Comparison of assemblies for *Vibrio cholerae*

[Table S 1 Comparison of contigs from CABOG runs on HiSeq data 3](#_Toc393208417)

[Table S 2 Comparison of scaffolds from CABOG runs on HiSeq data 5](#_Toc393208418)

[Table S 3 Comparison of contigs from CABOG runs on MiSeq data 7](#_Toc393208419)

[Table S 4 Comparison of scaffolds from CABOG runs on MiSeq data 9](#_Toc393208420)

[Table S 5 Comparison of contigs from MIRA runs on HiSeq data 11](#_Toc393208421)

[Table S 6 Comparison of contigs from MIRA runs on MiSeq data 13](#_Toc393208422)

[Table S 7 Comparison of contigs from MaSuRCA runs on HiSeq data 15](#_Toc393208423)

[Table S 8 Comparison of scaffolds from MaSuRCA runs on HiSeq data 17](#_Toc393208424)

[Table S 9 Comparison of contigs from MaSuRCA runs on MiSeq data 19](#_Toc393208425)

[Table S 10 Comparison of scaffolds from MaSuRCA runs on MiSeq data 21](#_Toc393208426)

[Table S 11 Comparison of contigs from Soapdenovo runs on HiSeq data 23](#_Toc393208427)

[Table S 12 Comparison of scaffolds from SOAPdenovo runs on HiSeq data 25](#_Toc393208428)

[Table S 13 Comparison of contigs from SOAPdenovo runs on MiSeq data 27](#_Toc393208429)

[Table S 14 Comparison of scaffolds from SOAPdenovo runs on MiSeq data 29](#_Toc393208430)

[Table S 15 Comparison of contigs from SPAdes runs on HiSeq data 31](#_Toc393208431)

[Table S 16 Comparison of scaffolds from SPAdes runs on HiSeq data 33](#_Toc393208432)

[Table S 17 Comparison of contigs from SPAdes runs on MiSeq data 35](#_Toc393208433)

[Table S 18 Comparison of scaffolds from SPAdes runs on MiSeq data 37](#_Toc393208434)

[Table S 19 Comparison of contigs from Velvet runs on HiSeq data 39](#_Toc393208435)

[Table S 20 Comparison of scaffolds from Velvet runs on HiSeq data 41](#_Toc393208436)

[Table S 21 Comparison of contigs from Velvet runs on MiSeq data 43](#_Toc393208437)

[Table S 22 Comparison of scaffolds from Velvet runs on MiSeq data 45](#_Toc393208438)

All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 200 bp)" and "Total length (>= 200 bp)"). Best result for each metric is written in bold. The first data column (Supplementary table Sx) refers to data obtained from GAGE-Bs supplementary material apr4[[1]](#footnote-1). All statistics for HiSeq data is listed in GAGE-Bs Supplementary Table S6 while statistics for MiSeq data is listed in GAGE-Bs Supplementary Table S7.

The values for N50/NA50 are written in kb to make it easier to compare to the values from the GAGE-B Supplementary Tables S6-S7. The second column marked as “Assembly file” refers to assemblies downloaded from <http://ccb.jhu.edu/gage_b/genomeAssemblies/index.html> with various assemblers. There’s a difference between two GAGE-B values (supplementary vs. assembly file) if the adjacent cells are colored light red. A cell is marked with the value N/A if there is no data available for that given metric. For the duplication ratio obtained from GAGE-Bs supplementary material, since the number seems to be rounded up to 1 decimal, it has been excluded from the comparison of best result unless all the other values exceed 1.0.

Each assembly contains two files used in the comparison, one with contigs and the other containing scaffolds. The only exception to this is assemblies computed by MIRA which only contains one file with contigs.

Table S 1 Comparison of contigs from CABOG runs on HiSeq data

The values for the first column (GAGE-B supplementary table S6) are obtained from GAGE-Bs supplementary material, while the next column (GAGE-B assembly file) represents data computed by running assemblies on QUAST 2.2. The last two columns represents assemblies reproduced with HiSeq data using CABOG 7.0/8.1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Assembly | GAGE-B supp | GAGE-B  (Cabog 7.0) | Cabog 7.0 | Cabog 8.1 |
| # contigs (>= 200 bp) | **127** | **127** | 144 | 519 |
| # contigs (>= 1000 bp) | N/A | **127** | 142 | 499 |
| Total length (>= 200 bp) | N/A | **3 855 600** | 3 849 141 | 3 708 859 |
| Total length (>= 1000 bp) | N/A | **3 855 600** | 3 848 582 | 3 698 786 |
| # contigs | N/A | **127** | 142 | 509 |
| Total length | N/A | **3 855 600** | 3 848 582 | 3 705 707 |
| Largest contig | N/A | **256 726** | 224 340 | 39 047 |
| Reference length | 4 033 464 | 4 033 464 | 4 033 464 | 4 033 464 |
| GC (%) | N/A | 47.53 | 47.54 | **47.94** |
| Reference GC (%) | N/A | **47.49** | **47.49** | **47.49** |
| N50 | 57 900 | **61 249** | 57 089 | 10 437 |
| NG50 | N/A | **57 883** | 55 585 | 9 568 |
| N75 | N/A | 29 636 | **30 707** | 6 201 |
| NG75 | N/A | **26 097** | 25 091 | 5 114 |
| L50 | N/A | **19** | 20 | 101 |
| LG50 | N/A | **21** | 22 | 117 |
| L75 | N/A | **43** | **43** | 217 |
| LG75 | N/A | **48** | 49 | 260 |
| # misassemblies | 33 | 20 | 10 | **8** |
| Misassembled contigs length | N/A | 385 832 | 94 087 | **39 006** |
| # local misassemblies | 12 | 11 | **7** | **7** |
| # unaligned contigs | 0 + ? part | **0 + 0 part** | **0 + 0 part** | 0 + 1 part |
| Unaligned contigs length | N/A | **0** | **0** | 52 |
| Genome fraction (%) | **96.6** | 95.623 | 95.361 | 91.790 |
| Duplication ratio | 1.0 | 1.007 | 1.006 | **1.002** |
| # N's per 100 kbp | N/A | **0.00** | **0.00** | **0.00** |
| # mismatches per 100 kbp | N/A | 17.01 | **7.72** | 9.53 |
| # indels per 100 kbp | N/A | 7.34 | 7.41 | **4.65** |
| # genes | N/A | **3 374 + 123 part** | 3 346 + 120 part | 2 970 + 384 part |
| Largest alignment | N/A | **256 726** | 224 338 | 39 047 |
| NA50 | 48 800 | **57 813** | 57 089 | 10 437 |
| NGA50 | N/A | 53 007 | **55 585** | 9 546 |
| NA75 | N/A | 28 288 | **28 862** | 6 069 |
| NGA75 | N/A | **25 832** | 25 091 | 5 039 |
| LA50 | N/A | 21 | **20** | 101 |
| LGA50 | N/A | **22** | **22** | 118 |
| LA75 | N/A | 46 | **44** | 218 |
| LGA75 | N/A | 51 | **49** | 262 |

Table S 2 Comparison of scaffolds from CABOG runs on HiSeq data

The values for the first column (GAGE-B supplementary table S6) are obtained from GAGE-Bs supplementary material, while the next column (GAGE-B assembly file) represents data computed by running assemblies on QUAST 2.2. The last two columns represents assemblies reproduced with HiSeq data using CABOG 7.0/8.1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Assembly | GAGE-B  supp | GAGE-B | Cabog 7.0 | Cabog 8.1 |
| # scaffold (>= 500 bp) | 108 | 108 | **94** | 183 |
| # scaffold (>= 1000 bp) | N/A | 108 | **94** | 183 |
| Total length (>= 500 bp) | N/A | **3 855 980** | 3 850 141 | 3 717 805 |
| Total length (>= 1000 bp) | N/A | **3 855 980** | 3 850 141 | 3 717 805 |
| # scaffold | N/A | 108 | **94** | 183 |
| Total length | N/A | **3 855 980** | 3 850 141 | 3 717 805 |
| Largest contig | N/A | 256 726 | **355 677** | 184 259 |
| Reference length | 4 033 464 | 4 033 464 | 4 033 464 | 4 033 464 |
| GC (%) | N/A | 47.53 | 47.54 | **47.93** |
| Reference GC (%) | N/A | **47.49** | **47.49** | **47.49** |
| N50 | 67 000 | 67 078 | **134 075** | 38 698 |
| NG50 | N/A | 67 009 | **102 138** | 35 387 |
| N75 | N/A | 34 758 | **52 902** | 19 302 |
| NG75 | N/A | 28 288 | **45 596** | 17 084 |
| L50 | N/A | 16 | **10** | 30 |
| LG50 | N/A | 17 | **11** | 34 |
| L75 | N/A | 36 | **24** | 64 |
| LG75 | N/A | 40 | **27** | 77 |
| # misassemblies | 34 | 21 | 11 | **10** |
| Misassembled contigs length | N/A | 531 956 | 183 425 | **169 129** |
| # local misassemblies | 24 | 23 | **13** | 38 |
| # unaligned contigs | 0 + ? part | **0 + 0 part** | **0 + 0 part** | **0 + 0 part** |
| Unaligned contigs length | N/A | **0** | **0** | **0** |
| Genome fraction (%) | **96.6** | 95.629 | 95.321 | 92.029 |
| Duplication ratio | 1.0 | **1.012** | 1.023 | 1.044 |
| # N's per 100 kbp | N/A | **9.85** | 25.97 | 189.33 |
| # mismatches per 100 kbp | N/A | 17.32 | **7.41** | 8.49 |
| # indels per 100 kbp | N/A | **9.00** | 18.75 | 65.49 |
| # genes | N/A | 3 380 + 117 part | **3 383 + 84 part** | 3 185 + 170 part |
| Largest alignment | N/A | 256 726 | **355 655** | 184 259 |
| NA50 | 53 200 | 63 201 | **134 075** | 38 209 |
| NGA50 | N/A | 62 046 | **102 118** | 34 659 |
| NA75 | N/A | 33 039 | **47 302** | 19 085 |
| NGA75 | N/A | 28 288 | **40 477** | 16 587 |
| LA50 | N/A | 18 | **10** | 31 |
| LGA50 | N/A | 19 | **11** | 35 |
| LA75 | N/A | 39 | **25** | 65 |
| LGA75 | N/A | 43 | **28** | 78 |

Table S 3 Comparison of contigs from CABOG runs on MiSeq data

The values for the first column (GAGE-B supplementary table S7) are obtained from GAGE-Bs supplementary material, while the next column (GAGE-B assembly file) represents data computed by running assemblies on QUAST 2.2. The last two columns represents assemblies reproduced with MiSeq data using CABOG 7.0/8.1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Assembly | GAGE-B supp | GAGE-B | Cabog 7.0 | Cabog 8.1 |
| # contigs (>= 200 bp) | 241 | 241 | **188** | 286 |
| # contigs (>= 1000 bp) | N/A | 241 | **188** | 286 |
| Total length (>= 200 bp) | N/A | **3 953 298** | 318 746 | 3 790 980 |
| Total length (>= 1000 bp) | N/A | **3 953 298** | 318 746 | 3 790 980 |
| # contigs | N/A | 241 | **188** | 286 |
| Total length | N/A | **3 953 298** | 318 746 | 3 790 980 |
| Largest contig | N/A | **140 691** | 5 646 | 106 586 |
| Reference length | 4 967 469 | 4 033 464 | 4 033 464 | 4 033 464 |
| GC (%) | N/A | 47.52 | 41.54 | **47.92** |
| Reference GC (%) | N/A | **47.49** | **47.49** | **47.49** |
| N50 | 32 800 | **33 710** | 1 707 | 24 969 |
| NG50 | N/A | **32 790** | N/A | 22 450 |
| N75 | N/A | **17 245** | 1 367 | 12 543 |
| NG75 | N/A | **16 374** | N/A | 10 823 |
| L50 | N/A | **34** | 69 | 47 |
| LG50 | N/A | **35** | N/A | 52 |
| L75 | N/A | **74** | 121 | 101 |
| LG75 | N/A | **77** | N/A | 116 |
| # misassemblies | 22 | 17 | **7** | 8 |
| Misassembled contigs length | N/A | 128 064 | **15 751** | 30 914 |
| # local misassemblies | 7 | 7 | **1** | 6 |
| # unaligned contigs | 1 + ? part | 1 + 2 part | 0 + 97 part | **0 + 1 part** |
| Unaligned contigs length | N/A | 5 249 | 7 706 | **49** |
| Genome fraction (%) | **97.8** | 96.968 | 7.639 | 93.765 |
| Duplication ratio | 1.0 | 1.016 | 1.011 | **1.009** |
| # N's per 100 kbp | N/A | **0.00** | **0.00** | **0.00** |
| # mismatches per 100 kbp | N/A | 8.16 | 36.35 | **6.95** |
| # indels per 100 kbp | N/A | **3.63** | 17.20 | 3.70 |
| # genes | N/A | **3 401 + 144 part** | 123 + 277 part | 3 286 + 166 part |
| Largest alignment | N/A | **140 691** | 5 582 | 106 586 |
| NA50 | 32 500 | **33 710** | 1 624 | 24 969 |
| NGA50 | N/A | **32 784** | N/A | 22 450 |
| NA75 | N/A | **17 241** | 1 288 | 12 543 |
| NGA75 | N/A | **16 374** | N/A | 10 092 |
| LA50 | N/A | **34** | 72 | 47 |
| LGA50 | N/A | **35** | N/A | 52 |
| LA75 | N/A | **74** | 128 | 101 |
| LGA75 | N/A | **77** | N/A | 117 |

Table S 4 Comparison of scaffolds from CABOG runs on MiSeq data

The values for the first column (GAGE-B supplementary table S7) are obtained from GAGE-Bs supplementary material, while the next column (GAGE-B assembly file) represents data computed by running assemblies on QUAST 2.2. The last two columns represents assemblies reproduced with MiSeq data using CABOG 7.0/8.1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Assembly | GAGE-B supp | GAGE-B | Cabog 7.0 | Cabog 8.1 |
| # scaffold (>= 500 bp) | 241 | 241 | **188** | 285 |
| # scaffold (>= 1000 bp) | N/A | 241 | **188** | 285 |
| Total length (>= 500 bp) | N/A | **3 953 298** | 318 746 | 3 791 000 |
| Total length (>= 1000 bp) | N/A | **3 953 298** | 318 746 | 3 791 000 |
| # scaffold | N/A | 241 | **188** | 285 |
| Total length | N/A | **3 953 298** | 318 746 | 3 791 000 |
| Largest contig | N/A | **140 691** | 5 646 | 106 586 |
| Reference length | 4 967 469 | 4 033 464 | 4 033 464 | 4 033 464 |
| GC (%) | N/A | 47.52 | 41.54 | **47.92** |
| Reference GC (%) | N/A | **47.49** | **47.49** | **47.49** |
| N50 | 32 800 | **33 710** | 1 707 | 24 969 |
| NG50 | N/A | **32 790** | N/A | 22 450 |
| N75 | N/A | **17 245** | 1 367 | 12 652 |
| NG75 | N/A | **16374** | N/A | 10 885 |
| L50 | N/A | **34** | 69 | 47 |
| LG50 | N/A | **35** | N/A | 52 |
| L75 | N/A | **74** | 121 | 100 |
| LG75 | N/A | **77** | N/A | 116 |
| # misassemblies | 22 | 17 | **7** | 9 |
| Misassembled contigs length | N/A | 128 064 | **15 751** | 46 637 |
| # local misassemblies | 7 | 7 | **1** | 6 |
| # unaligned contigs | 1 + ? part | 1 + 2 part | 0 + 97 part | **0 + 1 part** |
| Unaligned contigs length | N/A | 5 249 | 7 706 | **49** |
| Genome fraction (%) | **97.8** | 96.968 | 7.639 | 93.765 |
| Duplication ratio | 1.0 | 1.016 | 1.011 | **1.009** |
| # N's per 100 kbp | N/A | **0.00** | **0.00** | 0.53 |
| # mismatches per 100 kbp | N/A | 8.21 | 36.35 | **7.01** |
| # indels per 100 kbp | N/A | **3.63** | 17.20 | 3.70 |
| # genes | N/A | **3 401 + 144 part** | 123 + 277 part | 3 286 + 166 part |
| Largest alignment | N/A | **140 691** | 5 582 | 106 586 |
| NA50 | 32 500 | **33 710** | 1 624 | 24 969 |
| NGA50 | N/A | **32 784** | N/A | 22 450 |
| NA75 | N/A | **17 241** | 1288 | 12 543 |
| NGA75 | N/A | **16 374** | N/A | 10 092 |
| LA50 | N/A | **34** | 72 | 47 |
| LGA50 | N/A | **35** | N/A | 52 |
| LA75 | N/A | **74** | 128 | 101 |
| LGA75 | N/A | **77** | N/A | 117 |

Table S 5 Comparison of contigs from MIRA runs on HiSeq data

The values for the first column (GAGE-B supplementary table S6) are obtained from GAGE-Bs supplementary material, while the next column (GAGE-B assembly file) represents data computed by running assemblies on QUAST 2.2. The last column represents an assembly reproduced with HiSeq data using MIRA 3.4.0.

|  |  |  |  |
| --- | --- | --- | --- |
| Assembly | GAGE-B supp | GAGE-B | MIRA 3.4.0 |
| # contigs (>= 200 bp) | 728 | **733** | 1524 |
| # contigs (>= 1000 bp) | N/A | **115** | 946 |
| Total length (>= 200 bp) | N/A | **4 113 459** | 3 956 352 |
| Total length (>= 1000 bp) | N/A | **3 932 194** | 3 646 461 |
| # contigs | N/A | **152** | 1234 |
| Total length | N/A | **3 957 629** | 3 858 629 |
| Largest contig | N/A | **316 666** | 22 077 |
| Reference length | 4 033 464 | 4 033 464 | 4 033 464 |
| GC (%) | N/A | 47.50 | **47.73** |
| Reference GC (%) | N/A | **47.49** | **47.49** |
| N50 | **92 000** | **92 000** | 4 840 |
| NG50 | N/A | **92 000** | 4 642 |
| N75 | N/A | **56 477** | 2 597 |
| NG75 | N/A | **55 736** | 2 373 |
| L50 | N/A | **14** | 237 |
| LG50 | N/A | **14** | 255 |
| L75 | N/A | **27** | 502 |
| LG75 | N/A | **28** | 555 |
| # misassemblies | 89 | **24** | 45 |
| Misassembled contigs length | N/A | 558 799 | **125 492** |
| # local misassemblies | 15 | **9** | 10 |
| # unaligned contigs | 10 + ? part | **0 + 2 part** | 0 + 7 part |
| Unaligned contigs length | N/A | **173** | 615 |
| Genome fraction (%) | **99.7** | 97.925 | 94.917 |
| Duplication ratio | 1.0 | 1.016 | **1.009** |
| # N's per 100 kbp | N/A | **5.00** | 23.12 |
| # mismatches per 100 kbp | N/A | **10.38** | 17.37 |
| # indels per 100 kbp | N/A | **3.85** | 4.41 |
| # genes | N/A | **3 516 + 65 part** | 2 627 + 893 part |
| Largest alignment | N/A | **211 909** | 22 077 |
| NA50 | 87 100 | **89 505** | 4 812 |
| NGA50 | N/A | **89 505** | 4 596 |
| NA75 | N/A | **52 881** | 2 576 |
| NGA75 | N/A | **51 186** | 2 336 |
| LA50 | N/A | **15** | 241 |
| LGA50 | N/A | **15** | 260 |
| LA75 | N/A | **29** | 509 |
| LGA75 | N/A | **30** | 562 |

Table S 6 Comparison of contigs from MIRA runs on MiSeq data

The values for the first column (GAGE-B supplementary table S7) are obtained from GAGE-Bs supplementary material, while the next column (GAGE-B assembly file) represents data computed by running assemblies on QUAST 2.2. The last column represents an assembly reproduced with MiSeq data using MIRA 3.4.0.

|  |  |  |  |
| --- | --- | --- | --- |
| Assembly | GAGE-B supp | GAGE-B | MIRA 3.4.0 |
| # contigs (>= 200 bp) | 430 | 431 | **224** |
| # contigs (>= 1000 bp) | N/A | **96** | 131 |
| Total length (>= 200 bp) | N/A | **4 100 278** | 4 008 924 |
| Total length (>= 1000 bp) | N/A | **3 971 344** | 3 964 673 |
| # contigs | N/A | **154** | 162 |
| Total length | N/A | **4 011 692** | 3 986 805 |
| Largest contig | N/A | **450 326** | 383 237 |
| Reference length | 4 967 469 | 4 033 464 | 4 033 464 |
| GC (%) | N/A | 47.48 | **47.50** |
| Reference GC (%) | N/A | **47.49** | **47.49** |
| N50 | **112 900** | **112 926** | 108 646 |
| NG50 | N/A | **112 926** | 108 646 |
| N75 | N/A | **59 711** | 43 347 |
| NG75 | N/A | **56 632** | 42 311 |
| L50 | N/A | **9** | 11 |
| LG50 | N/A | **9** | 11 |
| L75 | N/A | **20** | 25 |
| LG75 | N/A | **21** | 26 |
| # misassemblies | 148 | 49 | **23** |
| Misassembled contigs length | N/A | 1 093 737 | **357 376** |
| # local misassemblies | 17 | 7 | **4** |
| # unaligned contigs | 20 + ? part | 5 + 10 part | **1 + 4 part** |
| Unaligned contigs length | N/A | 13 951 | 11 825 |
| Genome fraction (%) | **99.6** | 98.311 | 98.078 |
| Duplication ratio | 1.0 | 1.016 | **1.012** |
| # N's per 100 kbp | N/A | 1.74 | **0.70** |
| # mismatches per 100 kbp | N/A | 8.88 | **5.28** |
| # indels per 100 kbp | N/A | 4.06 | **2.96** |
| # genes | N/A | **3 559 + 64 part** | 3 534 + 84 part |
| Largest alignment | N/A | **450 326** | 383 237 |
| NA50 | **108 700** | **108 689** | 108 646 |
| NGA50 | N/A | 108 689 | **108 646** |
| NA75 | N/A | 52 144 | **39 171** |
| NGA75 | N/A | 52 144 | **38 713** |
| LA50 | N/A | **9** | 11 |
| LGA50 | N/A | **9** | 11 |
| LA75 | N/A | **22** | 27 |
| LGA75 | N/A | **22** | 28 |

Table S 7 Comparison of contigs from MaSuRCA runs on HiSeq data

The values for the first column (GAGE-B supplementary table S6) are obtained from GAGE-Bs supplementary material, while the next column (GAGE-B assembly file) represents data computed by running assemblies on QUAST 2.2. The last four column represents assemblies reproduced with HiSeq data using MaSuRCA 1.8.3/2.1.0 with k values 89 and 99.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Assembly | GAGE-B supp | GAGE-B | MaSuRCA 1.8.3 k89 | MaSuRCA  1.8.3 k99 | MaSuRCA  2.1.0 k89 | MaSuRCA  2.1.0 k99 |
| # contigs  (>= 200 bp) | **105** | **105** | 137 | 330 | 179 | 119 |
| # contigs  (>= 1000 bp) | N/A | 80 | 108 | 273 | 135 | **72** |
| Total length (>= 200 bp) | N/A | 3 994 617 | 3 994 843 | 3 876 806 | **4 013 483** | 194 673 |
| Total length  (>= 1000 bp) | N/A | 3 980 159 | 3 977 918 | 3 839 316 | **3 986 296** | 162 098 |
| # contigs | N/A | **91** | 121 | 314 | 164 | 108 |
| Total length | N/A | 3 989 162 | 3 988 448 | 3 870 114 | **4 007 678** | 190 023 |
| Largest contig | N/A | **555 664** | 356 964 | 117 961 | 211 069 | 4 811 |
| Reference length | 4 033 464 | 4 033 464 | 4 033 464 | 4 033 464 | 4 033 464 | 4 033 464 |
| GC (%) | N/A | 47.50 | 47.51 | **47.52** | 47.51 | 43.79 |
| Reference GC (%) | N/A | **47.49** | **47.49** | **47.49** | **47.49** | **47.49** |
| N50 | **241 600** | **241 604** | 108 823 | 35 282 | 93 473 | 2 355 |
| N75 | N/A | **85 155** | 57 567 | 13 936 | 43 374 | 1 312 |
| L50 | N/A | **6** | 9 | 32 | 14 | 29 |
| L75 | N/A | **13** | 23 | 79 | 30 | 56 |
| # misassemblies | 12 | 8 | 8 | 11 | 8 | **2** |
| Misassembled contigs length | N/A | 708 761 | 422 752 | 52 197 | 228 152 | **4 646** |
| # local misassemblies | 5 | 5 | 4 | 5 | 6 | **0** |
| # unaligned contigs | 0 + ? part | 0 + 2 part | 0 + 2 part | 0 + 2 part | 0 + 4 part | **0 + 0 part** |
| Unaligned contigs length | N/A | 399 | 399 | 104 | 475 | **0** |
| Genome  fraction (%) | **99.4** | 98.147 | 98.121 | 95.714 | 98.498 | 4.709 |
| Duplication ratio | 1.0 | 1.013 | 1.008 | **1.003** | 1.009 | 1.004 |
| # N's per 100 kbp | N/A | **0.00** | **0.00** | **0.00** | **0.00** | **0.00** |
| # mismatches  per 100 kbp | N/A | 68.61 | 68.93 | **9.51** | 50.57 | 30.01 |
| # indels  per 100 kbp | N/A | 6.90 | 6.62 | **4.14** | 5.64 | 10.53 |
| # genes | N/A | **3 573 + 46 part** | 3 552 + 65 part | 3 289 + 246 part | 3 541 + 87 part | 94 + 129 part |
| Largest alignment | N/A | **555 664** | 356 958 | 117 961 | 211 069 | 4 811 |
| NA50 | **236 400** | **236 373** | 90 882 | 35 282 | 81 529 | 2 208 |
| NA75 | N/A | **76 901** | 48 002 | 13 936 | 43 374 | 1 297 |
| LA50 | N/A | **6** | 10 | 32 | 15 | 30 |
| LA75 | N/A | **15** | 24 | 79 | 31 | 57 |

Table S 8 Comparison of scaffolds from MaSuRCA runs on HiSeq data

The values for the first column (GAGE-B supplementary table S6) are obtained from GAGE-Bs supplementary material, while the next column (GAGE-B assembly file) represents data computed by running assemblies on QUAST 2.2. The last four column represents assemblies reproduced with HiSeq data using MaSuRCA 1.8.3/2.1.0 with k values 89 and 99

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Assembly | GAGE-B supp | GAGE-B | MaSuRCA 1.8.3 k89 | MaSuRCA 1.8.3 k99 | MaSuRCA 2.1.0 k89 | MaSuRCA 2.1.0 k99 |
| # scaffold  (>= 500 bp) | **88** | **88** | **88** | 245 | 118 | 167 |
| # scaffold  (>= 1000 bp) | N/A | **77** | **77** | 215 | 99 | 143 |
| Total length  (>= 500 bp) | N/A | 3 989 222 | 3 990 088 | 3 875 030 | **4 008 981** | 3 969 261 |
| Total length  (>= 1000 bp) | N/A | 3 980 219 | 3 981 085 | 3 852 394 | **3 994 223** | 3 951 885 |
| # scaffold | N/A | **88** | **88** | 245 | 118 | 167 |
| Total length | N/A | 3 989 222 | 3 990 088 | 3 875 030 | **4 008 981** | 3 969 261 |
| Largest contig | N/A | 555 664 | 555 775 | 199 706 | **555 891** | 232 643 |
| Reference length | 4 033 464 | 4 033 464 | 4 033 464 | 4 033 464 | 4 033 464 | 4 033 464 |
| GC (%) | N/A | 47.50 | 47.50 | **47.51** | **47.51** | 47.44 |
| Reference  GC (%) | N/A | **47.49** | **47.49** | **47.49** | **47.49** | **47.49** |
| N50 | 246 500 | 246 505 | **246 836** | 46 624 | 133 199 | 65 767 |
| NG50 | N/A | 246 505 | **246 836** | 46 196 | 133 199 | 65 767 |
| N75 | N/A | **85 155** | **85 155** | 19 711 | 77 491 | 43 921 |
| NG75 | N/A | 79 323 | **79 351** | 16 912 | 77 491 | 42 523 |
| L50 | N/A | **6** | **6** | 23 | 8 | 18 |
| LG50 | N/A | **6** | **6** | 24 | 8 | 18 |
| L75 | N/A | **13** | **13** | 53 | 18 | 37 |
| LG75 | N/A | **14** | **14** | 59 | 18 | 38 |
| # misassemblies | 11 | **9** | 10 | 12 | 13 | 11 |
| Misassembled contigs length | N/A | 716 209 | **788 570** | 194 566 | 610 408 | 151 551 |
| # local misassemblies | 8 | 7 | 15 | 31 | 20 | **6** |
| # unaligned contigs | 0 + ? part | 0 + 2 part | 0 + 2 part | **0 + 1 part** | 0 + 4 part | **0 + 1 part** |
| Unaligned contigs length | N/A | 399 | 399 | **72** | 475 | 453 |
| Genome  fraction (%) | **99.3** | 98.147 | 98.152 | 95.804 | 98.498 | 96.353 |
| Duplication ratio | 1.0 | 1.013 | 1.013 | **1.009** | 1.010 | 1.027 |
| # N's  per 100 kbp | N/A | 1.50 | 17.54 | 40.26 | 23.45 | **1.01** |
| # mismatches per 100 kbp | N/A | 68.63 | 68.83 | 9.78 | 50.95 | **5.04** |
| # indels  per 100 kbp | N/A | 6.85 | 12.07 | 13.69 | 16.71 | **3.37** |
| # genes | N/A | **3 573 + 46 part** | 3 568 + 51 part | 3 314 + 226 part | 3 557 + 71 part | 3 480 + 60 part |
| Largest alignment | N/A | 555 664 | 555 729 | 199 686 | **555 891** | 232 643 |
| NA50 | 236 400 | 236 373 | **236 430** | 46 604 | 132 265 | 65 252 |
| NGA50 | N/A | 236 373 | **236 430** | 46 176 | 132 265 | 63 260 |
| NA75 | N/A | **76 901** | **76 901** | 19 307 | 71 696 | 43 921 |
| NGA75 | N/A | **76 901** | **76 901** | 16 019 | 70 982 | 42 523 |
| LA50 | N/A | **6** | **6** | 23 | 9 | 18 |
| LGA50 | N/A | **6** | **6** | 24 | 9 | 19 |
| LA75 | N/A | **15** | **15** | 54 | 19 | 37 |
| LGA75 | N/A | **15** | **15** | 61 | 20 | 38 |

Table S 9 Comparison of contigs from MaSuRCA runs on MiSeq data

The values for the first column (GAGE-B supplementary table S7) are obtained from GAGE-Bs supplementary material, while the next column (GAGE-B assembly file) represents data computed by running assemblies on QUAST 2.2. The last four column represents assemblies reproduced with MiSeq data using MaSuRCA 1.8.3/2.1.0 with k values 89 and 99.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Assembly | GAGE-B supp | GAGE-B | MaSuRCA 1.8.3 k89 | MaSuRCA  1.8.3 k99 | MaSuRCA  2.1.0 k89 | MaSuRCA  2.1.0 k99 |
| # contigs  (>= 200 bp) | **173** | **173** | 174 | **173** | 182 | 182 |
| # contigs  (>= 1000 bp) | N/A | **142** | 153 | **142** | 149 | 144 |
| Total length  (>= 200 bp) | N/A | **4 034 988** | 4 013 202 | **4 034 988** | 3 973 806 | 3 974 638 |
| Total length  (>= 1000 bp) | N/A | **4 014 349** | 3 999 858 | **4 014 349** | 3 952 666 | 3 951 385 |
| # contigs | N/A | **163** | 167 | **163** | 173 | 168 |
| Total length | N/A | **4 030 761** | 4 010 282 | **4 030 761** | 3 970 080 | 3 968 761 |
| Largest contig | N/A | **255 146** | 232 613 | **255 146** | 232 643 | 232 643 |
| Reference length | 4 967 469 | 4 033 464 | 4 033 464 | 4 033 464 | 4 033 464 | 4 033 464 |
| GC (%) | N/A | **47.47** | **47.47** | **47.47** | 47.44 | 47.44 |
| Reference  GC (%) | N/A | **47.49** | **47.49** | **47.49** | **47.49** | **47.49** |
| N50 | **76 100** | **76 131** | 61 317 | **76 131** | 62 007 | 65 767 |
| NG50 | N/A | **76 131** | 61 317 | **76 131** | 61 549 | 65 767 |
| N75 | N/A | **45 736** | 43 446 | **45 736** | 37 453 | 43 921 |
| NG75 | N/A | **45 736** | 43 446 | **45 736** | 35 688 | 42 523 |
| L50 | N/A | **16** | 20 | **16** | 20 | 18 |
| LG50 | N/A | **16** | 20 | **16** | 21 | 18 |
| L75 | N/A | **33** | 40 | **33** | 40 | 37 |
| LG75 | N/A | **33** | 40 | **33** | 41 | 38 |
| # misassemblies | 23 | **19** | 20 | **19** | 11 | 9 |
| Misassembled contigs length | N/A | 216 881 | **244 726** | 216 881 | 89 457 | 140 011 |
| # local misassemblies | 5 | **3** | 5 | **3** | 5 | 6 |
| # unaligned contigs | 0 + ? part | 0 + 3 part | 0 + 5 part | 0 + 3 part | 0 + 3 part | **0 + 1 part** |
| Unaligned contigs length | N/A | 811 | 892 | 811 | 599 | **453** |
| Genome  fraction (%) | **98.3** | 97.670 | 96.864 | 97.670 | 96.147 | 96.349 |
| Duplication ratio | 1.0 | **1.023** | 1.027 | **1.023** | 1.029 | 1.027 |
| # N's  per 100 kbp | N/A | **0.00** | **0.00** | **0.00** | **0.00** | **0.00** |
| # mismatches per 100 kbp | N/A | 6.27 | 6.27 | 6.27 | **5.03** | 5.04 |
| # indels per 100 kbp | N/A | 3.17 | 3.40 | 3.17 | **3.07** | 3.37 |
| # genes | N/A | **3 534 + 66 part** | 3 500 + 68 part | **3 534 + 66 part** | 3 474 + 59 part | 3 480 + 60 part |
| Largest alignment | N/A | **255 146** | 232 613 | **255 146** | 232 643 | 232 643 |
| NA50 | **71 600** | **76 131** | 60 468 | **76 131** | 61 330 | 65 252 |
| NGA50 | N/A | **76 131** | 58 460 | **76 131** | 60 667 | 63 260 |
| NA75 | N/A | **45 141** | 40 539 | **45 141** | 37 453 | 43 921 |
| NGA75 | N/A | **45 141** | 40 539 | **45 141** | 35 688 | 42 523 |
| LA50 | N/A | **16** | 20 | **16** | 20 | 18 |
| LGA50 | N/A | **16** | 21 | **16** | 21 | 19 |
| LA75 | N/A | **34** | 41 | **34** | 40 | 37 |
| LGA75 | N/A | **34** | 41 | **34** | 41 | 38 |

Table S 10 Comparison of scaffolds from MaSuRCA runs on MiSeq data

The values for the first column (GAGE-B supplementary table S7) are obtained from GAGE-Bs supplementary material, while the next column (GAGE-B assembly file) represents data computed by running assemblies on QUAST 2.2. The last four column represents assemblies reproduced with MiSeq data using MaSuRCA 1.8.3/2.1.0 with k values 89 and 99.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Assembly | GAGE-B supp | GAGE-B | MaSuRCA 1.8.3 k89 | MaSuRCA  1.8.3 k99 | MaSuRCA  2.1.0 k89 | MaSuRCA  2.1.0 k99 |
| # scaffold  (>= 500 bp) | **163** | **163** | 167 | **163** | 171 | 167 |
| # scaffold  (>= 1000 bp) | N/A | **142** | 153 | **142** | 148 | 143 |
| Total length  (>= 500 bp) | N/A | **4 030 761** | 4 010 282 | **4 030 761** | 3 970 120 | 3 969 261 |
| Total length  (>= 1000 bp) | N/A | **4 014 349** | 3 999 858 | **4 014 349** | 3 953 572 | 3 951 885 |
| # scaffold | N/A | **163** | 167 | **163** | 171 | 167 |
| Total length | N/A | **4 030 761** | 4 010 282 | **4 030 761** | 3 970 120 | 3 969 261 |
| Largest contig | N/A | **255 146** | 232 613 | **255 146** | 232 643 | 232 643 |
| Reference length | 4 967 469 | 4 033 464 | 4 033 464 | 4 033 464 | 4 033 464 | 4 033 464 |
| GC (%) | N/A | **47.47** | **47.47** | **47.47** | 47.44 | 47.44 |
| Reference  GC (%) | N/A | **47.49** | **47.49** | **47.49** | **47.49** | **47.49** |
| N50 | **76 100** | **76 131** | 61 317 | **76 131** | 62 007 | 65 767 |
| NG50 | N/A | **76 131** | 61 317 | **76 131** | 61 549 | 65 767 |
| N75 | N/A | **45 736** | 43 446 | **45 736** | 37 453 | 43 921 |
| NG75 | N/A | **45 736** | 43 446 | **45 736** | 35 688 | 42 523 |
| L50 | N/A | **16** | 20 | **16** | 20 | 18 |
| LG50 | N/A | **16** | 20 | **16** | 21 | 18 |
| L75 | N/A | **33** | 40 | **33** | 40 | 37 |
| LG75 | N/A | **33** | 40 | **33** | 41 | 38 |
| # misassemblies | 23 | 19 | 20 | 19 | 13 | **11** |
| Misassembled contigs length | N/A | 216 881 | **244 726** | 216 881 | 96 509 | 151 551 |
| # local misassemblies | 5 | **3** | 5 | **3** | 5 | 6 |
| # unaligned contigs | 0 + ? part | 0 + 3 part | 0 + 5 part | 0 + 3 part | 0 + 3 part | **0 + 1 part** |
| Unaligned contigs length | N/A | 811 | 892 | 811 | 599 | **453** |
| Genome  fraction (%) | **98.3** | 97.670 | 96.864 | 97.670 | 96.147 | 96.353 |
| Duplication ratio | 1.0 | **1.023** | 1.027 | **1.023** | 1.029 | 1.027 |
| # N's per 100 kbp | N/A | **0.00** | **0.00** | **0.00** | 1.01 | 1.01 |
| # mismatches  per 100 kbp | N/A | 6.27 | 6.30 | 6.27 | **5.03** | 5.04 |
| # indels  per 100 kbp | N/A | 3.17 | 3.40 | 3.17 | **3.07** | 3.37 |
| # genes | N/A | **3 534 + 66 part** | 3 500 + 68 part | **3 534 + 66 part** | 3 474 + 59 part | 3 480 + 60 part |
| Largest alignment | N/A | 255 146 | 232 613 | 255 146 | 232 643 | 232 643 |
| NA50 | 71 600 | **76 131** | 60 468 | **76 131** | 61 330 | 65 252 |
| NGA50 | N/A | **76 131** | 58 460 | **76 131** | 60 667 | 63 260 |
| NA75 | N/A | **45 141** | 40 539 | **45 141** | 37 453 | 43 921 |
| NGA75 | N/A | **45 141** | 40 539 | **45 141** | 35 688 | 42 523 |
| LA50 | N/A | **16** | 20 | **16** | 20 | 18 |
| LGA50 | N/A | **16** | 21 | **16** | 21 | 19 |
| LA75 | N/A | **34** | 41 | **34** | 40 | 37 |
| LGA75 | N/A | **34** | 41 | **34** | 41 | 38 |

Table S 11 Comparison of contigs from Soapdenovo runs on HiSeq data

The values for the first column (GAGE-B supplementary table S6) are obtained from GAGE-Bs supplementary material, while the next column (GAGE-B assembly file) represents data computed by running assemblies on QUAST 2.2. The last column represents an assembly reproduced with HiSeq data using SOAPdenovo2 with GapCloser.

|  |  |  |  |
| --- | --- | --- | --- |
| Assembly | GAGE-B supp | GAGE-B | SOAPdenovo2 |
| # contigs (>= 200 bp) | **139** | **139** | 462 |
| # contigs (>= 1000 bp) | N/A | **81** | 294 |
| Total length (>= 200 bp) | N/A | **3 945 991** | 3 931 705 |
| Total length (>= 1000 bp) | N/A | **3 919 085** | 3 851 746 |
| # contigs | N/A | **100** | 364 |
| Total length | N/A | **3 932 751** | 3 899 782 |
| Largest contig | N/A | **425 556** | 69 685 |
| Reference length | 4 033 464 | 4 033 464 | 4 033 464 |
| GC (%) | N/A | 47.52 | **47.53** |
| Reference GC (%) | N/A | 47.49 | 47.49 |
| N50 | 125 900 | **135 118** | 21 660 |
| NG50 | N/A | **125 939** | 20 717 |
| N75 | N/A | **65 780** | 12 597 |
| NG75 | N/A | **62 739** | 11 591 |
| L50 | N/A | **10** | 57 |
| LG50 | N/A | **11** | 60 |
| L75 | N/A | **21** | 115 |
| LG75 | N/A | **22** | 123 |
| # misassemblies | 26 | 15 | **2** |
| Misassembled contigs length | N/A | 1155029 | **9685** |
| # local misassemblies | 50 | 50 | **0** |
| # unaligned contigs | 5 + ? part | 1 + 1 part | **1 + 0 part** |
| Unaligned contigs length | N/A | 603 | **551** |
| Genome fraction (%) | **99.5** | 97.295 | 96.488 |
| Duplication ratio | 1.0 | 1.009 | **1.002** |
| # N's per 100 kbp | N/A | 0.00 | 0.00 |
| # mismatches per 100 kbp | N/A | 14.02 | **3.31** |
| # indels per 100 kbp | N/A | 5.58 | **2.85** |
| # genes | N/A | **3 479 + 50 part** | 3 274 + 189 part |
| Largest alignment | N/A | **425 699** | 69 685 |
| NA50 | 106 500 | **112 904** | 21 657 |
| NGA50 | N/A | **112 904** | 20 717 |
| NA75 | N/A | **62 739** | 12 597 |
| NGA75 | N/A | **62 256** | 11 588 |
| LA50 | N/A | **11** | 57 |
| LGA50 | N/A | **11** | 60 |
| LA75 | N/A | **22** | 115 |
| LGA75 | N/A | **23** | 123 |

Table S 12 Comparison of scaffolds from SOAPdenovo runs on HiSeq data

The values for the first column (GAGE-B supplementary table S6) are obtained from GAGE-Bs supplementary material, while the next column (GAGE-B assembly file) represents data computed by running assemblies on QUAST 2.2. The last column represents an assembly reproduced with HiSeq data using SOAPdenovo2 with GapCloser.

|  |  |  |  |
| --- | --- | --- | --- |
| Assembly | GAGE-B supp | GAGE-B | SOAPdenovo2 |
| # contigs (>= 500 bp) | **75** | **75** | 77 |
| # contigs (>= 1000 bp) | N/A | **58** | **58** |
| Total length (>= 500 bp) | N/A | **3 935 433** | 3 933 896 |
| Total length (>= 1000 bp) | N/A | **3 923 264** | 3 920 135 |
| # contigs | N/A | **75** | 77 |
| Total length | N/A | **3 935 433** | 3 933 896 |
| Largest contig | N/A | 574 497 | **574 502** |
| Reference length | 4 033 464 | 4 033 464 | 4 033 464 |
| GC (%) | N/A | **47.52** | 47.51 |
| Reference GC (%) | N/A | **47.49** | **47.49** |
| N50 | 181 100 | 200 529 | **200 760** |
| NG50 | N/A | 181 109 | **181 223** |
| N75 | N/A | 102 918 | **103 044** |
| NG75 | N/A | 102 918 | **103 044** |
| L50 | N/A | **6** | **6** |
| LG50 | N/A | **7** | **7** |
| L75 | N/A | **13** | **13** |
| LG75 | N/A | **13** | **13** |
| # misassemblies | 26 | 15 | **6** |
| Misassembled contigs length | N/A | 1 605 854 | **774 094** |
| # local misassemblies | **76** | **76** | 81 |
| # unaligned contigs | 1 + ? part | 1 + 1 part | **1 + 0 part** |
| Unaligned contigs length | N/A | 603 | **551** |
| Genome fraction (%) | **99.0** | 97.305 | 97.517 |
| Duplication ratio | 1.0 | **1.009** | 1.011 |
| # N's per 100 kbp | N/A | 56.11 | **31.6** |
| # mismatches per 100 kbp | N/A | 14.01 | **9.2** |
| # indels per 100 kbp | N/A | 5.58 | **5.05** |
| # genes | N/A | 3 480 + 50 part | **3 485 + 55 part** |
| Largest alignment | N/A | 574 463 | **574 661** |
| NA50 | 168 100 | 181 115 | **181 222** |
| NGA50 | N/A | 181 115 | **181 222** |
| NA75 | N/A | 92 025 | **92 227** |
| NGA75 | N/A | **92 025** | 81 094 |
| LA50 | N/A | **7** | **7** |
| LGA50 | N/A | **7** | **7** |
| LA75 | N/A | **14** | **14** |
| LGA75 | N/A | **14** | 15 |

Table S 13 Comparison of contigs from SOAPdenovo runs on MiSeq data

The values for the first column (GAGE-B supplementary table S7) are obtained from GAGE-Bs supplementary material, while the next column (GAGE-B assembly file) represents data computed by running assemblies on QUAST 2.2. The last column represents an assembly reproduced with MiSeq data using SOAPdenovo2 with GapCloser.

|  |  |  |  |
| --- | --- | --- | --- |
| Assembly | GAGE-B supp | GAGE-B | SOAPdenovo2 |
| # contigs (>= 200 bp) | **244** | **244** | 439 |
| # contigs (>= 1000 bp) | N/A | **147** | 237 |
| Total length (>= 200 bp) | N/A | **3 944 735** | 3 930 265 |
| Total length (>= 1000 bp) | N/A | **3 895 651** | 3 829 930 |
| # contigs | N/A | **186** | 327 |
| Total length | N/A | **3 924 635** | 3 892 100 |
| Largest contig | N/A | **246 179** | 98 351 |
| Reference length | 4 967 469 | 4 033 464 | 4 033 464 |
| GC (%) | N/A | 47.56 | **47.59** |
| Reference GC (%) | N/A | 47.49 | 47.49 |
| N50 | **71 400** | **71 357** | 29 551 |
| NG50 | N/A | **71 357** | 28 424 |
| N75 | N/A | **37 218** | 17 092 |
| NG75 | N/A | **34 893** | 16 323 |
| L50 | N/A | **15** | 39 |
| LG50 | N/A | **15** | 41 |
| L75 | N/A | **34** | 83 |
| LG75 | N/A | **36** | 90 |
| # misassemblies | 21 | 12 | **2** |
| Misassembled contigs length | N/A | 408 395 | **6 284** |
| # local misassemblies | 48 | 44 | **0** |
| # unaligned contigs | 4 + ? part | 2 + 3 part | **2 + 0 part** |
| Unaligned contigs length | N/A | 6 480 | **5 991** |
| Genome fraction (%) | **99.3** | 96.940 | 96.230 |
| Duplication ratio | 1.0 | 1.003 | **1.001** |
| # N's per 100 kbp | N/A | 0.00 | 0.00 |
| # mismatches per 100 kbp | N/A | 8.16 | **3.01** |
| # indels per 100 kbp | N/A | 4.37 | **2.60** |
| # genes | N/A | **3 442 + 85 part** | 3 336 + 122 part |
| Largest alignment | N/A | **246 179** | 98 351 |
| NA50 | 65 500 | **71 357** | 29 551 |
| NGA50 | N/A | **68 152** | 28 424 |
| NA75 | N/A | **37 218** | 17 092 |
| NGA75 | N/A | **32 753** | 16 323 |
| LA50 | N/A | **15** | 39 |
| LGA50 | N/A | **16** | 41 |
| LA75 | N/A | **34** | 83 |
| LGA75 | N/A | **37** | 90 |

Table S 14 Comparison of scaffolds from SOAPdenovo runs on MiSeq data

The values for the first column (GAGE-B supplementary table S7) are obtained from GAGE-Bs supplementary material, while the next column (GAGE-B assembly file) represents data computed by running assemblies on QUAST 2.2. The last column represents an assembly reproduced with MiSeq data using SOAPdenovo2 with GapCloser.

|  |  |  |  |
| --- | --- | --- | --- |
| Assembly | GAGE-B supp | GAGE-B | SOAPdenovo2 |
| # contigs (>= 500 bp) | **165** | **165** | 166 |
| # contigs (>= 1000 bp) | N/A | **136** | 137 |
| Total length (>= 500 bp) | N/A | 3 933 425 | **3 936 790** |
| Total length (>= 1000 bp) | N/A | 3 911 595 | **3 914 739** |
| # contigs | N/A | **165** | **166** |
| Total length | N/A | 3 933 425 | **3 936 790** |
| Largest contig | N/A | 246 179 | **246 328** |
| Reference length | 4 967 469 | 4 033 464 | 4 033 464 |
| GC (%) | N/A | **47.55** | 47.54 |
| Reference GC (%) | N/A | **47.49** | **47.49** |
| N50 | 91 900 | 91 942 | **92 055** |
| NG50 | N/A | 91 942 | **92 055** |
| N75 | N/A | 39 820 | **39 821** |
| NG75 | N/A | **38 903** | 38 797 |
| L50 | N/A | **14** | **14** |
| LG50 | N/A | **14** | **14** |
| L75 | N/A | **32** | **32** |
| LG75 | N/A | **33** | **33** |
| # misassemblies | 24 | 14 | **6** |
| Misassembled contigs length | N/A | 543 041 | **247 170** |
| # local misassemblies | 80 | **77** | 111 |
| # unaligned contigs | **1 + ? part** | 2 + 2 part | 2 + 0 part |
| Unaligned contigs length | N/A | 6 469 | **5 991** |
| Genome fraction (%) | **98.7** | 97.050 | 97.068 |
| Duplication ratio | 1.0 | **1.005** | **1.005** |
| # N's per 100 kbp | N/A | 92.34 | **19.74** |
| # mismatches per 100 kbp | N/A | 8.48 | **7.99** |
| # indels per 100 kbp | N/A | 5.52 | **4.09** |
| # genes | N/A | **3 443 + 86 part** | 3 432 + 100 part |
| Largest alignment | N/A | 246 179 | **246 327** |
| NA50 | 89 900 | 91 942 | **92 054** |
| NGA50 | N/A | 89 759 | **89 882** |
| NA75 | N/A | **39 451** | **39 451** |
| NGA75 | N/A | 38 469 | **38 550** |
| LA50 | N/A | **14** | **14** |
| LGA50 | N/A | **15** | **15** |
| LA75 | N/A | **32** | **32** |
| LGA75 | N/A | **34** | **34** |

Table S 15 Comparison of contigs from SPAdes runs on HiSeq data

The values for the first column (GAGE-B supplementary table S6) are obtained from GAGE-Bs supplementary material, while the next column (GAGE-B assembly file) represents data computed by running assemblies on QUAST 2.2. The last two columns represents assemblies reproduced with HiSeq data using SPAdes 2.30/2.5.0.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Assembly | GAGE-B supp | GAGE-B | Spades 2.3.0 | Spades 2.5.0 |
| # contigs (>= 200 bp) | 205 | 205 | 158 | **134** |
| # contigs (>= 1000 bp) | N/A | 121 | 83 | **69** |
| Total length (>= 200 bp) | N/A | 3 954 925 | **4 000 021** | 3 948 039 |
| Total length (>= 1000 bp) | N/A | 3 912 505 | **3 957 821** | 3 915 729 |
| # contigs | N/A | 162 | 126 | **98** |
| Total length | N/A | 3 939 988 | **3 987 612** | 3 935 605 |
| Largest contig | N/A | 214 711 | 369 300 | **536 022** |
| Reference length | 4 033 464 | 4 033 464 | 4 033 464 | 4 033 464 |
| GC (%) | N/A | **47.51** | **47.51** | **47.51** |
| Reference GC (%) | N/A | **47.49** | **47.49** | **47.49** |
| N50 | 77 100 | 83 518 | 137 662 | **225 872** |
| NG50 | N/A | 77 074 | 137 662 | **197 838** |
| N75 | N/A | 46 841 | 95 826 | **97 170** |
| NG75 | N/A | 44 636 | 95 826 | **95 888** |
| L50 | N/A | 16 | 9 | **6** |
| LG50 | N/A | 17 | 9 | **7** |
| L75 | N/A | 32 | 17 | **12** |
| LG75 | N/A | 34 | 17 | **13** |
| # misassemblies | 7 | **4** | **4** | 8 |
| Misassembled contigs length | N/A | 75 745 | **463 075** | 613 451 |
| # local misassemblies | 4 | **2** | 9 | 3 |
| # unaligned contigs | 8 + ? part | **1 + 1 part** | 3 + 0 part | 3 + 0 part |
| Unaligned contigs length | N/A | **1 261** | 2 685 | 2 685 |
| Genome fraction (%) | **99.6** | 97.439 | 98.611 | 97.468 |
| Duplication ratio | 1.0 | 1.007 | 1.007 | **1.006** |
| # N's per 100 kbp | N/A | **0.00** | 8.75 | **0.00** |
| # mismatches per 100 kbp | N/A | **5.19** | 6.64 | 6.77 |
| # indels per 100 kbp | N/A | **2.93** | 3.44 | 3.38 |
| # genes | N/A | 3 483 + 46 part | **3 571 + 30 part** | 3 519 + 21 part |
| Largest alignment | N/A | 214 707 | 369 299 | **536 022** |
| NA50 | 77 100 | 83 518 | 137 656 | **197 838** |
| NGA50 | N/A | 77 065 | 137 656 | **197 838** |
| NA75 | N/A | 46 841 | 95 507 | **95 888** |
| NGA75 | N/A | 44 636 | **95 507** | 92 090 |
| LA50 | N/A | 16 | 10 | **7** |
| LGA50 | N/A | 17 | 10 | **7** |
| LA75 | N/A | 32 | 18 | **14** |
| LGA75 | N/A | 34 | 18 | **15** |

Table S 16 Comparison of scaffolds from SPAdes runs on HiSeq data

The values for the first column (GAGE-B supplementary table S6) are obtained from GAGE-Bs supplementary material, while the next column (GAGE-B assembly file) represents data computed by running assemblies on QUAST 2.2. The last two columns represents assemblies reproduced with HiSeq data using SPAdes 2.30/2.5.0.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Assembly | GAGE-B supp | GAGE-B | Spades 2.3.0 | Spades 2.5.0 |
| # contigs (>= 500 bp) | 106 | 106 | 109 | **93** |
| # contigs (>= 1000 bp) | N/A | 95 | 72 | **64** |
| Total length (>= 500 bp) | N/A | 4 012 066 | **4 211 504** | 3 935 742 |
| Total length (>= 1000 bp) | N/A | 4 004 247 | **4 186 225** | 3 915 866 |
| # contigs | N/A | 106 | 109 | **93** |
| Total length | N/A | 4 012 066 | **4 211 504** | 3 935 742 |
| Largest contig | N/A | 246 251 | **572 859** | 536 022 |
| Reference length | 4 033 464 | 4 033 464 | 4 033 464 | 4 033 464 |
| GC (%) | N/A | 47.42 | 47.47 | **47.51** |
| Reference GC (%) | N/A | **47.49** | **47.49** | **47.49** |
| N50 | 98 300 | 98 274 | 225 904 | **343 954** |
| NG50 | N/A | 98 274 | 225 904 | **258 937** |
| N75 | N/A | 52 916 | **169 794** | 166 780 |
| NG75 | N/A | 52 916 | **171 917** | 97 170 |
| L50 | N/A | 13 | 6 | **5** |
| LG50 | N/A | 13 | **6** | **6** |
| L75 | N/A | 26 | 12 | **10** |
| LG75 | N/A | 26 | **11** | **11** |
| # misassemblies | 27 | 21 | **7** | 8 |
| Misassembled contigs length | N/A | **390 381** | 620 123 | 613 488 |
| # local misassemblies | 19 | 17 | 11 | **6** |
| # unaligned contigs | **1 + ? part** | **1 + 1 part** | 3 + 0 part | 3 + 0 part |
| Unaligned contigs length | N/A | **1 261** | 2 685 | 2 685 |
| Genome fraction (%) | **99.6** | 98.209 | 98.753 | 97.478 |
| Duplication ratio | 1.0 | 1.021 | 1.062 | **1.011** |
| # N's per 100 kbp | N/A | **0.00** | 9.00 | **0.00** |
| # mismatches per 100 kbp | N/A | 9.26 | 7.03 | **6.66** |
| # indels per 100 kbp | N/A | 8.00 | **3.72** | 3.87 |
| # genes | N/A | 3 544 + 46 part | **3 586 + 26 part** | 3 524 + 20 part |
| Largest alignment | N/A | 246 251 | **572 859** | 536 022 |
| NA50 | 94 800 | 95 858 | 214 751 | **246 236** |
| NGA50 | N/A | 94 762 | 214 751 | **246 236** |
| NA75 | N/A | 52 911 | **129 169** | 106 179 |
| NGA75 | N/A | 52 911 | **152 627** | 97 170 |
| LA50 | N/A | 13 | 7 | **6** |
| LGA50 | N/A | 14 | 7 | **6** |
| LA75 | N/A | 27 | 13 | **11** |
| LGA75 | N/A | 27 | **12** | **12** |

Table S 17 Comparison of contigs from SPAdes runs on MiSeq data

The values for the first column (GAGE-B supplementary table S7) are obtained from GAGE-Bs supplementary material, while the next column (GAGE-B assembly file) represents data computed by running assemblies on QUAST 2.2. The last two columns represents assemblies reproduced with MiSeq data using SPAdes 2.30/2.5.0.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Assembly | GAGE-B supp | GAGE-B | Spades 2.3.0 | Spades 2.5.0 |
| # contigs (>= 200 bp) | 1 475 | 1 475 | 1 480 | **786** |
| # contigs (>= 1000 bp) | N/A | 73 | 76 | **71** |
| Total length (>= 200 bp) | N/A | 4 449 438 | **4 450 408** | 4 209 962 |
| Total length (>= 1000 bp) | N/A | 3 990 044 | **3 991 331** | 3 908 698 |
| # contigs | N/A | **146** | 148 | 186 |
| Total length | N/A | 4 035 776 | **4 036 787** | 3 983 820 |
| Largest contig | N/A | **741 022** | **741 022** | 392 814 |
| Reference length | 4 967 469 | 4 033 464 | 4 033 464 | 4 033 464 |
| GC (%) | N/A | **47.37** | **47.37** | 47.34 |
| Reference GC (%) | N/A | 47.49 | 47.49 | 47.49 |
| N50 | **262 200** | **262 160** | **262 160** | 246 502 |
| NG50 | N/A | **262 160** | **262 160** | 215 247 |
| N75 | N/A | **136 216** | **136 216** | 105 135 |
| NG75 | N/A | **136 216** | **136 216** | 91 956 |
| L50 | N/A | **5** | **5** | 6 |
| LG50 | N/A | **5** | **5** | 7 |
| L75 | N/A | **11** | **11** | 12 |
| LG75 | N/A | **11** | **11** | 13 |
| # misassemblies | 7 | 5 | 5 | **4** |
| Misassembled contigs length | N/A | **386 647** | **385 887** | 533 480 |
| # local misassemblies | 6 | 4 | 4 | **1** |
| # unaligned contigs | 1 336 | **57 + 3 part** | **57 + 3 part** | 81 + 3 part |
| Unaligned contigs length | N/A | **41 368** | **41 368** | 56 002 |
| Genome fraction (%) | **99.6** | 98.643 | 98.752 | 97.350 |
| Duplication ratio | 1.0 | 1.004 | 1.004 | **1.001** |
| # N's per 100 kbp | N/A | 5.18 | 5.85 | **0.00** |
| # mismatches per 100 kbp | N/A | **4.88** | 4.90 | 6.88 |
| # indels per 100 kbp | N/A | 3.24 | **3.11** | 3.36 |
| # genes | N/A | **3 598 + 28 part** | 3 597 + 29 part | 3 505 + 25 part |
| Largest alignment | N/A | **741 022** | **741 022** | 392 814 |
| NA50 | 246 600 | **262 160** | **262 160** | 198 500 |
| NGA50 | N/A | **262 160** | **262 160** | 198 500 |
| NA75 | N/A | **106 531** | **106 531** | 91 956 |
| NGA75 | N/A | **106 531** | **106 531** | 91 956 |
| LA50 | N/A | **5** | **5** | 7 |
| LGA50 | N/A | **5** | **5** | 7 |
| LA75 | N/A | **12** | **12** | 14 |
| LGA75 | N/A | **12** | **12** | 14 |

Table S 18 Comparison of scaffolds from SPAdes runs on MiSeq data

The values for the first column (GAGE-B supplementary table S7) are obtained from GAGE-Bs supplementary material, while the next column (GAGE-B assembly file) represents data computed by running assemblies on QUAST 2.2. The last two columns represents assemblies reproduced with MiSeq data using SPAdes 2.30/2.5.0.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Assembly | GAGE-B supp | GAGE-B | Spades 2.3.0 | Spades 2.5.0 |
| # contigs (>= 500 bp) | **145** | **145** | 147 | 184 |
| # contigs (>= 1000 bp) | N/A | 72 | 75 | **69** |
| Total length (>= 500 bp) | N/A | 4 041 279 | **4 042 074** | 3 983 773 |
| Total length (>= 1000 bp) | N/A | 3 995 331 | **3 996 618** | 3 908 651 |
| # contigs | N/A | **145** | 147 | 184 |
| Total length | N/A | 4 041 279 | **4 042 074** | 3 983 773 |
| Largest contig | N/A | **741 022** | **741 022** | 392 814 |
| Reference length | 4 967 469 | 4 033 464 | 4 033 464 | 4 033 464 |
| GC (%) | N/A | **47.37** | **47.37** | 47.34 |
| Reference GC (%) | N/A | **47.49** | **47.49** | **47.49** |
| N50 | **262 200** | **262 160** | **262 160** | 258 677 |
| NG50 | N/A | **262 160** | **262 160** | 258 677 |
| N75 | N/A | **136 216** | **136 216** | 105 135 |
| NG75 | N/A | **136 216** | **136 216** | 105 135 |
| L50 | N/A | **5** | **5** | 6 |
| LG50 | N/A | **5** | **5** | 6 |
| L75 | N/A | **11** | **11** | 12 |
| LG75 | N/A | **11** | **11** | 12 |
| # misassemblies | 7 | **5** | **5** | 4 |
| Misassembled contigs length | N/A | **386 647** | **385 887** | 533 480 |
| # local misassemblies | 6 | 4 | 4 | **3** |
| # unaligned contigs | 57 + ? part | **57 + 3 part** | **57 + 3 part** | 81 + 3 part |
| Unaligned contigs length | N/A | **46 754** | **46 754** | 56 002 |
| Genome fraction (%) | **99.6** | 98.648 | 98.752 | 97.356 |
| Duplication ratio | 1.0 | 1.004 | 1.004 | **1.001** |
| # N's per 100 kbp | N/A | 5.30 | 5.84 | **0.00** |
| # mismatches per 100 kbp | N/A | 4.90 | 4.90 | 7.13 |
| # indels per 100 kbp | N/A | 3.24 | **3.11** | 3.46 |
| # genes | N/A | **3 599 + 27 part** | 3 598 + 28 part | 3 508 + 25 part |
| Largest alignment | N/A | **741 022** | **741 022** | 392 814 |
| NA50 | 246 600 | **262 160** | **262 160** | 215 247 |
| NGA50 | N/A | **262 160** | **262 160** | 215 247 |
| NA75 | N/A | **106 531** | **106 531** | 105 081 |
| NGA75 | N/A | **106 531** | **106 531** | 91 956 |
| LA50 | N/A | **5** | **5** | 7 |
| LGA50 | N/A | **5** | **5** | 7 |
| LA75 | N/A | **12** | **12** | 13 |
| LGA75 | N/A | **12** | **12** | 14 |

Table S 19 Comparison of contigs from Velvet runs on HiSeq data

The values for the first column (GAGE-B supplementary table S6) are obtained from GAGE-Bs supplementary material, while the next column (GAGE-B assembly file) represents data computed by running assemblies on QUAST 2.2. The last two columns represents assemblies reproduced with HiSeq data using Velvet 1.2.08/1.2.10.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Assembly | GAGE-B supp | GAGE-B | Velvet 1.2.08 | Velvet 1.2.10 |
| # contigs (>= 200 bp) | 261 | 261 | **246** | **246** |
| # contigs (>= 1000 bp) | N/A | 183 | **177** | **177** |
| Total length (>= 200 bp) | N/A | **3 931 494** | 3 903 763 | 3 903 763 |
| Total length (>= 1000 bp) | N/A | **3 894 214** | 3 870 711 | 3 870 711 |
| # contigs | N/A | 213 | **205** | **205** |
| Total length | N/A | **3 914 915** | 3 889 783 | 3 889 783 |
| Largest contig | N/A | 157 783 | **157 831** | **157 831** |
| Reference length | 4 033 464 | 4 033 464 | 4 033 464 | 4 033 464 |
| GC (%) | N/A | 47.50 | **47.64** | **47.64** |
| Reference GC (%) | N/A | **47.49** | **47.49** | **47.49** |
| N50 | 40 100 | 40 877 | **46 346** | **46 346** |
| NG50 | N/A | 40 085 | **41 373** | **41 373** |
| N75 | N/A | 22 754 | **22 927** | **22 927** |
| NG75 | N/A | 20 673 | **20 721** | **20 721** |
| L50 | N/A | 30 | **28** | **28** |
| LG50 | N/A | 31 | **30** | **30** |
| L75 | N/A | 62 | **59** | **59** |
| LG75 | N/A | 66 | **63** | **63** |
| # misassemblies | 9 | **4** | 9 | 9 |
| Misassembled contigs length | N/A | **55 189** | 127 292 | 127 292 |
| # local misassemblies | 9 | **8** | 9 | 9 |
| # unaligned contigs | 1 + ? part | 1 + 1 part | **1 + 0 part** | **1 + 0 part** |
| Unaligned contigs length | N/A | 576 | **549** | **549** |
| Genome fraction (%) | **99.4** | 97.038 | 96.340 | 96.340 |
| Duplication ratio | **1.0** | **1.011** | 1.012 | 1.012 |
| # N's per 100 kbp | N/A | **0.00** | **0.00** | **0.00** |
| # mismatches per 100 kbp | N/A | **4.47** | 5.89 | 5.89 |
| # indels per 100 kbp | N/A | 4.85 | **3.14** | **3.14** |
| # genes | N/A | 3 386 + 132 part | **3 392 + 121 part** | **3 392 + 121 part** |
| Largest alignment | N/A | 157 783 | **157 831** | **157 831** |
| NA50 | 39 500 | 40 877 | **42 805** | **42 805** |
| NGA50 | N/A | 39 468 | **41 373** | **41 373** |
| NA75 | N/A | **22 753** | 22 638 | 22 638 |
| NGA75 | N/A | **20 673** | 19 470 | 19 470 |
| LA50 | N/A | 30 | **28** | **28** |
| LGA50 | N/A | 31 | **30** | **30** |
| LA75 | N/A | 62 | **59** | **59** |
| LGA75 | N/A | 66 | **64** | **64** |

Table S 20 Comparison of scaffolds from Velvet runs on HiSeq data

The values for the first column (GAGE-B supplementary table S6) are obtained from GAGE-Bs supplementary material, while the next column (GAGE-B assembly file) represents data computed by running assemblies on QUAST 2.2. The last two columns represents assemblies reproduced with HiSeq data using Velvet 1.2.08/1.2.10.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Assembly | GAGE-B supp | GAGE-B | Velvet 1.2.08 | Velvet 1.2.10 |
| # contigs (>= 500 bp) | **85** | **85** | 104 | 104 |
| # contigs (>= 1000 bp) | N/A | **65** | 79 | 79 |
| Total length (>= 500 bp) | N/A | **3 940 257** | 3 903 678 | 3 903 678 |
| Total length (>= 1000 bp) | N/A | **3 927 326** | 3 886 781 | 3 886 781 |
| # contigs | N/A | **85** | 104 | 104 |
| Total length | N/A | **3 940 257** | 3 903 678 | 3 903 678 |
| Largest contig | N/A | **524 347** | 411 927 | 411 927 |
| Reference length | 4 033 464 | 4 033 464 | 4 033 464 | 4 033 464 |
| GC (%) | N/A | 47.50 | **47.63** | **47.63** |
| Reference GC (%) | N/A | **47.49** | **47.49** | **47.49** |
| N50 | **172 500** | **172 545** | 163 386 | 163 386 |
| NG50 | N/A | **172 545** | 157 209 | 157 209 |
| N75 | N/A | **92 099** | 75 755 | 75 755 |
| NG75 | N/A | **82 167** | 71 623 | 71 623 |
| L50 | N/A | **7** | 8 | 8 |
| LG50 | N/A | **7** | 9 | 9 |
| L75 | N/A | **14** | 18 | 18 |
| LG75 | N/A | **15** | 19 | 19 |
| # misassemblies | 13 | **10** | **10** | **10** |
| Misassembled contigs length | N/A | 395 034 | **273 985** | **273 985** |
| # local misassemblies | 132 | 129 | **123** | **123** |
| # unaligned contigs | 1 + ? part | **1 + 0 part** | **1 + 0 part** | **1 + 0 part** |
| Unaligned contigs length | N/A | **549** | **549** | **549** |
| Genome fraction (%) | **98.9** | 97.140 | 96.251 | 96.251 |
| Duplication ratio | 1.0 | **1.016** | **1.016** | **1.016** |
| # N's per 100 kbp | N/A | 547.10 | **515.80** | **515.80** |
| # mismatches per 100 kbp | N/A | **5.18** | 5.80 | 5.80 |
| # indels per 100 kbp | N/A | 9.44 | **5.28** | **5.28** |
| # genes | N/A | **3 400 + 128 part** | 3 379 + 130 part | 3 379 + 130 part |
| Largest alignment | N/A | **521 619** | 409 820 | 409 820 |
| NA50 | 171 500 | **171 505** | 163 055 | 163 055 |
| NGA50 | N/A | **171 505** | 156 829 | 156 829 |
| NA75 | N/A | **91 938** | 71 376 | 71 376 |
| NGA75 | N/A | **80 883** | 59 664 | 59 664 |
| LA50 | N/A | **7** | 8 | 8 |
| LGA50 | N/A | **7** | 9 | 9 |
| LA75 | N/A | **14** | 18 | 18 |
| LGA75 | N/A | **15** | 20 | 20 |

Table S 21 Comparison of contigs from Velvet runs on MiSeq data

The values for the first column (GAGE-B supplementary table S7) are obtained from GAGE-Bs supplementary material, while the next column (GAGE-B assembly file) represents data computed by running assemblies on QUAST 2.2. The last two columns represents assemblies reproduced with MiSeq data using Velvet 1.2.08/1.2.10.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Assembly | GAGE-B supp | GAGE-B | Velvet\_1.2.08 | Velvet\_1.2.10 |
| # contigs (>= 200 bp) | 201 | 201 | **179** | **179** |
| # contigs (>= 1000 bp) | N/A | 124 | **108** | **108** |
| Total length (>= 200 bp) | N/A | **3 958 933** | 3 905 232 | 3 905 232 |
| Total length (>= 1000 bp) | N/A | **3 921 645** | 3 872 950 | 3 872 950 |
| # contigs | N/A | 156 | **135** | **135** |
| Total length | N/A | **3 944 260** | 3 891 789 | 3 891 789 |
| Largest contig | N/A | **246 346** | **246 346** | **246 346** |
| Reference length | 4 967 469 | 4 033 464 | 4 033 464 | 4 033 464 |
| GC (%) | N/A | 47.52 | **47.69** | **47.69** |
| Reference GC (%) | N/A | **47.49** | **47.49** | **47.49** |
| N50 | 92 000 | 92 036 | **105 176** | **105 176** |
| NG50 | N/A | 92 036 | **105 176** | **105 176** |
| N75 | N/A | **43 442** | **43 442** | **43 442** |
| NG75 | N/A | **39 974** | 39 408 | 39 408 |
| L50 | N/A | 15 | **12** | **12** |
| LG50 | N/A | 15 | **12** | **12** |
| L75 | N/A | 31 | **26** | **26** |
| LG75 | N/A | 33 | **28** | **28** |
| # misassemblies | 12 | 14 | **5** | **5** |
| Misassembled contigs length | N/A | 749 414 | **28 731** | **28 731** |
| # local misassemblies | 7 | **2** | 3 | 3 |
| # unaligned contigs | 1 + ? part | 1 + 5 part | **1 + 2 part** | **1 + 2 part** |
| Unaligned contigs length | N/A | 6 703 | **6 297** | **6 297** |
| Genome fraction (%) | **99.5** | 97.563 | 96.234 | 96.234 |
| Duplication ratio | 1.0 | **1.007** | **1.007** | **1.007** |
| # N's per 100 kbp | N/A | **0.00** | **0.00** | **0.00** |
| # mismatches per 100 kbp | N/A | **4.09** | 4.23 | 4.23 |
| # indels per 100 kbp | N/A | **2.80** | 2.63 | 2.63 |
| # genes | N/A | **3 491 + 79 part** | 3 460 + 70 part | 3 460 + 70 part |
| Largest alignment | N/A | **246 346** | **246 346** | **246 346** |
| NA50 | 67 100 | 67 096 | **105 176** | **105 176** |
| NGA50 | N/A | 63 574 | **105 176** | **105 176** |
| NA75 | N/A | 40 333 | **43 442** | **43 442** |
| NGA75 | N/A | 39 408 | **39 408** | **39 408** |
| LA50 | N/A | 16 | **12** | **12** |
| LGA50 | N/A | 17 | **12** | **12** |
| LA75 | N/A | 35 | **26** | **26** |
| LGA75 | N/A | 37 | **28** | **28** |

Table S 22 Comparison of scaffolds from Velvet runs on MiSeq data

The values for the first column (GAGE-B supplementary table S7) are obtained from GAGE-Bs supplementary material, while the next column (GAGE-B assembly file) represents data computed by running assemblies on QUAST 2.2. The last two columns represents assemblies reproduced with MiSeq data using Velvet 1.2.08/1.2.10.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Assembly | GAGE-B supp | GAGE-B | Velvet 1.2.08 | Velvet 1.2.10 |
| # contigs (>= 500 bp) | 138 | 138 | **133** | **133** |
| # contigs (>= 1000 bp) | N/A | 113 | **108** | **108** |
| Total length (>= 500 bp) | N/A | **3 945 576** | 3 892 321 | 3 892 321 |
| Total length (>= 1000 bp) | N/A | **3 928 680** | 3 875 057 | 3 875 057 |
| # contigs | N/A | 138 | **133** | **133** |
| Total length | N/A | **3 945 576** | 3 892 321 | 3 892 321 |
| Largest contig | N/A | **246 346** | **246 346** | **246 346** |
| Reference length | 4 967 469 | 4 033 464 | 4 033 464 | 4 033 464 |
| GC (%) | N/A | 47.51 | **47.69** | **47.69** |
| Reference GC (%) | N/A | **47.49** | **47.49** | **47.49** |
| N50 | **110 000** | **109 996** | 105 176 | 105 176 |
| NG50 | N/A | **109 996** | 105 176 | 105 176 |
| N75 | N/A | **47 047** | 43 442 | 43 442 |
| NG75 | N/A | **46 531** | 39 408 | 39 408 |
| L50 | N/A | **11** | 12 | 12 |
| LG50 | N/A | **11** | 12 | 12 |
| L75 | N/A | **25** | 26 | 26 |
| LG75 | N/A | **26** | 28 | 28 |
| # misassemblies | 17 | 22 | **6** | **6** |
| Misassembled contigs length | N/A | 1 210 318 | **45 660** | **45 660** |
| # local misassemblies | 23 | 13 | **6** | **6** |
| # unaligned contigs | 1 + ? part | 1 + 3 part | **1 + 2 part** | **1 + 2 part** |
| Unaligned contigs length | N/A | 6 324 | **6 297** | **6 297** |
| Genome fraction (%) | **99.2** | 97.598 | 96.242 | 96.242 |
| Duplication ratio | 1.0 | **1.007** | **1.007** | **1.007** |
| # N's per 100 kbp | N/A | 5.58 | **1.28** | **1.28** |
| # mismatches per 100 kbp | N/A | **3.96** | 4.33 | 4.33 |
| # indels per 100 kbp | N/A | 4.88 | **2.65** | **2.65** |
| # genes | N/A | **3 492 + 78 part** | 3 459 + 71 part | 3 459 + 71 part |
| Largest alignment | N/A | **246 346** | **246 346** | **246 346** |
| NA50 | 92 000 | 75 901 | **105 176** | **105 176** |
| NGA50 | N/A | 67 096 | **105 176** | **105 176** |
| NA75 | N/A | 43 122 | **43 442** | **43 442** |
| NGA75 | N/A | **40 333** | 39 408 | 39 408 |
| LA50 | N/A | 13 | **12** | **12** |
| LGA50 | N/A | 14 | **12** | **12** |
| LA75 | N/A | 32 | **26** | **26** |
| LGA75 | N/A | 33 | **28** | **28** |

1. <http://bioinformatics.oxfordjournals.org/content/suppl/2013/05/10/btt273.DC1/GAGE-B_SupplementaryMaterial_Apr4.docx> [↑](#footnote-ref-1)