# Benchmarking of *de novo* assembly tools: SPAdes 3.9 vs Velvet 1.2

| | |
|---|---|
| **Report number** | #1 |
| **Responsible** | Pimlapas Leekitcharoenphon (DTU) and Maria Borowiak (BfR) |
| **Other partners/institutions involved** | - |
| **Benchmarking launched (date)** | May 2016 |
| **Deliverable due (date)** | Due: May 2016 Delivered: August 2016 |

## Purpose of the benchmarking exercise

The purpose of this benchmarking exercise was to evaluate and compare the performance of the mostly used *de novo* assembly tool, i.e. Velvet, and the newer introduced *de novo* assembly tool, SPAdes.

## Tools included in the benchmarking exercise

*De novo* assembly tools; Velvet 1.2 with default parameters (Assembler-1.2 implemented in the tool Bacterial Analysis Pipeline - Batch Upload (https://cge.cbs.dtu.dk/services/cge/)) and SPAdes 3.9 (http://cab.spbu.ru/software/spades/) with default parameters in careful mode. Both tools were run using different k-mer sizes and the assembled genome was set to pick up from the best k-mer size.

## Species and/or genomes included

50 *Salmonella enterica* subsp. *enterica* serovar Paratyphi B *d*Ta+ (S. Java) isolates were tested. DNA from bacterial cells was isolated from liquid cultures using the PureLink® Genomic DNA Mini Kit (Invitrogen, Carlsbad, CA, USA). Sequencing libraries were prepared with the Nextera XT DNA Sample Preparation Kit (Illumina, San Diego, CA, USA) according to the manufacturer's protocol. Paired-end sequencing was performed in 2 × 300 cycles on the Illumina MiSeq benchtop using the MiSeq Reagent v3 600-cycle Kit (Illumina). Further details related to the included genomes can be found at the end of this report and in Supplementary Table 2 (Annex B).

## Results

### *Overall assembly quality*

Sequencing raw data without trimming was assembled using either Velvet or SPAdes assembly tools. Analysis of contigs using ContigAnalyzer-1.0, implemented in the Bacterial Analysis Pipeline - Batch Upload (https://cge.cbs.dtu.dk/services/cge/), revealed that the mean number of contigs is lower and the mean N50 value (median contig size of a genomic assembly) is higher in the genomes assembled using SPAdes (see Table 1 and Figure 1). The observed mean genome size however is similar for both assembly types.

**Table 1: Assembly quality analysed using ContigAnalyzer-1.0**

| | | Spades | Velvet |
|---|---|---|---|
| **Contig number** | **mean** | 100 | 249 |
| | **min** | 51 | 144 |
| | **max** | 181 | 376 |
| | **sd** | 30 | 55 |
| **N50** | **mean** | 176,144 | 57,148 |
| | **min** | 53,662 | 26,926 |
| | **max** | 393,606 | 146,576 |
| | **sd** | 93,110 | 23,786 |
| **Assembled genome size** | **mean** | 4,924,464 | 4,872,591 |
| | **min** | 4,663,179 | 4,505,678 |
| | **max** | 5,076,872 | 5,027,353 |
| | **sd** | 101,043 | 121,670 |

To further assess the quality of the assemblies, the Multi Locus Sequence Type (MLST) and antibiotic resistance genes were analysed.

### *Results regarding MLST identification*

Analysis of the obtained assemblies regarding the Multi Locus Sequence Type (MLST) was performed using the tool MLST 1.6 (https://cge.cbs.dtu.dk/services/cge/). MLST types (based on the Enterobase scheme, https://enterobase.warwick.ac.uk) could be predicted in 100% of the SPAdes assembled and in 94% of the Velvet assembled genomes.

### *Results regarding the identification of resistance genes*

Antimicrobial resistance patterns derived from MIC values (obtained by broth microdilution method following CLSI guidelines, and using the EUCAST epidemiological ECOFFs; testing conditions applied to the individual samples depend on the year the isolate was collected and are listed in Supplementary Table 2 (Annex B)) were compared with the ResFinder2.1 (https://cge.cbs.dtu.dk/services/cge/) output (AMR genes detected) for *de novo* assembled sequence data (see Supplementary Table 2 (Annex B)).

*Concordance between genotypic and phenotypic resistance data (for detailed results see also Supplementary Table 2 (Annex B)):*

- In 35/50 cases the phenotypic resistance profile could be explained with genes found using Velvet as assembler.
- In 38/50 cases the phenotypic resistance profile could be explained with genes found using SPAdes as assembler.
- In 12/50 cases the phenotypic resistance profile could not be explained with genes found using either SPAdes or Velvet for assembly, one or more genotypic resistance determinants were missing.

- In 5/50 cases resistance genes conferring resistance to aminoglycosides which were not expected based on phenotypic resistance data were found in both genome assemblies.
- In 7/50 cases additional resistance genes which were not expected based on the phenotypic resistance profiles were found in the genomes assembled using SPAdes. This involves *aac(3)*-VIa-like genes (6 cases) and *erm(B)* (1 case).

### Conclusions

All in all, SPAdes assembled genomes showed longer contigs and therefore higher N50 values. This seems to lead to an improved detection of MLST genes. Moreover, “missing” resistance genes, i.e. those absent from genomes assembled using Velvet, could be identified when using SPAdes for genome assembly. Nevertheless, there is a huge number of cases where not all expected genetic resistance determinants were identified. This can be caused by loss of resistance plasmid during storage and culturing or emergence of unknown resistance mechanisms and chromosomal point mutations which could not be identified using the ResFinder2.1 tool. Additional identification of streptomycin resistance determinants, which were not expected based on phenotypic data, are likely to be caused by incorrectly determined MIC values or changes regarding break points and test panels. For better comparison of the data, isolates with contradicting phenotypical and genotypical results should be subjected to MIC retesting. In case of the *aac(3)*-VIa-like genes and the *erm(B)* that were detected in 7 SPAdes assembled genomes, further analysis of the respective contigs revealed that all of them showed a low coverage. These contigs might have been derived from the assembly of low level read contaminations from other samples which might have led to the false positive detection of genotypic resistance determinants. Including low coverage contigs caused by read contamination in the assembled genomes might be a disadvantage of SPAdes. Additional filters should be applied to remove low coverage contigs.

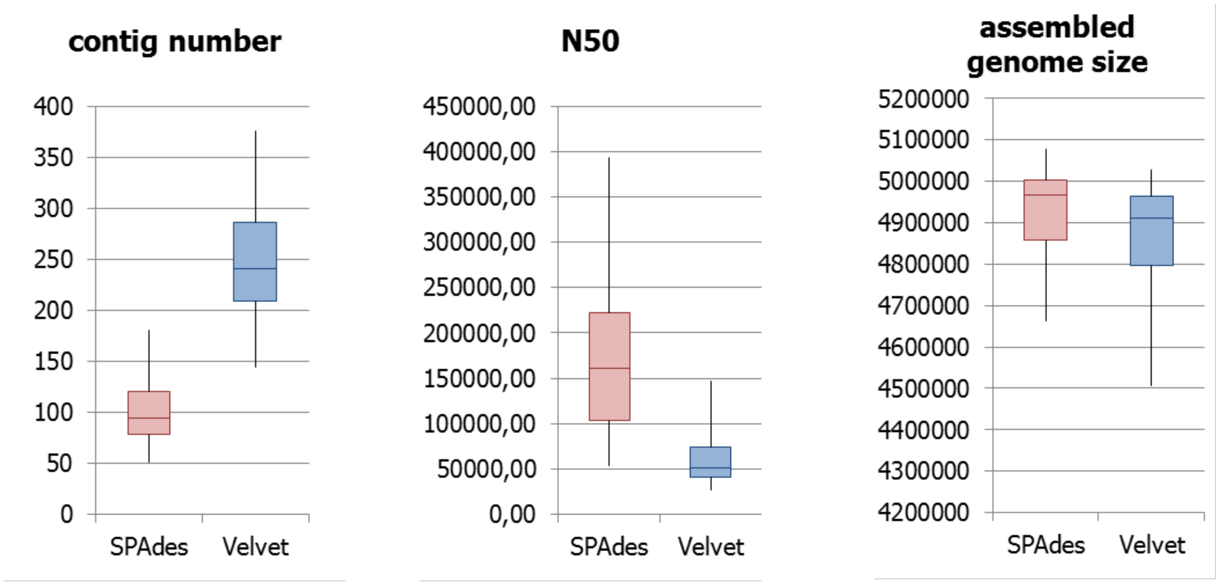

**Figure 1: Overall assembly quality.** Graphical representation of overall assembly quality parameters including contig numbers, N50 values and genome sizes of genomes assembled with either SPAdes or Velvet.

Supplementary Table - List of Strains (see also Supplementary Table 2 (Annex B)).

| | Spades | | | Velvet | | |
|---|---|---|---|---|---|---|
| sample_name | genome_size | contigs | n50 | genome_size | contigs | n50 |
| 03-02917 | 4674923 | 150 | 70852 | 4505678 | 360 | 26926 |
| 06-02242 | 4762839 | 157 | 88213 | 4633625 | 335 | 30290 |
| 07-01597 | 4663179 | 64 | 225719 | 4577464 | 213 | 51461 |
| 08-00436 | 4896492 | 118 | 119248 | 4797832 | 314 | 35933 |
| 08-00436 | 4967144 | 79 | 247068 | 4940435 | 222 | 85291 |
| 08-00844 | 4970846 | 91 | 213767 | 4941452 | 230 | 58722 |
| 08-00955 | 4965087 | 100 | 155361 | 4876611 | 278 | 43853 |
| 08-03422 | 4955841 | 120 | 137558 | 4876128 | 293 | 44300 |
| 09-02362 | 4871450 | 88 | 174043 | 4804866 | 227 | 53647 |
| 09-02946 | 5034312 | 91 | 225719 | 5027353 | 200 | 85169 |
| 09-02986 | 4954613 | 146 | 103875 | 4844786 | 337 | 30225 |
| 09-03610 | 4926660 | 97 | 164864 | 4918582 | 205 | 74734 |
| 09-04431 | 4962053 | 88 | 187927 | 4919965 | 239 | 51201 |
| 10-03145 | 4915801 | 181 | 53662 | 4754476 | 354 | 31818 |
| 10-03460 | 4818113 | 63 | 368622 | 4788341 | 346 | 34646 |
| 10-04072 | 4913494 | 122 | 82860 | 4833220 | 270 | 46531 |
| 10-04072 | 4909537 | 81 | 165445 | 4883184 | 192 | 76987 |
| 10-05043 | 4991716 | 172 | 68232 | 4888669 | 376 | 30963 |
| 11-01176 | 4782703 | 113 | 124638 | 4720850 | 271 | 44563 |
| 11-01525 | 4972448 | 92 | 184458 | 4962843 | 183 | 103705 |
| 11-02165 | 4966007 | 86 | 166565 | 4907581 | 242 | 44379 |
| 11-03654 | 5012273 | 83 | 173228 | 4986940 | 224 | 54113 |
| 11-03655 | 5011129 | 72 | 393606 | 4969447 | 222 | 56233 |
| 11-03656 | 5013701 | 86 | 206171 | 4995442 | 189 | 96509 |
| 11-04054 | 4897942 | 140 | 77220 | 4859167 | 290 | 40921 |
| 11-04056 | 4912808 | 90 | 165788 | 4873888 | 238 | 62293 |
| 11-04559 | 5014967 | 69 | 368674 | 5007664 | 144 | 146576 |
| 12-00555 | 5007211 | 115 | 94646 | 4855916 | 287 | 41914 |
| 12-01208 | 5016473 | 93 | 157181 | 4958028 | 248 | 48398 |
| 12-02541 | 4707937 | 128 | 93229 | 4634546 | 285 | 37900 |
| 12-02857 | 4774719 | 124 | 96314 | 4678889 | 302 | 35324 |
| 13-SA02194 | 4970145 | 75 | 385587 | 4943543 | 167 | 90284 |
| 13-SA02281 | 5008432 | 120 | 96736 | 4968954 | 303 | 38101 |
| 13-SA02283 | 4983075 | 68 | 253523 | 4968764 | 199 | 74230 |
| 13-SA02300 | 4986840 | 98 | 147698 | 4964091 | 248 | 46308 |
| 13-SA02435 | 4982663 | 104 | 121634 | 4949929 | 236 | 53586 |
| 13-SA02656 | 5076872 | 124 | 100656 | 5021983 | 285 | 45019 |
| 13-SA02788 | 4967735 | 80 | 192581 | 4948003 | 216 | 56954 |
| 14-SA00333 | 5010528 | 62 | 231654 | 4995680 | 210 | 83814 |
| 14-SA00775 | 4813906 | 109 | 103703 | 4772262 | 248 | 38773 |
| 14-SA00777 | 4987252 | 95 | 134015 | 4950980 | 259 | 51549 |
| 14-SA00918 | 4964052 | 96 | 121174 | 4914842 | 252 | 44954 |
| 14-SA01149 | 5013356 | 60 | 368866 | 4999641 | 185 | 93083 |
| 14-SA02536 | 5014878 | 69 | 275055 | 4998872 | 200 | 79253 |
| 14-SA02741 | 5009213 | 122 | 128575 | 4941385 | 287 | 47436 |
| 14-SA02860 | 4993807 | 116 | 131105 | 4961677 | 234 | 52422 |
| 15-SA00146 | 4776301 | 136 | 62450 | 4722696 | 267 | 35590 |
| 15-SA01434 | 4807642 | 67 | 172824 | 4795487 | 174 | 79619 |
| 15-SA01523 | 4805362 | 51 | 392833 | 4797115 | 175 | 82555 |
| 15-SA02829 | 4806710 | 64 | 231776 | 4789754 | 213 | 58324 |