersburg State University, St. Petersburg, Russia
[4] Department of Computer Science and Engineering, University of California, San Diego, USA
[5] E-mail address: ap@bioinf.spbau.ru

Project web site: http://bioinf.spbau.ru/en/spades
Source code available at: http://bioinf.spbau.ru/en/spades
Licence: GPLv2

Despite all the efforts high quality genome assembly is a complex task that so far remains unsolved. It is well known that majority of problems caused by repeats present in all genomes of any nature. The usage of multiple methods of genomic DNA isolation, different sequencing technologies and different types of genomic libraries for research projects introduces additional levels of complication to the genome assembly. The assembler tool SPAdes was originally developed at the St. Petersburg Academic University for the purpose of overcoming the complications associated with single-cell microbial data (uneven coverage and increased level of chimerical reads). The tool was able to successfully resolve these issues for Illumina reads and was recognized by the scientific community as one of the best assemblers working with both isolates and single-cell data. Even though the assembler was specifically designed to work solely with microbial genomes, scientists have tested the tool on a large number of different types of other data.

Their efforts and feedback have inspired us to extend the capabilities of SPAdes to include additional platforms (Ion Torrent, PacBio, Sanger), combinations of platforms, and to work with both paired-end and mate-pair libraries of different insert sizes. In this work we present novel features of SPAdes 3.1: hybrid assemblies including the combination of Illumina/IonTorrent with PacBio (or other long reads technologies), improved algorithms for scaffolding and repeat resolution, and an approach for mate-pair only assembly using new Illumina NexteraMP protocol.

We also have noticeably improved both BayesHammer and SPAdes performance. For example, new version of BayesHammer corrects data set of 100 Mbp diploid genome (25 Gb, 310M reads) in 16 hours instead of 90 (16 threads on server with Intel Xeon 2.27GHz processors). As to SPAdes, the main performance improvements were done in the exSPAnder repeat resolution module. On 60 Mbp repeat-rich genome repeat resolution step takes only 2 hours comparing to 78 hours for SPAdes 3.0.

SPAdes is openly available as source code and as pre-built Linux and Mac OS binaries. Additionally, you can use SPAdes on such online cloud services as DNAnexus and Illumina BaseSpace.