Error Correction of Illumina Sequencing Data

Foutcorrectie van sequencingdata van Illumina

Mahdi Heydari

Promotoren: prof. dr. ir. J. Fostier, prof. dr. Y. Van de Peer
Proefschrift ingediend tot het behalen van de graad van
Doctor in de ingenieurswetenschappen: computerwetenschappen

UNIVERSITEIT
GENT

Vakgroep Informatietechnologie
Voorzitter: prof. dr. ir. B. Dhoedt
Faculteit Ingenieurswetenschappen en Architectuur
Academiejaar 2018 - 2019

ISBN 978-94-6355-241-7
NUR 980, 922
Wettelijk depot: D/2019/10.500/49

Ghent University
Faculty of Engineering and Architecture
Department of Information Technology

Examination Board:

prof. dr. Filip De Turck (chair)
prof. dr. Jan Fostier (supervisor)
prof. dr. Yves Van de Peer (supervisor)
dr. Eric Rivals
prof. dr. Jo Vandesompele
prof. dr. Tim De Meyer
prof. dr. Veerle Fack
mr. Stephane Rombauts

Dissertation for acquiring the grade of
Doctor of Computer Science Engineering
Academic Year 2018-2019

Acknowledgments

"Tell me and I forget. Teach me and I remember. Involve me and I learn."

– Benjamin Franklin

Living life is like writing a book, and every action we take in our daily life is like writing a paragraph, a sentence, or even a word. At the end of the day, we already finished a page which cannot be rewritten again. When I started my Ph.D., I knew it would be a particular chapter of my life, and now I am writing the last pages of this chapter. So, it is time to step back and acknowledge the support of all those who stayed with me in this journey.

First, I would like to express my sincere gratitude to my supervisors: Prof. Yves Van de Peer and Prof. Jan Fostier to whom I am deeply grateful for giving me such a great opportunity. Definitely, this journey would not have been possible if it was not for their continuous support, critical reviews, constructive feedback, immense knowledge, and constant motivation.

Besides my advisors, I would like to thank Prof. Filip De Turck for accepting my request to chair my Ph.D. examination board, and the rest of my thesis committee: Dr Eric Rivals, Prof. Jo Vandesompele, Prof. Tim De Meyer, Prof. Veerle Fack, and Mr. Stephane Rombauts, for their insightful comments and encouragement, but also for the hard questions which helped me to widen my research from various perspectives.

I thank my fellow colleagues for their strong support and generous help with a special mention to Aranka, Arun, Camilo, David, Dries, Giles, Joeri, Louise, Lieven, Maarten, Mushthofa, Patricio, Razgar, Simon, Tom, and Yan. I would like to extend my sincere thanks to Aranka for the effort she made to translate the summary of this dissertation from English to Dutch and Giles for his constructive comments about the introduction chapter and his collaboration for publishing the presented research papers in this thesis.

I am thankful to the technical and administrative staff of IDLab, international admissions office, department of personnel and organization, and the Deans office of the faculty of engineering and architecture for helping me and carrying out all administrative formalities very smoothly from the very early days of my stay in Ghent till today. Thank you Mike Van Puyenbroeck, Sara Ysebaert, Martine

Buysse, Davinia Stevens, Karen Van Landeghem, Bernadette Becue, and Muriel Vervaeke.

Some special words of gratitude go to my former supervisor Mehdi Sadeghi and my friends who have been a role model in my life and always a significant source of support especially when things would get a bit discouraging: Saeed Omidi, Saeed Amiri, and Ali Golshani. Thanks, guys for always being there for me.

Luckily for me, all the challenges and stressful moments of my Ph.D. have been reduced thanks to great friends whom I met in Ghent. Particularly, I would like to name Mostafa, Maryam, Reza, Sahel, Alireza LP, Hemen, Fatemeh, and Nasrin for their generous help and support. Last, my board game buddies: Alireza, Amir, Babak, Dave, Ehsan, Foad, Javad, Hadi, and Moha.

There are many others not named here who have rendered their help for the accomplishment of this work, to them I express my gratefulness.

Last but not least, I would like to thank my family: my parents and my sisters for their endless love, support and encouragement during my Ph.D. and my life in general.

" The greatest enemy of knowledge is not ignorance, it is the illusion of knowledge.

– Stephen Hawking

Gent, June 2019
Mahdi Heydari

Table of Contents

Acknowledgments **i**

Samenvatting **xxiii**

Summary **xxvii**

1 Introduction **1-1**
- 1.1 DNA sequencing . . . 1-2
- 1.2 Illumina sequencing data . . . 1-5
- 1.3 Fastq file format . . . 1-8
- 1.4 Illumina sequencing errors and biases . . . 1-9
- 1.5 *De novo* assembly . . . 1-11
- 1.6 Error correction . . . 1-17
- 1.7 Assessment of the quality of genome assembly . . . 1-21
- 1.8 Sequence alignment . . . 1-23
  - 1.8.1 Read alignment . . . 1-25
  - 1.8.2 Graph alignment . . . 1-27
- 1.9 Research goals and outline . . . 1-28
- 1.10 Publications . . . 1-30
  - 1.10.1 Journal papers . . . 1-30
  - 1.10.2 Conference paper& abstracts & posters . . . 1-30

References . . . 1-32

2 Evaluation of the Impact of Illumina Error Correction Tools on de novo Genome Assembly **2-1**
- 2.1 Background . . . 2-2
- 2.2 Material and Methods . . . 2-5
  - 2.2.1 Error correction tools . . . 2-5
  - 2.2.2 Data . . . 2-6
  - 2.2.3 Error metrics . . . 2-6
  - 2.2.4 Evaluation of assembly results . . . 2-7
- 2.3 Results and Discussion . . . 2-7
  - 2.3.1 Ability of EC tools to correct sequencing errors . . . 2-7
  - 2.3.2 Ability of EC tools to improve genome assembly . . . 2-11
  - 2.3.3 Error rate versus assembly quality . . . 2-13

2.3.4 Time and space requirements . . . 2-19
2.4 Conclusions . . . 2-20
References . . . 2-26

3 Illumina error correction near highly repetitive DNA regions improves de novo genome assembly. 3-1
3.1 Introduction . . . 3-2
3.2 Methods . . . 3-5
3.2.1 Error correction tools . . . 3-5
3.2.2 Evaluation tools . . . 3-5
3.2.3 Data . . . 3-5
3.2.4 Targeted error correction . . . 3-6
3.3 Results . . . 3-12
3.3.1 Ability of EC tools to improve genome assembly . . . 3-12
3.3.2 Time and space requirements . . . 3-14
3.4 Discussion . . . 3-14
3.5 Conclusions . . . 3-16
References . . . 3-18

4 BrownieAligner: Accurate Alignment of Illumina Sequencing Data to de Bruijn Graphs 4-1
4.1 Background . . . 4-2
4.2 Methods . . . 4-4
4.2.1 Read alignment algorithm . . . 4-4
4.2.2 Implicit repeat resolution using a Markov model . . . 4-6
4.2.3 Choice of parameters: . . . 4-9
4.2.4 Graph aligner tools . . . 4-10
4.2.5 Data . . . 4-10
4.2.6 Evaluation metrics . . . 4-11
4.3 Results and discussion . . . 4-11
4.3.1 Alignment ratio . . . 4-11
4.3.2 Time and space requirements . . . 4-13
4.4 Conclusions . . . 4-15
References . . . 4-18

5 Conclusion and further directions 5-1
5.1 Discussion . . . 5-1
5.2 Future work . . . 5-2
5.2.1 BrownieAligner . . . 5-3
5.2.2 BrownieCorrector . . . 5-4
5.3 Current limitations and future perspective . . . 5-5
References . . . 5-8
. . . 9

A Supplementary Data: Evaluation of the Impact of Illumina Error Correction Tools on de novo Genome Assembly **A-1**
A.1 Error Correction Tool Parameter Settings . . . A-1
A.1.1 ACE . . . A-2
A.1.2 BayesHammer v. 3.7.1 . . . A-2
A.1.3 BFC v. r181 . . . A-2
A.1.4 BLESS 2 v. 1.02 . . . A-3
A.1.5 Blue v. 1.1.2 . . . A-3
A.1.6 Fiona v. 0.2.5 . . . A-4
A.1.7 Karect v. 1.0 . . . A-4
A.1.8 Lighter v. 1.1.0 . . . A-4
A.1.9 Musket v. 1.1 . . . A-5
A.1.10 RACER v. 1.0.1 . . . A-5
A.1.11 SGA-EC v. 0.10.14 . . . A-5
A.1.12 Trowel v. 0.2.0.4 . . . A-6
A.2 Data simulation . . . A-7
A.3 Error Metrics . . . A-7
A.3.1 Alignment ratio . . . A-7
A.3.2 EC gain . . . A-8
A.3.2.1 Accuracy comparison method . . . A-8
A.3.2.2 Real Data . . . A-9
A.3.2.3 Simulated Data . . . A-10
A.4 Assembly Result for Real Data . . . A-12
A.4.1 DISCOVAR . . . A-13
A.4.1.1 *B. dentium* . . . A-13
A.4.1.2 *E. coli str. K-12 substr. DH10B* . . . A-13
A.4.1.3 *E. coli str. K-12 substr. MG1655* . . . A-13
A.4.1.4 *S. enterica* . . . A-13
A.4.1.5 *P. aeruginosa* . . . A-13
A.4.1.6 *H. sapiens* Chr. 21 . . . A-13
A.4.1.7 *C. elegans* . . . A-13
A.4.1.8 *D. melanogaster* . . . A-13
A.4.2 IDBA . . . A-22
A.4.2.1 *B. dentium* . . . A-22
A.4.2.2 *E. coli str. K-12 substr. DH10B* . . . A-22
A.4.2.3 *E. coli str. K-12 substr. MG1655* . . . A-22
A.4.2.4 *S. enterica* . . . A-22
A.4.2.5 *P. aeruginosa* . . . A-22
A.4.2.6 *H. sapiens* Chr. 21 . . . A-22
A.4.2.7 *C. elegans* . . . A-22
A.4.2.8 *D. melanogaster* . . . A-22
A.4.3 SPAdes . . . A-31
A.4.3.1 *B. dentium* . . . A-31
A.4.3.2 *E. coli str. K-12 substr. DH10B* . . . A-31
A.4.3.3 *E. coli str. K-12 substr. MG1655* . . . A-31

A.4.3.4 *S. enterica* A-31
A.4.3.5 *P. aeruginosa* A-31
A.4.3.6 *H. sapiens* Chr. 21 . . . . . A-31
A.4.3.7 *C. elegans* A-31
A.4.3.8 *D. melanogaster* A-31
A.4.4 Velvet A-41
A.4.4.1 *B. dentium* A-41
A.4.4.2 *E. coli str. K-12 substr. DH10B* A-41
A.4.4.3 *E. coli str. K-12 substr. MG1655* A-41
A.4.4.4 *S. enterica* A-41
A.4.4.5 *P. aeruginosa* A-41
A.4.4.6 *H. sapiens* Chr. 21 . . . . . A-41
A.4.4.7 *C. elegans* A-41
A.4.4.8 *D. melanogaster* A-41
A.5 Memory and Runtime A-50
A.5.1 Real Data A-50
A.5.1.1 Memory A-50
A.5.1.2 Runtime A-50
A.5.2 Simulated Data A-51
A.5.2.1 Memory A-51
A.5.2.2 Runtime A-51

B Supplementary Data: Illumina error correction near highly repetitive DNA regions improves de novo genome assembly B-1
B.1 Parameter settings B-1
B.1.1 ACE B-1
B.1.2 BFC B-2
B.1.3 BLESS2 . . . . . B-2
B.1.4 Browniecorrector B-2
B.1.5 Karect B-2
B.1.6 RECKONER B-2
B.1.7 SPAdes B-2
B.1.8 Quast B-3
B.2 Data preparation B-3
B.2.1 Illumina real data B-3
B.2.2 Pacbio real data B-4
B.3 *k*-mer selection B-5
B.4 *k*-mer coverage B-8
B.5 Results B-8
B.5.1 Average improvement ratio of NGA50 . . . . . B-8
B.5.2 Choice of highly repetitive *k*-mer B-8
B.5.3 Choice of the number of iterations B-9
B.5.4 Full Quast report (contigs) B-12
B.5.4.1 D1 . . . . . B-12
B.5.4.2 D2 . . . . . B-12

B.5.4.3 D3 . . . B-12
B.5.4.4 D4 . . . B-12
B.5.4.5 D5 . . . B-12
B.5.4.6 D6 . . . B-12
B.5.4.7 D7 . . . B-12
B.5.4.8 D8 . . . B-13
B.5.4.9 D9 . . . B-13
B.5.5 Full Quast report (scaffolds) . . . B-23
B.5.5.1 D1 . . . B-23
B.5.5.2 D2 . . . B-23
B.5.5.3 D3 . . . B-23
B.5.5.4 D4 . . . B-23
B.5.5.5 D5 . . . B-23
B.5.5.6 D6 . . . B-23
B.5.5.7 D7 . . . B-23
B.5.5.8 D8 . . . B-24
B.5.5.9 D9 . . . B-24
B.5.6 Runtime and memory usage . . . B-43
References . . . B-44

C Supplementary Data: BrownieAligner: Accurate Alignment of Illumina Sequencing Data to de Bruijn Graphs C-1
C.1 Parameter Settings . . . C-1
C.1.1 BGREAT . . . C-1
C.1.2 BrownieAligner . . . C-2
C.1.3 deBGA . . . C-2
C.2 Simulated data preparation . . . C-3
C.3 Real data preparation . . . C-3
C.4 Evaluation Metric . . . C-3
C.4.1 Alignment ratio . . . C-3
C.5 Results . . . C-5
C.5.1 Simulated Data . . . C-5
C.5.1.1 BGREAT . . . C-6
C.5.1.2 BrownieAligner . . . C-6
C.5.1.3 deBGA . . . C-6
C.5.1.4 Choice of parameters . . . C-7
C.5.2 Real Data . . . C-8
C.5.3 Time and space requirements . . . C-9
C.5.3.1 Simulated data . . . C-9
C.5.3.2 Real data . . . C-10

List of Figures

1.1 A double-stranded DNA structure diagram. 1-3

1.2 This figure compares the throughput of three popular Illumina machines (MiSeq, HiSeq and NovaSeq) 1-5

1.3 This figure shows the Illumina sequencing steps: A. Library preparation: the DNA molecule is initially fragmented into smaller fragments. Then, in the ligation step, adapters are appended to both ends of each fragment. B. Cluster amplification step: these fragments are loaded into a flow cell where they attach to the surface of this cell. Then, a reverse strand is created through the polymerase process, which forms a bridge on the surface. Multiple bridges are shaped around the initial segments forming clonal clusters. C. Sequencing step: every cluster is now sequenced in parallel. In each round, the fluorescently labeled nucleotides are added, and the actual bases are determined based on the emitted color. D. Scientists use the collected sequencing data in different contexts and applications. For example, based on the overlaps between reads a *de novo* assembler attempts to construct the original sequence. . . . 1-7

1.4 This figure shows eight lines of an interleaved Fastq file consisting of four reads (two pairs). 1-8

1.5 This figure shows the nonuniform distribution of coverage along the reference genome. Even though the average coverage is 30, some regions are not covered enough. 1-10

1.6 Each read consists of three k-mers. Except for the middle k-mer in two reads which is identical, the other k-mers are unique and hence are shown in different colors. Each path in the de Bruijn graph that consists of two nodes represents an overlap between two k-mers in the reads and vice versa (e.g., AB, BC, DB and BE). However, the longer paths that span three or more nodes do not necessarily represent a valid overlap of k-mers in the reads (i.e., there is a connection between D to C because D is connected to B and B is connected to C). Therefore, all the overlaps in the data are present through some paths in the graph, but not all the paths in the graph show valid sequencing data. 1-13

1.7 This figure shows a single-stranded, compact representation of the de Bruijn graph constructed from a set of reads. Sequencing errors in the reads create spurious artifacts in the graph which are categorized as Tips, Bubbles, and Chimeric connections. 1-14

1.8 This figure shows the frequency of 31-mers in a dataset for three different reads. For each 31-mer, the frequency shows the number of reads in the dataset that contain 31-mer. The plot shows a sudden drop of the frequency of 31-mers at the end of the green line and in the middle of the orange line, which implies the presence of an error at the end and in the middle of corresponding reads. The blue line shows the 31-mer frequencies for an error-free read. . . . 1-18

1.9 The histogram shows a mixture of two distributions–erroneous 31-mers on the left and true 31-mers on the right side. This figure also shows that although the error rate in Illumina data is low, due to the high throughput nature of the data, there are many erroneous k-mers. 1-19

1.10 The picture shows a snapshot of aligned reads to a known reference. Indicated bases with black color are either sequencing errors or variants. Although there are enough reads to cover this region, too many errors in these reads reduces the actual coverage. 1-20

1.11 This figure shows two examples of contigs sets produced from the same dataset. Different metrics (N50, NG50, NA50 and NGA50) are computed for each set. 1-22

1.12 A schematic representation of three types of alignments: Global, Local and Overlap. 1-24

1.13 This figure shows a suffix tree and the equivalent suffix array for a short DNA sequence. An additional reserved character, a sentinel, which does not occur in the text, shows the end of the sequence. In this example, we have used the $ sign for this purpose. 1-27

2.1 Mismatches in read alignment. 2-8

2.2 SPAdes assemblies. 2-12

2.3 Fragmented assembly using corrected data. 2-14

2.4 Lost true 21-mers spectrum 2-15

2.5 Alignment of uncorrected and ACE-corrected reads in the neighborhood of a contig breakpoint 2-16

2.6 Alignment of uncorrected and corrected reads by Musket and Fiona in the neighborhood of a contig breakpoint: 2-17

2.7 Error correction with Karect resolves a breakpoint in the uncorrected data assembly. The first track (Ref) shows a part of the reference genome, which is assembled into a single contig from Karect-corrected reads. The second track (Uncorrected) shows the alignment of the uncorrected reads to the reference. The third track (Corrected Karect) uses these same alignment positions, but with the sequence content of reads corrected by Karect. The short overlap between the uncorrected reads is less than 21, i.e., a true 21-mer is missing from the uncorrected data. There are three reads which expand along this region but they contain some errors which are highlighted in purple. After error correction those three reads are partially cleaned which suffices to connect the two groups of reads. 2-18

2.8 Peak memory usage. 2-19

2.9 Runtime. 2-20

2.10 Runtime of DISCOVAR plus EC tools. 2-21

2.11 Peak memory usage of DISCOVAR and EC tools. 2-22

2.12 Runtime of SPAdes plus EC tools. 2-23

2.13 Peak memory usage of SPAdes and EC tools. 2-24

3.1 Overview of the first three steps of BrownieCorrector's pipeline. Read pairs for which one read contains a highly repetitive k-mer are extracted and clustered based on the sequence similarity between different read pairs. Each cluster is expected to contain reads that were derived from a single genomic regions. 3-7

3.2 The average quality score of bases in reads for different polymers and a group of randomly sampled reads. 3-7

3.3 While C is the initial coverage (top), the expected number of reads that fully cover a selected k-mer is C_k. Depending on the insert size and the insert size variability, the left and right flanking regions that are covered by the paired reads have a coverage of $C_k/2$ or lower. 3-8

3.4 The final step of the BrownieCorrector pipeline: (1) de Bruijn Graph is built from the uncorrected reads in a cluster. Uncorrected reads contain sequencing errors which result in the appearance of erroneous k-mers and subsequently erroneous nodes/arcs in the graph; (2) erroneous nodes (colored in red) are detected and removed from the graph based on coverage and graph toplogy. Such erroneous nodes often appear as tips or bubbles; (3) reads are aligned individually to the corrected graph and mismatches and indels in the reads are detected and fixed with the correct path in the graph. 3-10

3.5 Real example of a k-mer frequency spectrum that is a superposition of two distributions corresponding to real and erroneous k-mers, respectively. A model of two Poisson distributions is fit to the data using the expectation-maximization algorithm. The coverage cutoff is established at the intersection of the two distributions. 3-11
3.6 Peak memory usage. Peak memory usage of the EC tools. 3-14
3.7 Runtime. Runtime of the EC tools. 3-15
3.8 Alignment of BrownieCorrector-corrected, Reckoner-corrected and uncorrected paired reads in the neighborhood of a contig breakpoint: the first track contains part of the reference genome, which is assembled into a single contig from BrownieCorrector-corrected data but breaks into two contigs using Reckoner-corrected or uncorrected data. The second track (BrownieCorrector) shows the alignment of the BrownieCorrector-corrected reads. The only the reads in orange are corrected by BrownieCorrector. The third track (Reckoner) shows the alignment of the Reckoner-corrected reads. The forth track (Uncorrected) shows the alignment of uncorrected reads. Mismatches in the sequencing data are indicated with letters whereas an insertion is shown with a | sign and a deletion is shown with a - sign. 3-17

4.1 This figure shows the association between the de Bruijn graph and MM tables. On the left side, part of a de Bruijn graph is shown. True paths are depicted by blue lines. The numbers inside each node indicate the multiplicity of that node, i.e., the number of times the node's sequence is present in the reference genome. A table at each node guides the aligner based on previously observed nodes. The 2-MM and 3-MM tables of node A are shown on the right side. Based on the 2-MM table, reads that align to CA are guided to E as the continuation to node D is not allowed. However, the information in this table is insufficient to guide reads that align to BA since continuations to E and D are both valid. In contrast, the 3-MM table guides the reads that align to FBA to D, and GBA to E. The information in the final row in 3-MM table is redundant because it is also contained in the lower-order 2-MM table. 4-8
4.2 Peak memory Peak memory usage of the aligner tools for simulated datasets. 4-14
4.3 Runtime. Average runtime of tools to align 1M reads for the simulated datasets. 4-14
4.4 Runtime. The effect of branch and bound strategy on the running time of BrownieAligner. 4-15
4.5 Peak memory usage. Peak memory usage of the aligner tools for real datasets. 4-16
4.6 Runtime. Average runtime of tools to align 1M reads for the real datasets. 4-16

A.1 SPAdes assembly results for P. aeruginosa for both uncorrected and corrected data. Scaffolds with length NGAx or larger produce x% of the genome. A-37

B.1 The impact of changing the number of iterations in reads clustering on the quality of assembly in D1(*Homo sapiens* chr. 21). B-11
B.2 SPAdes assembly results for dataset D1 (*Homo sapiens* Chr. 21) for both uncorrected and corrected data. Contigs with length NGAx or larger produce x% of the genome. B-26
B.3 SPAdes assembly results for dataset D2 (*Homo sapiens* Chr. 14) for both uncorrected and corrected data. Contigs with length NGAx or larger produce x% of the genome. B-28
B.4 SPAdes assembly results for dataset D3 (*C. elegans*) for both uncorrected and corrected data. Contigs with length NGAx or larger produce x% of the genome. B-30
B.5 SPAdes assembly results for dataset D4 (*D. melanogaster*) for both uncorrected and corrected data. Contigs with length NGAx or larger produce x% of the genome. B-32
B.6 SPAdes assembly results for dataset D5 (*D. melanogaster*) for both uncorrected and corrected data. Contigs with length NGAx or larger produce x% of the genome. B-34
B.7 SPAdes assembly results for dataset D6 (*D. melanogaster*) for both uncorrected and corrected data. Contigs with length NGAx or larger produce x% of the genome. B-36
B.8 SPAdes assembly results for dataset D7 (*D. melanogaster*) for both uncorrected and corrected data. Contigs with length NGAx or larger produce x% of the genome. B-38
B.9 SPAdes assembly results for dataset D8 (*D. melanogaster*) for both uncorrected and corrected data. Contigs with length NGAx or larger produce x% of the genome. B-40
B.10 SPAdes assembly results for dataset D9 (*A. thaliana*) for both uncorrected and corrected data. Contigs with length NGAx or larger produce x% of the genome. B-42

List of Tables

1.1 An example of aligning three sequences. Mismatches are shown in different colors in respect to the first sequence, and gaps are indicated by a dash sign. 1-23

2.1 List of EC tools evaluated in this paper. The algorithmic approach is either k-mer spectrum based ('k-mer') or multiple sequence alignment based ('MSA'). Tools can be further classified according to data structure and heuristics used. Some tools are able to correct insertions or deletions. In their accompanying publication, all tools were assessed directly on their ability to reduce error rate, either on the read or base level. Most tools did not use assembly analyses with modern assemblers in their evaluation. SPAdes was used for the evaluation of BayesHammer, but no comparison was made with assembly results from uncorrected data. 2-4
2.2 Real datasets used for the evaluation of EC tools. 2-5
2.3 Accuracy comparison of EC tools in terms of EC gain, percentage of corrected errors, and number of newly introduced errors per Mbp of read data. 2-9
2.4 NGA50 of respectively contigs (top) and scaffolds (bottom) assembled by SPAdes before and after error correction. Arrows in the table are based on their value relative to the NGA50 value obtained from uncorrected data as follows: $\downarrow\downarrow$ $<$ -10% $<\downarrow<$ 0% $<$ $\uparrow<$ +10% $<\uparrow\uparrow$ 2-10

3.1 Real datasets used for the evaluation of the error correction tools. . 3-6
3.2 NGA50 of respectively contigs (top) and scaffolds (bottom) assembled by SPAdes before and after error correction. 3-12
3.3 NGA50 of respectively contigs (top) and scaffolds (bottom) assembled by SPAdes after error correction by both BrownieCorrector and Karect. 3-13

4.1 Artificial datasets used for the evaluation of graph aligner tools. . . 4-10
4.2 Real datasets used for the evaluation of graph aligner tools. 4-11
4.3 Accuracy comparison of graph aligner tools in terms of correct alignment of reads to the graph on simulated data. 4-12

4.4 Accuracy evaluation of BrownieAlignerNoMM and BrownieAligner on the subset of the simulated reads that align to a path of at least two nodes in the graph. 4-12
4.5 Accuracy comparison of graph aligner tools in terms of correct alignment of reads to the graph on real data. 4-13

A.1 Scaffold N50 of the SPAdes assembly from BFC-corrected reads . A-2
A.2 Scaffold N50 of the SPAdes assembly from BLESS 2-corrected reads A-3
A.3 Scaffold N50 of the SPAdes assembly from Blue-corrected reads . A-3
A.4 Scaffold N50 of the SPAdes assembly from Karect-corrected reads A-4
A.5 Scaffold N50 of the SPAdes assembly from Lighter-corrected reads A-5
A.6 Scaffold N50 of the SPAdes assembly from Musket-corrected reads A-5
A.7 Scaffold N50 of the SPAdes assembly from Trowel-corrected reads A-6
A.8 Percentage of reads that mapped with 0 mismatches (%). A-7
A.9 Percentage of reads that do not align with $<$10 mismatches. A-8
A.10 Detailed confusion matrices for real data. A-9
A.11 Accuracy comparison in terms of EC gain, percentage of corrected errors, and number of errors introduced per Mbp in simulated data. A-10
A.12 Detailed confusion matrices in simulated data A-11
A.13 Assembly quality metrics for *B. dentium* A-14
A.14 Assembly quality metrics for *E. coli str. K-12 substr. DH10B* . . A-15
A.15 Assembly quality metrics for *E. coli str. K-12 substr. MG1655* . . A-16
A.16 Assembly quality metrics for *S. enterica* A-17
A.17 Assembly quality metrics for *P. aeruginosa* A-18
A.18 Assembly quality metrics for *H. sapiens* Chr. 21 . . . . . A-19
A.19 Assembly quality metrics for *C. elegans* A-20
A.20 Assembly quality metrics for *D. melanogaster* A-21
A.21 Assembly quality metrics for *B. dentium* A-23
A.22 Assembly quality metrics for *E. coli str. K-12 substr. DH10B* . . A-24
A.23 Assembly quality metrics for *E. coli str. K-12 substr. MG1655* . . A-25
A.24 Assembly quality metrics for *S. enterica* A-26
A.25 Assembly quality metrics for *P. aeruginosa* A-27
A.26 Assembly quality metrics for *H. sapiens* Chr. 21 . . . . . A-28
A.27 Assembly quality metrics for *C. elegans* A-29
A.28 Assembly quality metrics for *D. melanogaster* A-30
A.29 Assembly quality metrics for *B. dentium* A-32
A.30 Assembly quality metrics for *E. coli str. K-12 substr. DH10B* . . A-33
A.31 Assembly quality metrics for *E. coli str. K-12 substr. MG1655* . . A-34
A.32 Assembly quality metrics for *S. enterica* A-35
A.33 Assembly quality metrics for *P. aeruginosa* A-36
A.34 Assembly quality metrics for *H. sapiens* Chr. 21 . . . . . A-38
A.35 Assembly quality metrics for *C. elegans* A-39
A.36 Assembly quality metrics for *D. melanogaster* A-40
A.37 Assembly quality metrics for *B. dentium* A-42

A.38 Assembly quality metrics for *E. coli str. K-12 substr. DH10B* . . A-43
A.39 Assembly quality metrics for *E. coli str. K-12 substr. MG1655* . . A-44
A.40 Assembly quality metrics for *S. enterica* A-45
A.41 Assembly quality metrics for *P. aeruginosa* A-46
A.42 Assembly quality metrics for *H. sapiens* Chr. 21 . . . . . A-47
A.43 Assembly quality metrics for *C. elegans* A-48
A.44 Assembly quality metrics for *D. melanogaster* A-49
A.45 Memory usage of EC tools (GB) A-50
A.46 Runtime of EC tools (min) A-50
A.47 Peak memory usage of EC tools. A-51
A.48 Runtime of the EC tools. A-51

B.1 The top-5 most frequent 15-mers in each dataset. B-6
B.2 The top-5 most frequent 15-$mers$ in the beginning or end of assembled contigs in different dataset. B-7
B.3 The improvement rate of NGA50 values for contigs and scaffolds upon the uncorrected data for different EC tools B-9
B.4 Two highly repetitive k-mers used in this study. The number of corrected and total number of reads in each dataset is compared. . B-10
B.5 NGA50 of respectively contigs (top) and scaffolds (bottom) assembled by SPAdes before and after error correction. Reads that contain a 15-mer poly (C/G) are corrected by BrownieCorrector. . B-11
B.6 NGA50 of respectively contigs (top) and scaffolds (bottom) assembled by SPAdes before and after error correction. Reads that contain a 15-mer poly (AC/TG) are corrected by BrownieCorrector. B-11
B.7 Assembly quality metrics for D1 . . . . . B-14
B.8 Assembly quality metrics for D2 . . . . . B-15
B.9 Assembly quality metrics for D3 . . . . . B-16
B.10 Assembly quality metrics for D4 . . . . . B-17
B.11 Assembly quality metrics for D5 . . . . . B-18
B.12 Assembly quality metrics for D6 . . . . . B-19
B.13 Assembly quality metrics for D7 . . . . . B-20
B.14 Assembly quality metrics for D8 . . . . . B-21
B.15 Assembly quality metrics for D9 . . . . . B-22
B.16 Assembly quality metrics for D1 . . . . . B-25
B.17 Assembly quality metrics for D2 . . . . . B-27
B.18 Assembly quality metrics for D3 . . . . . B-29
B.19 Assembly quality metrics for D4 . . . . . B-31
B.20 Assembly quality metrics for D5 . . . . . B-33
B.21 Assembly quality metrics for D6 . . . . . B-35
B.22 Assembly quality metrics for D7 . . . . . B-37
B.23 Assembly quality metrics for D8 . . . . . B-39
B.24 Assembly quality metrics for D9 . . . . . B-41
B.25 Peak memory (GB) usage of the aligners on real data. B-43

B.26 Run time (min) of the aligners on real data B-43

C.1 Accuracy evaluation of graph aligners on simulated data C-5
C.2 Accuracy evaluation of BGREAT on simulated data for different values of k. C-6
C.3 Accuracy evaluation of BrownieAligner on simulated data for different values of k. C-6
C.4 Accuracy evaluation of deBGA on simulated data for different values of k. C-6
C.5 Accuracy evaluation of BrownieAligner on simulated data for different values of maxOrder. C-7
C.6 Accuracy evaluation of BrownieAligner on simulated data for different values of minLikelihoodRatio. C-7
C.7 Accuracy evaluation of BrownieAligner on simulated data for different values of minChainCov. C-7
C.8 Accuracy comparison of graph aligners on real data C-8
C.9 Accuracy evaluation of BrownieAlignerNoMM and BrownieAligner on the subset of the real data that are corrected by DFS Algorithm. C-8
C.10 Peak memory (GB) usage of the aligners on simulated data C-9
C.11 Run time (min) of the aligners on simulated data C-9
C.12 Effect of the branch and bound strategy on the run time (min) of BrownieAligner on simulated data C-9
C.13 Peak memory (GB) usage of the aligners on real data. C-10
C.14 Run time (min) of the aligners on real data C-10

List of Acronyms

bp	Base pair
BFS	Breadth-first search
BWA	Burrows Wheeler Aligner
BWT	Burrows Wheeler Transform
CPU	Central Processing Unit
DBG	De Bruijn Graph
DFS	Depth-first search
DNA	Deoxyribonucleic acid
DNA-seq	DNA sequencing
EC	Error correction
FGS	First generation sequencing
FN	False negative
FP	False positive
GA	Genome Analyzer
GB	Gigabyte
GHz	Gigahertz
Indel	Insertion or deletion
Kb	Kilobase
Kbp	Kilobase pair
Mbp	Megabase pair
MEMs	Maximal exact matches
MM	Markov model
MSA	Multiple sequence alignment

NGS	Next Generation Sequencing
NP	Nondeterministic Polynomial time
OLC	Overlap Layout Consensus
PacBio	Pacific Biosciences
PCR	Polymerase chain reaction
PSA	Pairwise Sequence Alignment
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
SGS	Second Generation Sequencing
SMRT	Single molecule real time
SNP	Single Nucleotide Polymorphism
TGS	Third Generation Sequencing
TN	True negative
TP	True positive
TPR	True positive rate

Samenvatting

Illumina-sequeneringsplatformen zijn alomtegenwoordig. De nieuwste generatie Illumina-machines is in staat terabasen aan sequeneringsdata the genereren per experiment. Deze data wordt gekarakteriseerd door een korte *read*-lengte (100-300 bp) en een hoge accuraatheid (1-2% foutenmarge). Substitutiefouten komen het vaakst voor, terwijl inserties en deleties zeldzamer zijn. Ondanks de lage foutenmarge, lijdt de output van Illumina-platformen aan bias uit verschillende bronnen, zo komen fouten bijvoorbeeld vaker voor op het einde van reads. Daarenboven zijn de reads niet uniform verdeeld over het genoom, waardoor bepaalde regio's van het genoom niet of nauwelijks gecoverd zijn terwijl andere regio's extreem hoog gecoverd zijn. Illumina-data wordt gebruikt in verscheidene contexten, zoals klinische diagnostiek, gepersonaliseerde medicijnen en landbouw, maar ook in therapeutisch, forensisch en fundamenteel onderzoek, zelfs in sociologie. Zo kan men bijvoorbeeld korte reads aligneren op een referentiegenoom en genetische varianten onderzoeken in de hoop dat dit de oorzaak van bepaalde beperkingen of ziektes onthult. Wanneer men geen referentiegenoom ter beschikking heeft, wordt het genoom *de novo* geassembleerd om zo de originele DNA-sequentie te onthullen. Ook in de aanwezigheid van een referentiegenoom kan de-novo-assemblage nuttig zijn om grotere structurele varianten zoals herschikkingen, translocaties of inversies te detecteren. Het assembleren van DNA is echter geen eenvoudig probleem en wordt gecategoriseerd als NP-compleet. Deze taak wordt verder bemoeilijkt door de aanwezigheid van sequeneringsfouten, variabiliteit in coverage en de korte lengte van de reads.

Foutcorrectie- (FC-)tools proberen sequeneringsfouten op te sporen en te corrigeren om zo assemblagemethodes te voorzien van correctere inputdata en de kwaliteit van de resulterende assemblage te verbeteren. Vreemd genoeg worden FC-tools echter niet geëvalueerd aan de hand van hun verbetering van de kwaliteit van de-novo-genoomassemblage. Vaker wordt hun vermogen om fouten te corrigeren an sich als evaluatiecriterium gebruikt. Wij onderzochten de impact van FC-tools op de-novo-genoomassemblage. In deze studie concludeerden we dat moderne assemblagetools zoals SPAdes of DISCOVAR weinig baat hebben bij een foutcorrectie vooraf, zelfs al verminderen de FC-tools het aandeel sequeneringsfouten significant zonder veel nieuwe fouten toe te voegen. Een mogelijke oorzaak hiervan is dat state-of-the-art-assemblagetools zelf in staat zijn sequeneringsfouten op te spo-

ren door middel van een correctieprocedure op de de Bruijn-graaf (DBG). Fouten die gemaakt zijn in het begin of einde van genomische reads, komen bijvoorbeeld vaak in de graaf voor als *tips*, knopen die een doodlopend pad vormen. Fouten die gemaakt worden in het midden van genomische reads, daarentegen, komen voor als *bubbels*, parallelle paden, in een de Bruijn-graaf. Dit soort artefacten kunnen gemakkelijk opgespoord worden door assemblagetools. Immers, een knoop die een tip of bubbel vormt en weinig gecoverd is door reads, heeft een hogere kans om een sequeneringsfout voor te stellen en kan dus uit de graaf verwijderd worden. Bijgevolg is de grote meerderheid aan sequeneringsfouten onschuldig voor de genoomassemblage en vormt slechts een kleine fractie van de sequeneringsfouten een probleem. Fouten die zich voordoen in laag gecoverde regio's zijn een voorbeeld van problematische fouten omdat assemblagetools in deze situatie vaak niet in staat zijn het overlappen van reads te bepalen. Een tweede probleem doet zich voor wanneer een sequeneringsfout in een bepaalde regio resulteert in een k-meer dat bestaat in een andere genomische regio. Een sequeneringsfout van deze aard heeft een chimerische, valse verbinding tussen knopen in de de Bruijn-graaf tot gevolg. Tijdens ons onderzoek naar de performantie van FC-tools bemerkten we dat deze tools niet altijd in staat zijn dit soort fouten te corrigeren. Meer bepaald vonden we dat FC-tools vaak slecht om kunnen gaan met regio's met een lage read-coverage die zich in de buurt van zeer frequente *repeats* bevinden. FC-tools gaan verkeerdelijk uit van de aanwezigheid van sequeneringsfouten bij lage read-coverage, terwijl de repetitieve elementen de FC-tools aanzetten tot inconsistente substituties. Als gevolg hiervan is het mogelijk dat twee reads die afkomstig zijn van dezelfde genomische regio nog steeds geen overlap vertonen na correctie. De onderliggende reden voor deze fouten is het individueel corrigeren van de reads. Bijgevolg zijn FC-tools niet in staat de kwaliteit van de assemblage te verbeteren, maar introduceren ze zelfs nieuwe fouten in de data.

In deze studie bieden we een antwoord op het hierboven gestelde probleem. We introduceren BrowieCorrector; dit is een gerichte foutcorrectie-tool die enkel focust op het corrigeren van reads die extreem repetitieve patronen bevatten. Dit soort reads vormt immers een grotere uitdaging voor zowel assemblagetools als foutcorrectie-tools. BrownieCorrector begint met het extraheren van reads die zowel een lage complexiteit als een extreem repetitief patroon, zoals een poly-A/T, delen. Vervolgens wordt de volledige sequentie van de read alsook *paired-end* informatie gebruikt om de read-paren in homogene groepen te clusteren. Hiertoe gebruiken we het *Louvain community detection*-algoritme om de reads te clusteren op basis van sequentiegelijkheid. De reads worden per cluster gecorrigeerd zodat de correctie van alle reads in een groep consistent is. Omdat we kleine groepen van reads in een cluster onafhankelijk van andere clusters corrigeren en niet de volledige dataset, zijn we in staat een kleinere k-meer grootte te gebruiken zonder te lijden onder chimerische verbindingen in de graaf. Daarenboven worden reads op een meer consistente wijze gecorrigeerd aangezien reads uit een cluster verondersteld worden afkomstig te zijn uit dezelfde genomische regio. Om de performantie van BrownieCorrector te evalueren, vergeleken we met verschillende andere state-of-the-art-FC-tools door middel van zes echte Illumina-datasets afkomstig van ver-

schillende eukaryote genomen. We onderzochten bovendien ook de impact van foutcorrectie op hybride assemblage, hierbij worden gecorrigeerde Illumina-reads ondersteund door PacBio-data. Onze resultaten bevestigen dat BrownieCorrector in staat is de kwaliteit van genoomassemblage te verbeteren. In de meeste gevallen resulteert de correctie met BrownieCorrector in de beste assemblage, zelfs al worden slechts 2% van de reads gecorrigeerd.

Binnen de clusters worden de reads in drie stappen door BrownieCorrector gecorrigeerd. Ten eerste wordt de met de reads geassocieerde de Bruijn-graaf, opgebouwd. Vervolgens worden de knopen en bogen die corresponderen met foute k-meren verwijderd door middel van de typische graafcorrectieprocedures zoals het verwijderen van tips en detecteren van bubbels. Ten slotte worden alle reads in de cluster opnieuw gealigneerd op de gecorrigeerde DBG. Om de reads te aligneren op de DBG, stellen we BrownieAligner voor. Dit is nieuwe software die ontwikkeld en geïmplementeerd werd om korte Illumina-reads te aligneren op een DBG. Deze taak is computationeel veel duurder dan het aligeneren van reads op een lineair referentiegenoom en is zelfs gekend als een NP-compleet probleem. Als oplossing stellen we een seed-and-extendmethodologie voor waarbij k-meermatches als seed gebruikt worden. Indien geen k-meer-matches gevonden worden, gebruiken we maximale exacte matches. Gegeven een seed, zal ons algoritme alle takken van de zoekboom verkennen totdat een optimaal gealigneerd pad gevonden wordt. Om de zoekruimte te verkleinen terwijl we toch optimale resultaten garanderen, stellen we een aantal branch-and-boundtechnieken voor. Daarenboven stellen we hogere-orde Markov Modellen (MM) voor om het aligneren tegen paden in de DBG die geen echte deelsequenties van het referentiegenoom voorstellen, te vermijden. BrownieAligner werd toegepast op zowel synthetische als echte datasets. Deze tool presteert algemeen beter dan state-of-the-arttools op het vlak van accuraatheid, terwijl hij gelijkaardige tijd- en geheugenvereisten heeft. Onze resultaten bevestigen dat het gebruik van hogere-orde MM's in Brownie-Aligner de accuraatheid verbeteren, terwijl branch-and-boundalgoritmen de looptijd verlagen.

Samengevat behandelden we een veelvoorkomend probleem bij Illumina-FC-tools die gebruikt worden bij het pre-processen van sequeneringsdata alvorens over te gaan tot genoomassemblage. Hierbij hebben we te maken met problematische genomische regio's die zowel FC-tools als assemblagetools uitdagen. In deze studie stellen we BrownieCorrector voor om sequeneringsfouten in de reads te corrigeren die zich voornamelijk voordoen in deze problematische regio's en zo de algemene kwaliteit van de assemblage te verbeteren. Wij zijn van mening dat het voor toekomstige FC-tools voordeliger is om te focussen op problematische regio's in plaats van een voldoende foutcorrectie over het volledige genoom te beogen. Nieuwe, complexe algoritmen die mogelijks meer CPU-cycli nodig hebben, kunnen ontwikkeld en geïmplementeerd worden om reads uit zulke regio's te corrigeren. Deze algoritmen zullen ook de informatie bevat in paired-end reads moeten uitbuiten om een consistentere foutcorrectie te bekomen. Om deze redenen verwachten we dat de door ons voorgestelde techniek in de toekomst verder zal gebruikt en ontwikkeld worden in dezelfde of een andere context. Daaren-

boven introduceerden we tevens een nieuw graafalignerings-algoritme dat korte Illumina-reads kan afbeelden op een de Bruijn-graaf. Wij zijn van mening dat reads van andere soorten data zoals 10x-Genomics of zelf langere PacBio-reads kunnen gealigneerd worden op een de Bruijn-graaf door gebruik te maken van dezelfde branch-and-boundtechniek. Het is tevens mogelijk het voorgestelde MM te gebruiken om repeats op te lossen in een standalone-applicatie voor scaffolding. Zowel BrownieCorrector als BrownieAligner werden geschreven en C++ en staan vrij ter beschikking.

Summary

Currently, Illumina sequencing platforms are widely used. With the latest generation of Illumina machines, it is possible to generate terabases of sequencing data per run. These data are characterized by a relatively short read length (100-300 bp) and a high accuracy (1-2% error rate). Substitution errors are most common whereas insertions and deletions occur less frequently. Despite the low error rate, data produced on Illumina platforms suffer from various sources of bias. For example, there is a higher rate of error toward the end of the reads and the distribution of reads across the genome is not uniform. As a result, some regions in the genome have no or only poor coverage whereas other regions are covered higher than average. Illumina data can be exploited in diverse contexts such as clinical diagnostics, fundamental research, therapeutic discovery, personalized medicine, forensics, agriculture and even in sociology. In particular, one may want to align short reads to a reference genome and perform variant calling, thus revealing the underlying cause of certain disabilities or diseases. Alternatively, in the absence of a reference genome, one needs to perform *de novo* genome assembly to retrieve the original DNA sequence. Even with a known reference genome, *de novo* genome assembly can be useful to discover larger structural variations such as rearrangements, translocations or inversions. However, DNA assembly is a difficult problem and categorized as NP-complete. The presence of errors, coverage variability and the short read length further complicate this task.

Error correction (EC) tools try to detect and correct the sequencing errors to provide assembly methods with cleaner input data and hence improve the quality of the resulting assemblies. Peculiarly, most EC tools were not evaluated on their ability to enhance the quality of *de novo* genome assembly with modern assemblers, but rather directly on their ability to correct sequencing errors. In this study, we assessed the impact of EC tools on the *de novo* genome assembly. We found that although EC tools significantly reduce the fraction of sequencing errors without introducing too many new errors, modern assemblers like SPAdes or DISCOVAR do not benefit much from this pre-correction step. The reason for this is that state-of-the-art assembly tools also detect sequencing errors through a correction procedure on the de Bruijn graph (DBG). For instance, errors that occur at the end (or beginning) of reads mostly appear as tips or dead-end nodes, and errors in the middle of reads appear as parallel paths or bubbles in the DBG. These artifacts can

easily be detected by the assembler as well: a tip or bubble node that is not covered by many reads most likely represents a sequencing error and needs to be removed from the graph. Therefore, the vast majority of sequencing errors are harmless to the assembly process. Only a relatively small fraction of sequencing errors is truly problematic, e.g., those that occur in regions with shallow read coverage. Such errors may render the assembler unable to identify overlap between reads in that region. Another issue is when a sequencing error gives rise to a spurious chimeric connection between nodes in the DBG. This occurs when a sequencing error in one context yields a true k-mer from a different genomic location. We investigated the performance of EC tools, and noticed that they are sometimes unable to identify and correct these types of errors. In particular, we observed that regions with low read coverage and in the vicinity of highly frequent repeats are often difficult for EC tools to handle. Due to the shallow coverage, EC tools incorrectly assume the presence of sequencing errors. The repeated elements, on the other hand, cause EC tools to apply inconsistent substitutions. Hence, two reads that originate from the same genomic location might still not overlap after error correction. The underlying reason is that reads are often corrected individually. Consequently, EC tools sometimes deteriorate assembly results due to newly introduced errors.

In this study, we provided an answer to the above-raised problem. We introduced BrownieCorrector, a targeted EC tool that focuses only on the correction of reads that contain highly repetitive patterns. BrownieCorrector first extracts the reads that share a low-complexity, highly repetitive pattern such as a poly (A/T). Then it uses the entire read sequence as well as the paired-end read information to cluster read pairs in homogeneous groups. We used the Louvain community detection algorithm to cluster the reads based on sequence similarity. Reads within each cluster are assumed to originate from the same genomic location and are corrected per cluster, thus achieving a consistent correction for all reads within each cluster. Correcting smaller groups of reads in each cluster independently from other clusters, instead of trying to correct the entire dataset as a whole, allows us to effectively use a small k-mer size without suffering much from chimeric connections in the graph. Additionally, since each cluster is assumed to contain reads from a single genomic region, reads are corrected consistently. To evaluate the performance of BrownieCorrector, we compared it with other state-of-the-art EC tools using six real Illumina datasets from different eukaryotic genomes. Additionally, we investigate the impact of error correction on the hybrid assembly where the corrected Illumina reads are supplemented with PacBio data. Our results confirm that BrownieCorrector improves the quality of genome assembly and leads to the best assembly in most cases, despite the fact that it corrects fewer than 2% of the reads.

BrownieCorrector corrects the reads in each cluster in three steps. It first constructs the associated DBG, then performs typical graph cleaning procedures such as tip-clipping and bubble detection to remove erroneous nodes and arcs which represent erroneous k-mers in the data. Finally, the reads in that cluster are aligned back to the cleaned DBG. To align reads to the graph, we propose BrownieAligner, new software that is designed and implemented to align short Illumina reads to the

DBG graph. Compared to aligning reads to a linear reference genome, this task is computationally more expensive and also known to be NP-complete. We propose a seed-and-extend methodology where seeds correspond to either k-mer matches or maximal exact matches in case no k-mer matches can be found. Given a seed, our algorithm explores all branches of the tree until the optimal alignment path is found. We propose the use of branch-and-bound techniques to reduce the search space while still guaranteeing optimal results. Additionally, we propose higher-order Markov Models (MM) to avoid the alignment against paths in the DBG that do not represent actual subsequences of the original reference genome. BrownieAligner is applied to both synthetic and real datasets. It generally outperforms other state-of-the-art tools in terms of accuracy, while having similar runtime and memory requirements. Our results confirm that using the higher-order MM in BrownieAligner improves the accuracy, while the branch-and-bound algorithm reduces the runtime.

In conclusion, we addressed a prevalent issue in Illumina EC tools which are used for the pre-processing of sequencing data prior to genome assembly. There are problematic regions in the genome that are more challenging for both EC tools and assemblers to deal with. In this study, we proposed BrownieCorrector which corrects sequencing errors within reads particularly originating from these regions and improves the overall quality of assembly. We believe that for the future EC tools, it is more worthwhile to focus on the problematic regions rather than having a fair genome-wide error correction. New complex algorithms which perhaps need more CPU cycles can be designed and implemented to correct reads from these regions. These algorithms need to exploit the paired-end read information to have a more consistent error correction. Therefore, we expect our proposed technique to be further used and evolved in the same way or other contexts in future. Furthermore, we introduced a new graph aligner that maps short Illumina reads to the DBG. We think that reads from other types of data like 10x-Genomics, or even longer reads from PacBio can be aligned to the DBG using the same branch-and-bound technique. Moreover, it is possible to use the suggested MM to resolve repeats in a standalone scaffolder application. Both BrownieCorrector and BrownieAligner are written in C++ and are available for the community.

1 Introduction

"... knowledge of sequences could contribute much to our understanding of living matter.[1] *"*

In this introduction, we are first going to describe the process of DNA sequencing in a chronological order. We will then review three generations of sequencing technologies and the current state-of-the-art in this field. Our primary focus is on the Illumina platforms, the most commonly used sequencers nowadays. Following this, we will describe the different types of sequencing errors and biases in the Illumina data, the reason why these occur and their potential consequences in downstream applications such as assemblers. With a main focus on the de Bruijn graph based assemblers, we shortly review the existing challenges and concepts in de novo assembly. Later, we will explain the different underlying approaches in error correction tools which intend to reduce the number of errors in the sequencing data. Next, we are going to briefly introduce sequence alignment, particularly short read alignment, where Illumina reads are mapped to a given reference genome. In the read alignment, the reference is often a linear sequence, yet in the graph alignment Illumina reads are mapped to a nonlinear reference given in the format of the de Bruijn graph. Finally, we will provide an overview of the different chapters in this thesis together with a list of publications authored or co-authored throughout this research period.

⋆ ⋆ ⋆

[1] Frederick Sanger

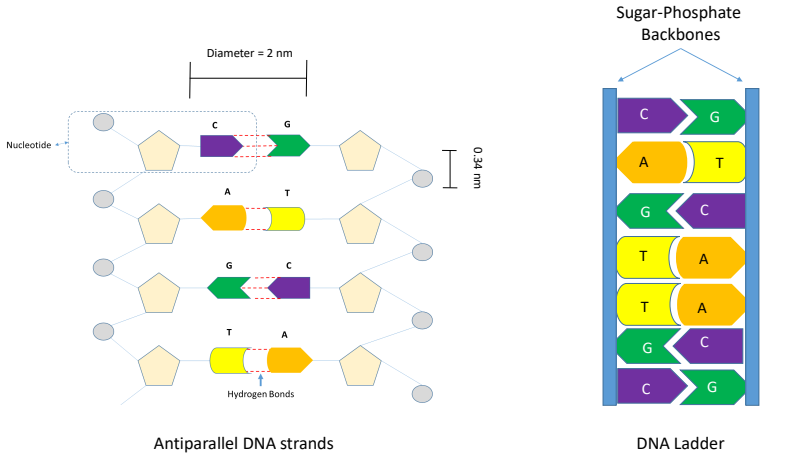
1.1 DNA sequencing

All living species such as animals or plants are composed of cells; some organisms have a single cell, and some more complex multicellular organisms have over millions or even billions of cells. Cells are the smallest living units that can reproduce themselves. Each cell has an identical copy of the DNA constituting the genome. The DNA sequence of every organism's genome is the unique blueprint of their development that carries and transfers genetic information. DNA is responsible for the functioning, growth and reproduction of every organism. In 1953, James D. Watson and Francis Crick discovered the three-dimensional structure of the DNA molecule [1–3]. The DNA molecule has a double-helix shape comprised of two complementary strands curled around each other like a twisting ladder. Each strand is made up of a chain of nucleotides. Each nucleotide is composed of a phosphate group, a sugar, and one of four nitrogen-containing nucleobases (Adenine [A], Guanine [G], Cytosine [C], and Thymine [T]). The nucleotides are connected to each other based on the covalent bonds between the sugar of one nucleotide and the phosphate of the next. The nitrogenous bases of the two strands are bound together, based on the base-pairing rules (A with T and C with G). Hydrogen bonds join the two strands together, which results in double-stranded DNA (see Fig. 1.1). The genetic information is encoded into these strands, and the two strands have the same biological data. DNA has a dynamic structure which makes it able to curl into tight loops. It can be a very long molecule containing millions to hundreds of millions of nucleotides. For example, the human genome is comprised of almost three billion nucleotides and its longest chromosome, chromosome number 1, has about 220 million base pairs, which is nearly 85 mm long when straightened. All the differences and similarities between all living organisms are encoded in the DNA based on the number and order of these bases.

Establishing the precise order of nucleotides within a DNA molecule is called DNA sequencing. It can refer to a technique or technology that determines the order of the four bases, Adenine (A), Thymine (T), Cytosine (C), and Guanine (G), in a given strand of DNA. The advent of DNA sequencing was a turning point in the history of biological and medical science. Nowadays, sequencing data are used in a wide variety of applications within diverse contexts such as clinical diagnostics, fundamental research, therapeutic discovery, personalized medicine, forensics, agriculture, and even in sociology [4]. Since the discovery of DNA, constant attempts have been made to propose better methods to do the sequencing in a shorter time, with a lower cost, and higher accuracy in higher volume. The history of DNA sequencing techniques is typically divided into three chapters, which are known as First-, Second-, and Third-Generation Sequencing, which are described in the following paragraphs [5–9].

First-generation sequencing (FGS) started in 1977 when Frederick Sanger and

Figure 1.1 A double-stranded DNA structure diagram.



his colleagues proposed the Sanger or dideoxy sequencing method. Later, in 1980, Sanger was awarded his second Nobel prize in chemistry for this outstanding achievement. Sanger sequencing was the most commonly used sequencing technique for many years, and even nowadays it is still being used for certain small-scale projects. In Sanger sequencing, initially, the target DNA needs to be prepared as a single strand. Then, this template DNA is supplied with a mixture of all four regular (deoxy) nucleotides in large quantities, and in smaller quantity a mixture of all four dideoxynucleotides. Each dideoxynucleotide is labeled with a "tag" that emits a distinct color based on its attached nucleotide. The chain polymerization halts when, by chance, a dideoxy nucleotide base is inserted instead of a regular base. In the next step, segments with different lengths and the same starting position are sorted in the increasing order of length size. The end positions of these segments are known. Consequently, each segment can be sequenced by looking at the last bases of its smaller segments. The ratio of the regular nucleotides to the dideoxy type determines the expected average length of the reads. With a sufficiently high ratio, several hundred nucleotides can be added to the DNA strands. Typically, Sanger reads are around 1 kb in length.

Sanger opened the first window into the world of sequencing, yet his method was too expensive and also too slow to sequence many whole genomes. Later, several new techniques were proposed and implemented and have become commercially available since 2000, which is known as massively parallel sequencing, Second Generation Sequencing (SGS) or Next-Generation Sequencing (NGS). SGS emerged to address FGS's shortcomings like high cost and low throughput. In SGS, thousands of different strands are sequenced in a parallel fashion,

which tremendously increases the throughput and allows us to sequence the entire genome at once. SGS is also known to use repeated wash-and-scan cycles for the sequencing. The scanning cycle starts with fragmenting the genome into small segments. Because thousands of copies of one segment need to be sequenced simultaneously, the amplification of the initial segment is performed using the polymerase chain reaction (PCR) technique[2]. Then, segments are loaded on the surface of the sequencing panel where chemical reactions occur between segments and the added nucleotides. This process builds the reverse-complement strand for each segment. As a result of each reaction, a distinct color is emitted each time, and a susceptible camera specifies different colors and translates them into the four corresponding bases. In the wash step, previously added nucleotides are removed by a chemical reaction to start a new phase.

The best-known examples of SGS platforms are 454, Illumina Genome Analyzer (GA), Illumina HiSeq, Illumina MiSeq, and Illumina NovaSeq. The HiSeq machine is still one of the most popular sequencing machines. The latest series can generate over 5 billion paired-end reads of length (2×150 bp [3]) within less than four days. However, Illumina MiSeq can generate 25 million paired-end reads of length (up to 2×300 bp) in a day. Finally, Illumina NovaSeq machines can generate 20 billion paired-end reads of length (2×250 bp) in less than two days. By comparison with the FGS, SGS machines generate reads faster, cheaper, and in higher throughput, but the read length is shorter. Fig.1.2[4] compares the throughput of three popular Illumina machines. In the next section, we will discuss the properties of the Illumina sequencing data in more detail.

Third-generation sequencing (TGS), which is usually known as single-molecule sequencing, appeared to break the limitations of the SGS, such as the short read length and eliminate the slow process of PCR amplification phase which introduces biases in the DNA sequences [10]. There are two main types of this technology: First, those that observe synthesis of a single molecule of DNA (e.g., Pacific Biosciences Single-molecule real-time sequencing (SMRT) technology); second, those that observe bases while passing through a nanopore.

SMRT reduces the sequencing time from days to hours because it does not need the scanning and washing steps or the PCR amplification. In contrast to the previous technology, the phospholinked nucleotides carry their fluorescent label on the terminal phosphate rather than the base. Therefore, the enzyme cleaves away the fluorescent label as part of the incorporation process leaving behind a complementary strand of DNA.

Typically, nanopore technologies distinguish bases by observing their effect

[2]PCR amplification is not a necessary step in Illumina DNA sequencing. The Illumina machine needs a minimum amount or concentration of DNA to operate efficiently; if that is supplied by the user, the PCR phase can be eliminated.

[3]Here, $2\times$ means that reads are paired.

[4]The data is collected by the time of writing this thesis, May 2019.

Figure 1.2 This figure compares the throughput of three popular Illumina machines (MiSeq, HiSeq and NovaSeq)

Sequencing System	MiSeq	HiSeq	NovaSeq
Model		4000	6000
Output per run	15 Gb	1.5 Tb	1Tb- 6 Tb

while they are passing through an electrical current. In this way, the DNA strands are mixed with copies of a processive enzyme. The size of these strands can be determined by the user (approaching 1 Mb according to the Nanoporetech [11]). The enzyme ratchets the DNA strand through the nanopore one base at a time (the user can control the speed). As the DNA moves through the pore, the nucleotides in the strand being processed create a characteristic disruption in the electrical current. This nanopore signal can be used to define the order of bases on that DNA strand. When a nanopore has processed a complete read, it will start a new one.

Popular TGS platforms are PacBio and Oxford Nanopore. In contrast to the SGS, they generate reads exceeding in length several kilobases, with fast and more straightforward sample preparation. However, the throughput is decreased, and the error rate increased from nearly 1% in the SGS to 15%, which are most notably indels.

1.2 Illumina sequencing data

Illumina generates sequencing data in high throughput with a low financial cost[5]. It has been estimated that over 90% of sequencing data worldwide are generated on Illumina platforms. The Illumina reads are characterized by a relatively short read length ($\leq$ 300 bp) and a high accuracy (1- 2% errors, mostly substitutions). Illumina dye sequencing or DNA sequencing is a method used to specify the series of base pairs in DNA. Bruno Canard and Simon Sarfati at the Pasteur Institute

[5]The price to sequence 1 Mbp of data with Illumina MiSeq is approximately 0.1 dollar while it costs 2400 dollars to sequence the same amount of data with Sanger.

in Paris proposed the idea for the first time. Later, it was developed by Shankar Balasubramanian and David Klenerman at Cambridge University. The technique can be used in a wide variety of genomic analyses from whole-genome and regional sequencing to transcriptome analysis, metagenomics, small RNA discovery, and methylation profiling.

Illumina sequencing has four steps: library preparation, cluster amplification, sequencing, and analysis. An overview of the sequencing process is shown in Fig. 1.3. The first step begins with a purified DNA molecule. The DNA can be randomly fragmented into smaller pieces using transposase enzymes or mechanically by acoustic shearing, then adapters are added to both ends of each fragment. For the amplification phase, the library is loaded into a surface of the flow cell, which is filled with *oligonucleotides* (short nucleotide sequences). Adapters at the end of fragments find their reverse-complement oligos and bind from both sides forming bridges in the surface of the panel. Then, bridges are duplicated repeatedly based on the reverse-complement matching rules. Through the bridge amplification process, which is done in parallel, multiple identical copies of the initial fragment are produced.

Once an adequate number of identical fragments are produced, the complementary strands are washed off the flow and the sequencing by synthesizing starts. In this step, primers attach to the forward strands add fluorescently tagged nucleotides to the DNA strand. A reversible terminator is on every nucleotide to prevent multiple additions in one round. Using the four-color chemistry[6], each of the four bases has its own unique emission color, and the machine detects the added base after each round. When a certain number of bases are scanned from one side of the bridge, the same process starts from the other side to scan the reverse read. This time, the forward strands are washed off, and the machine detects bases from the opposite strand base per base. The sequencing is performed in millions of clusters in parallel while each cluster contains hundreds of identical copies of DNA fragments. The newly identified sequencing reads are used in a wide variety of data analyses, ranging from *de novo* assembly, single nucleotide polymorphism (SNP) detection, phylogenetic inference, and metagenomic studies.

[6]There is an alternative system that uses only two colors to determine the bases in the sequencing step. In this way, each of the four different combinations of two colors specifies one base. For example, by using the two colors of green and red, one can determine the bases as follows: A (red, green), T (green, green), C (red, red) and G (no color, no color). By employing this technique recently in NexSeq and NovaSeq Illumina machines, the sequencing time was reduced. However, this method results in higher pollution of the light signals over time which makes it more difficult to distinguish the bases and further to interpret the base quality. Besides, when there is no signal to detect, the machine can mistakenly record a G base. This can lead to a bias toward G base in the sequencing data.

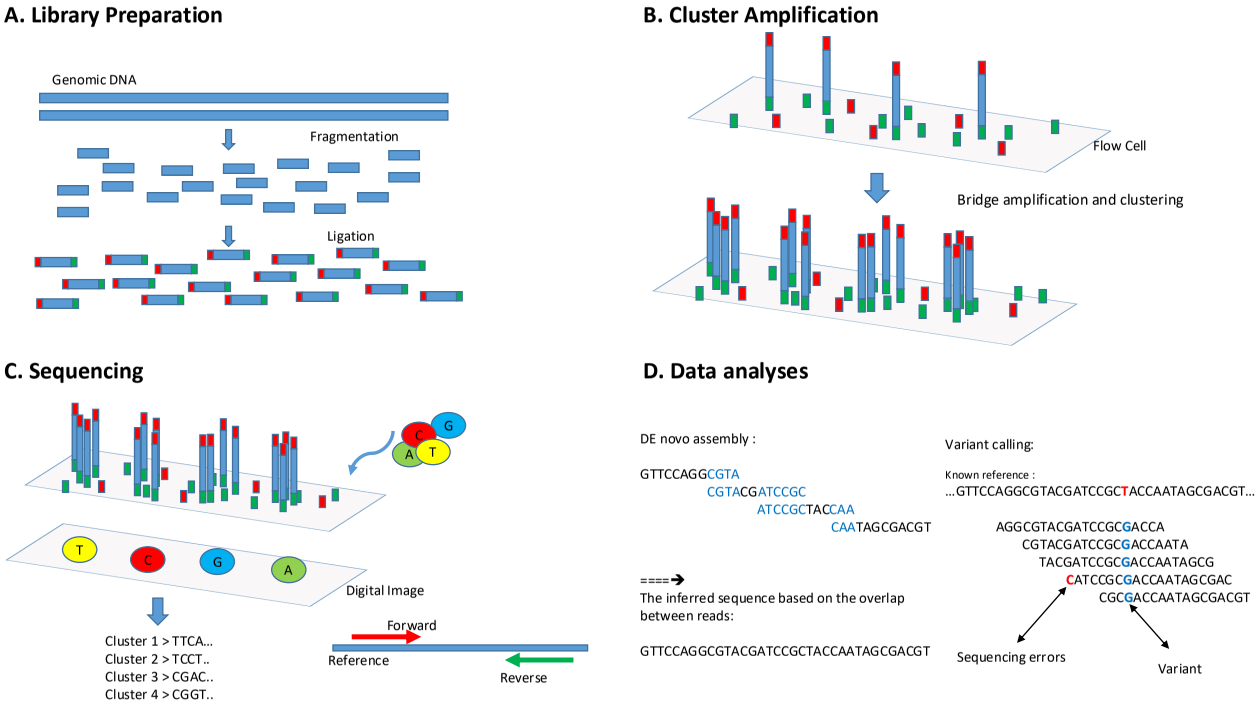


Figure 1.3: This figure shows the Illumina sequencing steps:
A. Library preparation: the DNA molecule is initially fragmented into smaller fragments. Then, in the ligation step, adapters are appended to both ends of each fragment.
B. Cluster amplification step: these fragments are loaded into a flow cell where they attach to the surface of this cell. Then, a reverse strand is created through the polymerase process, which forms a bridge on the surface. Multiple bridges are shaped around the initial segments forming clonal clusters.
C. Sequencing step: every cluster is now sequenced in parallel. In each round, the fluorescently labeled nucleotides are added, and the actual bases are determined based on the emitted color.
D. Scientists use the collected sequencing data in different contexts and applications. For example, based on the overlaps between reads a de novo assembler attempts to construct the original sequence.

1.3 Fastq file format

Illumina sequencing data are normally provided in FASTQ file format. In this format, both the biological sequence and its per base quality score are given. Both the sequence and the quality scores are encoded with ASCII characters. A FASTQ file uses four lines per sequence entry (see Fig. 1.4).

1. line 1: always begins with an "@" character followed by a sequence-unique identifier.
2. line 2: Sequence letters.
3. line 3: Begins with a "+" and is sometimes followed by the sequence identifier.
4. line 4: The quality values for the sequence in line 2. The 2 and 4 have the same length; for each base in the second line, there is an equivalent quality score in the fourth line.

Figure 1.4 This figure shows eight lines of an interleaved Fastq file consisting of four reads (two pairs).

Quality scores are produced during the sequencing run for every base based on the observable properties of clusters such as intensity profiles and signal-to-noise ratios. A quality score (Q-score), also known as a Phred score, is an integer value predicting the estimated probability of an error. The higher the Phred score for a base, the more reliable that base is, and the less likely it is to be incorrect. With P, the error probability and Q the Phred score:

$$P = 10^{-Q/10}$$

$$Q = -10\, log_{10}(P)$$

For example, the probability of a base with a quality score of Q40 being incorrect is 10^{-4}, or for a base with a quality score of Q30, one base call in 1,000 is

predicted to be incorrect. Quality scores are encoded to ASCII characters ranging from "!", which represents the lowest quality (Q0) to "K", which represents the highest quality (Q42).[7]

1.4 Illumina sequencing errors and biases

In the idealized case of perfect DNA sequencing, the sequencer distributes the reads uniformly across the genome without sequencing errors or coverage variability. However, all existing sequencing machines, including Illumina, fail to reach this target to some extent. Coverage bias is a deviation from the uniform distribution of reads across the genome. Similarly, error bias is a deviation from the expectation of uniform mismatch, insertion, and deletion rates in reads across the genome [12]. Various reasons have been considered to be the primary source of errors, coverage variability, and biases in Illumina sequencing machines such as the secondary structure, the folding effects of inverted repeats, or phenomena like phasing, crosstalk, fading, or T accumulation [13, 14]. An inverted repeat is a single-stranded sequence of nucleotides followed downstream by its reverse complement. Inverted repeats provoke the formation of a secondary structure and result in a delay in nucleotide elongation on both sides of the strand. A phasing event means that an error in one base can affect the rest of the bases in the read. This can result in a higher number of errors toward the end of the reads. Crosstalk refers to the overlap between the illumination of two channels (C with A or G with T). T accumulation occurs because the fluorophores used for thymine are not always appropriately washed after each iteration. Finally, the fading event is due to the low intensity of the fluorescent signal in each cycle. In a recent comprehensive study [15], errors in Illumina sequencing data generated from multiple platforms (GA, HiSeq, and MiSeq) have been thoroughly investigated, and different types of biases are reported which are summarized in the following paragraphs.

The dominant error type is the substitution error, and between the three platforms, GA has the highest substitution rate while HiSeq has the lowest rate. In a paired-end reads library, bases in the second read are roughly twice as likely to be erroneous in comparison with the bases from the first read. The proportion of errors associated with the four different types of original nucleotides were also investigated. Accordingly, T has the highest substitution rate in GA and HiSeq. In MiSeq, the error rates for the four nucleotides in the first read are comparable, but in the second read, a higher rate of error when the original base needs to be A or T was observed. From the opposite angle, a bias toward G is recorded to be

[7] There are two formats to encode these quality scores to ASCII characters, with different offsets: 33 or 64. The older Illumina software uses offset 64, and the quality scores start from "@", which denotes Q0. However, in the more recent Illumina software, the offset 33 is used and the quality scores start from "!" (Q0).

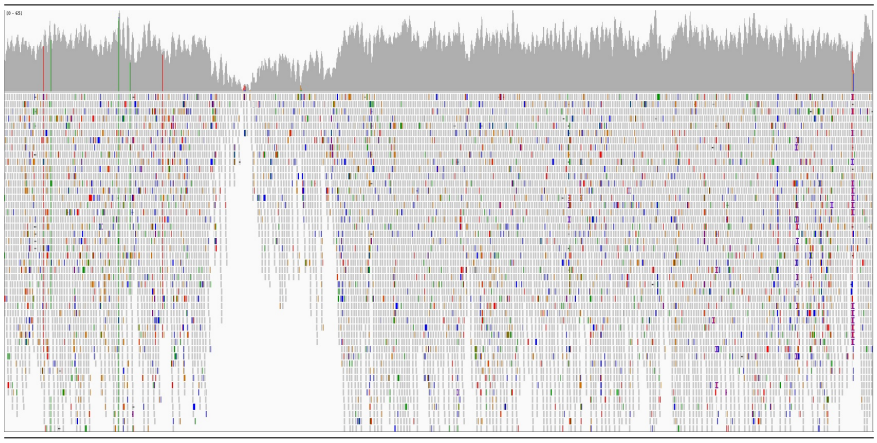
introduced falsely in HiSeq datasets.

Errors have been analyzed based on the prior motifs (3-mers). For the substitution errors, among 64 possible motifs, “CGG” and “GGG” are reported as the most frequent ones in the three platforms. For example, in HiSeq datasets, on average 9.5% and 10.0% of all substitution errors in R1 and R2 are preceded by “GGG”. In general, the most frequent motifs in all platforms end in “GG”. On the other hand, the top two recorded motifs associated with the deletion errors are “AAA” and “TTT” and for the insertions, “CCC” is also recorded frequently.

It is also interesting to see how the quality score reflects the sequencing errors. Based on this study, a quality score can characterize the majority of substitution errors. However, errors in the second read show a higher correlation with their quality score (69% of the R1 substitution and 86% of the R2 substitution have a quality score below 20). The quality score for the indels is meaningless where only 19% of the R1 and 35% R2 indel errors have a quality score below 20.

The coverage is defined as the number of times that each base is expected to be sequenced on average. Another challenging bias with the Illumina sequencing data is that the coverage is not uniform (see Fig. 1.5). This means that either some regions of the genome are not covered by any reads or poorly covered whereas some regions are covered more than average. The low coverage regions are more problematic because they can result in missing SNPs or a more fragmented assembly.

Figure 1.5 This figure shows the nonuniform distribution of coverage along the reference genome. Even though the average coverage is 30, some regions are not covered enough.



1.5 *De novo* assembly

Because of the limitation in the technology, there is no sequencing machine that can identify the entire DNA sequence in one run. Therefore, the complementary phase of DNA sequencing is DNA assembly in which different segments of the initial sequence are put together to reconstruct the original sequence. The complexity of solving such a problem is comparable with the complexity of solving a jigsaw puzzle without any map. Theoretically, genome assembly is defined as an act of putting short DNA sequences, which are called reads, together to reconstruct the original large and complex DNA sequence. It can be done in two ways: reference based, and *de novo*. In the first approach, reads are aligned to the known reference genome from the same or evolutionary closely related species. In the *de novo* approach, the sequence is built from scratch without the aid of the reference data. The reference-based approach is easier and even with low coverage data results in less fragmented contigs. However, the result is biased to the old reference, and it may not reflect all structural variants in the new genome. Therefor, in the absence of an accurate reference genome, or with a highly rearranged genome, the *de novo* approach is often used.

In general, de novo assemblers of NGS-reads perform in three steps: contig assembly, scaffolding, and gap filling. In the first step, the longest possible consensus sequences are obtained using either implicit or explicit multiple sequence alignment between reads. These sequences do not contain any gaps and are called contigs. In the second step, based on the (pair-end/mate-pair) reads information and the insert size, contigs are linearly ordered, which results in a set of scaffolds. Each scaffold is a series of contigs that are connected to each other either directly or with a gap in between. The sizes of these gaps are estimated based on the insert size of the (pair-end/mate-pair) reads. In the third step, the gaps between contigs in each scaffold are carefully filled. This can be done by aligning the reads to the edges of contigs to find the potential short overlaps between them. If such overlap is not found, the gap is filled with a series of N bases. In this section, we mainly focus on the first step, which explains different approaches for the contig assembly and existing challenges in this field.

There are two main algorithmic strategies to assemble the genome *de novo*: overlap-layout-consensus (OLC) or by using the de Bruijn graph [16]. Assemblers in the first category perform in three steps: First, they find the existing overlaps between reads to build an initial overlap graph (O). In this graph, each read represents a node and there is a directed edge between two nodes if they share sufficient overlap. In the second step (layout), they simplify the overlap graph to a nonredundant one, i.e., edges that can be inferred from others are removed. Then, they extract all the contigs, nonbranching nodes followed by each other (L). In the third step, the reads are aligned back to the contigs to find the most likely nucleotide

sequence, which is called the consensus sequence (C). The most time-consuming step is building the overlap graph. All reads need to be verified to see if they share sufficient overlap. However, to avoid the quadratic number of calculations for all binary combinations of reads, hash tables or prefix trees are used to suggest those pairs of reads that are more likely to have such overlap. In general, because of the high runtime demand for the sequence alignment in high-throughput data, this strategy is often used for low throughput long read technologies like Celera [17] assembler for 454, Sanger and Canu [18] assembler for PacBio data.

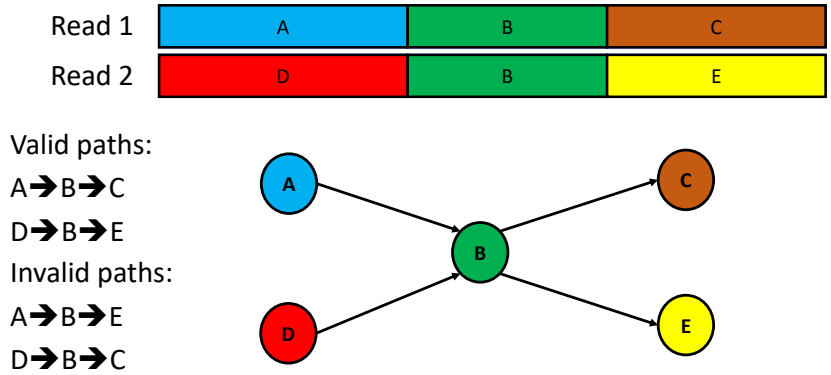
On the other hand, a vast majority of assemblers are based on the de Bruijn graph, such as Velvet [19], ALLPATHS-LG [20], IDBA [21], MaSuRCA [22], SPAdes [23], and SOAPdenovo [24]. They are often used to assemble the high-throughput Illumina short reads. The name de Bruijn graph comes from the name of a Dutch mathematician, Nicolaas de Bruijn, who introduced this graph. He employed the de Bruijn graph to solve the superstring problem [25]. The subject of this problem is to find the shortest circular superstring that contains all possible substrings of length k (k-mers). Later, his proposed graph was used in bioinformatics; in that context, k-mers refer to all possible substrings (of length k) from a read obtained through DNA sequencing. The de Bruijn graph refers to a directed graph where nodes correspond to k-mers and arcs represent an overlap of $k-1$ nucleotides between nodes. By another definition, two k-mers are connected in this graph only if they overlap in one of the input reads. In this way, the graph has fewer spurious arcs [26].

In the compact version of the de Bruijn graph, chains of nonbranching nodes merge into a single node which is called *unitig*. The marginal information contained by a single node is the last base of its corresponding k-mer. Therefore, when consecutive nodes are merged, the series of those last bases represent the sequence of the final merged node. To make sure that the overlaps between opposite strands of reads are also taken into account, the graph is built and maintained symmetrically such that for every node A there is a twin node A'. The sequence represented by A' is the reverse complement of A and if there is an arc between node A and node B, there is also an arc from B' to A' in the opposite direction. Any change to any node or arc implicitly applies to its twin as well. When the de Bruijn graph is built from sequencing data, all k-mers and their overlaps are present in the input data. Each k-mer occurs in a unique node of the graph, and the original sequence can be found as some path through the graph. The de Bruijn graph can thus be seen as a compact multiple sequence alignment representation of the input reads.

The benefit of the de Bruijn graph over the overlap graph is its simplicity and speed, but also it is not dedicated just for reads, a mix of short reads, long reads, or even preassembled contigs with different lengths can be used to build the graph. On the other hand, the main drawback of the de Bruijn graph is the loss of informa-

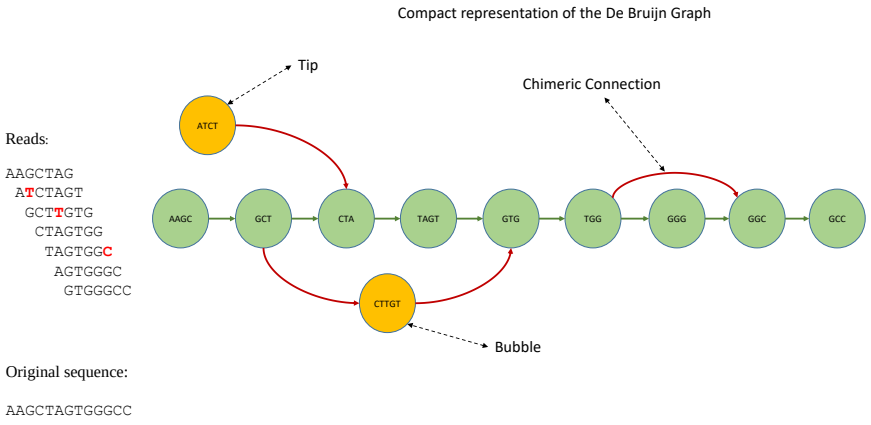
tion caused by decomposing the initial sequencing data (such as reads) into paths of k-mers. In addition, although all paths of size two in the de Bruijn graph represent an overlap between two k-mers in the sequencing data, not all the paths in the graph that span three or more nodes correspond to a valid sequence of k-mers in the data (see Fig.1.6).

Figure 1.6 Each read consists of three k-mers. Except for the middle k-mer in two reads which is identical, the other k-mers are unique and hence are shown in different colors. Each path in the de Bruijn graph that consists of two nodes represents an overlap between two k-mers in the reads and vice versa (e.g., AB, BC, DB and BE). However, the longer paths that span three or more nodes do not necessarily represent a valid overlap of k-mers in the reads (i.e., there is a connection between D to C because D is connected to B and B is connected to C). Therefore, all the overlaps in the data are present through some paths in the graph, but not all the paths in the graph show valid sequencing data.



Errors in the sequencing data appear as spurious artifacts in a de Bruijn graph. Fig. 1.7 displays a compact representation of a de Bruijn graph which is constructed from a set of reads that contain some errors. For simplicity, the graph is single-stranded. In general, three different types of topological structures can be created in the graph due to the sequencing errors in the data: Tips, Bubbles, and Chimeric connections [27]. Tips are created due to the presence of errors in less than k bp from the start (or end) of a read. In Fig. 1.7, an error in the second base of the second read creates a tip in the graph. Bubbles, on the other hand, can be created either by an error in the middle of a read with a margin of k bp, or if two tips share a spurious overlap. In Fig. 1.7, an error in the fourth base of the third read creates a bubble in the graph. Chimeric connections appear in the graph if a sequencing error in one context is a valid sequencing datum in another context. For example, in Fig. 1.7, GGC is an erroneous k-mer in the fifth read but a valid

Figure 1.7 This figure shows a single-stranded, compact representation of the de Bruijn graph constructed from a set of reads. Sequencing errors in the reads create spurious artifacts in the graph which are categorized as Tips, Bubbles, and Chimeric connections.



k-mer in the sixth one. Unlike tips and bubbles that are spurious nodes, chimeric connections are spurious links in the graph and are more difficult to be detected. A link that is not supported by many reads can be a chimeric connection. However, attention should be paid to avoid labeling a true link that connects two nodes in a low coverage region as a chimeric connection.

To clean the de Bruijn graph from the erroneous nodes, tips and bubbles need to be detected and removed. This cleaning step simplifies the graph because more consecutive nodes merge together. Tips can be easily identified in the graph, nodes that do not have any (out-going /in-going) arc are tips. For a given node, we can determine if the node is a tip or not in $O(1)$. Therefore, the complexity of enumerating all the tips in the graph is in the order of $O(n)$, where n is the number of nodes in the graph. However, not all the tips are erroneous, those that are supported by many reads, or those that do not have any alternative path, could be the start or end of a true path in the graph, and hence should not be removed. For example, node GCC in Fig. 1.7 is a tip because it does not have any out-going arc; however, there is not an alternative path for it, and apparently, it is the end of a true path. On the other hand, ATCT is an erroneous tip because: first it is supported by only one read; second, there is an alternative path for it (AGC $\rightarrow$ GCT) with higher coverage (i.e., AGC occurs in one read, and GCT occurs in two reads. Therefore, on average, each k-mer in this node is covered by 1.5 reads. On the other hand, ATC and TCT occur in only one read and hence the average coverage is 1).

Finding bubbles in the graph is computationally more expensive than finding tips. To enumerate all the bubbles in a given graph, all nodes that have more than one out-going arc should be examined to see if two separate paths which start from that node cross at some point. The Depth-First-Search (DFS)[8] algorithm can be applied to search for any mutual node in two separate paths. The complexity of the DFS in the de Bruijn graph is in the order of $O(4^d)$, where d is the maximum depth of traversing the graph (i.e., due to limited resources, such as memory or CPU cycles, one needs to keep track of the set of all previously explored nodes in the graph to stop the search when the number of visited nodes exceeds a predefined limit). Again, not all the bubbles are erroneous nodes. Repeats in the data can create true bubbles in the graph that need to be preserved. Multiple criteria can be applied to check if one of the two parallel paths is erroneous. For example, their corresponding sequence should be very similar, they should have almost the same length, and one of them should have a low coverage. If all constraints are satisfied, then the one with the lower coverage more likely represents a sequencing error and can be removed.

The value of k that is used to build the de Bruijn graph directly impacts its topological structure. The proper value of k depends on different factors such as the genome size, repetitiveness of the genome, average length of the reads, or the coverage. With an extremely small value of k, the graph becomes too complex and dense with too many spurious chimeric connections. However, it guarantees to preserve all existing overlaps between the reads even within the low coverage regions. On the other hand, with an extremely large value of k, the expected number of chimeric connections decreases but also the resulting graph becomes too sparse and disconnected, particularly within the poorly covered regions. Therefore, some assemblers like SPAdes and IDBA do not use a single value of k; instead, they use multiple values of k starting from a smaller size like 21. Then, they iteratively increase k to a larger value like 51 to get rid of chimeric connections as much as possible.

There are two approaches to represent a de Bruijn graph: Hamiltonian and Eulerian. In the Hamiltonian approach, which was explained earlier, the k-mers are the nodes, whereas in the Eulerian approach they are the edges. In the Hamiltonian approach, the sequences are assembled by finding a Hamiltonian path in

[8]Depth-first search (DFS) and breadth-first search (BFS) are two commonly used algorithms to traverse or search a graph data structure. DFS starts from an arbitrary or a given node in a graph (or the root node in the tree). It explores recursively the leftmost child (the order of exploring nodes can vary) as far as possible along each branch. When it explores a leaf (a node without a child), or when there is no other unexplored child, it backtracks to its parent node. The search continues until no other unexplored nodes remain in the graph. BFS explores all the children of a node, before moving to the node at the next depth level. The time complexity of both algorithms is similar to $O(b^m)$ where b is the branching factor and m is the maximum depth. However, the advantage of DFS over BFS is that it requires less memory. DFS needs to store only a single path from the root to the leaf and the remaining unexplored sibling nodes for each node in the path $O(bm)$. However, BFS has to keep the whole search space in memory $O(b^m)$ [28].

the graph that traverses all the nodes but only once. This problem is in the set of NP-complete problems[9] and hence there is not yet a polynomial solution for it. In contrast, the time required to find an Eulerian path in the graph is roughly proportional to the number of edges in the graph if such a path exists. A directed graph has an Eulerian path if and only if all the nodes in the graph have the same in-degree and out-degree. In theory, a de Bruijn graph that is built from a set of k-mers comes from a circular sequence where each k-mer that appears exactly once in that sequence has an Eulerian path. In practice, not all the k-mers in a genome occur once, those that are located in the repetitive regions occur more than once. With a known multiplicity (M) of a k-mer, one can solve this problem by establishing M arcs, instead of one, to represent that k-mer. However, that information is not given a priori and predicting the true multiplicity of a k-mer is another challenging problem. In addition, due to the coverage gap, some k-mers of the original sequence are missing, which may lead to a disconnectivity in the graph. Besides, because of the sequencing errors in the data, some false k-mers are present in the graph which are not actually present in the initial sequence. Therefore, in practice, finding an Eulerian path could be as difficult as finding a Hamiltonian cycle in the de Bruijn graph, although there is a common belief that for the larger scale genomes, the Eulerian graphs like SPAdes, EULER, or MaSuRCA perform better than a Hamiltonian de Bruijn graph like Velvet and SOAPdenovo [16, 30]

In general, there are three main challenges for the *de novo* assembly: repeats in the genome, nonuniform distribution of reads along the genome, which sometimes results in coverage gaps, and errors in the sequencing data. Sometimes, there is a mix of these scenarios, which makes the problem even more difficult to deal with. For example, the presence of sequencing errors in a low coverage region of the genome. At one extreme, consider a genomic location that is covered by only two

[9] The Nondeterministic Polynomial (NP) is a set of computational decision problems that for a given instance of size n and an answer A, the number of steps needed to verify A is in the polynomial order of n. For example, finding the smallest element in a given list $p_1, p_2, \ldots, p_n$ is an NP problem. Because if someone claims that p_i is the smallest element, we can verify this statement by comparing p_i to all other elements in the list in n steps. The Hamiltonian cycle problem is another example of an NP problem where a connected graph G with n nodes is given, the goal is to find a cycle that traverses every node once and finally returns to the starting node. In this case also, if someone presents an answer to the problem, for instance, a list of nodes $C = v_1, v_2, \ldots, v_i, v_{i+1}, \ldots, v_n, v_1$, we can check in n steps, if C contains all the nodes in the graph and if there is always a link between v_i and v_{i+1} also v_n and v_1. Therefore, the similarity between these two examples of problems is that the correctness of a given answer for both problems can be verified in polynomial time, and hence, they are both *NP* problems. However, the difference is, for the first problem, we can propose an algorithm to find the smallest element of a given list in polynomial time (in $O(n)$), but for the second, we cannot yet propose a polynomial time algorithm to find a Hamiltonian cycle. However, it has not been proved that finding such an algorithm is impossible. These types of problems where we do not know yet if they are solvable in polynomial time, but the correctness of a given answer can be verified in polynomial time, are called *NP-complete*. To solve these types of problems, approximation, randomized, or heuristic approaches can be used. These methods do not return the exact or optimal solution because finding the optimal solution takes a lot of time. Instead, they return a reasonably good solution(s), but in a shorter time [29].

reads with an overlap of size k. In case that two associated k-mers are identical, most likely both of them are correct, the assembler can establish a connection between two sides; otherwise, the assembler has no way to recover this connection. To cope with these three challenges, using longer size reads can help with the first one. Increasing the coverage depth can reduce coverage gaps, and doing a careful error correction can alleviate the last problem while an aggressive error correction can make the situation even worse.

1.6 Error correction

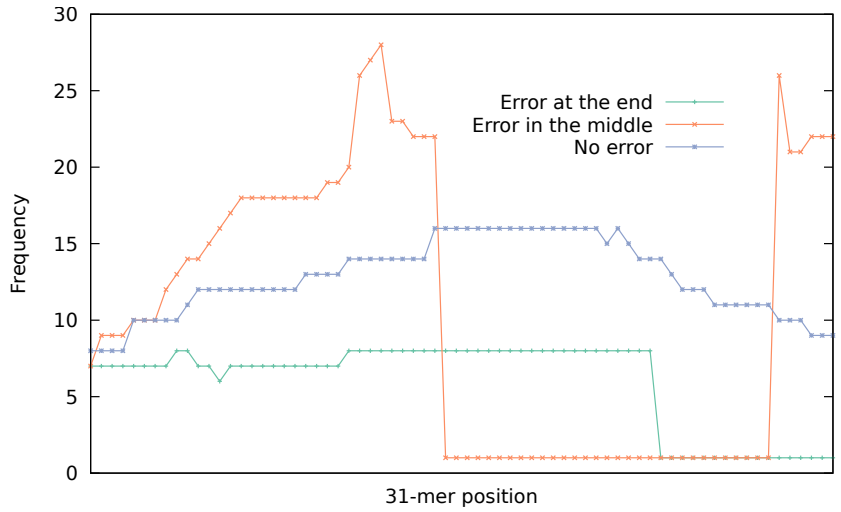
Although Illumina sequencing data have a lower rate of errors compared with the other platforms, due to the high throughput nature of the data there are many errors in the sequenced data. Identifying sequencing errors from true biological variants in the absence of a reference genome is a challenging task in bioinformatics.

Many tools have been introduced in recent years to identify and correct sequencing errors in reads: Quake [31], ACE [32], BFC [33], BLESS 2 [34], Blue [35], Fiona [36], Karect [37], Pollux [38], QuorUM [39], RECKONER [40], and SGA-EC [41]. In general, the underlying algorithms of these EC tools can be categorized into two main groups: based on k-mer spectrum (e.g., Quake, ACE) or based on multiple sequence alignment (MSA) (e.g., Karect, Fiona).

If an error occurs in a read, k-mers that contain that base become erroneous. A simple way to guess if a k-mer is erroneous is by looking at the frequency of that k-mer in the whole dataset. Generally speaking, the frequency of each true k-mer in a dataset correlates with the depth of coverage of that dataset. While k-mers from repeat-rich regions are more frequent, low-frequency k-mers are more likely to be erroneous. Figure 1.8 shows the frequency of all 31-mers in a typical dataset whose depth of coverage is 33 for three different reads. The orange line shows the frequency of 31-mers where an error occurs in the middle of a read. In this case, 31 consecutive 31-mers become erroneous and consequently the frequency drops suddenly in the middle. Because there are some true 31-mers at the end of the read, the frequency rises again at the end. The green line shows the frequency of a read where an error is at the end of the read. In addition, the blue line shows the frequency of k-mers in a read without any error. Although the blue line shows a small fluctuation of coverage for different k-mers, there is not such a sudden drop or jump.

A k-mer histogram of a typical dataset shows a mixture of two distributions– the coverage of erroneous k-mers on the left side, and the coverage of true k-mers on the right side. For example, Fig. 1.9 shows the 31-mer coverage histogram of the above-mentioned dataset. The k-mer spectrum-based tools operate on the level of individual k-mers. First, all the k-mers in the datasets and their frequencies are determined. Second, based on the frequency histogram of all the k-mers

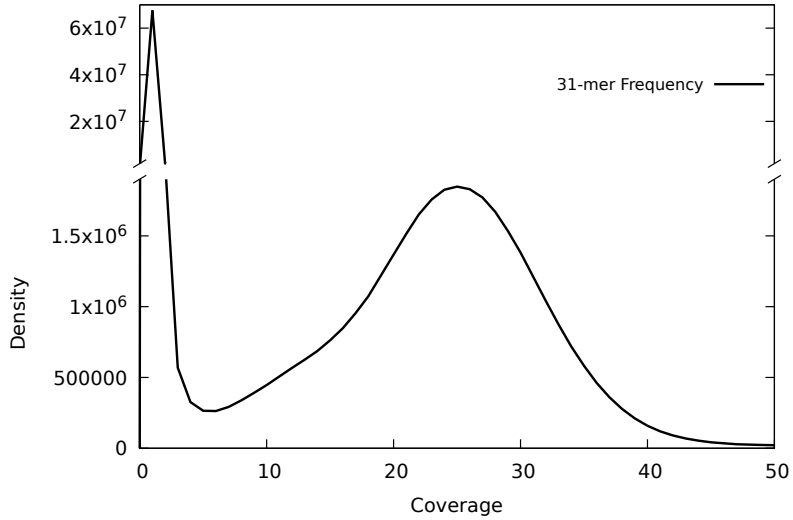
Figure 1.8 This figure shows the frequency of 31-mers in a dataset for three different reads. For each 31-mer, the frequency shows the number of reads in the dataset that contain 31-mer. The plot shows a sudden drop of the frequency of 31-mers at the end of the green line and in the middle of the orange line, which implies the presence of an error at the end and in the middle of corresponding reads. The blue line shows the 31-mer frequencies for an error-free read.



(Fig. 1.9) a threshold is determined to separate the true k-mers from the erroneous ones. Third, the k-mers that are labeled as erroneous are transformed to the most similar true k-mer using a minimum edit distance approach. Determination of the threshold value is the most challenging task in this strategy. Setting the threshold to a lower value causes more erroneous k-mers not to be detected, while setting it to a higher value may classify more true k-mers as erroneous.

Assuming that two distributions are Poisson, one approach to determine the threshold value is finding the intersection point of the two distributions. To understand which data point belongs to which distribution and to estimate the parameter λ of each distribution, the Expectation Maximization (EM) algorithm can be used. EM is an iterative algorithm that can be used to estimate unknown parameters in statistical models of data. For example, knowing that the data are a mixture of two Poisson distributions, all we have to find is λ_e (the mean for the erroneous k-mers) and λ_c (the mean for the valid k-mers). If we already knew which data points belong to which distribution, we could estimate the λ by computing the mean. Because that is not given, it starts with initial random assignments for the λ_e and λ_c parameters. Using these random parameters, it computes the posterior, the probability that shows to what extent each data point belongs to each distribution. Then,

Figure 1.9 The histogram shows a mixture of two distributions–erroneous 31-mers on the left and true 31-mers on the right side. This figure also shows that although the error rate in Illumina data is low, due to the high throughput nature of the data, there are many erroneous k-mers.

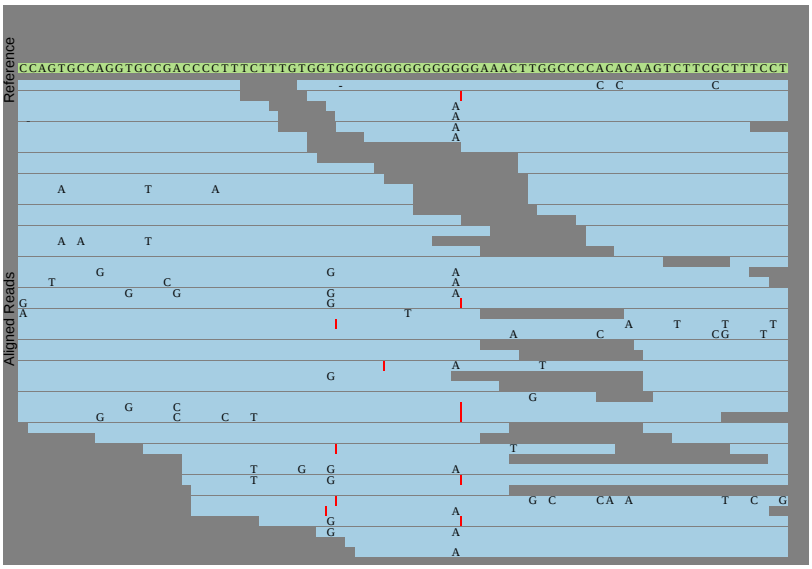


it updates the previous estimations for λ_e and λ_c based on the weighted average of data points where the weights are the posterior values. The new estimated parameters are used to create a better membership assignment for each data point. The process continues until the algorithm converges to a fixed point.

The main drawback of the k-spectrum approach is dealing with low-coverage regions or coverage gaps in the dataset. Because the coverage is not uniform, there are some regions in the data that are not covered enough by the reads and hence true k-mers in these regions look erroneous because their frequency is less than the threshold. An EC tool that corrects reads at the level of k-mers cannot take into account the context in which that k-mer occurs (e.g., the frequency of other k-mers in that read or other adjacent reads). The actual coverage in some region of the dataset can also drop if reads in that region contain too many errors (see Fig. 1.10). For example, in the vicinity of homopolymeric regions, a notable increase in the error rate has been reported. In this case, although there might be enough reads to cover that region, there are not enough reads that overlap with a sufficient length without any errors. To alleviate this problem, Blue which is a k-spectrum-based EC tool, proposed to identify an erroneous k-mer in the context of the read in which it occurs. Blue showed that employing this technique leads to a more accurate error correction compared with its former EC tools. This approach is somehow similar to the method which was introduced earlier in CRAC [42] to

identify genetic variations from sequencing errors in RNA sequencing data.

Figure 1.10 The picture shows a snapshot of aligned reads to a known reference. Indicated bases with black color are either sequencing errors or variants. Although there are enough reads to cover this region, too many errors in these reads reduces the actual coverage.



In contrast, MSA-based tools operate on the level of reads. First, reads that are assumed to represent overlapping genomic regions are clustered together, and a consensus is obtained through multiple alignments of those reads. Second, reads are corrected according to the consensus alignment. The main drawback of the EC tools in this group is that they require a substantial amount of memory and runtime to complete.

While all EC tools can be classified into either of these two groups, there is still a great diversity in the implementation details, heuristics, and data structures they use (bloom filter, hash table, suffix tree,...). For example, BLESS2 and BFC rely on Bloom filters. Racer [43] and Blue use the hash table. Fiona, SGA-EC, and ACE use different types of the suffix tree, suffix array, or k-mer trie, while Karect uses the partially ordered graph as the primary data structure.

1.7 Assessment of the quality of genome assembly

Different assemblers use various algorithms and techniques to assemble the reads resulting in different contigs and scaffolds. Doing the error correction of reads before the assembly can also affect the result, here the goal is to improve the quality of assembly. However, the assessment of the quality of one (or more) set of contigs obtained from an assembly process is challenging especially for unsequenced species. Using available reference genomes to evaluate the quality of a new assembly of finished genomes is helpful. However, there is not a perfect common reference genome that can be used as a gold standard because different strains of the same species or different individuals can vary at both the nucleotide and the structural level due to the rearrangement events. Despite this, the reference-based evaluation of assemblies of species is often performed to capture the performance of different assemblers or EC tools.

One way of evaluating the quality of an assembly is by looking at the contiguity of contigs. In this way, having fewer but larger contigs is more privileged. In this regard, given a set of assembled contigs that are sorted in decreasing order of their length, the N50 metric is defined as the sequence length of the shortest contig at 50% of the total length of the assembled contigs. This metric is useful especially in the absence of a close reference genome. However, a careless assembler may get a higher N50 score by concatenating two contigs which are not actually beside each other in the genome. When the reference genome is available, we can first break the contigs into smaller maximal segments that are continuously aligned to the reference genome. Then, we can obtain the N50 of these segments, and this new metric is called NA50. There is another issue with N50, assume two assemblers, the first one only outputs the larger contigs, preferably the more confident ones, and the second one reports all the contigs. The first assembler obtains a higher N50 score even though it covers a smaller fraction of the genome because the N50 takes into account 50% of the total length of the assembled contigs. To incorporate the genome coverage we can compute N50 in respect to the 50% of the genome size and not the sum of the contigs' lengths. In this way, we can use NG50 instead of N50 and NGA50 instead of NA50. Therefore, the NGA50 is the characteristic length of assembled contigs that can be contiguously aligned to the reference genome and it is arguably the most commonly used metric to evaluate assembly quality. Fig. 1.11 shows a toy example of computing N50, NA50, NG50 and NGA50 for two different sets of contigs which are obtained from the same dataset by two different assemblers. Here, even though the N50 of the first contig set is higher, the second assembler is more accurate (contigs can align to the reference genome with less fragmentation) and covers a greater fraction of the genome. Therefore, the second assembler has a higher NGA50. Quast [44] is a leading assembly comparison software that evaluates assemblies both with and without a

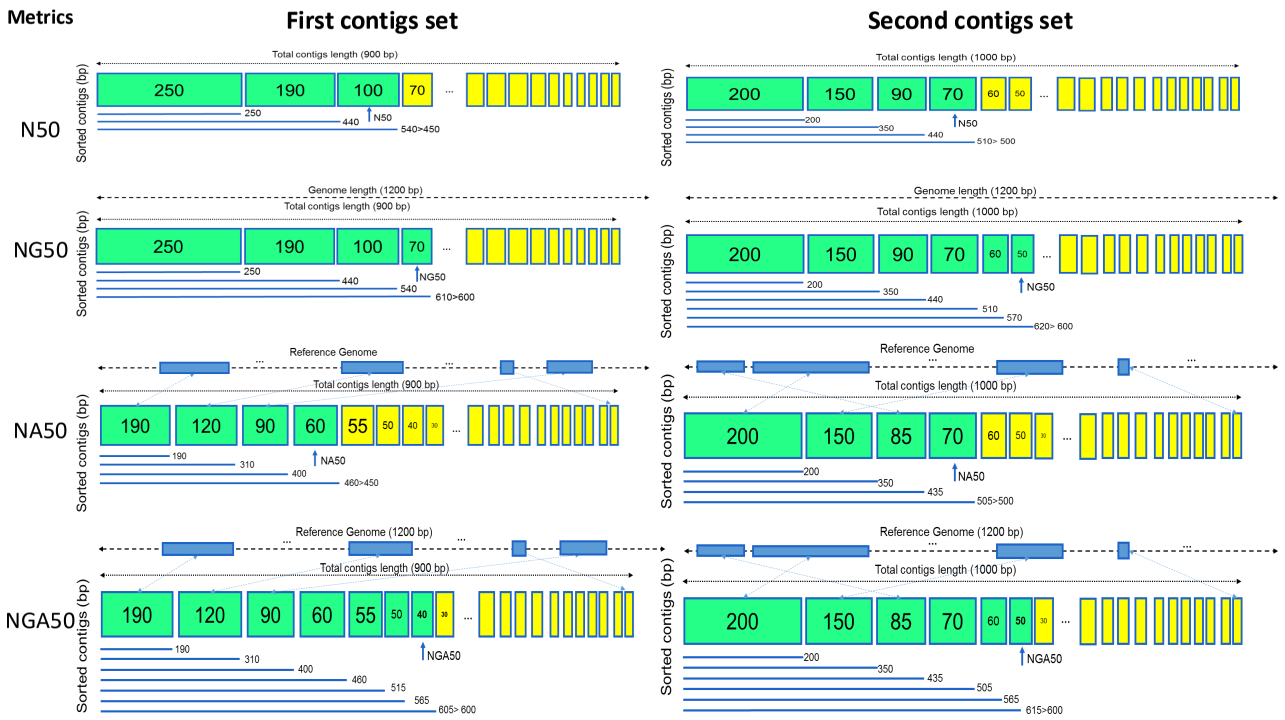


Figure 1.11: This figure shows two examples of contigs sets produced from the same dataset. Different metrics (N50, NG50, NA50 and NGA50) are computed for each set.

Table 1.1 An example of aligning three sequences. Mismatches are shown in different colors in respect to the first sequence, and gaps are indicated by a dash sign.

A	C	C	C	T	A	-	T	G
A	C	G	C	T	-	C	-	-
-	C	G	C	T	A	C	-	G

reference genome. It produces comprehensive reports including summary tables, plots and different metrics such as N50 and NGA50.

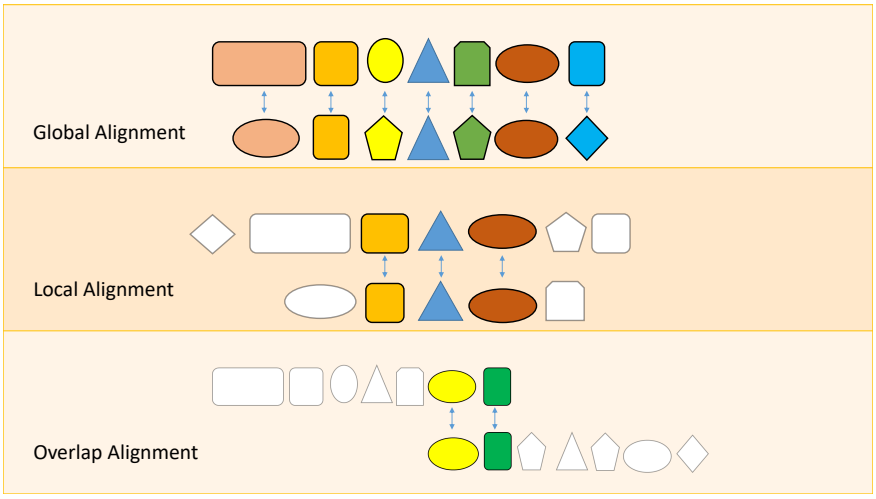
1.8 Sequence alignment

The sequence alignment in bioinformatics defines a way of arranging sequences of DNA, RNA or protein to distinguish similar regions between those sequences. These similarities can be a sign of evolutionary, functional or structural relationships. An alignment of $n > 1$ DNA sequences (not necessarily with the same length) can be represented by an n-row matrix such that i-row contains the characters of the ith sequence. Matches and mismatches are in the same columns and an indel (insertion or deletion) is indicated by a dash sign (-) at the side of the deleted character. Consequently, the characters in each row appear in the same order of the corresponding sequence, but not necessarily adjacently. Furthermore, no column of the alignment matrix contains dashes in all rows (see Table 1.1). Defining the best alignment can vary in different contexts and types of data; however, alignments with a fewer mismatches or indels are often more desired. Therefore, from the mathematical point of view, finding the best alignment can be defined as the maximization problem given a set of constraints to penalize a mismatch or an indel with negative scores and rewarding a match with a positive score.

Typically, computational methods to do the sequence alignment consist of three distinct approaches: global, overlap, and local alignment. In the global, the goal is to align the entire length of all sequences by penalizing the unaligned part by a negative gap penalty score. In overlap or semi-global alignment, the start gap in one sequence and one end gap in the other may be ignored. In contrast, in the local alignment, the goal is to align arbitrary-length segments of the sequences with no penalization for the end gaps. Local alignment is useful for finding similar motifs or conserved regions between divergent sequences. On the other hand, the global alignment is usually used for the alignment of similar sequences with roughly the same size. Fig 1.12 shows a schematic representation of aligning two sequences in these three ways.

From a different angle and based on the number of input sequences, sequence alignment methods can be categorized into two groups: Pairwise Sequence Align-

Figure 1.12 A schematic representation of three types of alignments: Global, Local and Overlap.



ment (PSA) and MSA. The PSA method is used to find the best-matching alignment (global or local) between two sequences at a time. The popular approach to get the best pairwise alignment is still the dynamic programming solution that was introduced by Saul B. Needleman and Christian D. Wunsch in 1970 [45], which can be applied to DNA, RNA, or protein sequences. The Needleman-Wunsch algorithm is sometimes called the optimal matching solution because it guarantees to find the best alignment under the given constraints (e.g., match score, mismatch score, and gap penalty). In this way, an optimal sequence alignment of two given sequences is recursively computed using the optimal alignments of smaller subsequences (i.e., the dynamic programming solves the original problem by dividing it into smaller independent subproblems). A common extension to the standard algorithm which has a linear gap cost is called the affine gap model. In this version, there are two different types of gap penalties: one for opening a gap and one for extending a current gap. Usually, the gap extension is less penalized than the gap opening.

In contrast, in MSA methods the aim is to find the best alignment between three or more sequences. The purpose of this alignment is often to find conserved regions across a group of sequences hypothesized to be evolutionarily related. From the theoretical point of view, the dynamic programming algorithm can be applied to find an optimal alignment between more than two sequences as well, however, computationally, this problem is not tractable and finding an optimal solution is known to be an NP-complete problem [46]. Therefore, over 100 methods have been proposed to deal with the complexity of this problem by an alterna-

tive approximate algorithm in the past decades [47] such as CLUSTAL W [48], MUSCLE [49], or T-Coffee [50]. The most common approach in these aligners is building a guide tree based on the pairwise alignment between all sequences. Iteratively, the most similar sequences in the reference tree are aligned and then replaced by a consensus sequence. The main drawback of this approach is that any errors made in any of the earlier steps can propagate through to the final alignment.

1.8.1 Read alignment

Read alignment is one of the fundamental problems in bioinformatics and it is a prerequisite step in several genome analysis pipelines, such as genetic variant calling, reference-based genome assembly, or personalized medicine [51]. It is estimated that over 60 short-read mappers are available, which mostly were published after 2008 [52]. Read alignment can be formulated as finding all substrings m of a set of reference sequences R for a given query sequence set Q that respect certain constraints. The constraints can depend on the specific type of data.

Dealing with short Illumina data, most of the current aligner tools like BWA [53] or Bowtie [54] are designed to align reads to a linear reference genome. A common strategy in these aligners is a seed-and-extend model. It is assumed that an accurate alignment between a read and the reference genome should contain an exact match [10] of a size at least k. First, seeds such as maximal exact matches between a read and the reference sequence are identified. Those seeds indicate candidate positions in the reference genome from which the read originated. In the second step, each seed is extended to the left and right until a full read alignment is obtained while maximizing a well-defined objective function like similarity score as used in the Needleman-Wunsch algorithm. Certain constraints, such as maximum number of mismatches can be considered in the extension phase to have an early stop mechanism which reduces the search space. Finally, those alignments whose similarity score is higher than some predefined threshold are reported.

The challenging task for the short read aligner is the seeding or pattern matching part. In this regard, they need a fast and memory-efficient data structure that is capable of returning all occurrences of a given query in the reference genome. A simple strategy could be using a k-mer index table. For each k-mer, it stores the positions in the reference genome that it occurs. However, many of these k-mers differ in only one nucleotide and allocating a separate entry for each k-mer is not memory efficient because it cannot take advantage of the overlaps between k-mers and hence it stores redundant information. To reduce the memory consumption and accelerate the search, Minimap [55] generates a hash key for each k-mer and then stores it in the table. The size of k can contribute to the efficiency of the

[10] The idea can be extended for maximal inexact matches where a few mismatches are allowed, which is more applicable to the data with a higher rate of errors.

aligner using this approach. For example, the value of k is a trade-off between accuracy and speed. With higher values, it can produce the results faster, while smaller values yield more accurate results. However, using a single k may not be sufficient to search every pattern. Assume a read that has many errors and does not have a valid k-mer that can be found in the k-mer table. Consequently, this read cannot be mapped to the reference. One can suggest using a full-text k-mer index, i.e., having k-mer index for every value of k, up to the length of the reference. This would enable us to search for any query of any size up to the reference size; however, this requires storing $O(n^2)$ locations in the memory which is not practically affordable.

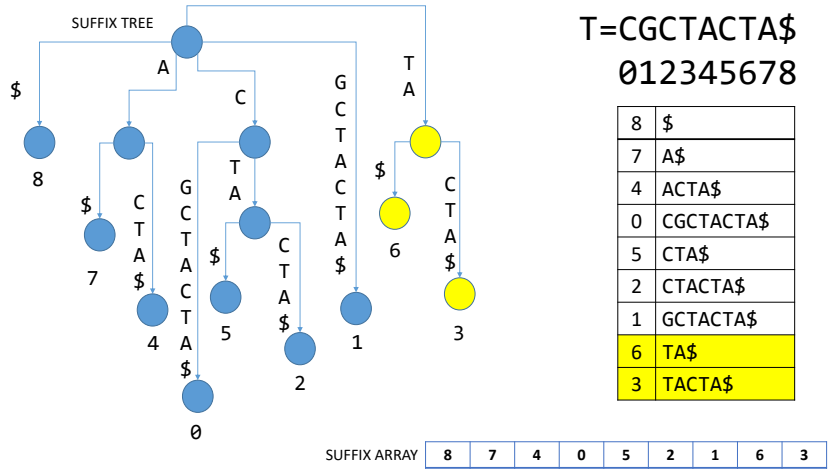
There are other alternative data structures like suffix trie [11], (also known as a prefix tree), suffix tree or suffix array that is more space efficient and can be used for searching queries of any size. These data structures contain all suffixes of a given string T of size n. Because any substring of T is a prefix of some suffix, it enables us to check whether a given pattern p is a substring of T in an efficient way. For example, suffix trie containing all the suffixes of T as their keys. A suffix tree is a suffix trie with two differences. First, parent nodes with only one child node are merged. Second, each leaf (a prefix) is labeled with the index where that suffix starts in the text. Therefore, a suffix tree can tell us not only if p exists in T but also it can point to the positions in T where p occurs. Searching a query p of length m in the tree takes $O(m)$ comparisons [56]. However, a tree needs to store all the suffixes somewhere in the memory. A suffix array is a sorted array of all suffixes of string T [57]. Instead of storing all the suffixes, it only stores the starting position of each suffix in that array and hence it is more space efficient. Because the array is sorted, we can search a given pattern p in this array in $O(m\ log(n))$ [12]. Fig. 1.13 shows an example of a suffix tree and a suffix array which is built from a toy example sequence. Even though the suffix array uses less memory, the searching process is slower. Enhanced suffix array is an improved version of suffix array with additional tables that reproduce the full functionality of suffix trees. Searching a pattern p in an enhanced suffix array is of the order of $O(m)$ [58]. FM index [59] is another popular indexing algorithm which is based on Burrows–Wheeler Transform (BWT). BWT is a reversible method that transforms a given text T to F, wherein F, similar characters tend to appear consequently [60]. This method was initially used for the compression but it also enables us to search for a pattern p in T in linear time, in respect to the length of p and independent of the length of T in $O(m)$. In addition, using the compression feature of this method, the memory usage in the FM index is less than the suffix array. That explains the popularity and the high speed of some aligners like BWA-mem and Bowtie, both

[11] A trie (pronounced try) gets its name from re**trie**val

[12] Searching in a sorted array with binary search requires $log(n)$ probes into the array, and each probe can also take $O(m)$ comparisons between p and the suffix.

of which use FM index for the indexing purpose.

Figure 1.13 This figure shows a suffix tree and the equivalent suffix array for a short DNA sequence. An additional reserved character, a sentinel, which does not occur in the text, shows the end of the sequence. In this example, we have used the $ sign for this purpose.



Similar to the genome assembly problem, there are two challenges for every read aligner: the presence of sequencing errors in the data and repeats in the reference genome. A good read aligner should be able to distinguish between a sequencing error and genetic variation in the dataset [42]. In addition, because of the repeats in the reference genome, there might be multiple locations in the reference genome for a read to align. To tackle these two problems, the quality scores or the paired read information can be used to guide the alignment procedure.

1.8.2 Graph alignment

Although most of the aligners align short reads to a linear reference, for certain applications the reference genome can be given as a de Bruijn graph. For instance, a read assembler can align reads to the assembly graph to resolve the repeats or to do the scaffolding and have a more precise result instead of aligning reads to a list of independent linear contigs. In the absence of such an accurate aligner, SPAdes [23] keeps track of the reads' information in their corresponding nodes and paths during the graph construction and later in graph correction; this may demand a substantial amount of memory and runtime. In addition, in the context of pan-genomics, processing reads with respect to a graph representation of multiple references rather than individual linear sequences is crucial [61]. As another example, in a metagenomics study, reads that are sequenced from different unknown

species can be aligned to a de Bruijn graph that is built from multiple genomes. Such a tool can let us better study closely related species in an integrated manner which may be useful for the subsequent analyses such as rearrangement or structural variant detection studies or phylogenetic inference [62, 63]. Furthermore, in hybrid error correction, long reads are aligned to a graph which is built from short reads [64, 65]. In addition, a graph aligner can be used for the sequencing error correction purpose. This can be done in three steps: first, the graph is built from a set of uncorrected reads, then spurious artifacts in the graph such as tips and bubbles are removed. Finally, uncorrected reads are aligned back to the cleaned graph. Recently, two standalone tools have been proposed to align short Illumina reads to de Bruijn graphs: BGREAT [66] and deBGA [67].

For the alignment of reads to the graph, the same seed-and-extend approach can be used similarly to the linear reference genome. The seeding phase is more straightforward than before because the graph is a compact representation of all the k-mers in the data and each k-mer occurs only once in the graph. However, searching for an optimal path in the graph is computationally more expensive than a linear reference because the search space can grow exponentially in the length of the read. For example, given a seed, a naive approach could be exploring all possible nodes at a branching point in the graph (e.g., depth-first search (DFS) or breadth-first search (BFS)). To avoid searching all the nodes, BGREAT and deBGA have an early stop mechanism. In this approach, exploring a new node stops if the number of mismatches exceeds a certain threshold. This strategy reduces the search space, yet does not guarantee to return an optimal solution. The second challenge arises based on the fact that not all paths in the graph necessarily correspond to a substring of the reference genome, i.e., while every two consecutive nodes represent a substring of the reference genome, paths of three or more connected nodes do not necessarily represent a subsequence of the reference genome. Besides these two challenges, the two previously mentioned difficulties in the *de novo* assembly, repeats and sequencing errors, make the task of graph alignment even more complex. In addition, if the graph is built from sequencing data that may contain errors instead of a reference genome, then each erroneous k-mer in data results in up to k erroneous nodes or creates a chimeric connection in the graph that needs to be identified and ignored in the alignment procedure.

1.9 Research goals and outline

The rest of this dissertation is comprised of papers published within the scope of my Ph.D. research. These publications present a complete, detailed overview of the work performed. A brief outline of each chapter and the connection between them are provided here.

Chapter 2 reviews the current state-of-the-art methods and applications that

are designed and implemented to correct Illumina sequencing errors. The key idea behind these EC tools is that downstream applications such as *de novo* genome assemblers can benefit from reduced error rate in data and hence result in better assembly quality. This chapter evaluates the impact of these tools on the *de novo* genome assembly. As a result, we observe that modern assemblers do not benefit from this precorrection. However, EC tools suffer from poor performance in certain sequence contexts such as regions with low coverage or areas that contain highly repetitive elements or low-complexity subsequences. Reads overlapping such regions are often ill-corrected in an inconsistent manner, leading to breakpoints in the resulting assemblies that are not present in assemblies obtained from uncorrected data.

Chapter 3 provides an answer to the above-raised problem. BrownieCorrector is a targeted EC tool that we introduce in this chapter. The main novelty of BrownieCorrector is that instead of correcting all the reads, it merely focuses on the correction of reads that contain highly repetitive patterns, which are more difficult to handle for both assemblers and error correction tools. BrownieCorrector uses the entire read sequence as well as the paired-end read information to cluster read pairs in homogeneous groups. The Louvain [13]community detection algorithm [68] is applied to an undirected weighted graph whose nodes represent paired-end reads while an edge between two nodes denotes their similarity score. Reads in each cluster are corrected independently such that a consistent correction is achieved for all reads within each cluster. The performance of BrownieCorrector is compared with other error correction tools using six Illumina and two Pacbio datasets from different eukaryotic genomes. Although BrownieCorrector corrects less than 2% of the reads, it leads to the best assembly results in most cases.

BrownieCorrector corrects the reads in a particular cluster in three steps. It first constructs the associated de Bruijn graph, then performs typical graph cleaning procedures such as tip clipping and bubble detection to remove erroneous nodes and arcs that represent erroneous k-mers in the data. Finally, the reads in a cluster are aligned back to the cleaned de Bruijn graph using BrownieAligner, which is discussed next.

Chapter 4 presents BrownieAligner, a graph aligner tool which is designed and implemented to align short Illumina reads to the de Bruijn graph. BrownieAligner

[13] Generally speaking, communities in a network are described as sets of nodes that tend to group. While members of each community are densely connected, links between nodes from different communities are sparse. The Louvain method is a modularity-based community detection algorithm. Modularity degree measures the density of edges inside the communities to edges outside communities and the value is between –1 and 1. Optimizing this value results in the best community detection; however, solving such a problem requires enumerating all possible groups of nodes which is not tractable in practice. Therefore, Louvain uses a heuristic technique which repeatedly performs two steps. First, it finds small communities by locally optimizing modularity. Second, it builds a new network whose nodes are the communities in the first step. By repeating these two steps multiple times, the hierarchy of communities is produced. The time complexity of this method is O(n log n).

uses the seed-and-extend paradigm, which is a typical approach in read-alignment tools. Given a seed, the algorithm greedily explores all branches of the graph until an optimal alignment path is found. To reduce the search space, it computes the upper bounds to the alignment score for each branch and discards the branch if it cannot improve the best solution found so far. Furthermore, by using a two-pass alignment strategy and a higher-order Markov model[14], paths in the de Bruijn graph that do not represent a subsequence in the original reference genome are discarded from the search procedure. BrownieAligner is applied to both synthetic and real datasets. It generally outperforms other state-of-the-art tools in terms of accuracy, while having similar runtime and memory requirements.

Finally, chapter 5 concludes the thesis and outlines some potential future work.

1.10 Publications

1.10.1 Journal papers

- **Heydari, M.**, Miclotte, G., Demeester, P., Van de Peer, Y. & Fostier, J. "Evaluation of the impact of Illumina error correction tools on de novo genome assembly". *BMC Bioinformatics*, 18(1): 374, Aug. 2017.
- **Heydari, M.**, Miclotte, G., Van de Peer, Y. & Fostier, J. "BrownieAligner: accurate alignment of Illumina sequencing data to de Bruijn graphs". *BMC Bioinformatics*, 19(1): 311, Sep. 2018.
- **Heydari, M.**, Miclotte, G., Van de Peer, Y. & Fostier, J. "Illumina error correction near highly repetitive DNA regions improves de novo genome assembly."*BMC Bioinformatics* 19(1):311, June. 2019.
- Miclotte, G., **Heydari, M.**, Demeester, P., Rombauts,S. , Van de Peer, Y. & Fostier, J. " Jabba: hybrid error correction for long sequencing reads". *Algorithms for Molecular Biology*, 11(1), May. 2016.

1.10.2 Conference paper& abstracts & posters

- Miclotte, G., **Heydari, M.**, Demeester, P., Rombauts,S. , Van de Peer, Y. & Fostier, J. " Jabba: hybrid error correction for long sequencing reads".

[14] A Markov Model, named after Andrey Markov, is a stochastic model which is used to model randomly changing systems. It is assumed that the future is independent of the past if you know the present. It means the future only depends on the current state. However, we can build a memory for the states by using a higher-order Markov model. In this way, the future is dependent not only on the current state but also on the last n visited states. Mathematically speaking:
$P(x_i|x_{i-1}, x_{i-2}, \ldots, x_1) = P(x_i|x_{i-1}, \ldots, x_{i-n})$
Additional history can let us predict the next state more accurately. However, by increasing the order, the number of parameters we need to estimate grows exponentially and the predicted value would be more specific and loses generality.

23e Annual International Conference on Intelligent Systems for Molecular Biology, Abstracts., 2015.

- **Heydari, M.**, Miclotte, G., & Fostier, J. "Brownie : correcting second generation sequencing errors using de Bruijn graphs". *ISMB/ECCB, Dublin , poster.*, 2015.
- **Heydari, M.**, Miclotte, G., & Fostier, J. "Brownie : correcting second generation sequencing errors using de Bruijn graphs". *ECCB, THE HAGUE , poster.*, 2016.
- **Heydari, M.**, Miclotte, G., Van de Peer, Y & Fostier, J. "BrownieAligner: Accurate Alignment of Illumina Sequencing Data to de Bruijn Graphs". *ECCB 2018 Athens, poster.*, 2018.

References

[1] James D. Watson and Francis Crick. *Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid*. Nature, 171(4356):737–738, apr 1953. 1-2

[2] Rosalind Franklin and Raymond Gosling. *Molecular Configuration in Sodium Thymonucleate*. Nature, 171(4356):740–741, apr 1953. 1-2

[3] Maurice Wilkins, Alex Stokes, and Herbert Rees Wilson. *Molecular Structure of Nucleic Acids: Molecular Structure of Deoxypentose Nucleic Acids*. Nature, 171(4356):738–740, apr 1953. 1-2

[4] Robert Williams and Matthias Wienroth. *Social and Ethical Aspects of Forensis Genetics*. Forensic Sci Review, 29(2):413–425, 2017. 1-2

[5] Jay Shendure and Hanlee Ji. *Next-generation DNA sequencing*. Nature Biotechnology, 26(10):1135–1145, oct 2008. 1-2

[6] Eric. E. Schadt, Steve Turner, and Andrew. Kasarskis. *A window into third-generation sequencing*. Human Molecular Genetics, 19(R2):R227–R240, sep 2010. 1-2

[7] Lin Liu, Yinhu Li, Siliang Li, Ni Hu, Yimin He, Ray Pong, Danni Lin, Lihua Lu, and Maggie Law. *Comparison of Next-Generation Sequencing Systems*. Journal of Biomedicine and Biotechnology, 2012:1–11, 2012. 1-2

[8] James M. Heather and Benjamin Chain. *The sequence of sequencers: The history of sequencing DNA*. Genomics, 107(1):1–8, jan 2016. 1-2

[9] Mehdi Kchouk, Jean Francois Gibrat, and Mourad Elloumi. *Generations of Sequencing Technologies: From First to Next Generation*. Biology and Medicine, 09(03), 2017. 1-2

[10] Jerzy K. Kulski. *Next-Generation Sequencing — An Overview of the History, Tools, and "Omic" Applications*. In Next Generation Sequencing - Advances, Applications and Challenges. InTech, jan 2016. 1-4

[11] Miten Jain, Sergey Koren, Karen H Miga, Josh Quick, Arthur C Rand, Thomas A Sasani, John R Tyson, Andrew D Beggs, Alexander T Dilthey, Ian T Fiddes, Sunir Malla, Hannah Marriott, Tom Nieto, Justin Grady, Hugh E Olsen, Brent S Pedersen, Arang Rhie, Hollian Richardson, Aaron R Quinlan, Terrance P Snutch, Louise Tee, Benedict Paten, Adam M Phillippy, Jared T Simpson, Nicholas J Loman, and Matthew Loose. *Nanopore sequencing and assembly of a human genome with ultra-long reads*. Nature Biotechnology, 36(4):338–345, jan 2018. 1-5

[12] Michael G. Ross, Carsten Russ, Maura Costello, Andrew Hollinger, Niall J. Lennon, Ryan Hegarty, Chad Nusbaum, and David B. Jaffe. *Characterizing and measuring bias in sequence data*. Genome Biology, 14(5):R51+, May 2013. 1-9, 2-2, 3-6

[13] Martin Kircher, Udo Stenzel, and Janet Kelso. *Improved base calling for the Illumina Genome Analyzer using machine learning strategies*. Genome Biology, 10(8):R83, 2009. 1-9

[14] Kensuke Nakamura, Taku Oshima, Takuya Morimoto, Shun Ikeda, Hirofumi Yoshikawa, Yuh Shiwa, Shu Ishikawa, Margaret C. Linak, Aki Hirai, Hiroki Takahashi, Md. Altaf-Ul-Amin, Naotake Ogasawara, and Shigehiko Kanaya. *Sequence-specific error profile of Illumina sequencers*. Nucleic Acids Research, 39(13):e90–e90, 2011. 1-9, 3-2

[15] Melanie Schirmer, Rosalinda D'Amore, Umer Z. Ijaz, Neil Hall, and Christopher Quince. *Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data*. BMC Bioinformatics, 17(1), mar 2016. 1-9, 3-3

[16] Jang il Sohn and Jin-Wu Nam. *The present and future of de novo whole-genome assembly*. Briefings in Bioinformatics, page bbw096, oct 2016. 1-11, 1-16

[17] Eugene W. Myers, Granger G. Sutton, Art L. Delcher, Ian M. Dew, Dan P. Fasulo, Michael J. Flanigan, Saul A. Kravitz, Clark M. Mobarry, Knut H. J. Reinert, Karin A. Remington, Eric L. Anson, Randall A. Bolanos, Hui-Hsien Chou, Catherine M. Jordan, Aaron L. Halpern, Stefano Lonardi, Ellen M. Beasley, Rhonda C. Brandon, Lin Chen, Patrick J. Dunn, Zhongwu Lai, Yong Liang, Deborah R. Nusskern, Ming Zhan, Qing Zhang, Xiangqun Zheng, Gerald M. Rubin, Mark D. Adams, and J. Craig Venter. *A Whole-Genome Assembly of Drosophila*. Science, 287(5461):2196–2204, 2000. 1-12

[18] Sergey Koren, Brian P. Walenz, Konstantin Berlin, Jason R. Miller, Nicholas H. Bergman, and Adam M. Phillippy. *Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation*. Genome Research, 27(5):722–736, mar 2017. 1-12

[19] Daniel R. Zerbino and Ewan Birney. *Velvet: Algorithms for de novo short read assembly using de Bruijn graphs*. Genome Research, 18(5):821–829, feb 2008. 1-12, 2-2, 2-5, 3-3, 3-5, 3-10

[20] Sante Gnerre, Iain MacCallum, Dariusz Przybylski, Filipe J. Ribeiro, Joshua N. Burton, Bruce J. Walker, Ted Sharpe, Giles Hall, Terrance P. Shea, Sean Sykes, Aaron M. Berlin, Daniel Aird, Maura Costello, Riza Daza,

Louise Williams, Robert Nicol, Andreas Gnirke, Chad Nusbaum, Eric S. Lander, and David B. Jaffe. *High-quality draft assemblies of mammalian genomes from massively parallel sequence data*. Proceedings of the National Academy of Sciences, 108(4):1513–1518, 2011. 1-12

[21] Yu Peng, Henry C M Leung, S M Yiu, and Francis Y L Chin. *IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth.* Bioinformatics, 28(11):1420–8, June 2012. 1-12

[22] Aleksey V. Zimin, Guillaume Marçais, Daniela Puiu, Michael Roberts, Steven L. Salzberg, and James A. Yorke. *The MaSuRCA genome assembler*. Bioinformatics, 29(21):2669–2677, aug 2013. 1-12

[23] Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, Sergey I. Nikolenko, Son Pham, Andrey D. Prjibelski, Alexey V. Pyshkin, Alexander V. Sirotkin, Nikolay Vyahhi, Glenn Tesler, Max A. Alekseyev, and Pavel A. Pevzner. *SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing*. Journal of Computational Biology, 19(5):455–477, may 2012. 1-12, 1-27, 2-5, 3-3

[24] Ruibang Luo, Binghang Liu, Yinlong Xie, Zhenyu Li, Weihua Huang, Jianying Yuan, Guangzhu He, Yanxiang Chen, Qi Pan, Yunjie Liu, Jingbo Tang, Gengxiong Wu, Hao Zhang, Yujian Shi, Yong Liu, Chang Yu, Bo Wang, Yao Lu, Changlei Han, David W Cheung, Siu-Ming Yiu, Shaoliang Peng, Zhu Xiaoqian, Guangming Liu, Xiangke Liao, Yingrui Li, Huanming Yang, Jian Wang, Tak-Wah Lam, and Jun Wang. *SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler*. GigaScience, 1(1), dec 2012. 1-12, 2-4

[25] Nicolaas Govert de Bruijn. *A combinatorial problem*. Proceedings of the Section of Sciences of the Koninklijke Nederlandse Akademie van Wetenschappen te Amsterdam, 49(7):758–764, 1946. 1-12

[26] Bastien Cazaux, Thierry Lecroq, and Eric Rivals. *Linking indexing data structures to de Bruijn graphs: Construction and update*. Journal of Computer and System Sciences, July 2016. 1-12

[27] Daniel R. Zerbino. *Genome assembly and comparison using de Bruijn graphs*. PhD dissertation, Darwin College, 2009. 1-13

[28] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall Press, Upper Saddle River, NJ, USA, 3rd edition, 2009. 1-15

[29] Michael R. Garey and David S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., New York, NY, USA, 1979. 1-16

[30] Phillip E C Compeau, Pavel A Pevzner, and Glenn Tesler. *How to apply de Bruijn graphs to genome assembly*. Nature Biotechnology, 29(11):987–991, nov 2011. 1-16, 4-2

[31] David R Kelley, Michael C Schatz, and Steven L Salzberg. *Quake: quality-aware detection and correction of sequencing errors.* Genome Biol., 11(11):R116, 2010. 1-17, 2-3, 3-3

[32] Siavash Sheikhizadeh and Dick de Ridder. *ACE: accurate correction of errors using K-mer tries.* Bioinformatics, 31(19):3216–8, 2015. 1-17, 2-3, 3-3, 3-6, 5-4

[33] Heng Li. *BFC: correcting Illumina sequencing errors.* Bioinformatics, 31(17):2885–7, 2015. 1-17, 2-3, 3-3

[34] Yun Heo, Anand Ramachandran, Wen-Mei Hwu, Jian Ma, and Deming Chen. *BLESS 2: accurate, memory-efficient and fast error correction method.* Bioinformatics, 32(15):2369–2371, mar 2016. 1-17, 2-3, 3-3

[35] Paul Greenfield, Konsta Duesing, Alexie Papanicolaou, and Denis C Bauer. *Blue: correcting sequencing errors using consensus and context.* Bioinformatics, 30(19):2723–32, 2014. 1-17, 2-3, 2-5, 3-3, 3-6

[36] Marcel H Schulz, David Weese, Manuel Holtgrewe, Viktoria Dimitrova, Sijia Niu, Knut Reinert, and Hugues Richard. *Fiona: a parallel and automatic strategy for read error correction.* Bioinformatics, 30(17):i356–63, 2014. 1-17, 2-3, 3-3

[37] Amin Allam, Panos Kalnis, and Victor Solovyev. *Karect: accurate correction of substitution, insertion and deletion errors for next-generation sequencing data.* Bioinformatics, 31(July):3421–3428, July 2015. 1-17, 2-3, 3-3, 3-6, 5-4

[38] Eric Marinier, Daniel G. Brown, and Brendan J. McConkey. *Pollux: platform independent error correction of single and mixed genomes*. BMC Bioinformatics, 16(1):10, 2015. 1-17, 2-3, 3-3

[39] Guillaume Marcais, James A. Yorke, and Aleksey Zimin. *QuorUM: An error corrector for Illumina reads*. PLoS One, 10(6):1–13, 2015. 1-17, 2-3, 3-3

[40] Maciej Długosz and Sebastian Deorowicz. *RECKONER: read error corrector based on KMC*. Bioinformatics, page btw746, 2017. 1-17, 3-3

[41] Jared T. Simpson. *Exploring genome characteristics and sequence quality without a reference*. Bioinformatics, 30(9):1228–1235, jan 2014. 1-17

[42] Nicolas Philippe, Mikal Salson, Thérèse Commes, and Eric Rivals. *CRAC: an integrated approach to the analysis of RNA-seq reads*. Genome Biology, 14(3):R30, 2013. 1-19, 1-27

[43] Lucian Ilie and Michael Molnar. *RACER: Rapid and accurate correction of errors in reads.* Bioinformatics, 29(19):2490–3, 2013. 1-20, 2-3, 3-3

[44] Alexey Gurevich et al. *QUAST: quality assessment tool for genome assemblies.* Bioinformatics, 29(8):1072–5, April 2013. 1-21, 2-11, 3-5

[45] Saul B. Needleman and Christian D. Wunsch. *A general method applicable to the search for similarities in the amino acid sequence of two proteins*. Journal of Molecular Biology, 48(3):443–453, mar 1970. 1-24, 3-4, 4-4, 5-3

[46] Maria Chatzou, Cedrik Magis, Jia-Ming Chang, Carsten Kemena, Giovanni Bussotti, Ionas Erb, and Cedric Notredame. *Multiple sequence alignment modeling: methods and applications*. Briefings in Bioinformatics, 17(6):1009–1023, nov 2015. 1-24

[47] Carsten Kemena and Cedric Notredame. *Upcoming challenges for multiple sequence alignment methods in the high-throughput era*. Bioinformatics, 25(19):2455–2465, jul 2009. 1-25

[48] Julie D. Thompson, Desmond G. Higgins, and Toby J. Gibson. *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice*. Nucleic Acids Research, 22(22):4673–4680, 1994. 1-25

[49] Robert C. Edgar. *MUSCLE: multiple sequence alignment with high accuracy and high throughput*. Nucleic Acids Research, 32(5):1792–1797, mar 2004. 1-25

[50] Cédric Notredame, Desmond G Higgins, and Jaap Heringa. *T-coffee: a novel method for fast and accurate multiple sequence alignment 1 1Edited by J. Thornton*. Journal of Molecular Biology, 302(1):205–217, sep 2000. 1-25

[51] Hao Ye, Joe Meehan, Weida Tong, and Huixiao Hong. *Alignment of Short Reads: A Crucial Step for Application of Next-Generation Sequencing Data in Precision Medicine*. Pharmaceutics, 7(4):523–541, nov 2015. 1-25

[52] Nuno A. Fonseca, Johan Rung, Alvis Brazma, and John C. Marioni. *Tools for mapping high-throughput sequencing data*. Bioinformatics, 28(24):3169–3177, oct 2012. 1-25

[53] Heng Li and Richard Durbin. *Fast and accurate short read alignment with BurrowsWheeler transform*. Bioinformatics, 25(14):1754–1760, 05 2009. 1-25, 2-7, 3-5, 4-2

[54] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome*. Genome Biology, 10(3):R25, 2009. 1-25

[55] Heng Li. *Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences*. Bioinformatics, 32(14):2103–2110, 2016. 1-25

[56] Peter Weiner. *Linear Pattern Matching Algorithms*. In Proceedings of the 14th Annual Symposium on Switching and Automata Theory (Swat 1973), SWAT '73, pages 1–11, Washington, DC, USA, 1973. IEEE Computer Society. 1-26

[57] Udi Manber and Gene Myers. *Suffix Arrays: A New Method for On-line String Searches*. In Proceedings of the First Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '90, pages 319–327, Philadelphia, PA, USA, 1990. Society for Industrial and Applied Mathematics. 1-26

[58] Mohamed Ibrahim Abouelhoda, Stefan Kurtz, and Enno Ohlebusch. *Replacing suffix trees with enhanced suffix arrays*. Journal of Discrete Algorithms, 2(1):53 – 86, 2004. The 9th International Symposium on String Processing and Information Retrieval. 1-26

[59] Paolo Ferragina and Giovanni Manzini. *Opportunistic data structures with applications*. In Proceedings 41st Annual Symposium on Foundations of Computer Science, pages 390–398, Nov 2000. 1-26

[60] Michael Burrows and David Wheeler. *A block-sorting lossless data compression algorithm*. Technical report, 1994. 1-26

[61] The Computational Pan-Genomics Consortium. *Computational pan-genomics: status, promises and challenges*. Briefings in Bioinformatics, 19(1):118–135, 10 2016. 1-27

[62] Birte Kehr, Kathrin Trappe, Manuel Holtgrewe, and Knut Reinert. *Genome alignment with graph data structures: a comparison*. BMC Bioinformatics, 15(1):99, 2014. 1-28

[63] Mingjie Wang, Yuzhen Ye, and Haixu Tang. *A de Bruijn Graph Approach to the Quantification of Closely-Related Genomes in a Microbial Community*. Journal of Computational Biology, 19(6):814–825, jun 2012. 1-28

[64] Leena Salmela and Eric Rivals. *LoRDEC: accurate and efficient long read error correction*. Bioinformatics, 30(24):3506–3514, aug 2014. 1-28, 3-4

[65] Giles Miclotte, Mahdi Heydari, Piet Demeester, Stephane Rombauts, Yves Van de Peer, Pieter Audenaert, and Jan Fostier. *Jabba: hybrid error correction for long sequencing reads*. Algorithms for Molecular Biology, 11(1), may 2016. 1-28, 3-4

[66] Antoine Limasset, Bastien Cazaux, Eric Rivals, and Pierre Peterlongo. *Read mapping on de Bruijn graphs*. BMC Bioinformatics, 17(1), jun 2016. 1-28, 4-3, 4-4, 5-3

[67] Bo Liu, Hongzhe Guo, Michael Brudno, and Yadong Wang. *deBGA: read alignment with de Bruijn graph-based seed and extension*. Bioinformatics, 32(21):3224–3232, jul 2016. 1-28, 4-3, 5-3

[68] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. *Fast unfolding of communities in large networks*. Journal of Statistical Mechanics: Theory and Experiment, 2008(10):P10008, 2008. 1-29, 3-4, 3-8

2

Evaluation of the Impact of Illumina Error Correction Tools on de novo Genome Assembly

"... Somewhere, something incredible is waiting to be known. . .[1] *"*

⋆ ⋆ ⋆

Mahdi Heydari, Giles Miclotte, Piet Demeester, Yves Van de Peer and Jan Fostier

Published in BMC Bioinformatics 18(1): 374, Aug. 2017 *In this chapter, we review the contemporary approaches to correct sequencing errors in Illumina data, and how that affect the genome assemblers.. . .*

Abstract

Recently, many standalone applications have been proposed to correct sequencing errors in Illumina data. The key idea is that downstream analysis tools such as *de novo* genome assemblers benefit from a reduced error rate in the input data.

[1] Carl Sagan

Surprisingly, a systematic validation of this assumption using state-of-the-art assembly methods is lacking, even for recently published methods.

For twelve recent Illumina error correction tools (EC tools) we evaluated both their ability to correct sequencing errors and their ability to improve *do novo* genome assembly in terms of contig size and accuracy.

We confirm that most EC tools reduce the number of errors in sequencing data without introducing many new errors. However, we found that many EC tools suffer from poor performance in certain sequence contexts such as regions with low coverage or regions that contain short repeated or low-complexity sequences. Reads overlapping such regions are often ill-corrected in an inconsistent manner, leading to breakpoints in the resulting assemblies that are not present in assemblies obtained from uncorrected data. Resolving this systematic flaw in future EC tools could greatly improve the applicability of such tools.

2.1 Background

Modern Illumina systems generate sequencing data with very high throughput and low financial cost. Illumina estimates that over 90% of sequencing data worldwide are generated on Illumina platforms. This data is characterized by a relatively short read length (100-300 bp) and a high accuracy (1-2% errors, mostly substitutions) [1]. Data generated on Illumina platforms suffers from various sources of bias, most notably a higher number of sequencing errors towards the 3'-end of the reads and a non-uniform distribution of reads across the genome [2].

Despite its short read length, Illumina data is often used for *de novo* genome assembly, sometimes complemented by data generated through other platforms. Most short-read assemblers first generate a de Bruijn graph from the input reads [3]. This graph represents all k-mers that occur in the input reads and the overlap between them. As such, de Bruijn graphs are used to efficiently establish the overlap between individual reads. The original genomic sequence is then represented as some path through the de Bruijn graph.

The presence of sequencing errors significantly complicates this task: a single sequencing error in a read results in up to k erroneous k-mers in the de Bruijn graph. These k-mers create artifacts in the de Bruijn graph such as spurious dead ends, parallel paths and chimeric connections [4]. Despite the low error rate, erroneous k-mers can vastly outnumber true k-mers, challenging the identification of the original sequence. To reduce the number of erroneous k-mers, trimming tools can be used as a primary solution to discard parts of each input read that have a per-base quality score below a user-defined threshold. However, this further reduces the read length and might aggravate the coverage bias.

Error correction tools (EC tools) on the other hand, try to identify and correct the sequencing errors. Often, this is achieved by generating a k-mer coverage

spectrum from the input data and replacing poorly covered (and hence likely erroneous) k-mers by similar k-mers with a higher coverage. Sometimes, this process is further guided by using the per-base quality scores. Many standalone read error correction algorithms and implementations have been proposed for Illumina data, including ACE [5], BayesHammer [6], BFC [7], BLESS [8], BLESS 2 [9], Blue [10], EC [11], Fiona [12], Karect [13], Lighter [14], Musket [15], Pollux [16], Quake [17], QuorUM [18], RACER [19], SGA-EC [20] and Trowel [21]. For a comprehensive overview of the characteristics of these EC tools and those for other sequencing platforms, we refer to [22].

The key idea is that the prior application of EC tools on raw Illumina sequencing data provides assembly methods with cleaner input data and hence improves the quality of assembly both in terms of reduced fragmentation (i.e., longer contigs or scaffolds) and higher accuracy of the resulting assemblies. As a secondary goal, the prior use of EC tools may reduce the memory usage and the runtime of the assembly tool. This is useful when assembling larger genomes, a task that is typically quite resource-intensive.

Surprisingly, most EC tools are not evaluated on their ability to improve the quality of *de novo* genome assembly with modern assemblers, but rather directly on their ability to correct sequencing errors. Using simulated Illumina data, such an evaluation is straightforward as error-free data is known. In that case, the *error correction gain*, a metric that expresses to what degree the error rate is reduced, is used to describe the performance of EC tools. With real Illumina data, the error correction performance is typically assessed through the use of a read mapper: both corrected and uncorrected reads are aligned to their corresponding reference genome and various performance metrics are derived to express the reduction in mismatches in the respective alignments. EC tools that result in more aligned reads and/or alignments with fewer mismatches are assumed to be superior.

We argue that a lower average error-rate in the input data does not necessarily lead to better assembly results. First, the vast majority of sequencing errors are benign to the assembly process. For example, consider a sequencing error that gives rise to one or more erroneous k-mers that otherwise do not exist in the sequenced genome. In the de Bruijn graph, such sequencing error causes a spurious dead end or a short parallel path. These graph artifacts are easily detected and corrected for by many assembly tools assuming the corresponding true k-mers occur with sufficient coverage in the input reads. Only a relatively small fraction of sequencing errors is truly problematic, for example when they give rise to erroneous k-mers that do exist elsewhere in the genome. These errors thus give rise to spurious 'chimeric' connections between nodes in the de Bruijn graph that are otherwise distantly located in the original sequence. As such, they may result in misassemblies and/or shorter contig sizes. A second class of problematic errors are those that occur in regions with very low coverage. Such errors may render the assembly

tool unable to detect overlap between reads because no k-mers are shared. Overall, an EC tool that is able to correct all benign sequencing errors and not a single problematic sequencing error might exhibit a high error correction gain but will not substantially improve the assembly process. Second, EC tools might introduce new errors in the sequence data. If such events are rare and unbiased, they may not pose a great threat to the assembly process. However, if EC tools systematically make the same mistake in a given context, the genome assembler may not be able to recover from this error.

Most state-of-the-art genome assembly tools have built-in algorithms to detect and handle sequencing errors, either directly or implicitly through a correction procedure on the de Bruijn graph. The prior use of standalone EC tools thus only makes sense if they outperform these built-in error correction algorithms. Table 2.1 lists for every EC tool the accuracy analyses that were performed in the accompanying publication. Even though all tools were evaluated for their ability to reduce sequencing errors, their ability to improve the genome assembly process is either lacking or performed with older assembly tools. Also, recent review papers on EC tools [23, 24] did not contain such analyses.

Table 2.1 List of EC tools evaluated in this paper. The algorithmic approach is either k-mer spectrum based ('k-mer') or multiple sequence alignment based ('MSA'). Tools can be further classified according to data structure and heuristics used. Some tools are able to correct insertions or deletions. In their accompanying publication, all tools were assessed directly on their ability to reduce error rate, either on the read or base level. Most tools did not use assembly analyses with modern assemblers in their evaluation. SPAdes was used for the evaluation of BayesHammer, but no comparison was made with assembly results from uncorrected data.

EC tool	Algorithm	Data structure	Indel support	Accuracy analysis	Assembly analysis	Year
ACE	k-mer	k-mer trie		read level	-	2015
BayesHammer	k-mer	hamming graph		read level	SPAdes	2013
BFC	k-mer	bloom filter		read level	Velvet, ABySS [25]	2015
BLESS 2	k-mer	bloom filter		read level	Gossamer [26]	2016
Blue	k-mer	hash table	✓	read level	Velvet	2014
Fiona	MSA	suffix tree	✓	base level	-	2014
Karect	MSA	partially-ordered graph	✓	read, base level	Velvet, SGA, Celera [27]	2015
Lighter	k-mer	bloom filter		read level	Velvet	2013
Musket	k-mer	bloom filter		base level	SGA	2013
RACER	k-mer	hash table		read level	-	2013
SGA-EC	MSA	suffix array		read level	SGA	2012
Trowel	k-mer	hash table		read, base level	Velvet, SOAPdenovo [28]	2014

In this paper, we review twelve recently published EC tools. We compiled a benchmark suite of eight public datasets sequenced from organisms with a genome size ranging from 2 to 116 Mbp and assessed the performance of the different EC tools both on their potential to correct the sequencing errors and on their ability to improve assembly results using four assemblers (DISCOVAR [29], IDBA [30],

SPAdes [31] and Velvet [4]). We discuss the impact on the resulting assembly quality and investigate systematic errors in some of the EC tools. Finally, computational efficiency (memory usage and runtime) of the different EC tools is discussed. Note that the effect of error correction for other applications such as variant calling is beyond the scope of this paper.

Table 2.2 Real datasets used for the evaluation of EC tools.

Abbr.	Organism	Reference ID	Genome size	Cov.	Sequencing platform	Read length	Trimmed reads	Dataset ID	Ref.
D1	*Bifidobacterium dentium*	Nc013714.1	2.6 Mbp	373 X	Illumina MiSeq	251 bp		SRR1151311	[23]
D2	*Escherichia coli K-12 DH10B*	NC010473	4.5 Mbp	418 X	Illumina MiSeq	150 bp		Ill. Data library	[10]
D3	*Escherichia coli K-12 MG1655*	NC000913	4.5 Mbp	612 X	Illumina GAII	100 bp		ERA000206	[10]
D4	*Salmonella enterica*	NC011083.1	4.7 Mbp	97 X	Illumina MiSeq	239 bp	✓	SRR1206093	[23]
D5	*Pseudomonas aeruginosa*	ERR330008	6.1 Mbp	169 X	Illumina MiSeq	120 bp	✓	ERR330008	[10]
D6	*Homo sapiens* Chr. 21	HG19	45.2 Mbp	29 X	Illumina HiSeq	100 bp		Ill. Data library	[10]
D7	*Caenorhabditis elegans*	WS222	97.6 Mbp	58 X	Illumina HiSeq	101 bp		SRR543736	[23]
D8	*Drosophila melanogaster*	Release 5	116.4 Mbp	52 X	Illumina HiSeq	100 bp		SRR823377	[23]

2.2 Material and Methods

2.2.1 Error correction tools

Twelve state-of-the-art (published in 2012 or later) EC tools for Illumina data were included in this review and listed in Table 2.1. We were unable to produce corrected reads with QuorUM and EC and hence these tools were excluded in this study.

EC tools have been classified according to their underlying algorithmic principles in several review papers [22, 23, 32]. In Table 2.1, tools were classified according to their main algorithmic approach: k-mer spectrum based or multiple sequence alignment (MSA) based. The k-mer spectrum based tools operate on the level of individual k-mers. First, the complete set of k-mers that occur in the input data and their corresponding frequency is determined. Second, reads that contain rarely occurring k-mers are assumed to contain sequencing errors and are modified, using a minimum edit distance strategy, such that these k-mers are replaced by similar, more frequently occurring k-mers. In contrast, MSA-based tools operate on the level of reads. First, reads that are assumed to represent overlapping genomic regions are clustered together and a consensus is obtained through multiple alignment. Second, reads are corrected according to the consensus alignment. While all EC tools considered in this review rely on either of these two approaches, there is still a great diversity in the specific implementation heuristics and data structures (bloom filter, hash table, suffix tree, ...).

Most tools require users to specify a k-mer length to be used during the error correction procedure. The optimal value can differ from one dataset to another, depending on the coverage, genome size and error distribution. This optimal value

was empirically obtained by running the EC tool multiple times with different k-mer sizes and selecting the k-mer size that yields the most contiguous SPAdes assembly results as measured in terms of N50. This optimal value was used to produce the results of Table 2.4. For all other tables and figures, the default or recommended k-mer size was used for all datasets. Parameters and settings are provided in App. A.1. All tools support multithreading, and with the exception of ACE and RACER, the number of parallel threads can be specified. Those tools were run with 32 threads. Runtime and peak memory usage were measured with the GNU 'time -v' command. We recorded elapsed (wall clock) time and peak resident memory usage. All tools were run on a machine with four Intel(R) Xeon(R) E5-2698 v3 @ 2.30 GHz CPUs (64 cores in total) and 256 GB of memory.

2.2.2 Data

Tools are benchmarked on eight datasets for which both a high quality reference genome and real Illumina data are publicly available (see Table 2.2). Genome sizes range from 2 Mbp (*Bifidobacterium dentium*) to 116 Mbp (*Drosophila melanogaster*) while read coverage varies from 29 X to 612 X. Data is produced by the Illumina HiSeq, MiSeq and GAII platforms with read lengths varying between 100 bp and 251 bp. Two of the datasets have a variable read length due to read trimming, all other datasets have fixed read lengths.

To assess the performance of tools on simulated data, synthetic Illumina reads for the same set of organisms were generated using ART [33]. The same coverage and read lengths were used as for the real data (App. A.2). ART also generates a corresponding set of error-free reads, which greatly facilitates the evaluation of EC tools on synthetic data.

2.2.3 Error metrics

The error rate is the ratio of the total number of sequencing errors (substitutions or indels) and the number of nucleotides in the input data. Error correction performance is measured as follows: true positives (TP) correspond to corrected errors; true negatives (TN) correspond to initially correct bases left untouched; false positives (FP) correspond to newly introduced errors; false negatives (FN) correspond to unidentified errors. The error correction gain (EC gain) is defined as:

$$\text{EC gain} = \frac{\text{TP} - \text{FP}}{\text{TP} + \text{FN}}.$$

The EC gain measures the degree in which the error rate is reduced. A gain of 100% means all errors were corrected and no new errors were introduced. The sensitivity (true positive rate – TPR) is defined as follows:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

2.2.4 Evaluation of assembly results

To assess the impact of error correction on *de novo* assembly results, the following assemblers were used: DISCOVAR, IDBA, SPAdes and Velvet. All four assemblers have built-in error correction functionality. Velvet, IDBA and SPAdes remove erroneous k-mers through the identification of parallel paths ('bubbles' and 'tips') in the de Bruijn graph. SPAdes and IDBA iteratively increase the k-mer size. This way, they take advantage of shorter k-mers for a sensitive detection of overlap between reads and of longer k-mers for dealing with repeat resolution. DISCOVAR uses a different methodology: for each read, a group of 'true friends' is determined. These are reads that share a k-mer with the read and that do not have a high quality base difference with the read. DISCOVAR then corrects each read based on the consensus sequence obtained from the multiple sequence alignment of its true friends.

We investigated the underlying causes of suboptimal assembly results after error correction. MUMmer [34] was used to align contigs, and to check if the contig has no structural misassemblies. In order to determine the k-mer frequencies Jellyfish [35] was used.

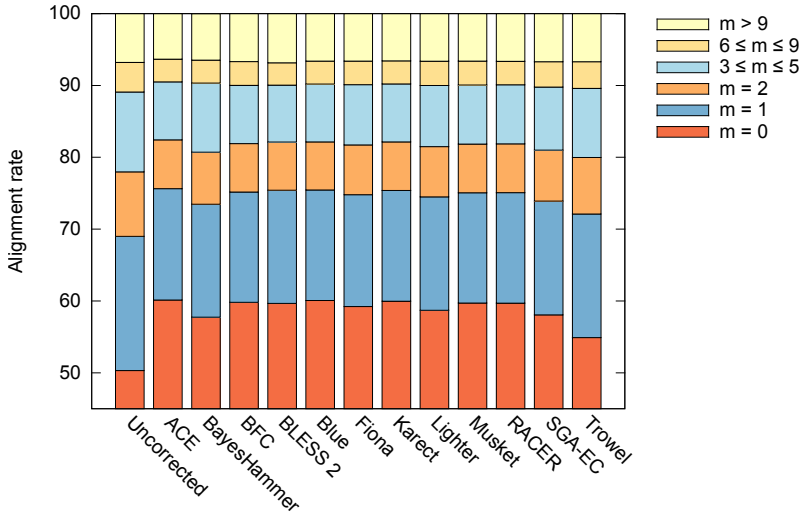
2.3 Results and Discussion

2.3.1 Ability of EC tools to correct sequencing errors

In order to estimate the reduction in error rate through the use of EC tools, both uncorrected and corrected data were aligned to the corresponding reference genome using BWA [36]. For all datasets D1-D8 and EC tools, the fraction of reads that align with respectively $m = 0$ and $m > 9$ mismatches is reported in App. A.3.1. All EC tools are able to substantially reduce the number of mismatches required for read alignment. This is especially true for bacterial genomes, where often $>$95% of the corrected reads show perfect alignment with the reference. In contrast, for larger genomes, this is typically in the range of 60-80%. Error correction also reduces the fraction of highly erroneous reads (i.e., reads that require more than 9 mismatches to align), albeit to varying degrees. For the largest dataset D8 (*D. melanogaster*), Fig. 2.1 provides a more detailed breakdown of the number of mismatches m required for read alignment. Initially, about 50% of the uncorrected reads perfectly align. ACE shows the highest increase of this figure to 60.14%. ACE also has the lowest percentage of highly erroneous reads.

After applying error correction to a read, there is no guarantee that BWA will again align that read to the same genomic location. Therefore, this evaluation metric might favor overly aggressive EC tools that transform reads into similar reads that do exist in the genome, but that do not represent the actual sequenced genomic region. Therefore, in an alternative evaluation metric, we assume that the

Figure 2.1 Mismatches in read alignment.



Classification of (un)corrected reads for *D. melanogaster*, based on the number of mismatches in their alignment to the reference genome.

error-free read is represented by the segment of the reference genome to which the uncorrected read aligns. Uncorrected reads that can not be mapped to the reference genome are excluded from this evaluation. As BayesHammer and BLESS 2 do not provide a one-to-one correspondence between input and output, they are not included in this evaluation.

Table 2.3 shows the EC gain, the percentage of corrected errors and the number of newly introduced errors per Mbp of read data for each of the eight datasets. Detailed confusion matrices are provided in App. A.3.2.2. Major differences in EC gain can now be observed between the different EC tools.

All EC tools perform much better on the smaller bacterial genomes (D1-D5), than on the larger eukaryotes (D6-D8). For all datasets, Karect shows the highest number of true positives (errors that were successfully corrected) and the lowest number of false negatives (uncorrected errors). With the exception of dataset D7 (*C. elegans*) and D8 (*D. melanogaster*), Karect also has the lowest number of false positives (newly introduced errors). Overall, Karect has the highest error correction gain for all datasets.

For most datasets, BFC, SGA-EC and Trowel correct significantly fewer sequencing errors compared with other EC tools. BFC and SGA-EC appear to be conservative as they introduce only a small number of new errors. In contrast, ACE, Racer and Trowel often introduce a significant amount of new errors. Note that for dataset D7, the EC gain of ACE is negative, indicating a higher number

Table 2.3 Accuracy comparison of EC tools in terms of EC gain, percentage of corrected errors, and number of newly introduced errors per Mbp of read data.

	D1	D2	D3	D4	D5	D6	D7	D8
Error correction gain (%)								
ACE	96.3	97.9	98.7	96.2	91.1	41.7	-3.3	25.9
BFC	78.7	84.3	80.2	81.4	78.6	52.8	63.3	24.1
Blue	98.5	98.8	98.7	96.7	95.4	51.1	65.2	28.8
Fiona	87.4	94.6	97.5	85.5	91.4	55.0	65.8	29.8
Karect	99.4	99.8	99.7	98.5	98.2	63.1	75.5	34.3
Lighter	85.4	93.8	92.5	80.1	84.6	45.7	50.3	21.7
Musket	91.3	93.6	93.4	88.0	87.1	49.5	59.2	23.5
RACER	92.3	94.4	97.0	88.3	94.0	17.4	32.6	22.3
SGA-EC	55.3	67.2	45.5	53.1	65.2	48.7	60.6	23.0
Trowel	38.4	49.4	38.8	40.5	46.8	13.2	1.1	10.5
Percentage of corrected errors (sensitivity)								
ACE	97.7	98.5	99.2	98.0	97.0	61.3	73.8	34.5
BFC	78.8	84.4	80.2	81.4	78.7	54.1	63.8	24.7
Blue	98.7	99.3	99.1	97.0	95.7	59.9	70.6	31.4
Fiona	87.5	94.8	97.7	85.5	91.7	60.6	71.7	31.5
Karect	99.4	99.9	99.7	98.5	98.2	64.4	76.7	35.5
Lighter	85.5	94.0	92.7	80.2	86.3	48.9	59.1	24.3
Musket	91.3	93.6	93.4	88.1	87.3	52.9	65.3	26.4
RACER	92.9	95.8	98.2	89.0	94.8	59.2	68.2	34.0
SGA-EC	55.3	67.2	45.5	53.1	65.3	50.4	61.3	23.2
Trowel	39.0	49.9	43.4	40.9	47.6	23.6	31.2	11.8
Number of errors introduced per Mbp								
ACE	44	23	40	151	194	1217	2375	1123
BFC	2	3	7	2	3	83	15	73
Blue	8	20	30	31	10	547	167	341
Fiona	2	7	14	6	9	347	183	218
Karect	0	1	3	1	1	80	36	157
Lighter	2	6	14	8	56	202	273	332
Musket	1	2	5	3	6	214	190	383
RACER	21	62	97	58	27	2603	1097	1524
SGA-EC	1	3	6	2	3	105	22	24
Trowel	21	26	376	41	25	647	930	172

Table 2.4 NGA50 of respectively contigs (top) and scaffolds (bottom) assembled by SPAdes before and after error correction. Arrows in the table are based on their value relative to the NGA50 value obtained from uncorrected data as follows: ⇊ < -10% < ↓ < 0% < ↑ < +10% < ⇈ .

Tools	D1	D2	D3	D4	D5	D6	D7	D8
				Contig NGA50				
Uncorrected	397 392	92 570	119 253	231 409	264 881	8 559	6 429	50 484
ACE	397 392 =	92 570 =	125 608 ↑	231 409 =	264 881 =	8 771 ↑	3 143 ⇊	28 679 ⇊
BayesHammer	397 392 =	92 344 ↓	132 564 ⇈	231 409 =	264 881 =	9 075 ↑	6 540 ↑	53 534 ↑
BFC	397 392 =	92 570 =	132 876 ⇈	231 409 =	264 881 =	9 375 ↑	6 389 ↓	49 185 ↓
BLESS 2	397 392 =	92 570 =	119 265 ↑	231 409 =	264 881 =	7 975 ↓	3 047 ⇊	23 814 ⇊
Blue	397 392 =	92 708 ↑	132 876 ⇈	231 409 =	289 353 ↑	7 628 ⇊	6 191 ↓	50 486 ↑
Fiona	397 392 =	92 611 ↑	119 253 =	231 409 =	264 881 =	9 224 ↑	5 346 ⇊	45 472 ↓
Karect	397 392 =	92 611 ↑	132 876 ⇈	231 409 =	264 881 =	9 865 ⇈	6 392 ↓	54 132 ↑
Lighter	397 392 =	92 570 =	132 564 ⇈	231 409 =	289 353 ↑	9 609 ⇈	6 423 ↓	50 440 ↓
Musket	397 392 =	92 566 ↓	132 876 ⇈	231 409 =	264 881 =	9 293 ↑	6 170 ↓	46 377 ↓
RACER	397 392 =	92 523 ↓	112 393 ↓	231 409 =	264 881 =	7 336 ⇊	3 244 ⇊	21 538 ⇊
SGA-EC	397 392 =	92 344 ↓	119 255 ↑	231 409 =	264 881 =	9 296 ↑	6 435 ↑	52 105 ↑
Trowel	397 392 =	92 344 ↓	119 335 ↑	231 409 =	264 881 =	7 808 ↓	6 389 ↓	48 357 ↓
				Scaffold NGA50				
Uncorrected	397 392	97 353	132 876	231 409	289 353	8 829	6 472	60 554
ACE	397 392 =	97 353 =	133 713 ↑	231 409 =	264 881 ↓	9 190 ↑	3 158 ⇊	35 392 ⇊
BayesHammer	397 392 =	97 353 =	133 309 ↑	231 409 =	264 881 ↓	9 443 ↑	6 576 ↑	58 570 ↓
BFC	397 392 =	97 353 =	133 088 ↑	231 409 =	264 881 ↓	9 664 ↑	6 419 ↓	59 613 ↓
BLESS 2	397 392 =	97 353 =	132 876 =	231 409 =	264 881 ↓	8 441 ↓	3 073 ⇊	35 638 ⇊
Blue	397 392 =	97 288 ↓	133 309 ↑	231 409 =	289 353 =	7 841 ⇊	6 183 ↓	61 289 ↑
Fiona	397 392 =	97 353 =	132 876 =	231 409 =	264 881 ↓	9 491 ↑	5 385 ⇊	54 188 ⇊
Karect	397 392 =	97 353 =	133 058 ↑	231 409 =	264 881 ↓	10 302 ⇈	6 446 ↓	62 304 ↑
Lighter	397 392 =	97 353 =	133 309 ↑	231 409 =	289 353 =	9 955 ⇈	6 468 ↓	59 697 ↓
Musket	397 392 =	97 353 =	133 088 ↑	231 409 =	264 881 ↓	9 502 ↑	6 219 ↓	55 842 ↓
RACER	397 392 =	97 353 =	132 876 =	231 409 =	264 881 ↓	7 603 ⇊	3 266 ⇊	23 783 ⇊
SGA-EC	397 392 =	97 353 =	132 876 =	231 409 =	264 881 ↓	9 640 ↑	6 483 ↑	60 636 ↑
Trowel	397 392 =	97 353 =	132 876 =	231 409 =	264 881 ↓	8 107 ↓	6 435 ↓	57 078 ↓

of sequencing errors after error correction than in the uncorrected data: ACE successfully corrects about 10.8 million errors but introduces almost 11.3 million new errors.

For comparison, *artificial* data was generated for the eight genomes using the same read length and coverage as the corresponding real datasets. Data was corrected using identical settings as before. The confusion matrix and derived metrics can be unambiguously constructed for artificial data since the true, error-free read is known (see App. A.3.2.3). BFC now shows the highest gain for four datasets, while Karect and Fiona each have the highest gain for two datasets. The numbers indicate that EC tools perform much better on artificial data than on real data. This is due to the fact that simulated data are produced according to simplified models that may fail to capture the intricacies of real data.

2.3.2 Ability of EC tools to improve genome assembly

To evaluate the effect of error correction on *de novo* genome assembly, both uncorrected and corrected reads were assembled using respectively DISCOVAR, IDBA, SPAdes and Velvet. The resulting assemblies were evaluated using QUAST [37] and detailed reports for all combinations of assemblers and EC tools are provided in App. A.4 for reference. We found that SPAdes and DISCOVAR consistently produced higher quality contigs than Velvet and IDBA. We were unable to produce assemblies with DISCOVAR using the reads that were corrected by Trowel and Fiona[2]. Therefore, only SPAdes assemblies are discussed in detail in the remainder of this section.

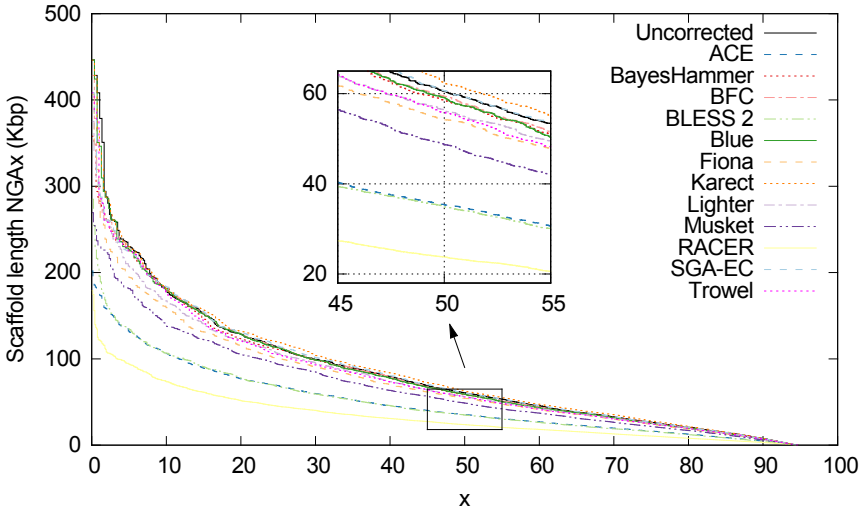
Table 2.4 shows the contig and scaffold NGA50 values for all eight datasets and EC tools. For the EC tools that allow the k-mer size to be specified, the optimal value of k was used (see App. A.1). The NGA50 represents the characteristic length of the assembled contigs/scaffolds that can be contiguously aligned to the reference genome. These contigs/scaffolds thus contain no major structural assembly errors and a higher NGA50 hence implies a less fragmented assembly. For smaller genome sizes (datasets D1-D5), the prior application of EC tools often does not significantly influence the scaffold NGA50. For dataset D3, many tools are able to improve the contig NGA50, sometimes significantly. Remarkably, for dataset D5 (*P. aeruginosa*) most EC tools lead to a somewhat lower scaffold NGA50 compared to the assembly result obtained from uncorrected data. However, the NGAx plot of this dataset reveals no major differences in assembly quality between corrected and uncorrected reads (see App. A.4.3.5). For the larger genomes, the use of EC tools does occasionally improve assembly results, especially on dataset D6 (Human, chr. 21) where eight out of twelve EC tools lead to a

[2]Trowel and Fiona manipulate the quality scores of the bases in the sequencing data besides correcting sequencing errors. However, the new scores are not in the acceptable range of DISCOVAR.

higher scaffold NGA50. On the largest datasets D7 and D8 however, error correction may significantly deteriorate the assembly quality. In some cases, the NGA50 obtained is less than half of the corresponding value on uncorrected data.

Especially for dataset D8 (*D. melanogaster*), the prior use of different EC tools results in a large variability in assembly quality (see Fig. 2.2). Only Blue, Karect and SGA-EC improve the NGA50 for this dataset. In contrast, error correction with ACE, BLESS 2, Fiona or RACER leads to significantly shorter scaffolds. Additionally, a lower percentage of the genome was found to be covered by scaffolds and a higher rate of insertions, deletions and mismatches was observed (see App. A.4).

Figure 2.2 SPAdes assemblies.



SPAdes assembly results for *D. melanogaster* for (un)corrected data. Scaffolds with length NGAx or larger contain x% of the genome.

At this point it should be stressed that error correction does consistently lead to substantially better assembly results for Velvet or IDBA. However, in our hands, the NGA50 values obtained with Velvet or IDBA were much lower than with SPAdes or DISCOVAR. Even after error correction, Velvet and IDBA yield significantly shorter contigs than SPAdes or DISCOVAR. From this we conclude that the built-in error correction procedures in Velvet and IDBA are less accurate than those in SPAdes and DISCOVAR.

2.3.3 Error rate versus assembly quality

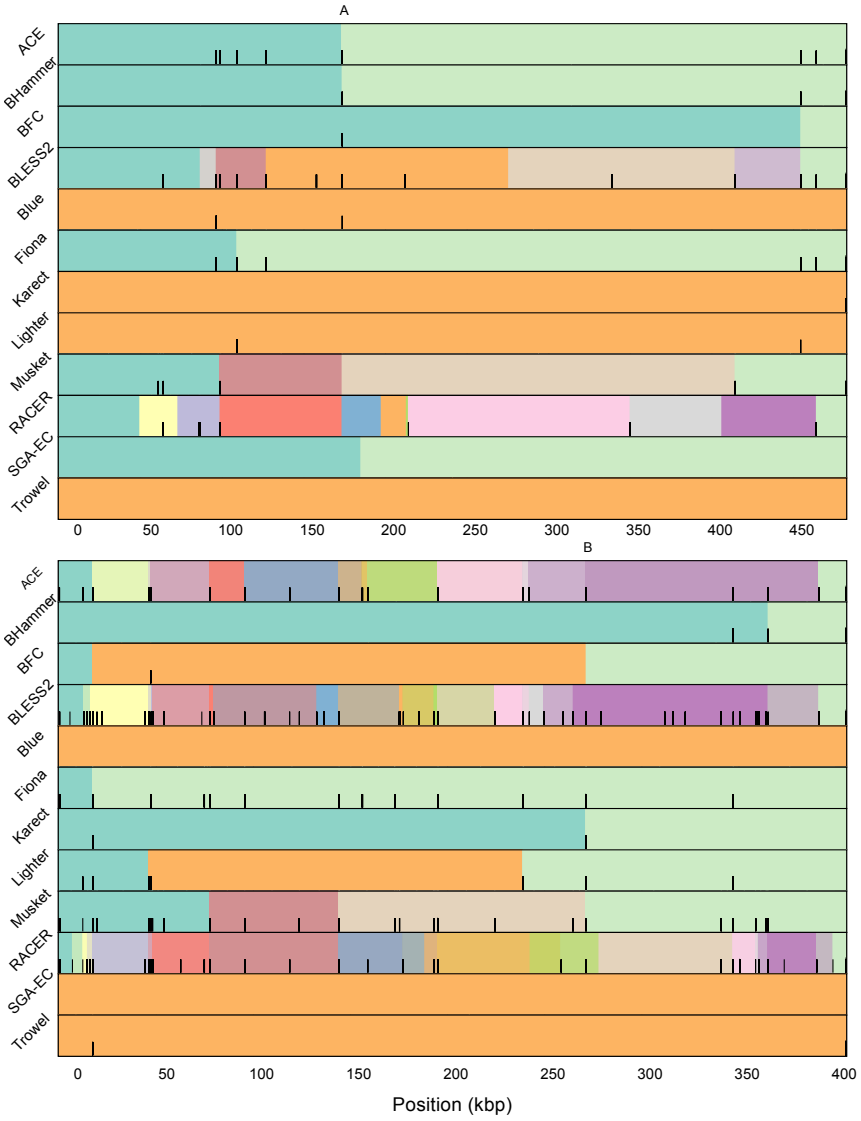
Even though EC tools almost always reduce the error rate in the input data, they do not necessarily lead to better assemblies. In order to better understand these contrasting observations, we investigated why the use of corrected data can lead to a more fragmented assembly. For the largest dataset (D8), the two largest contigs (>400 kbp each) that were correctly assembled from uncorrected data were selected. The corresponding (shorter) contigs obtained from assemblies on corrected data were aligned to these contigs and visualized in Fig. 2.3. With the exception of Trowel, all error correction tools lead to a more fragmented assembly of at least one of these contigs. Breakpoints, i.e., endpoints of the shorter contigs, caused by error correction do not appear to occur at random positions. Rather, different EC tools often cause breakpoints at the same positions. For example, in Fig. 2.3, the breakpoints marked as 'A' and 'B' each occur in four cases.

In order to identify the mechanisms that cause breakpoints, the k-mer spectrum of both corrected and uncorrected data along the two contigs was examined. In this section, $k = 21$ is used throughout, as it corresponds to the smallest k-mer size that is used to establish overlap between individual reads by the multi-k SPAdes assembler. In Fig. 2.3, black bars visualize the locations of 'lost true 21-mers', i.e., 21-mers that do exist in the reference sequence (hence 'true') and also do exist in the uncorrected data but that are no longer present in the corrected data (hence 'lost'). Lost true k-mers hence refer to those k-mers that were systematically, but erroneously removed during error correction. In many cases, lost true 21-mers occur in the direct vicinity of breakpoints, indicating a possible causal relationship between lost true 21-mers and these breakpoints (see Fig. 2.3).

To varying degrees, all EC tools suffer from lost true k-mers. For dataset D8, Fig. 2.4 shows the 21-mer spectrum of the uncorrected data, along with the lost true 21-mer spectrum for the individual EC tools. Unsurprisingly, true k-mers are almost exclusively lost when their corresponding coverage in the uncorrected data is low. Indeed, a lower than expected coverage is an important feature for EC tools to select candidate errors. Trowel and SGA-EC appear most conservative in terms of lost true k-mers: almost no true 21-mers that occur >2 times are removed. In contrast, ACE, BLESS 2, Musket and RACER remove a significant number of true 21-mers, some of which occur >10 times in the initial data. These EC tools lead to a more fragmented assembly, which becomes especially evident for the second biggest contig (cfr. Fig. 2.3).

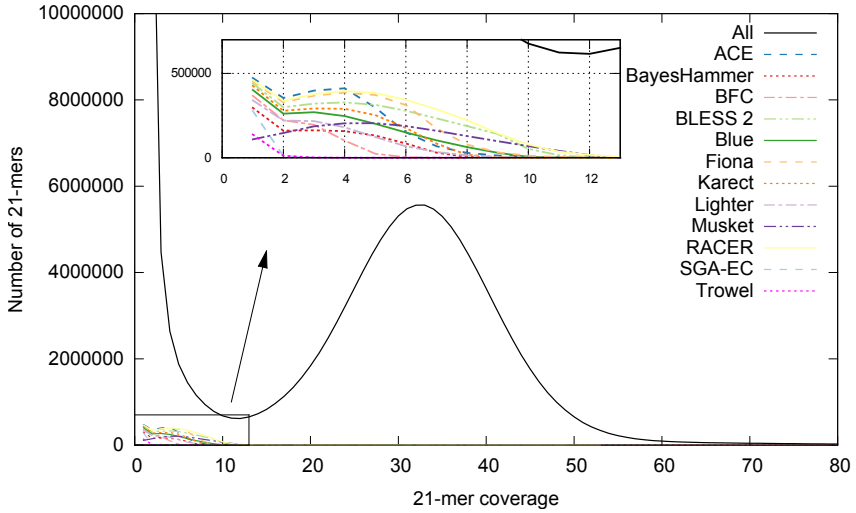
In principle, a lost true k-mer should not necessarily lead to a breakpoint. If all reads that initially contain the lost true k-mer(s) are modified in a consistent manner, the assembler will still be able to correctly identify the overlap between those reads and the lost true k-mers would appear as mismatches in the resulting assembly. In practice, the lost true k-mers will likely be replaced by k-mers that actually occur elsewhere in the genome and the genome assembler will be chal-

Figure 2.3 Fragmented assembly using corrected data.



Contigs assembled from corrected data are aligned to the largest (top) and second largest (bottom) contig obtained from uncorrected data. Different colors denote different contigs. Black bars indicate the location of lost true k-mers in the contigs. This indicates a possible causal relationship between lost true k-mers and the breakpoints in the assemblies of corrected data.

Figure 2.4 Lost true 21-mers spectrum



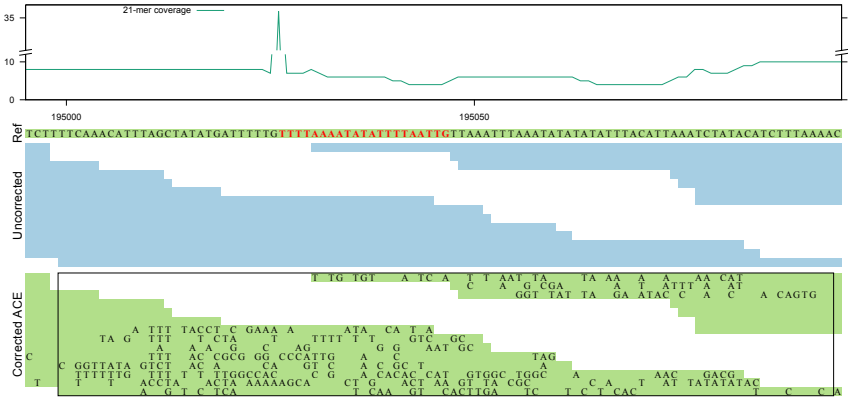
For dataset D8, this figure shows the 21-mer spectrum of the uncorrected data, along with the lost true 21-mer spectrum for all EC tools. EC tools erroneously remove low frequency true 21-mers during error correction.

lenged by a spurious repeat that it may or may not be able to resolve. Vice versa, not all breakpoints due to error correction are directly related lost true k-mers. The ill-correction of reads could potentially only lead to a decrease in coverage without losing the true k-mer in all reads. This can still result in a breakpoint.

In practice however, we find that breakpoints due to error correction are often related to lost true k-mers (cfr. Fig. 2.3). Further inspection revealed that true k-mers are typically lost in regions that suffer from poor coverage in the direct vicinity of a local coverage peak. Often, such sudden increase in coverage is caused by the presence of a short repeated element. For example, Fig. 2.5 shows a genomic region with low k-mer coverage (around 7 X) that contains a repeated k-mer with coverage 35. This repeated k-mer also occurs in other reads that originate from different genomic locations. We can therefore assume that the EC tool makes erroneous decisions based on the sequence content of these reads. In this example, ACE makes a large number of substitutions in originally error-free reads causing 75 consecutive lost true k-mers. Clearly, the error correction procedure is not performed in a consistent manner for all reads, rendering the assembler unable to detect overlap between these reads and ultimately leading to a breakpoint. For the same reasons, BLESS 2 and RACER also break at this specific location.

As a second example, Fig. 2.6 shows a short 22 bp long AT repeat with very high coverage (nearly 14 000 X), in a genomic region with otherwise low coverage.

Figure 2.5 Alignment of uncorrected and ACE-corrected reads in the neighborhood of a contig breakpoint



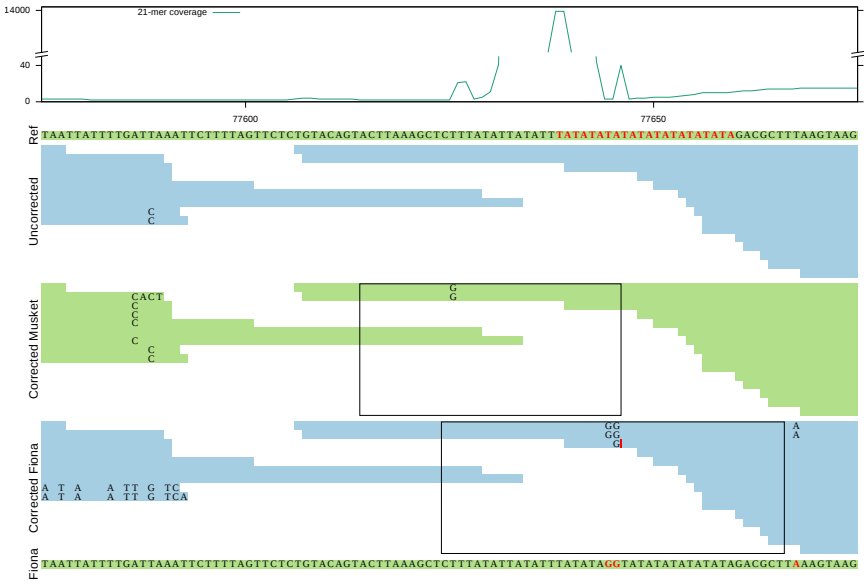
The first track shows the 21-mer coverage of the uncorrected data. The second track (Ref) contains part of the reference genome, which is assembled into one contig from uncorrected data. A repeated 21-mer is indicated in red. The third track (Uncorrected) shows the alignment of the uncorrected, but error-free reads to the reference. The fourth track (Corrected) uses these same alignment positions, but with the sequence content of the corrected reads. Newly introduced errors are indicated by a character in the reads. The rectangle in the fourth track indicates 75 overlapping 21-mers that are lost as a result of erroneous error correction.

Musket introduces a new error in two out of four overlapping reads. Within this specific context, these substitutions cause a number of true k-mers to be lost. More importantly, because the error correction is not performed in an identical manner across all four reads overlapping this locus, the overlap is broken and a breakpoint is introduced. Similarly, due to the same AT repeat, Fiona introduces errors that result in a number of lost true k-mers. In this case however, the newly introduced errors result in mismatches in the assembled sequence rather than a breakpoint.

From these examples, the limitations of k-mer spectrum based error correction tools become evident. Due to their primary focus on individual k-mers, they do not take into account the surrounding context in which the k-mer occurs. Because these tools correct reads individually, different corrections may be applied to different reads even though the reads overlap the same genomic region. This may render de Bruijn graph assemblers unable to detect overlap between those reads. In that respect, error correction tools that rely on multiple sequence alignments (MSA) are in principle less susceptible to this kind of error. As overlapping reads are clustered and aligned, the error correction is systematic across those reads. MSA-based tools indeed yield higher NGA50 values on average.

These results demonstrate that evaluating error correction tools directly on

Figure 2.6 Alignment of uncorrected and corrected reads by Musket and Fiona in the neighborhood of a contig breakpoint:



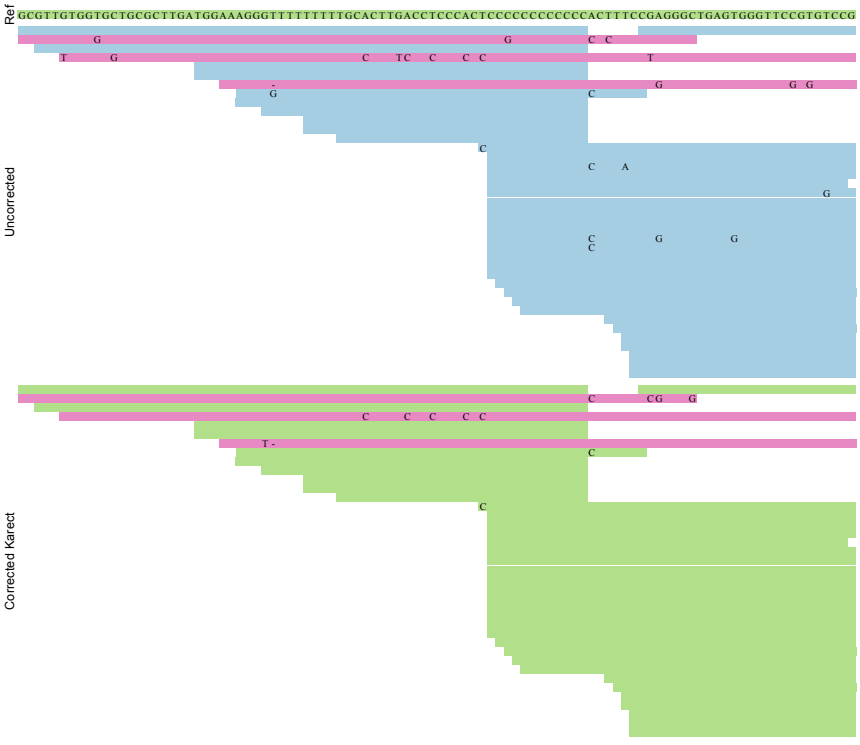
Lost true k-mer can result in two different scenarios. The first track shows the 21-mer coverage of the uncorrected data. The second track (Ref) shows a part of the reference genome, which is assembled into one contig from uncorrected data. A frequently occurring AT-repeat is indicated in red. The third track (Uncorrected) shows the alignment of the uncorrected reads to the reference. The fourth and the fifth tracks (Corrected Musket and Corrected Fiona) use these same alignment positions, but with the sequence content of corrected reads by Musket and Fiona. The sixth track is the assembled contig from corrected reads by Fiona. The rectangles indicate the regions in corrected reads by Musket and Fiona that no longer contain any true 21-mers. The coverage is low around an 'AT' repeat with coverage 13750x in the uncorrected data. Musket incorrectly changed two bases, breaking the connection between two groups of reads. In contrast, in the Fiona-corrected reads, the connection is not lost. Instead the lost true k-mers in Fiona appear as mismatches in the assembled contig.

their ability to reduce error rate has significant limitations as there is often no clear correlation between such metrics and the ability to improve assembly. For example, on datasets D8, ACE ranked fourth in terms of gain and showed the highest number of corrected reads that align error-free to the reference genome. Yet, ACE-corrected reads do not lead to good assembly results on this dataset.

We should emphasize that error correction is not always destructive: EC tools can improve the quality of assembly in certain cases. For example, even though

Karect also suffers from a significant number of 'lost true k-mers' (see Fig. 2.4), the tool leads to the highest NGA50 values in many cases (see Table 2.4). Again for dataset D8, we selected the longest contig (>500 kbp) that was correctly assembled from corrected data by Karect and aligned the corresponding (shorter) contigs obtained from assemblies on uncorrected data. A specific case where Karect removes errors that subsequently lead to the correct connection between two contigs is shown in Fig. 2.7.

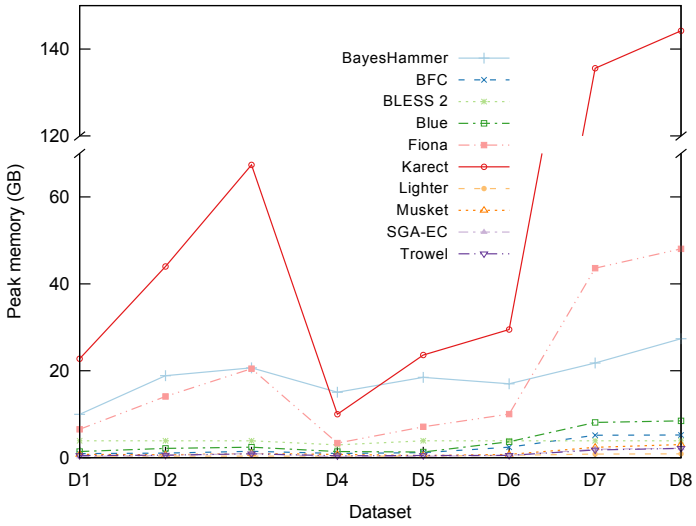
Figure 2.7 Error correction with Karect resolves a breakpoint in the uncorrected data assembly. The first track (Ref) shows a part of the reference genome, which is assembled into a single contig from Karect-corrected reads. The second track (Uncorrected) shows the alignment of the uncorrected reads to the reference. The third track (Corrected Karect) uses these same alignment positions, but with the sequence content of reads corrected by Karect. The short overlap between the uncorrected reads is less than 21, i.e., a true 21-mer is missing from the uncorrected data. There are three reads which expand along this region but they contain some errors which are highlighted in purple. After error correction those three reads are partially cleaned which suffices to connect the two groups of reads.



2.3.4 Time and space requirements

Fig. 2.8 and 2.9 show the memory usage and runtime of the EC tools (see App. A.5.1 for detailed tables). Since it is not possible to specify the number of threads for ACE and RACER, they were omitted. For all datasets, BayesHammer, Fiona and Karect use significantly more memory than other tools while BayesHammer, Fiona, Karect, Musket, and SGA-EC have a relatively high runtime. In general, we note that all tools that rely on multiple sequence alignments require more resources. The tools that rely on Bloom filters (BLESS 2, Lighter and BFC) are both memory efficient and fast.

Figure 2.8 Peak memory usage.

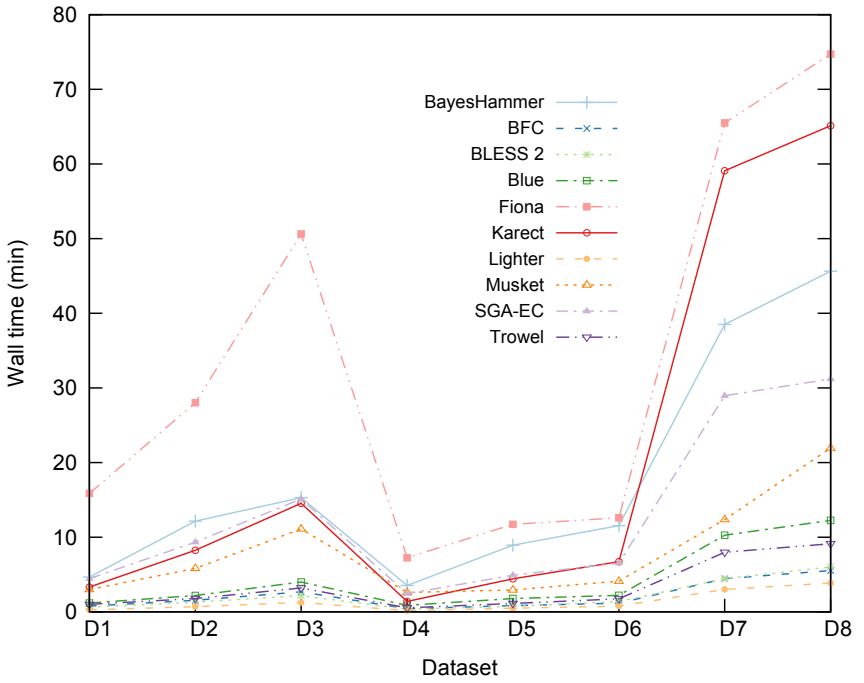


Peak memory usage of the EC tools.

Given the reduced error in the input data, we evaluate the potential of error correction tools to reduce the peak memory usage and/or runtime of the assembly process itself. Since error correction is computationally intensive, this may be an important aspect of error correction tools. Peak memory usage and runtime were measured for all assemblies with SPAdes and DISCOVAR. The runtime of DISCOVAR shows no decrease after error correction (Fig. 2.10), while the peak memory usage decreases slightly (Fig. 2.11). Conversely, the runtime of SPAdes does decrease after error correction (Fig. 2.12), but the peak memory usage does not (Fig. 2.13).

The peak memory usage and runtime tables for artificial data show that Lighter and SGA-EC are again among the most memory-efficient tools, while Karect and

Figure 2.9 Runtime.



Runtime of the EC tools.

Fiona consume more memory than any other tools. Lighter is the fastest tool followed by BLESS 2 in all the cases (App. A.5.2).

2.4 Conclusions

The performance of different EC tools was compared using two approaches: the ability of EC tools to correct sequencing errors in Illumina data, and the effects of those corrections on the resulting *de novo* genome assembly quality. We found that EC tools correct a significant fraction of sequencing errors. However, state-of-the-art Illumina assemblers do not always appear to benefit from this. The assembly results for eight different datasets with SPAdes and DISCOVAR show that the prior application of EC tools often does not lead to a significant increase in NGA50, and in fact may result in a lower NGA50. Many erroneous corrections occur in regions that have low read coverage and in the vicinity of highly frequent repeats. Due to the low coverage, error correction tools incorrectly assume the presence of sequencing errors. The repeated elements on the other hand cause erroneous substitutions to be applied. A too aggressive and/or inconsistent transformation of

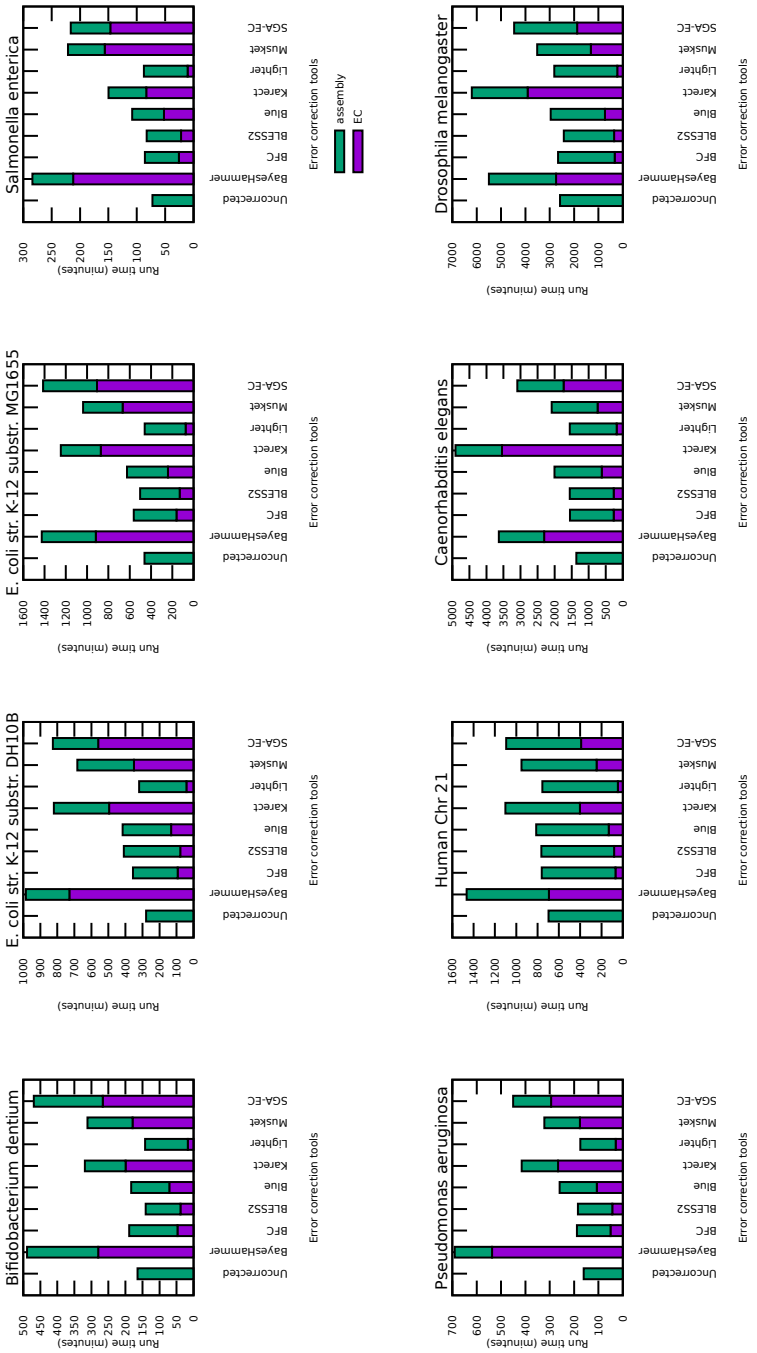


Figure 2.10: Runtime of DISCOVAR plus EC tools.

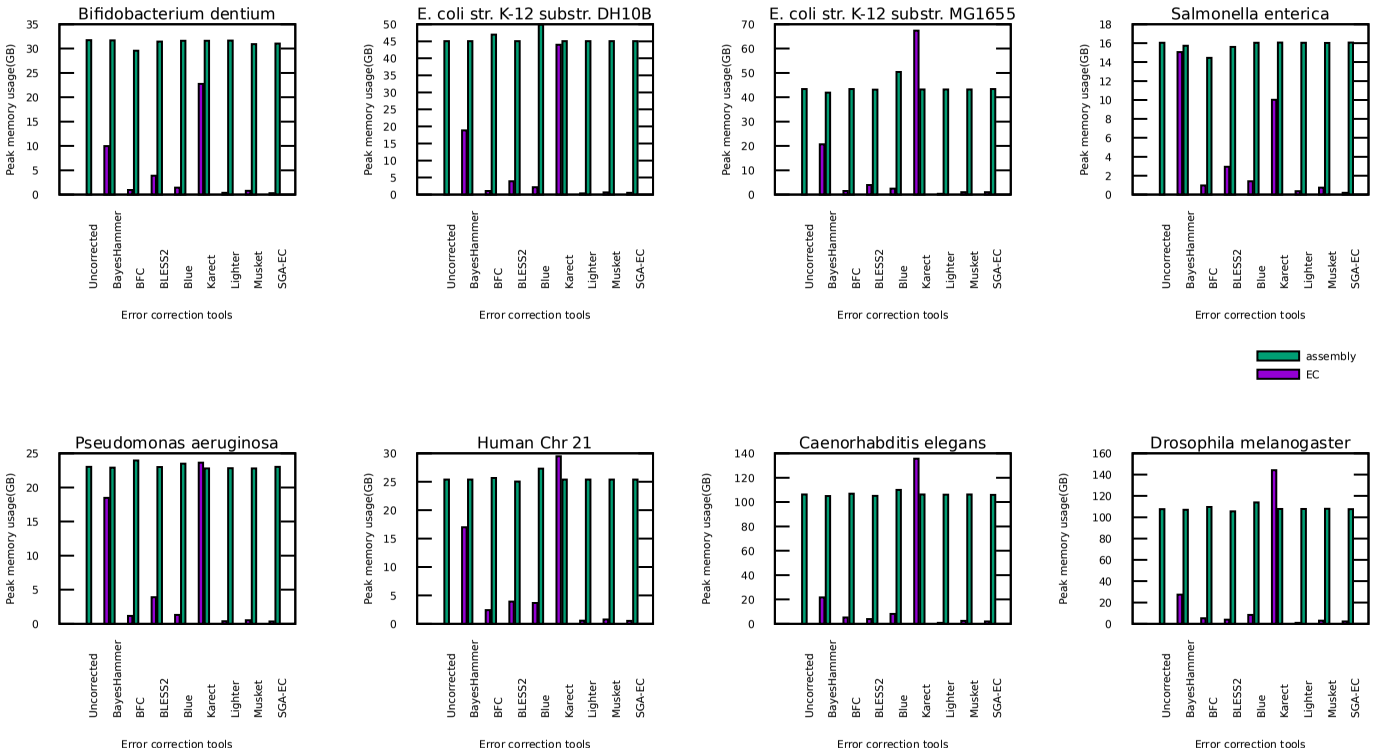


Figure 2.11: Peak memory usage of DISCOVAR and EC tools.

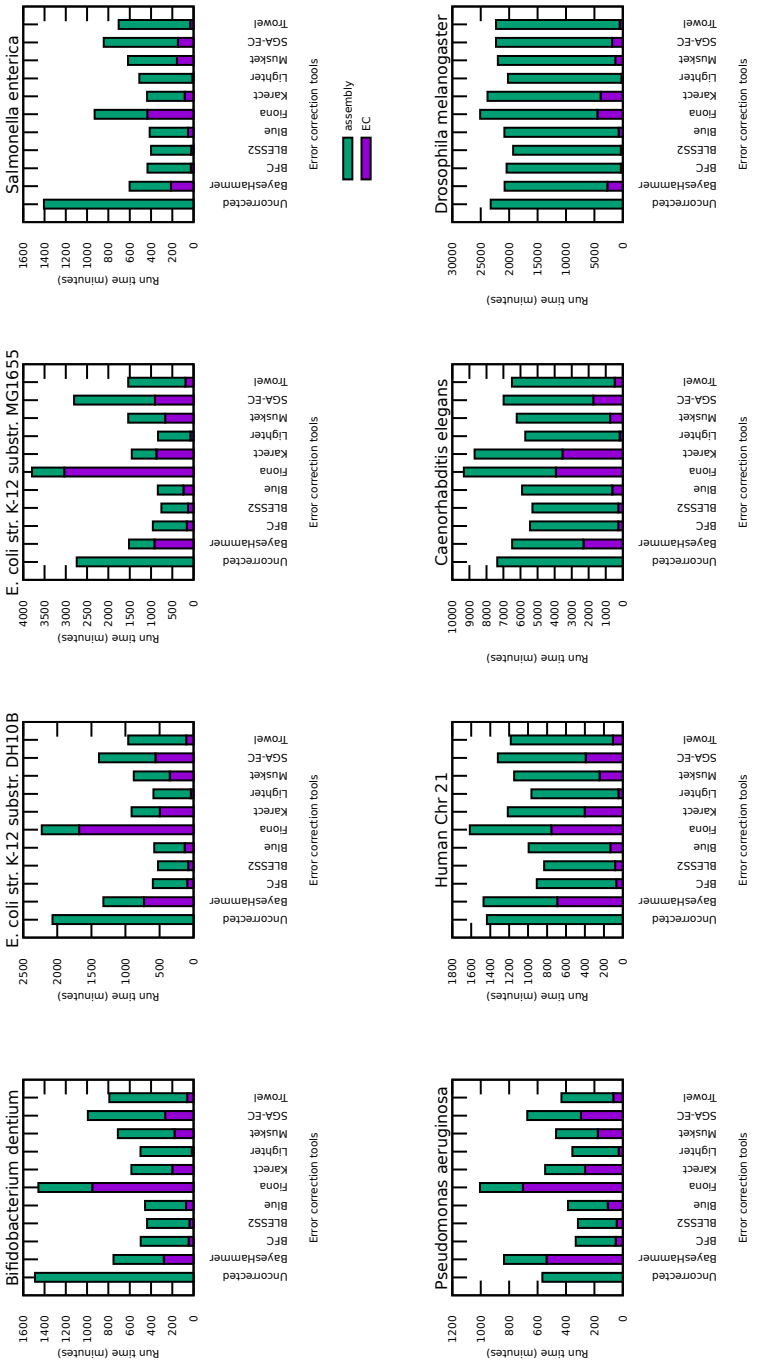


Figure 2.12: Runtime of SPAdes plus EC tools.

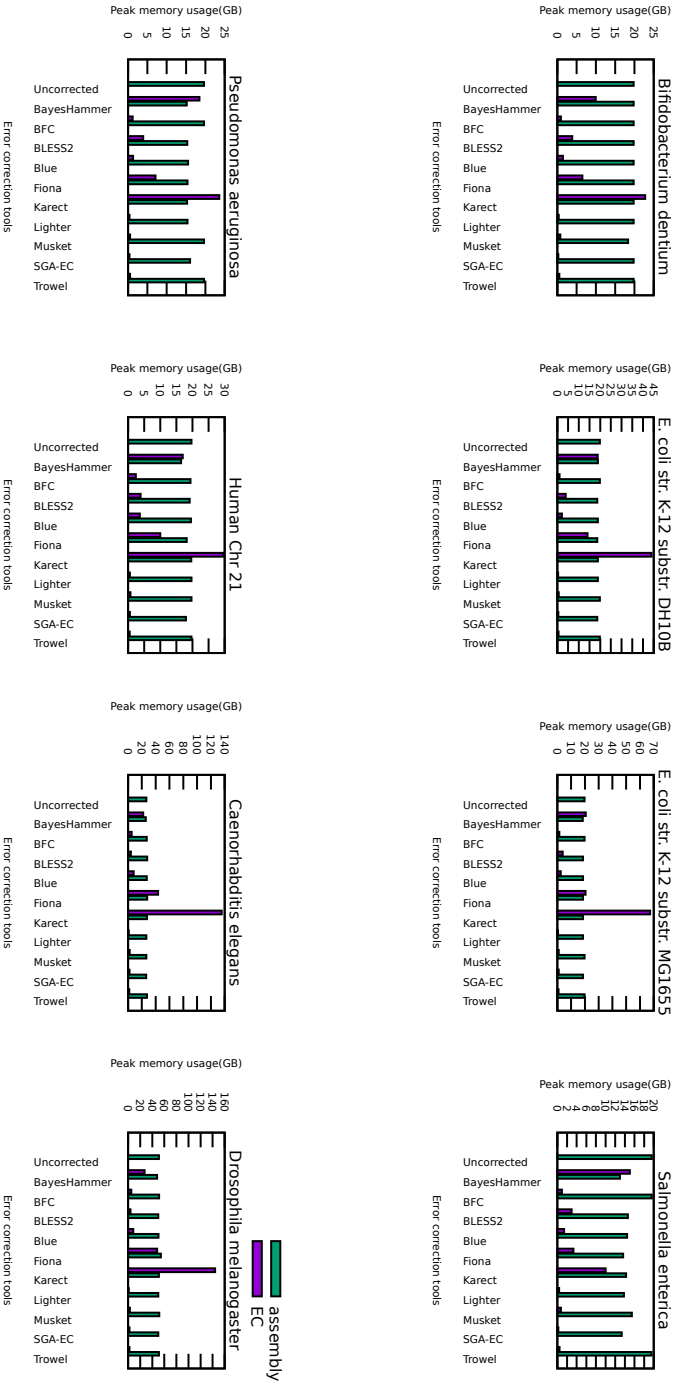


Figure 2.13: Peak memory usage of SPAdes and EC tools.

such reads in such region may lead to loss of information from which no recovery is possible during the assembly process. This inevitably leads to an increased assembly fragmentation. Additionally, the prior use of EC tools does not lead to a major decrease in overall runtime and/or memory requirements compared with the assembly from uncorrected data.

From a methodological point of view, multiple sequence alignment (MSA) based methods might have an advantage over methods that operate on isolated k-mers. MSA-based methods take multiple reads into account when applying substitutions and hence appear to make more consistent corrections across overlapping reads.

We recommend future EC tools to be primarily evaluated on their ability to improve assembly results using state-of-the-art assemblers and sufficiently large datasets. Only a relatively small fraction of sequencing errors are truly impacting the assembly process. It is the behavior of the error correction tool on precisely these cases that will ultimately determine its degree of success.

References

[1] André E Minoche, Juliane C Dohm, and Heinz Himmelbauer. *Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems.* Genome Biol., 12(11):R112, January 2011. 2-2, 3-2, 4-2

[2] Michael G. Ross, Carsten Russ, Maura Costello, Andrew Hollinger, Niall J. Lennon, Ryan Hegarty, Chad Nusbaum, and David B. Jaffe. *Characterizing and measuring bias in sequence data*. Genome Biology, 14(5):R51+, May 2013. 1-9, 2-2, 3-6

[3] Phillip E. Compeau, Pavel A. Pevzner, and Glenn Tesler. *How to apply de Bruijn graphs to genome assembly.* Nature biotechnology, 29(11):987–991, November 2011. 2-2

[4] Daniel R Zerbino and Ewan Birney. *Velvet: algorithms for de novo short read assembly using de Bruijn graphs.* Genome Res., 18(5):821–9, May 2008. 1-12, 2-2, 2-5, 3-3, 3-5, 3-10

[5] Siavash Sheikhizadeh and Dick de Ridder. *ACE: accurate correction of errors using K-mer tries.* Bioinformatics, 31(19):3216–8, October 2015. 1-17, 2-3, 3-3, 3-6, 5-4

[6] Sergey I Nikolenko, Anton I Korobeynikov, and Max a Alekseyev. *BayesHammer: Bayesian clustering for error correction in single-cell sequencing.* BMC Genomics, 14 Suppl 1(Suppl 1):S7, January 2013. 2-3, 3-3

[7] Heng Li. *BFC: correcting Illumina sequencing errors.* Bioinformatics, 31(17):2885–7, September 2015. 1-17, 2-3, 3-3

[8] Yun Heo et al. *BLESS: bloom filter-based error correction solution for high-throughput sequencing reads.* Bioinformatics, 30(10):1354–62, May 2014. 2-3, 3-3

[9] Yun Heo, Anand Ramachandran, Wen-mei Hwu, Jian Ma, and Deming Chen. *Genome analysis BLESS 2 : Accurate , memory-efficient , and fast error correction method*. pages 1–3, 2016. 1-17, 2-3, 3-3

[10] Greenfield et al. *Blue: correcting sequencing errors using consensus and context.* Bioinformatics, 30(19):2723–32, October 2014. 1-17, 2-3, 2-5, 3-3, 3-6

[11] Subrata Saha and Sanguthevar Rajasekaran. *EC: an efficient error correction algorithm for short reads*. BMC Bioinformatics, 16(Suppl 17):S2, 2015. 2-3

[12] Marcel H Schulz, David Weese, Manuel Holtgrewe, Viktoria Dimitrova, Sijia Niu, Knut Reinert, and Hugues Richard. *Fiona: a parallel and automatic strategy for read error correction.* Bioinformatics, 30(17):i356–63, September 2014. 1-17, 2-3, 3-3

[13] Amin Allam et al. *Karect: accurate correction of substitution, insertion and deletion errors for next-generation sequencing data.* Bioinformatics, 31(21):3421–28, July 2015. 1-17, 2-3, 3-3, 3-6, 5-4

[14] Li Song, Liliana Florea, and Ben Langmead. *Lighter: fast and memory-efficient sequencing error correction without counting*. Genome Biol., 15(11):509, November 2014. 2-3, 3-3

[15] Yongchao Liu, Jan Schröder, and Bertil Schmidt. *Musket: a multistage k-mer spectrum-based error corrector for Illumina sequence data.* Bioinformatics, 29(3):308–15, February 2013. 2-3, 3-3

[16] Eric Marinier, Daniel G. Brown, and Brendan J. McConkey. *Pollux: platform independent error correction of single and mixed genomes*. BMC Bioinformatics, 16(1):10, 2015. 1-17, 2-3, 3-3

[17] David R Kelley et al. *Quake: quality-aware detection and correction of sequencing errors.* Genome Biol., 11(11):R116, January 2010. 1-17, 2-3, 3-3

[18] Guillaume Marcais, James A. Yorke, and Aleksey Zimin. *QuorUM: An error corrector for Illumina reads*. PLoS One, 10(6):1–13, 2015. 1-17, 2-3, 3-3

[19] Lucian Ilie and Michael Molnar. *RACER: Rapid and accurate correction of errors in reads.* Bioinformatics, 29(19):2490–3, October 2013. 1-20, 2-3, 3-3

[20] JT Simpson and Richard Durbin. *Efficient de novo assembly of large genomes using compressed data structures*. Genome Res., pages 549–556, 2012. 2-3, 3-3

[21] Eun-Cheon Lim, Jonas Müller, Jörg Hagmann, Stefan R Henz, Sang-Tae Kim, and Detlef Weigel. *Trowel: a fast and accurate error correction module for Illumina sequencing reads.* Bioinformatics, 30(22):3264–5, November 2014. 2-3, 3-3

[22] Andy S. Alic, David Ruzafa, Joaquin Dopazo, and Ignacio Blanquer. *Objective review of de novo stand-alone error correction methods for NGS data*. Wiley Interdiscip. Rev. Comput. Mol. Sci., 6(April), 2016. 2-3, 2-5

[23] Xiao Yang, Sriram P Chockalingam, and Srinivas Aluru. *A survey of error-correction methods for next-generation sequencing.* Brief. Bioinform., 14(1):56–66, January 2013. 2-4, 2-5

[24] Michael Molnar and Lucian Ilie. *Correcting Illumina data.* Brief. Bioinform., 16(4):588–99, July 2015. 2-4

[25] Jared T. Simpson, Kim Wong, Shaun D. Jackman, Jacqueline E. Schein, Steven J M Jones, and Inanç Birol. *ABySS: A parallel assembler for short read sequence data*. Genome Res., 19(6):1117–1123, 2009. 2-4

[26] Thomas Conway, Jeremy Wazny, Andrew Bromage, Justin Zobel, and Bryan Beresford-smith. *Gossamer - A resource-efficient de novo assembler*. Bioinformatics, 28(14):1937–1938, 2012. 2-4

[27] Jason R. Miller, Arthur L. Delcher, Sergey Koren, Eli Venter, Brian P. Walenz, Anushka Brownley, Justin Johnson, Kelvin Li, Clark Mobarry, and Granger Sutton. *Aggressive assembly of pyrosequencing reads with mates*. Bioinformatics, 24(24):2818–2824, 2008. 2-4

[28] Ruibang Luo, Binghang Liu, Yinlong Xie, Zhenyu Li, Weihua Huang, Jianying Yuan, Guangzhu He, Yanxiang Chen, Qi Pan, Yunjie Liu, Jingbo Tang, Gengxiong Wu, Hao Zhang, Yujian Shi, Yong Liu, Chang Yu, Bo Wang, Yao Lu, Changlei Han, David W. Cheung, Siu-Ming Yiu, Shaoliang Peng, Zhu Xiaoqian, Guangming Liu, Xiangke Liao, Yingrui Li, Huanming Yang, Jian Wang, Tak-Wah Lam, and Jun Wang. *SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler*. GigaScience, 1(1):18, 2012. 1-12, 2-4

[29] Neil I Weisenfeld, Shuangye Yin, Ted Sharpe, Bayo Lau, Ryan Hegarty, Laurie Holmes, Brian Sogoloff, Diana Tabbaa, Louise Williams, Carsten Russ, Chad Nusbaum, S Eric, Iain Maccallum, and David B Jaffe. *Comprehensive variation discovery in single human genomes* . 46(12):1350–1355, 2015. 2-4

[30] Yu Peng, Henry C. M. Leung, S. M. Yiu, and Francis Y. L. Chin. *IDBA – A Practical Iterative de Bruijn Graph De Novo Assembler*, pages 426–440. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. 2-4, 3-5

[31] Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey a Gurevich, Mikhail Dvorkin, Alexander S Kulikov, Valery M Lesin, Sergey I Nikolenko, Son Pham, Andrey D Prjibelski, Alexey V Pyshkin, Alexander V Sirotkin, Nikolay Vyahhi, Glenn Tesler, Max a Alekseyev, and Pavel a Pevzner. *SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing.* J. Comput. Biol., 19(5):455–77, May 2012. 1-12, 1-27, 2-5, 3-3

[32] David Laehnemann, Arndt Borkhardt, and Alice Carolyn McHardy. *Denoising DNA deep sequencing data-high-throughput sequencing errors and their correction.* Brief. Bioinform., 17(1):154–79, January 2016. 2-5

[33] Weichun Huang, Leping Li, Jason R Myers, and Gabor T Marth. *ART: a next-generation sequencing read simulator.* Bioinformatics, 28(4):593–4, February 2012. 2-6, 4-10

[34] Arthur L. Delcher, Simon Kasif, Robert D. Fleischmann, Jeremy Peterson, Owen White, and Steven L. Salzberg. *Alignment of whole genomes*. Nucleic Acids Res., 27(11):2369–2376, 1999. 2-7

[35] Guillaume Marcais and Carl Kingsford. *A fast, lock-free approach for efficient parallel counting of occurrences of k-mers.* Bioinformatics, 27(6):764–70, March 2011. 2-7

[36] Heng Li and Richard Durbin. *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 25(14):1754–60, July 2009. 1-25, 2-7, 3-5, 4-2

[37] Alexey Gurevich et al. *QUAST: quality assessment tool for genome assemblies.* Bioinformatics, 29(8):1072–5, April 2013. 1-21, 2-11, 3-5

3

Illumina error correction near highly repetitive DNA regions improves de novo genome assembly.

"... Failure is the condiment that gives success its flavor. [1] *"*

⋆ ⋆ ⋆

Mahdi Heydari, Giles Miclotte, Yves Van de Peer and Jan Fostier

Published in BMC Bioinformatics 19(1): 311, June. 2019

In this chapter, we introduce BrownieCorrector, a targeted error correction tool for highly repetitive regions...

Abstract

Several standalone error correction tools have been proposed to correct sequencing errors in Illumina data in order to facilitate de novo genome assembly. However, in a recent survey, we showed that state-of-the-art assemblers often did not

[1] Truman Capote

benefit from this pre-correction step. We found that many error correction tools introduce new errors in reads that overlap highly repetitive DNA regions such as low-complexity patterns or short homopolymers, ultimately leading to a more fragmented assembly.

We propose BrownieCorrector, an error correction tool for Illumina sequencing data that focuses on the correction of only those reads that overlap short DNA patterns that are highly repetitive in the genome. BrownieCorrector extracts all reads that contain such a pattern and clusters them into different groups using a community detection algorithm that takes into account both the sequence similarity between overlapping reads and their respective paired-end reads. Each cluster holds reads that originate from the same genomic region and hence each cluster can be corrected individually, thus providing a consistent correction for all reads within that cluster.

BrownieCorrector is benchmarked using six real Illumina datasets for different eukaryotic genomes. The prior use of BrownieCorrector improves assembly results over the use of uncorrected reads in all cases. In comparison with other error correction tools, BrownieCorrector leads to the best assembly results in most cases even though less than 2% of the reads within a dataset are corrected. Additionally, we investigate the impact of error correction on hybrid assembly where the corrected Illumina reads are supplemented with PacBio data. Our results confirm that BrownieCorrector improves the quality of hybrid genome assembly as well. BrownieCorrector is written in standard C++11 and released under GPL license. BrownieCorrector relies on multithreading to take advantage of multi-core/multi-CPU systems. The source code is available at :
https://github.com/biointec/browniecorrector.

3.1 Introduction

Illumina platforms generate accurate sequencing data with high throughput at a low financial cost. It is estimated that more than 90% of sequencing data worldwide are generated by Illumina platforms. These data are characterized by a relatively short read length (100-300 bp) and low error rate (1-2% errors). Despite this relatively high accuracy, Illumina data suffers from different kinds of biases, most notably a higher number of sequencing errors towards the end of the reads. The most common errors are substitutions whereas insertions and deletions are less common and particularly occur in homopolymers [1]. Phenomena like crosstalk, phasing, fading or T accumulation can be major sources of errors in Illumina sequencing machines [2].

Due to its cost-efficiency and high accuracy, Illumina data is frequently used for *de novo* genome assembly, sometimes complemented by data generated through other platforms (e.g. Pacific Biosciences, Oxford Nanopore). Short-read assem-

blers typically rely on the de Bruijn graph (DBG) data structure in which overlap between reads is established in a computationally efficient manner through the identification of shared k-mers. Yet, the presence of sequencing errors challenges *de novo* genome assembly tools: sequencing errors result in erroneous nodes and arcs in the DBG, often classified as 'tips' (dead ends), 'bubbles' (parallel paths) and 'chimeric connections' (spurious connections) [3]. As a single sequencing error leads to up to k erroneous k-mers in the DBG, true nodes in the DBG are vastly outnumbered by erroneous nodes. These artifacts highly complicate the task of identifying the path in the graph that represents the original genomic sequence.

Trimming tools are sometimes used as a primary solution to exclude parts of the input data with a lower quality score. However, this further reduces the read length and aggravates the coverage bias. Additionally, indels are often not associated with a low quality score [4], rendering it difficult to remove them by trimming reads. Recently, a number of standalone error correction (EC) tools have been proposed which aim to identify and correct errors in sequencing data: ACE [5], BayesHammer [6], BFC [7], BLESS [8], BLESS 2 [9], Blue [10], Fiona [11], Karect [12], Lighter [13], Musket [14], Pollux [15], Quake [16], QuorUM [17], RACER [18], RECKONER [19], SGA-EC [20] and Trowel [21]. The key idea is that the prior application of EC tools to raw Illumina data provides a cleaner input dataset to the assemblers and subsequently leads to improved assemblies.

However, in a recent survey [22], we showed that state-of-the-art assemblers such as SPAdes [23] and Discovar [24] did not benefit much from this pre-correction step. In fact, the prior use of EC tools was often found to deteriorate assembly results. Most of the EC tools successfully detect and correct a large fraction of sequencing errors, however, most of these errors are harmless to the assembly process as they are properly handled by the assembly tools as well. Specifically, the vast majority of sequencing errors lead to short spurious dead ends or short parallel paths which are easily identified and removed from the DBG based on graph topology and coverage considerations. On the other hand, in certain genomic contexts, EC tools have difficulties identifying sequencing errors and might even introduce new errors. In turn, this may result in misassemblies or assembly breakpoints, leading to shorter contigs/scaffolds. In [22], we reported that misassemblies and breakpoints often occur in two regions: (i) genomic regions with low read coverage where the EC tools incorrectly transform true k-mers into similar k-mers with higher coverage and (ii), the direct vicinity of short, highly repetitive patterns such as homopolymers. We found that EC tools often modify reads that overlap such pattern in an inconsistent manner.

We introduce BrownieCorrector, an EC tool for Illumina sequencing data that focuses solely on the correction of (paired-end) reads that overlap highly repetitive patterns. BrownieCorrector performs four steps: (i) selection of a repetitive k-mer, (ii) read extraction, (iii) read clustering and (iv) per-cluster read error correction.

Initially, it selects a highly repetitive k-mer such as a short poly(A/T) pattern and identifies all paired-end reads for which one of the paired reads contains that k-mer. Next, using a community detection algorithm, it clusters the read pairs such that each cluster contains read pairs that overlap with the same genomic region. As a similarity score for the clustering algorithm, BrownieCorrector computes the overlap alignment score (a variation of the Needleman-Wunsch alignment score [25]) between different read pairs. The read clustering problem is expressed as a community detection problem in graph theory [26]. The Louvain community detection algorithm [27] is applied to an undirected weighted graph whose nodes represent paired-end reads while an edge between two nodes denotes their similarity score. Edges exist only in case the similarity score exceeds a threshold. Hence, the graph is generally sparse. In order to have a robust clustering, BrownieCorrector repeats the community detection process multiple times with different initialization conditions and identifies stable community cores in the network [28]. These cores contain read pairs that were often clustered together in the different runs of the community detection algorithm. The reads are corrected for each cluster separately. From the (paired-end) reads in a particular cluster, BrownieCorrector first constructs the associated DBG. It then performs typical graph cleaning procedures such as tip clipping and bubble detection to remove erroneous nodes and arcs, taking into account both the graph topology and the k-mer frequency (i.e., the number of reads that support each node/arc). Finally, the reads in a cluster are aligned back to the cleaned DBG using BrownieAligner [29]. A similar approach has been already employed for the correction of long reads in LoRDEC [30] and Jabba [31] which has been shown to work effectively even for those errors prone sequencing technologies. This way, sequencing errors are identified and corrected in a consistent manner for all reads within a cluster. This procedure is repeated for all clusters individually.

Correcting smaller groups of reads in each cluster independently from other clusters has a number of advantages over tools that try and correct the entire dataset: first, a small k-mer size (for example $k = 15$) can be used to construct the DBG of each cluster. This allows to establish overlap between individual reads with increased sensitivity without suffering from chimeric connections. This is particularly relevant for low-coverage regions. Second, as each cluster is expected to contain reads from a single genomic region, reads are corrected in a consistent manner.

Note that only a small fraction of pairs are corrected using BrownieCorrector. Read pairs that do not contain highly repetitive k-mer are not modified. The rationale is that state-of-the-art genome de novo assembly tools handle such reads very well. To the best of our knowledge, BrownieCorrector is the first EC tool that uses the paired-read read information in the error correction process, whereas other error correction tools correct reads or even k-mers individually.

3.2 Methods

3.2.1 Error correction tools

The performance of BrownieCorrector is compared with the state-of-the-art EC tools which are all published in 2015 or later: ACE, BLESS 2, BFC, Karect and RECKONER. All tools were run on a machine with four Intel(R) Xeon(R) E5-2698 v3 @ 2.30 GHz CPUs (64 cores in total) and 256 GB of memory. All tools support multi-threading and were run with 64 threads. BLESS 2 failed to finish with 64 cores in some data sets, hence we used 32 cores to get the corrected reads. For all results the default or recommended parameters are used. Parameters and settings are provided in App. B.1 Elapsed (wall clock) time and peak resident memory were measured with the GNU *time* command.

3.2.2 Evaluation tools

SPAdes is a universal *de novo* genome assembler which removes erroneous k-mers through the identification of bubbles and tips in multisized DBGs. In a recent comprehensive study [22], SPAdes is compared to DISCOVAR [24], IDBA [32] and Velvet [3], and it was shown that SPAdes produces longer and more accurate contigs/scaffolds than other assemblers, both with and without pre-correcting reads. SPAdes works with Illumina single-end, paired-end and mate-pair read data and can effectively be used for hybrid assembly where reads from other platforms such as Ion Torrent, PacBio, Oxford Nanopore are also provided. Therefore, in this study, SPAdes is used to evaluate the impact of error correction on *de novo* genome assembly results. SPAdes is provided with a standalone EC tool (BayesHammer) that can apply error correction to the input reads prior to the actual assembly process. All assembly results in this work were obtained without the use of BayesHammer by providing the –only-assembler flag to SPAdes in all cases. Note however that the assembly module within SPAdes also applies error correction procedures directly on the de Bruijn graphs. The resulting assemblies were evaluated using QUAST [33]. In order to determine k-mer frequencies Jellyfish [34] is used. To align reads to the reference genome BWA [35] is used.

3.2.3 Data

Tools are evaluated on six real Illumina eukaryotic datasets for which a high-quality reference genome is available: human chromosomes 14 and 21, two different datasets for fruit fly (*Drosophila melanogaster*), one nematode (*Caenorhabditis elegans*) and one plant organism (*Arabidopsis thaliana*) (see Table 2.2). Genome sizes range from 45.2 Mbp (Homo sapiens chr. 21) to 135 Mbp (*A. thaliana*) while read coverage varies between 29× and 67×. All datasets have fixed read lengths.

Table 3.1 Real datasets used for the evaluation of the error correction tools.

Abbr.	Organism	Reference ID	Reference size	Platform	Insert size mean	Insert size STD	Cov.	Number of Reads	Read length mean	Ref.	Dataset ID
R1	*Homo sapiens* chr. 21	HG19	40 Mbp	Illumina	312	14	33×	13 486 136	100 bp	[10, 22]	Ill. Data library
R2	*Homo sapiens* chr. 14	HG14	104 Mbp	Illumina	158	17	35×	36 504 800	101 bp	[12]	GAGE
R3	*Caenorhabditis elegans*	WS222	97 Mbp	Illumina	173	16	58×	57 721 732	101 bp	[5, 22, 36]	SRR543736
R4	*Drosophila melanogaster*	Release 5	116 Mbp	Illumina	281	92	52×	63 014 762	100 bp	[5, 22, 36]	SRR823377
R5	*Drosophila melanogaster*	Release 5	116 Mbp	Illumina	598	39	64×	75 938 276	101 bp	[5, 36]	SRR988075
R6	*Arabidopsis thaliana*	TAIR10	116 Mbp	Illumina	477	18	72×	93 429 346	90 bp	[37]	SRR1174256
P1	*Drosophila melanogaster*	Release 5	116 Mbp	PacBio	n/a	n/a	10×	169 923	7374 bp	[38]	SRR1204466
P2	*Arabidopsis thaliana*	TAIR10	116 Mbp	PacBio	n/a	n/a	13×	187 292	8298 bp	[38]	SRR1284707

In addition, two publicly available PacBio datasets for *D. melanogaster* and *A. thaliana* are used to evaluate the impact of EC tools on hybrid assembly (See App. B.2).

Note the absence of bacterial datasets. As also observed in [22], error correction often does not have a significant impact on the assembly quality for such small genomes.

3.2.4 Targeted error correction

The targeted error correction pipeline has four main steps (Fig. 3.1). The first step is the k-mer selection procedure. Our experimental investigation shows that most of the breakpoints in the assembled contigs occur in the direct vicinity of low-complexity k-mers such as poly(A/T) or poly(C/G) (see App. B.3). There are two main reasons for this. Firstly, these k-mers are highly repetitive in the datasets. For example, the poly(A/T) 15-mer has the highest frequency among all 15-mers in 3 out of 6 datasets (see App. B.3). Such highly repetitive k-mers form hubs in the DBG through which a vast number of reads pass. They represent the central node in a densely connected subgraph of the DBG for which the resolution of the true path is very complex. Secondly, it has been observed in Illumina sequencing data that GC-rich or GC-depleted regions such as homopolymers are more prone to sequencing errors, especially insertions and deletions [39]. Fig. 3.2 shows that the average quality scores of reads that contain a poly(A/T) or poly(C/G) pattern are much lower than average. As such, those reads generally contain more sequencing errors than average. Particularly, in dataset D2, the average quality score of bases in reads that contain a poly(A/T) pattern is 20, whereas the average quality score for regular reads is 31. This means that a base of a read that contains a poly(A/T) sequence is about 10 times more likely to be erroneous than average. Therefore, it is very difficult for the assembler to establish a connection between reads in these regions which explains why the produced contigs by SPAdes often end with a poly(A/T) k-mer (see Table 2 in App. B.3). Reads with other kinds of low-complexity repeats appear less susceptible for an excessive number of sequencing errors.

In this paper we correct only read pairs for which one of the reads contains a poly(A/T) 15-mer or longer. The effect of our proposed error correction procedure

Figure 3.1 Overview of the first three steps of BrownieCorrector's pipeline. Read pairs for which one read contains a highly repetitive k-mer are extracted and clustered based on the sequence similarity between different read pairs. Each cluster is expected to contain reads that were derived from a single genomic regions.

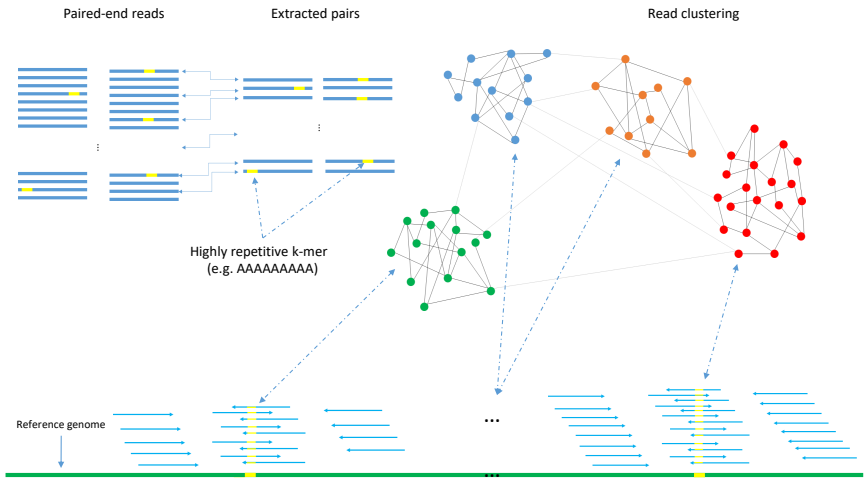
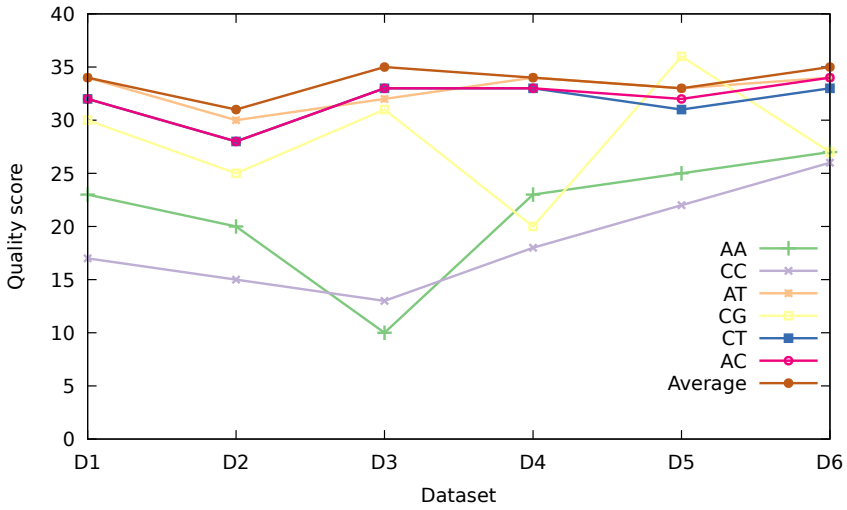


Figure 3.2 The average quality score of bases in reads for different polymers and a group of randomly sampled reads.

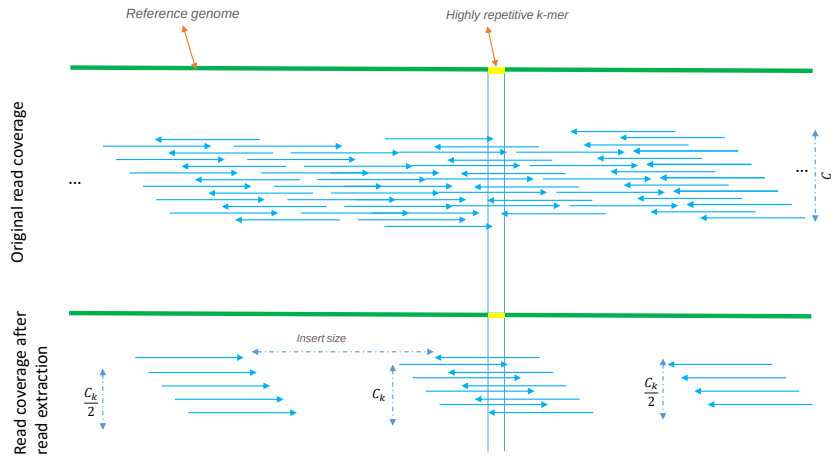


for other examples of low-complexity k-mers also has been investigated and the corresponding results are reported in supplementary data.

The second step in the pipeline is the read extraction. Reads that contain a

specific k-mer can easily be extracted in a single pass over the dataset. The expected number of reads that fully cover a k-mer occurrence in the genome can be computed as follows. Let C denote the coverage, i.e., the average number of reads that covers any base of that k-mer occurrence. Some of these reads will not cover the entire k-mer or might contain sequencing errors and hence they will not be extracted in this step. The expected number of extracted reads C_k that fully overlap a k-mer in the genome is given by $C_k = \frac{l-k+1}{l} C(1-e)^k$ where l is the read length and e denotes the error rate (see App. B.4). Reads are extracted in pairs and since these paired reads can occur on either side of the k-mer, the expected number of reads covering the flanking regions is $C_k/2$. Due to fragment length (insert size) variability, these paired reads might be more spread out over the flanking regions. Fig. 3.3 provides a schematic representation of the coverage distribution after read extraction where the expected number of pairs in one cluster is C_k.

Figure 3.3 While C is the initial coverage (top), the expected number of reads that fully cover a selected k-mer is C_k. Depending on the insert size and the insert size variability, the left and right flanking regions that are covered by the paired reads have a coverage of $C_k/2$ or lower.



The third step in the pipeline is read pair clustering. The idea is to partition the read pairs into distinct clusters in such a way that all read pairs within a cluster originate from the same genomic region. The expected number of read pairs in each cluster is C_k. BrownieCorrector uses the Louvain community detection algorithm, a very fast and memory efficient hierarchical graph clustering algorithm [27]. It is based on the greedy maximization of modularity and can

handle large-scale networks ($N > 10^8$) [40]. The Louvain community detection algorithm takes as input a graph where nodes represent read pairs and where arcs between nodes represent the similarity score between read pairs. This similarity score is obtained by computing the pairwise overlap alignment score. The overlap alignment score represents the highest alignment score between a prefix of one sequence and the suffix of another, hence, trailing and leading gaps in the alignment are not penalized. Note that not only the sequence similarity between the two reads that contain the repetitive k-mer is taken in account, but also potential overlap between their respective paired reads. We found the information contained in the paired reads to be valuable to obtain robust and homogeneous clusters.

Computing the overlap alignment score between all pairs of reads has a quadratic time dependency on the number of read pairs and can hence be time-consuming for a large number of pairs. In BrownieCorrector, the read alignment score is only computed between read pairs that share at least one non-repeated k-mer, i.e., a k-mer for which the coverage is about C_k. This heuristic avoids the computation of alignment scores between read pairs with apparent low sequence similarity. This also means that the input graph for the community clustering algorithm is generally very sparse.

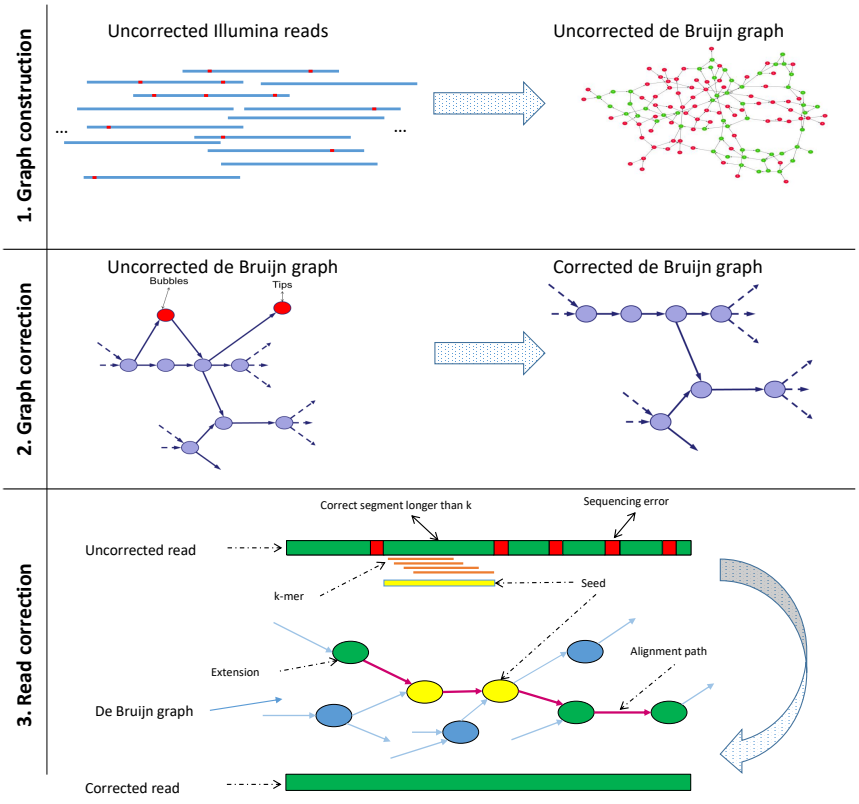
The Louvain community detection algorithm outputs clusters for which the nodes in each cluster are densely connected while having only relatively few connections between nodes that belong to different clusters. The algorithm is non-deterministic which means that in every run, it may output different clusters. In order to reduce the impact of non-deterministic behavior of the algorithm and improve the robustness of the clusters, BrownieCorrector repeats the clustering procedure several times. The stable core communities are then established as explained in [28].

The final step of the pipeline involves correcting the reads for each cluster independently. This step has three stages (see Fig. 3.4): i) construction of the DBG from input sequences; ii) correction of the DBG based on topology and coverage considerations; iii) correction of the input reads by aligning them to the corrected DBG.

i) Reads are first assembled in a DBG. Given a user-specified value for k, all k-mers are extracted from the reads and a DBG is constructed. To reduce memory requirements, k-mers are encoded by $2k$ bits and stored in a memory-efficient hash map with only 2 bits overhead per entry. Overlap between k-mers is encoded by 8 bits: 4 bits to indicate if the k-mers can be left-extended with A, C, T or G and similarly 4 bits to represent right overlap. Linear paths in the graph are contracted to bigger nodes (unitigs) and various statistics such as length (number of k-mers in a node), average k-mer coverage (average number of reads that cover a k-mer in the node) are computed for each node.

ii) Whereas k-mer spectrum-based EC methods such as Quake identify errors

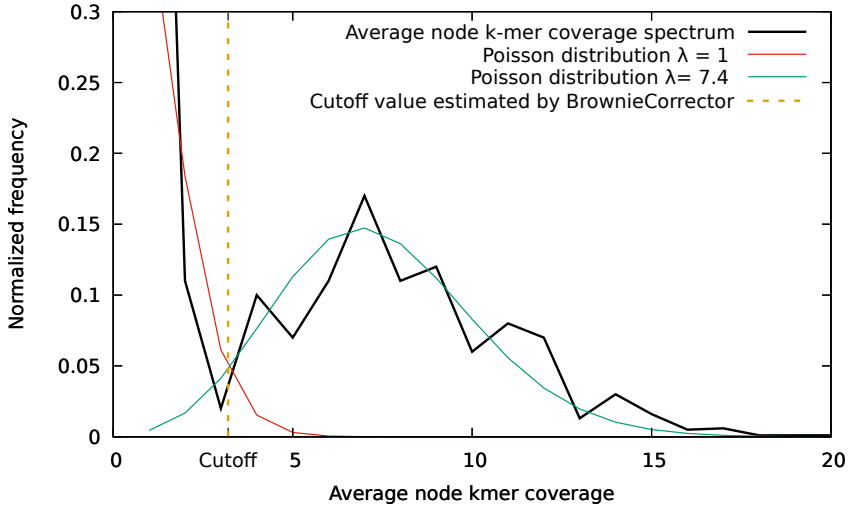
Figure 3.4 The final step of the BrownieCorrector pipeline: (1) de Bruijn Graph is built from the uncorrected reads in a cluster. Uncorrected reads contain sequencing errors which result in the appearance of erroneous k-mers and subsequently erroneous nodes/arcs in the graph; (2) erroneous nodes (colored in red) are detected and removed from the graph based on coverage and graph toplogy. Such erroneous nodes often appear as tips or bubbles; (3) reads are aligned individually to the corrected graph and mismatches and indels in the reads are detected and fixed with the correct path in the graph.



based on (relative) k-mer abundances, erroneous nodes in the DBG are identified by BrownieCorrector based on graph topology and coverage considerations, as conceptually described by Zerbino and Birney [3]. For example, a true k-mer with a low abundance might be incorrectly classified as erroneous when judging solely on k-mer spectrum. By taking into account the context in which the *k-mer* occurs, it could, for example, become clear that this k-mer is part of a linear path in the DBG and that no parallel path exists with higher coverage. As such, the k-mer can be correctly classified as a true k-mer. Vice versa, an erroneous k-mer with a higher

abundance can be detected because of topology considerations: either because the k-mer is part of a dead-end in the graph (a tip) or because it forms a path parallel to the correct sequence path. BrownieCorrector adopts a conservative, multi-round approach, avoiding the removal of true nodes as much as possible. A tip or a bubble is labeled as an erroneous node and will be removed if its length is less than the *maxErrorNodeLen* value and its average node k-mer coverage is less than the *cutoff* value. The value of *maxErrorNodeLen* is set to $avgReadLen - k$ where *avgReadLen* is the average read length and k is the k-mer size. The histogram of average node k-mer coverage for all the nodes in DBG shows a mixture of two distributions: one that represents erroneous nodes and one that represents correct nodes (see Fig. 3.5). Using the expectation-maximization algorithm, a mixture of two Poisson distributions is fit: a distribution of erroneous nodes with mean λ_e and a distribution of correct nodes with mean λ_c. BrownieCorrector computes the k-mer $cutoff$ value at the intersection point of the two distributions.

Figure 3.5 Real example of a k-mer frequency spectrum that is a superposition of two distributions corresponding to real and erroneous k-mers, respectively. A model of two Poisson distributions is fit to the data using the expectation-maximization algorithm. The coverage cutoff is established at the intersection of the two distributions.



iii) The original reads are aligned back to the corrected DBG using a seed-and-extend paradigm. In case a read contains at least one true k-mer, this k-mer is used as a seed that uniquely maps the read to a certain node in the DBG. A depth-first search on the graph is performed to align both ends of the read beyond the seed(s). Pairwise alignments are used to find the optimal alignment path. Branch-and-bound conditions are used to limit the search space. We refer to [29] for a

Table 3.2 NGA50 of respectively contigs (top) and scaffolds (bottom) assembled by SPAdes before and after error correction.

Tools	D1	D2	D3	D4	D5	D6	D7	D8	D9
					Contig NGA50				
Uncorrected	10 876	5 451	6 325	50 833	35 924	40 802	80 752	85 003	65 138
ACE	11 375	8 475	3 116	29 126	20 032	34 273	55 391	65 163	62 161
BFC	11 672	9 488	6 307	49 089	27 365	40 910	77 526	78 985	64 709
BLESS2	9 183	7 737	2 969	25 133	17 133	29 968	61 609	60 574	55 639
BrownieCorrector	**13 334**	**11 015**	**6 328**	52 152	**38 670**	**45 400**	83 397	**88 877**	**71 788**
Karect	12 507	10 103	6 295	**54 106**	29 286	41 391	**85 226**	85 881	68 873
Reckoner	9 154	6 440	6 281	41 977	26 296	39 605	58 176	71 724	56 734
					Scaffold NGA50				
Uncorrected	11 377	5 668	6 419	60 714	59 591	41 833	96 381	109 785	84 659
ACE	12 135	8 597	3 143	35 425	40 860	39 895	62 981	93 602	83 138
BFC	12 294	9 698	6 392	59 124	54 093	41 818	91 577	110 748	82 101
BLESS2	10 034	7 909	3 012	34 856	36 316	38 431	73 377	86 526	74 447
BrownieCorrector	**14 155**	**11 570**	**6 420**	61 474	**65 174**	**46 678**	96 385	118 192	**96 916**
Karect	13 528	10 298	6 377	**63 400**	59 526	42 256	**101 753**	**124 215**	90 661
Reckoner	9 670	6 509	6 354	47 781	50 834	40 779	67 061	99 419	71 646

more detailed description of the read-to-graph alignment procedure.

3.3 Results

3.3.1 Ability of EC tools to improve genome assembly

Table 2.4 shows the contig and scaffold NGA50 values for nine datasets and the different EC tools. The NGA50 denotes the characteristic length of the assembled contigs/scaffolds that can be contiguously aligned to the reference genome. These contigs/scaffolds thus contain no major structural assembly flaws and a higher NGA50 hence implies a better quality assembly. The first six columns show the assembly results for the Illumina datasets (D1=R1,..., D6=R6), while the last three columns refer to the hybrid assembly of Illumina and PacBio datasets (D7=R4+P1, D8=R5+P1 and D9=R6+P2).

Overall, BrownieCorrector shows the best performance and has the highest contig/scaffold NGA50 in 13 out of 18 cases while Karect has the highest NGA50 in the 5 remaining cases. In those cases, BrownieCorrector is second best. Pre-correcting reads with BrownieCorrector leads to improved assembly results in for all datasets. The other EC tools (ACE, BLESS 2, BFC and Reckoner) show mixed results. D2 is the only dataset for which all EC tools improve the contig/scaffold NGA50 over the use of uncorrected data. For datasets D3 and D5, all EC tools except BrownieCorrector deteriorate the assembly results. For some EC tools this leads to significantly shorter contigs/scaffolds. In 12 out of 18 cases, the contig and scaffold NGA50 values obtained from uncorrected data are among the top

Table 3.3 NGA50 of respectively contigs (top) and scaffolds (bottom) assembled by SPAdes after error correction by both BrownieCorrector and Karect.

Tools	D1	D2	D3	D4	D5	D6	D7	D8	D9
					Contig NGA50				
BrownieCorrector	13 334	11 015	**6 328**	52 152	**38 670**	45 400	83 397	**88 877**	71 788
Karect	12 507	10 103	6 295	54 106	29 286	41 391	85 226	85 881	68 873
Karect+BrownieCorrector	**13 526**	**12 409**	6 297	**56 046**	30 557	**45 423**	**89 065**	87 822	**74 620**
					Scaffold NGA50				
BrownieCorrector	14 155	11 570	**6 420**	61 474	**65 174**	**46 678**	96 385	118 192	96 916
Karect	13 528	10 298	6 377	63 400	59 526	42 256	101 753	124 215	90 661
Karect+BrownieCorrector	**14 613**	**12 795**	6 380	**65 857**	62 706	46 332	**103 872**	**126 449**	**104 037**

3 highest values (though often below those of BrownieCorrector and Karect). It shows that SPAdes, which uses advanced paired and multi-sized de Bruijn graphs, uses accurate built-in error correction algorithms in the assembly process as well.

The results indicate that BrownieCorrector and Karect are the only reliable EC tools that perform consistently across different datasets. Table 3.3 shows the contig/scaffold NGA50 when applying both BrownieCorrector and Karect to the Illumina data. The idea is that BrownieCorrector first corrects only reads with a highly repetitive k-mer and that Karect corrects the other reads. We observe that the combined effect of the two error correction tools further raises the assembly quality except for the cases where Karect already performs poorly. This indicates that both tools are complementary to some degree. The improvements in NGA50 over the use of uncorrected data (averaged over all datasets) shows that the combined use of BrownieCorrector and Karect leads to the highest positive impact on the quality of contigs/scaffolds (+21%/+25%) while BrownieCorrector (+18%/+19%), Karect (+11%/+15%), and BFC (+5%/+7%) are the second, third and forth best tool, respectively. On the other hand, BLESS2 (-25%/-19%), ACE (-17%/-14%) and Reckoner(-11%/-10%) deteriorate the quality of assembly on average (see App. B.5.1).

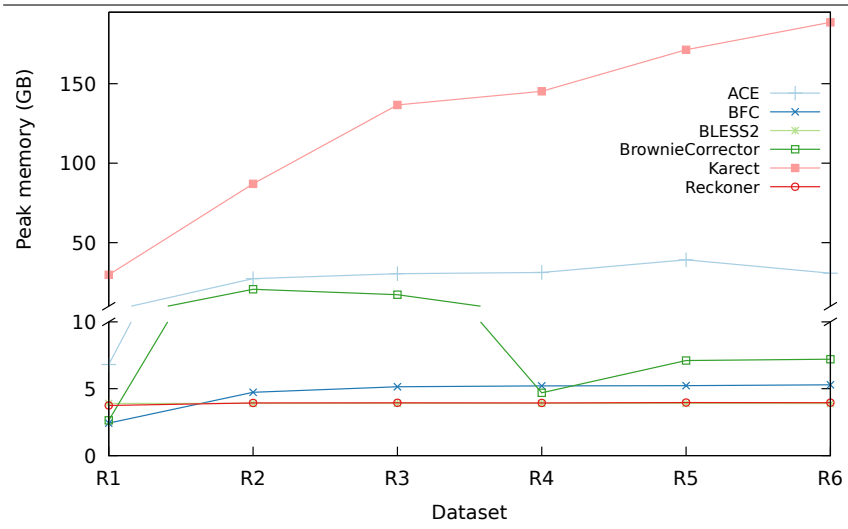
We additionally investigated the impact of error correction by using other highly repetitive k-mers. This time a poly(C/G) pattern was utilized to extract the reads. The results show that correcting these reads with BrownieCorrector has a smaller impact on the assembly quality except for dataset D3 in which the NGA50 of both contigs and scaffolds is higher than the values in Table 2.4 (see App. B.5.2). This can be explained by the fact that for most datasets the poly(C/G) k-mer is much less frequent than poly(A/T) pattern and hence SPAdes benefits less from the error correction of those reads. The correction of reads with a poly(AC/GT) 15-mer does not lead to improved assemblies, even though a poly(AC/GT) 15-mer is more frequent than a poly(C/G). This is because reads that contain these poly (AC/GT) k-mers do not suffer from the error bias that can be observed in reads containing a poly(A/T) or poly(C/G) pattern. Finally, we examined the impact of

the number of iterations in the stable core detection procedure on the final assembly result in D1. The default value for this parameter is 20, which is compared to 1 (when the stable core detection is disabled), 5, 10 and 30. The result shows that using the stable core improves the accuracy, however, BrownieCorrector is robust and performs good as well for other values (see App. B.5.3). The detailed Quast reports for all datasets and EC tools for both contigs and scaffolds are provided in App. B.5.4 and App. B.5.5.

3.3.2 Time and space requirements

Fig. 3.6 shows the memory usage of the EC tools (see App. B.5.6 for detailed tables). Reckoner, BLESS 2, and BFC are the most memory-efficient tools; memory usage of ACE and BrownieCorrector is comparable and Karect has the highest memory requirements. Fig. 3.7 compares the runtime of the different EC tools for each dataset. Reckoner, BLESS 2, and BFC are the fastest tools whereas ACE, Karect, and BrownieCorrector are somewhat slower. Generally speaking, Reckoner, BLESS 2 and BFC are fast and memory efficient.

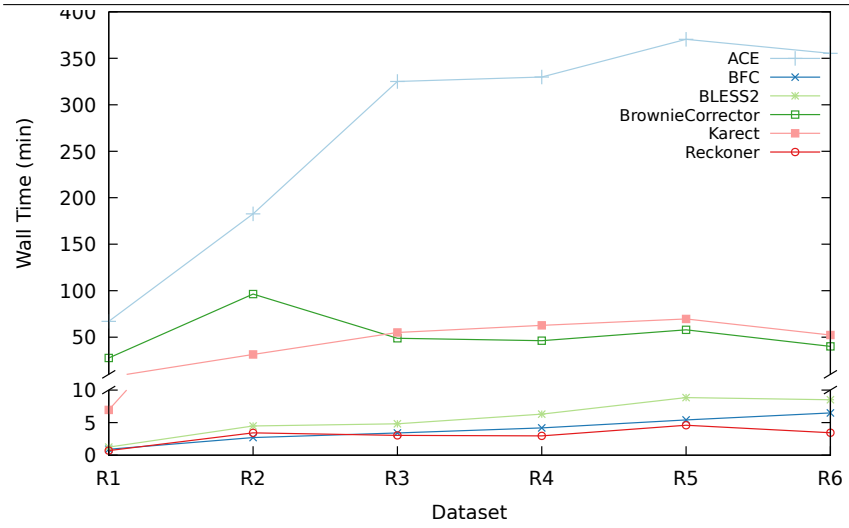
Figure 3.6 Peak memory usage. Peak memory usage of the EC tools.



3.4 Discussion

Although BrownieCorrector corrects only a small fraction of the reads (less than 2%, see App. B.5.2), results show that it performs well for a diverse set of organisms and even for relatively low coverage data (33×). The only parameter that can

Figure 3.7 Runtime. Runtime of the EC tools.



negatively affect the performance of BrownieCorrector is a larger standard deviation of fragment length (insert size). In that case, there is less overlap between paired reads and the identification of homogeneous clusters is more challenging. For example, BrownieCorrector performs worse than Karect in datasets D4 and D7 which is due to the fact that the standard deviation for the R4 Illumina dataset is 92, which is relatively high compared to the other datasets.

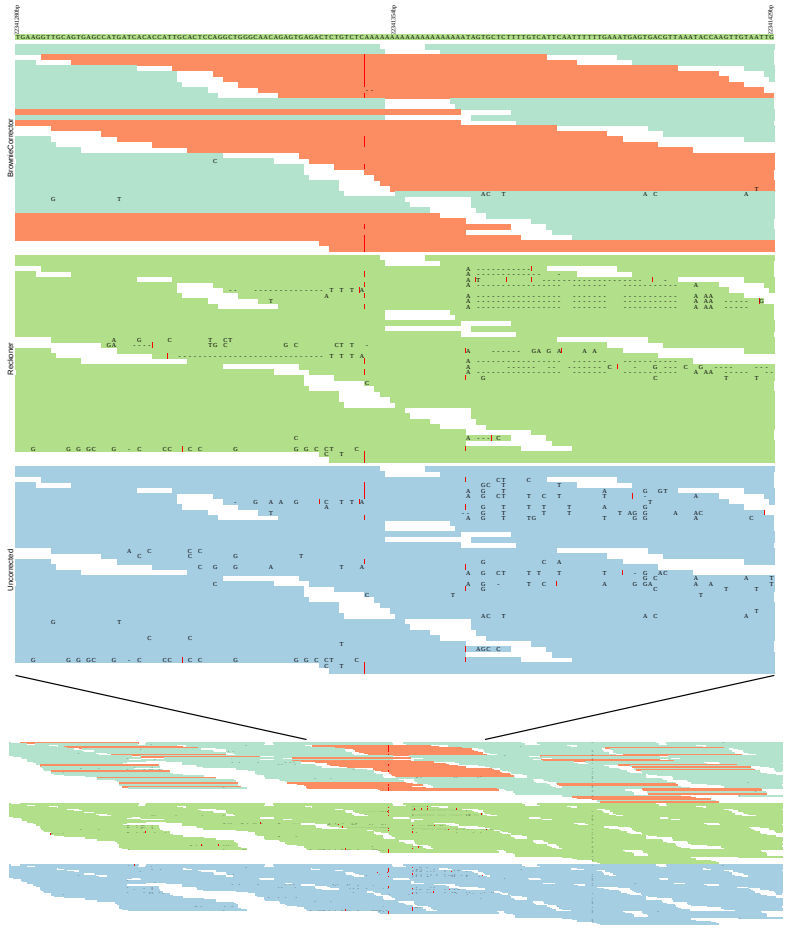
The main advantage of BrownieCorrector over other tools lies in its use of paired-end read information. Fig. 3.8 shows a specific case for dataset D1 where the use of BrownieCorrector resolves a breakpoint near a poly(A/T) pattern that occurs when using uncorrected or Reckoner-corrected data. To create this figure, all uncorrected reads were aligned to the reference genome using BWA and the read pairs that overlap the breakpoint were extracted. Next, the corresponding reads corrected by both BrownieCorrector and Reckoner were obtained. BrownieCorrector corrects only the reads that contain a poly(A/T) 15-mer (shown in orange). Although the average error rate in Illumina sequencing data is around (1-2%), we observe a much higher error rate in the vicinity of the poly(A/T) 15-mer. This is already confirmed by the low average quality scores of reads that contain poly(A/T) patterns (see Table 2 in App. B). This high error rate renders SPAdes unable to correctly bridge the breakpoint. Also EC tools that do not exploit the paired-read information are likely to correct these highly erroneous reads in an inconsistent manner as exemplified for Reckoner. In contrast, using the paired reads, BrownieCorrector can still correctly cluster and correct these low-quality reads.

3.5 Conclusions

We propose BrownieCorrector, a targeted error correction tool that corrects Illumina sequencing errors in paired-end reads that contain highly-repetitive patterns such as short homopolymers. Such reads form densely connected subnetworks in the de Bruijn graph, which, in the presence of sequencing errors, are difficult to resolve, ultimately leading to a fragmented assembly. BrownieCorrector uses the entire read sequence as well as the paired-end read information to cluster read pairs in homogeneous groups, where the paired-end reads in each group originate from the same genomic region. Reads in each cluster are corrected independently such that a consistent correction is achieved for all reads within each cluster. Despite the fact that BrownieCorrector corrects only a small fraction of the input reads, results indicate it outperforms other error correction tools in terms of contiguity of the assembled contigs and scaffolds. This observation lends support to the idea that error correction tools should focus their efforts on the correction of 'difficult' sequencing errors. Indeed, the utility of error correction tools lies in their ability to improve the quality of downstream applications. We believe that for future EC tools, it is ultimately more beneficial to try and correct problematic regions really well, rather than designing a method that performs well across the entire genome but fails to produce consistent corrections for certain regions. By limiting the application of these algorithms, which perhaps need more CPU cycles, to these specific regions, the computational cost can still be kept under control. Such algorithms likely need to exploit the paired-end read information to ensure a consistent error correction.

We also investigated the impact of BrownieCorrector in a hybrid genome assembly setup where Illumina sequencing data is combined with PacBio data. Our results show that the use of BrownieCorrector-corrected Illumina reads along with PacBio data leads to better assembly results in this case as well. One of the advantages of BrownieCorrector's pipeline is its modularity where each step can be replaced by a method of choice. For example, the Louvain community detection algorithm can easily be replaced by another clustering algorithm, other EC tools can be used to correct clusters or different metrics can be used to infer the similarity score between pairs of reads. We believe this flexibility allows the pipeline to further evolve in the future.

Figure 3.8 Alignment of BrownieCorrector-corrected, Reckoner-corrected and uncorrected paired reads in the neighborhood of a contig breakpoint: the first track contains part of the reference genome, which is assembled into a single contig from BrownieCorrector-corrected data but breaks into two contigs using Reckoner-corrected or uncorrected data. The second track (BrownieCorrector) shows the alignment of the BrownieCorrector-corrected reads. The only the reads in orange are corrected by BrownieCorrector. The third track (Reckoner) shows the alignment of the Reckoner-corrected reads. The forth track (Uncorrected) shows the alignment of uncorrected reads. Mismatches in the sequencing data are indicated with letters whereas an insertion is shown with a | sign and a deletion is shown with a - sign.



References

[1] André E Minoche, Juliane C Dohm, and Heinz Himmelbauer. *Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems.* Genome Biol., 12(11):R112, 2011. 2-2, 3-2, 4-2

[2] Kensuke Nakamura, Taku Oshima, Takuya Morimoto, Shun Ikeda, Hirofumi Yoshikawa, Yuh Shiwa, Shu Ishikawa, Margaret C. Linak, Aki Hirai, Hiroki Takahashi, Md. Altaf-Ul-Amin, Naotake Ogasawara, and Shigehiko Kanaya. *Sequence-specific error profile of Illumina sequencers*. Nucleic Acids Research, 39(13):e90–e90, 2011. 1-9, 3-2

[3] Daniel R Zerbino and Ewan Birney. *Velvet: algorithms for de novo short read assembly using de Bruijn graphs.* Genome Res., 18(5):821–9, 2008. 1-12, 2-2, 2-5, 3-3, 3-5, 3-10

[4] Melanie Schirmer, Rosalinda D'Amore, Umer Z. Ijaz, Neil Hall, and Christopher Quince. *Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data*. BMC Bioinformatics, 17(1):125, 2016. 1-9, 3-3

[5] Siavash Sheikhizadeh and Dick de Ridder. *ACE: accurate correction of errors using K-mer tries.* Bioinformatics, 31(19):3216–8, 2015. 1-17, 2-3, 3-3, 3-6, 5-4

[6] Sergey I Nikolenko, Anton I Korobeynikov, and Max a Alekseyev. *BayesHammer: Bayesian clustering for error correction in single-cell sequencing.* BMC Genomics, 14 Suppl 1(Suppl 1):S7, 2013. 2-3, 3-3

[7] Heng Li. *BFC: correcting Illumina sequencing errors.* Bioinformatics, 31(17):2885–7, 2015. 1-17, 2-3, 3-3

[8] Yun Heo, Xiao-Long Wu, Deming Chen, Jian Ma, and Wen-Mei Hwu. *BLESS: bloom filter-based error correction solution for high-throughput sequencing reads.* Bioinformatics, 30(10):1354–62, 2014. 2-3, 3-3

[9] Yun Heo, Anand Ramachandran, Wen-Mei Hwu, Jian Ma, and Deming Chen. *BLESS 2: accurate, memory-efficient and fast error correction method.* Bioinformatics, 32(15):2369–2371, 2016. 1-17, 2-3, 3-3

[10] Paul Greenfield, Konsta Duesing, Alexie Papanicolaou, and Denis C Bauer. *Blue: correcting sequencing errors using consensus and context.* Bioinformatics, 30(19):2723–32, 2014. 1-17, 2-3, 2-5, 3-3, 3-6

[11] Marcel H Schulz, David Weese, Manuel Holtgrewe, Viktoria Dimitrova, Sijia Niu, Knut Reinert, and Hugues Richard. *Fiona: a parallel and automatic strategy for read error correction.* Bioinformatics, 30(17):i356–63, 2014. 1-17, 2-3, 3-3

[12] Amin Allam, Panos Kalnis, and Victor Solovyev. *Karect: accurate correction of substitution, insertion and deletion errors for next-generation sequencing data.* Bioinformatics, 31(July):3421–3428, July 2015. 1-17, 2-3, 3-3, 3-6, 5-4

[13] Li Song, Liliana Florea, and Ben Langmead. *Lighter: fast and memory-efficient sequencing error correction without counting*. Genome Biology, 15(11):509, 2014. 2-3, 3-3

[14] Yongchao Liu, Jan Schröder, and Bertil Schmidt. *Musket: a multistage k-mer spectrum-based error corrector for Illumina sequence data.* Bioinformatics, 29(3):308–15, 2013. 2-3, 3-3

[15] Eric Marinier, Daniel G. Brown, and Brendan J. McConkey. *Pollux: platform independent error correction of single and mixed genomes*. BMC Bioinformatics, 16(1):10, 2015. 1-17, 2-3, 3-3

[16] David R Kelley, Michael C Schatz, and Steven L Salzberg. *Quake: quality-aware detection and correction of sequencing errors.* Genome Biol., 11(11):R116, 2010. 1-17, 2-3, 3-3

[17] Guillaume Marcais, James A. Yorke, and Aleksey Zimin. *QuorUM: An error corrector for Illumina reads*. PLoS One, 10(6):1–13, 2015. 1-17, 2-3, 3-3

[18] Lucian Ilie and Michael Molnar. *RACER: Rapid and accurate correction of errors in reads.* Bioinformatics, 29(19):2490–3, 2013. 1-20, 2-3, 3-3

[19] Maciej Długosz and Sebastian Deorowicz. *RECKONER: read error corrector based on KMC*. Bioinformatics, 33(7):1086–1089, 2017. 1-17, 3-3

[20] J. T. Simpson and R. Durbin. *Efficient de novo assembly of large genomes using compressed data structures*. Genome Research, 22(3):549–556, 2012. 2-3, 3-3

[21] Eun-Cheon Lim, Jonas Müller, Jörg Hagmann, Stefan R Henz, Sang-Tae Kim, and Detlef Weigel. *Trowel: a fast and accurate error correction module for Illumina sequencing reads.* Bioinformatics, 30(22):3264–5, 2014. 2-3, 3-3

[22] Mahdi Heydari, Giles Miclotte, Piet Demeester, Yves Van de Peer, and Jan Fostier. *Evaluation of the impact of Illumina error correction tools on de novo genome assembly*. BMC Bioinformatics, 18(1):374, 2017. 3-3, 3-5, 3-6

[23] Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, Sergey I. Nikolenko, Son Pham, Andrey D. Prjibelski, Alexey V. Pyshkin, Alexander V. Sirotkin, Nikolay Vyahhi, Glenn Tesler, Max A. Alekseyev, and Pavel A. Pevzner. *SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing*. Journal of Computational Biology, 19(5):455–477, 2012. 1-12, 1-27, 2-5, 3-3

[24] Neil I Weisenfeld, Shuangye Yin, Ted Sharpe, Bayo Lau, Ryan Hegarty, Laurie Holmes, Brian Sogoloff, Diana Tabbaa, Louise Williams, Carsten Russ, Chad Nusbaum, Eric S Lander, Iain MacCallum, and David B Jaffe. *Comprehensive variation discovery in single human genomes*. Nature Genetics, 46(12):1350–1355, 2014. 3-3, 3-5

[25] Saul B. Needleman and Christian D. Wunsch. *A general method applicable to the search for similarities in the amino acid sequence of two proteins*. Journal of Molecular Biology, 48(3):443–453, 1970. 1-24, 3-4, 4-4, 5-3

[26] Santo Fortunato. *Community detection in graphs*. Physics Reports, 486(3-5):75–174, 2010. 3-4

[27] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. *Fast unfolding of communities in large networks*. Journal of Statistical Mechanics: Theory and Experiment, 2008(10):P10008, 2008. 1-29, 3-4, 3-8

[28] Massoud Seifi, Ivan Junier, Jean-Baptiste Rouquier, Svilen Iskrov, and Jean-Loup Guillaume. *Stable Community Cores in Complex Networks*. In Ronaldo Menezes, Alexandre Evsukoff, and Marta C. González, editors, Complex Networks, pages 87–98. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. 3-4, 3-9

[29] Mahdi Heydari, Giles Miclotte, Yves Van de Peer, and Jan Fostier. *BrownieAligner: accurate alignment of Illumina sequencing data to de Bruijn graphs*. BMC Bioinformatics, 19(1):311, sep 2018. 3-4, 3-11, B-8

[30] Leena Salmela and Eric Rivals. *LoRDEC: accurate and efficient long read error correction*. Bioinformatics, 30(24):3506–3514, 2014. 1-28, 3-4

[31] Giles Miclotte, Mahdi Heydari, Piet Demeester, Stephane Rombauts, Yves Van de Peer, Pieter Audenaert, and Jan Fostier. *Jabba: hybrid error correction for long sequencing reads*. Algorithms for Molecular Biology, 11(1):10, 2016. 1-28, 3-4

[32] Yu Peng, Henry C. M. Leung, S. M. Yiu, and Francis Y. L. Chin. *IDBA – A Practical Iterative de Bruijn Graph De Novo Assembler*, pages 426–440. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. 2-4, 3-5

[33] Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. *QUAST: quality assessment tool for genome assemblies.* Bioinformatics, 29(8):1072–5, 2013. 1-21, 2-11, 3-5

[34] Guillaume Marçais and Carl Kingsford. *A fast, lock-free approach for efficient parallel counting of occurrences of k-mers*. Bioinformatics, 27(6):764–770, 2011. 3-5

[35] Heng Li and Richard Durbin. *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 25(14):1754–60, 2009. 1-25, 2-7, 3-5, 4-2

[36] Michael Molnar and Lucian Ilie. *Correcting Illumina data*. Briefings in Bioinformatics, 16(4):588–599, 2014. 3-6

[37] Nicolas Dierckxsens, Patrick Mardulyn, and Guillaume Smits. *NOVOPlasty: de novo assembly of organelle genomes from whole genome data*. Nucleic Acids Research, 45(4):gkw955, oct 2016. 3-6

[38] Kristi E Kim, Paul Peluso, Primo Babayan, P. Jane Yeadon, Charles Yu, William W Fisher, Chen-Shan Chin, Nicole A Rapicavoli, David R Rank, Joachim Li, David E. A Catcheside, Susan E Celniker, Adam M Phillippy, Casey M Bergman, and Jane M Landolin. *Long-read, whole-genome shotgun sequence data for five model organisms*. Scientific Data, 1:140045, 2014. 3-6

[39] Michael G. Ross, Carsten Russ, Maura Costello, Andrew Hollinger, Niall J. Lennon, Ryan Hegarty, Chad Nusbaum, and David B. Jaffe. *Characterizing and measuring bias in sequence data*. Genome Biology, 14(5):R51+, 2013. 1-9, 2-2, 3-6

[40] Zhao Yang, René Algesheimer, and Claudio J. Tessone. *A Comparative Analysis of Community Detection Algorithms on Artificial Networks*. Scientific Reports, 6(1):30750, 2016. 3-9

4

BrownieAligner: Accurate Alignment of Illumina Sequencing Data to de Bruijn Graphs

"... The computer was born to solve problems that did not exist before. ...[1] *"*

⋆ ⋆ ⋆

Mahdi Heydari, Giles Miclotte, Yves Van de Peer and Jan Fostier

Published in BMC Bioinformatics 19(1): 311, Sep. 2018

In this chapter, we introduce BrownieAligner which is designed and implemented to align short Illumina reads to the de Bruijn Graphs...

Abstract

Aligning short reads to a reference genome is an important task in many genome analysis pipelines. This task is computationally more complex when the reference

[1] Bill Gates

genome is provided in the form of a de Bruijn graph instead of a linear sequence string.

We present a branch and bound alignment algorithm that uses the seed-and-extend paradigm to accurately align short Illumina reads to a graph. Given a seed, the algorithm greedily explores all branches of the tree until the optimal alignment path is found. To reduce the search space we compute upper bounds to the alignment score for each branch and discard the branch if it cannot improve the best solution found so far. Additionally, by using a two-pass alignment strategy and a higher-order Markov model, paths in the de Bruijn graph that do not represent a subsequence in the original reference genome are discarded from the search procedure.

BrownieAligner is applied to both synthetic and real datasets. It generally outperforms other state-of-the-art tools in terms of accuracy. Our results show that using the higher-order Markov model in BrownieAligner improves the accuracy, while the branch and bound algorithm reduces runtime. BrownieAligner is written in standard C++11 and released under GPL license. BrownieAligner relies on multithreading to take advantage of multi-core/multi-CPU systems. The source code is available at: https://github.com/biointec/brownieAligner

4.1 Background

Modern Illumina machines produce sequencing data with a high throughput at a low financial cost. Reads generated by this platform are relatively short (100-300 bp) and have a relatively low error rate (1-2% errors) [1]. A key data structure to represent and manipulate these data in many bioinformatics applications is the de Bruijn graph. It has been used in different contexts, ranging from *de novo* genome assembly [2], transcriptome assembly [3], metagenomics [4], variant calling and structural variation detection [5].

The de Bruijn graph is a directed graph where nodes correspond to k-mers and edges represent an overlap of $k-1$ nucleotides between nodes. When the de Bruijn graph is built from sequencing data and all k-mers and their overlaps are present in the input data, the original sequence can be found as some path through the graph. The de Bruijn graph can thus be seen as a compact multiple sequence alignment representation of the input reads.

Aligning reads to a reference genome is a prerequisite step in many genome analysis pipelines. The vast majority of read alignment software aligns short reads to a linear reference genome [6, 7]. A common strategy in these aligners is a "seed-and-extend" paradigm. First, seeds such as maximal exact matches between a read and the reference sequence are identified. Those seeds indicate candidate positions in the reference genome from which the read originated. In the second step, each seed is extended to the left and right until a full read alignment is obtained and the

alignments with statistically significant similarity are reported [8].

For certain applications, the reference genome may be provided as a de Bruijn graph rather than a linear sequence. For example in the scaffolding phase of a short read assembler, reads can be aligned to the assembly graph [9]. Additionally, for genome identification of reads with an unknown origin in a metagenomics study, reads can be aligned to a de Bruijn graph that is built from multiple genomes. Recently, two standalone tools have been proposed to align short Illumina reads to de Bruijn graphs: BGREAT [10] and deBGA [11].

In order to align reads to a graph representation of the reference genome, the same seed-and-extend approach can be used. While the seeding phase is straightforward, the extension phase is computationally more expensive when dealing with graphs. Given a seed, a brute-force approach would be an exhaustive search in the graph (e.g. depth-first search (DFS) or breadth-first search (BFS)), exploring all possible branches of the tree until the best alignment is found that covers the entire read. However, the number of visited nodes can grow exponentially in the length of the read. Assuming a four-letter DNA alphabet, each node has up to four outgoing arcs. Therefore, to align a read of size l, up to 4^{l-k} nodes need to be explored in the worst-case scenario. While most of the reads never reach this upper bound, it shows that aligning reads to the graph can potentially be intractable, especially in repetitive regions where the graph contains many branches. To tackle this problem, BGREAT and deBGA have an early stop mechanism, which stops exploring nodes when the number of mismatches exceeds a certain threshold. This strategy reduces the search space but potentially fails to return the optimal solution.

A second complication that arises when aligning reads to a de Bruijn graph is that paths in the graph do not necessarily correspond to a substring of the reference genome. Although two connected nodes in the de Bruijn graph always correspond to two consecutive k-mers in the reference genome, paths of three or more connected nodes do not necessarily correspond to a chain of k-mers in the reference genome. Therefore, aligning reads to such paths would reduce the overall accuracy of the alignment procedure.

In this paper, we introduce BrownieAligner to align short Illumina reads to a de Bruijn graph. Even though for most practical applications, a de Bruijn graph would be constructed from sequencing data, we assume in this paper that it is built from a known reference genome, thus yielding a complete and error-free de Bruijn graph. This allows us to focus on the accuracy of the actual alignment algorithms unimpeded by superimposed noise from the graph structure itself. For read alignment, the seed-and-extend paradigm is used. We propose additional strategies to narrow down the search space and avoid the alignment to paths in the graph that do not correspond to sequences in the reference genome. First, the exhaustive DFS is augmented with a branch and bound algorithm. For each branch, an upper bound is computed to the alignment score that could be obtained in that

branch. The branch is discarded from the search procedure if it cannot improve the best solution found so far. In order to rapidly find candidate solutions with a high score, the DFS greedily prioritizes towards the node that appears best. Secondly, we propose to annotate the graph with information about the paths that do exist in the reference data. This is modeled as a higher-order Markov model (MM). A priori, this information is not present in the de Bruijn graph. We thus propose to perform the alignment in two passes: one alignment pass to train the MM, and a second alignment pass that is guided by the MM to obtain the final alignments. Using this MM improves the overall alignment accuracy. This procedure is similar to the strategy used in STAR to perform spliced alignment of RNA-seq reads: in a first alignment round the aligners learns new splice sites; in the second round, the final alignments are obtained [12].

4.2 Methods

4.2.1 Read alignment algorithm

In our de Bruijn graph representation, linear paths of connected k-mers are contracted to unitigs. Nodes thus represent sequences of length k or larger. The problem of finding an optimal read alignment in a graph can be formalized as finding the optimal walk in that graph. A walk is an alternating list $v_0, e_1, v_1, \ldots, e_w, v_w$ of nodes and edges such that, for $1 \leq i \leq w$, edge e_i has endpoints v_{i-1} and v_i. In a de Bruijn graph there is at most one edge between two nodes. Therefore the walk can unambiguously be represented as a chain of nodes $v_0, v_1, \ldots, v_w$. It has been shown that given a de Bruijn graph G, a read r and a cost function f, finding an optimal chain in G for r that minimizes the cost function is an NP-complete problem [10].

Given a read, the first step of finding such optimal chain is finding at least one node of that chain (seeding). Then, by traversing the graph to the left and right, we can find a chain that maximizes a well-defined objective function (extension). BrownieAligner attempts to maximize the similarity score as used in the Needleman-Wunsch algorithm [13]. Therefore a chain that has the highest similarity score to the input read is assumed to be the optimal chain. The advantage of this approach is that it can deal with both substitution errors as well as insertions and deletions. In contrast, the Hamming distance, which is for example used in BGREAT, can only deal with substitutions. In the following, the similarity score of a chain to a given read is defined as the similarity score of the sequence represented by that chain to the read.

BrownieAligner first generates a hash table index of the graph's k-mers (default: $k = 31$) to accelerate the seed-finding procedure. Given an input read as a query, it iterates over all k-mers of that read and returns a seed for all k-mers that

exist in the graph. Seeds that are contiguous in both the read and the graph are merged and sorted according to seed-length. The length of a seed is defined as the number of k-mers in that seed.

Depending on the k-mer size, read length and the error distribution, it is possible that no exact k-mer seeds can be found in some reads. In that case, maximal exact matches (MEMs) between the read and the unitigs of the graph are found using the essaMEM library [14]. Those MEMs are necessarily shorter than k nucleotides.

The extension phase is straightforward when the entire read is contained within a single unitig. In this case, the seed can be extended to the left and to the right within a single node and the alignment score is easily obtained. However, it is also possible that extending the seed moves the alignment across an edge, into an adjacent unitig. In this case, the aligner should decide at each branching point along which nodes to continue.

Our graph alignment algorithm at branching points is shown in Algorithm 1. The input of this algorithm is: a de Bruijn graph G, the unaligned part of the read s and final node v of the seed. The goal of this algorithm is to find a path in G with the highest similarity score among all possible paths in the graph starting from v to the input string s. Without loss of generality, consider the case of a seed extension to the right.

The algorithm always considers nodes with a higher priority score first. PQ denotes a priority queue whose elements are paths from the root v. The priority of an element is the similarity score of that element to (a prefix of) s. The algorithm keeps extending a path until a full alignment with s is obtained. $bestPath$ is then updated with $currPath$ if the current path has a higher similarity to s than $bestPath$. The algorithm terminates when there are no items left in the queue.

For w a path in G, s a sequence of size l, let $f(w, s)$ be the alignment score between w and a prefix of s, let $f_{id}(s)$ be the alignment score of s to itself, and let $f_{max}(w, s)$ be the maximal similarity score between any path in the graph G that starts with w, and s. In our case, $f_{id}(s) = ml$, where m is the score for a match. Then, given an alignment between a path w in G and a prefix $s[1 : n]$ of s:

$$\begin{aligned} f_{max}(w, s) &\leq f(w, s[1 : n]) + f_{id}(s[n + 1 : l]) \\ &= f(w, s[1 : n]) + m(l - n). \end{aligned}$$

This bound is used to a priori discard subtrees in G in which no path exists with a score that is higher than the best complete alignment found so far. The greedy heuristic of prioritizing extension of the highest scoring paths, combined with this branch and bound strategy narrows down the search space, while still resulting in the optimal alignment between G and s.

Algorithm 1 `Graph Alignment`

```
Input: Graph G
Input: String s                          ▷ unaligned suffix of read
Input: Node v                            ▷ final node of the seed
 1: global variables
 2:    PriorityQueue pq
 3:    Path bestPath                     ▷ best path so far
 4:    Path currPath                     ▷ current path
 5: end global variables
 6: currPath.append(v)
 7: currPath.updateScore(s)
 8: pq.push(currPath)
 9: while pq ≠ ∅ do
10:    currPath ← pq.pop()
11:    if currPath.len = s.len then
12:       if currPath.score > bestPath.score then
13:          bestPath ← currPath
14:       end if
15:       continue                       ▷ fully aligned
16:    end if
17:    if currPath.maxScore ≤ bestPath.score then
18:       continue                       ▷ branch-and-bound
19:    end if
20:    for node in currPath.outNodes do
21:       if pathExtensionIsValid(currPath, node) then
22:          newPath ← currPath
23:          newPath.append(node)
24:          newPath.updateScore(s)
25:          pq.push(newPath)
26:       end if
27:    end for
28: end while
29: return bestPath
```

4.2.2 Implicit repeat resolution using a Markov model

Even though all subsequences of the reference genome can be represented as a contiguous path in the de Bruijn graph, the opposite is not true. In particular, not all paths in the graph that span 3 or more nodes correspond to a subsequence of the reference genome. When extending a path in the alignment process of an individual read, a validation is performed, as shown in Algorithm 1 (line 21). This validation relies on a higher-order (≥ 2) Markov model (MM) and allows skipping paths in the graph that do not occur in the genome. At each branching point, we

take into account the topology of the graph and implicitly perform a consensus alignment between all the reads from that genomic region.

To do this, it is necessary to train the model by aligning all reads to the graph and clustering them by genomic region to which they align. However, the model needs to take into account that the data has sequencing errors and the read coverage is not uniform across the genome. The presence of sequencing errors in reads may result in a wrong alignment of reads to the graph. Therefore, the model needs to distinguish between spurious paths which appear to exist in the reference genome because of misaligned reads and true paths. On the other hand, a true path for which the corresponding sequence exists in the genome might not be observed due to a lack of coverage. The probability of observing a true path in the data is smaller for longer paths. The following section describes how BrownieAligner implicitly resolves repeats and guides the aligner using a higher-order MM.

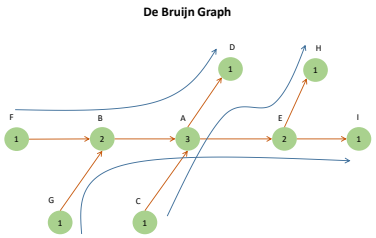
Markov models have been used as a robust statistical framework to model real-time processes in marketing, text analysis, bioinformatics, network analysis, weather forecasting, etc. [15]. One of these applications commonly used in text editors is word prediction. It predicts the most likely word of the user by considering the previously typed words based upon a model that is trained on a broad set of training data [16]. Similarly, in the graph alignment of an individual read, by looking at previously visited nodes, and the information of other reads, the aligner predicts the true path among all possible paths, and prevents alignment against false paths that do not exist in the reference genome.

Formally, an $n-$order Markov Model (n-MM) in the de Bruijn graph G is defined by:

- A set of states $S = \{s_1, \ldots, s_m\}$, in which each state represents a path of n nodes $(v_1, \ldots, v_n)$ in G.
- A set of transition probabilities p_{ij} representing the probability of extending state path s_i with node v_j.

The transition probability between a state path and a node is used to specify whether the path of length $n + 1$ exists in the reference genome. We are therefore not particularly interested in the actual values of the transition probabilities; we are merely interested in distinguishing the transitions that do *not* occur ($p_{ij} = 0$) from those that *do* occur ($p_{ij} > 0$). Transitions with zero probability correspond to paths that span $n + 1$ nodes in the graph that do not represent a true sequence in the reference genome and can hence be skipped during the read alignment step. The first-order MM is memoryless in that sense that the prediction of the next node only depends on the current node. Any edge in de Bruijn graph represents a valid overlap between two k-mers in the reference genome. Thus, a 1-MM is not informative in this regard. As shown in Fig. 4.1, the higher-order MM tables can be useful to guide the alignment procedure at branching nodes.

Figure 4.1 This figure shows the association between the de Bruijn graph and MM tables. On the left side, part of a de Bruijn graph is shown. True paths are depicted by blue lines. The numbers inside each node indicate the multiplicity of that node, i.e., the number of times the node's sequence is present in the reference genome. A table at each node guides the aligner based on previously observed nodes. The 2-MM and 3-MM tables of node A are shown on the right side. Based on the 2-MM table, reads that align to CA are guided to E as the continuation to node D is not allowed. However, the information in this table is insufficient to guide reads that align to BA since continuations to E and D are both valid. In contrast, the 3-MM table guides the reads that align to FBA to D, and GBA to E. The information in the final row in 3-MM table is redundant because it is also contained in the lower-order 2-MM table.



2-MM		3-MM	
From	To	From	To
BA	{D,E}	FBA	{D}
CA	{E}	GBA	{E}
		CA	{E}

To derive a n-MM table, reads are aligned to the graph in a first alignment pass using Algorithm 1 but without any restrictions on alignment path, i.e., without the if-statement in line 21. The goal of this alignment pass is merely to train the MM. Aligned reads imply paths in the graph and all observed paths or subpaths of length $n + 1$ are used to populate the MM table. The first n nodes of such path define a state s while the final *head* node denotes a possible continuation. The table consists of all observed states s and the corresponding frequencies of all observed continuations. These frequencies are then converted to probabilities.

However, there can be two types of errors in this process: (1) observing an invalid alignment path because of a misalignment (due to sequencing errors), and (2) missing valid alignment paths because of a lack of coverage.

To minimize the first type of error, we test for each observed path whether its frequency $freq$ corresponds to the expected frequency using the following two hypotheses:

- H_0 The multiplicity of the path is zero
- H_1 The multiplicity of the path is at least one

The multiplicity of a path in the graph indicates the number of times that the corresponding sequence appears in the reference genome. Two Poisson distributions are used to model the frequency of observed paths with multiplicity zero

($\lambda = 1$) and multiplicity one ($\lambda = C_M$). Paths for which $likelihoodRatio = \frac{P(freq|H_0)}{P(freq|H_1)} \geq minLikelihoodRatio$ are pruned from the list of eligible paths. Here, $likelihoodRatio$ is a measure of the degree of certainty of the decision of eliminating a path from the eligible set. The higher this value, the higher the certainty. However, setting a too high value for $minLikelihoodRatio$ reduces the ability of the model to avoid false paths.

To minimize the second type of error, the Markov model is only used for paths for which the expected number of reads covering this path $C_M \geq minChainCov$. Here, $minChainCov$ is a second user-defined threshold. Higher values of this parameter reduce the risk of making the second type of error, but again, a too high value reduces the applicability of the Markov model. Given the sequencing coverage c, the read length l, and a path P that implies a sequence of length M and multiplicity 1 in the reference genome, the expected coverage C_M of P is then given by the following formula:

$$C_M = \frac{l - M + 1}{l} c \tag{4.1}$$

Proof: First, consider a read covering a path of sequence size $M - 1$. Second, extend this path with one base, without loss of generality, to the left. For the read to cover this extended path, its start position has to be strictly before the start of the original path. The probability of this is $\frac{l-M+1}{l-M+2}$. Hence the following recurrence relation holds:

$$C_M = \frac{l - M + 1}{l - M + 2} C_{M-1}.$$

Solving the recurrence relation leads to

$$C_M = \frac{l - M + 1}{l} C_1.$$

Additionally, $C_1 = c$ by definition. This concludes the proof.

After constructing the MM tables for different orders as outline before, all reads are again aligned to the graph in a second alignment pass, this time guided by the MM. Different orders are required because a low order might not be sufficiently informative to guide the read alignment whereas a high order might not attain the coverage requirements. In this second alignment pass, the alignment of a read no longer solely depends on the identity between the read and the sequence implied by the alignment path. Rather, the collective information of all other reads is used to identify the true paths in the graph and thus obtain a higher alignment accuracy.

4.2.3 Choice of parameters:

The scoring system in Algorithm 1 has match, mismatch and gap scores of respectively $+1$, -1 and -3. The maximum MM order ($maxOrder$) is 10. The values of $minLikelihoodRatio$ and $minChainCov$ are set respectively to 10^5 and 10.

Table 4.1 Artificial datasets used for the evaluation of graph aligner tools.

Abbr.	Organism	Reference ID	Genome size	Repeated 31-mers (%)	Sequencing platform	Cov.	Read length
S1	*Escherichia coli K-12 DH10B*	NC010473	4.5 Mbp	3.2	Illumina HiSeq 2500	25	150 bp
S2	*Escherichia coli K-12 DH10B*	NC010473	4.5 Mbp	3.2	Illumina HiSeq 2000	50	100 bp
S3	*Homo sapiens* Chr. 21	HG19	45.2 Mbp	4.3	Illumina HiSeq 2500	25	150 bp
S4	*Homo sapiens* Chr. 21	HG19	45.2 Mbp	4.3	Illumina HiSeq 2000	50	100 bp
S5	*Drosophila melanogaster*	Release 5	116.4 Mbp	1.1	Illumina HiSeq 2500	25	150 bp
S6	*Drosophila melanogaster*	Release 5	116.4 Mbp	1.1	Illumina HiSeq 2000	50	100 bp

4.2.4 Graph aligner tools

The performance of BrownieAligner is compared with the state-of-the-art graph aligners BGREAT and deBGA. A de Bruijn graph is first constructed from the reference genome, followed by the alignment of reads to the corresponding graph. BrownieAligner and deBGA have the functionality to construct the de Bruijn graph. BGREAT does not support this feature, therefore we used BCALM [17] to construct the de Bruijn graph for BGREAT. A drawback of deBGA is that it only accepts a reference genome as an input and not a graph in general. Therefore, it cannot be used as a graph aligner tool to align reads against the assembly graph. BGREAT and BrownieAligner report corrected reads, i.e., the corresponding sequences from the reference genome after aligning reads, in the same order and file format as the input reads. In contrast, deBGA returns the alignment results in SAM format. Therefore, we developed sam2Alignment script, which is used to produce the corrected read from the reference genome based on the SAM entry. Three versions of BrownieAligner are provided and evaluated in this paper. BrownieAligner is the main tool and benefits from both (1) the greedy branch and bound algorithm and (2) MM repeat resolution. The first feature is disabled in BrownieAlignerNoBB and the second one is disabled in BrownieAlignerNoMM. For all results the default or recommended k-mer sizes are used. Parameters and settings are provided in App. C.1.

All tools were run on a machine with four Intel(R) Xeon(R) E5-2698 v3 @ 2.30 GHz CPUs (64 cores in total) and 256 GB of memory. All tools support multithreading and run with 32 threads. Elapsed (wall clock) time and peak resident memory were measured with the GNU *time* command.

4.2.5 Data

The performance of the three tools was measured on six artificial datasets (see Table 4.1). For three high-quality reference genomes (*E. coli str. K-12 substr. DH10B*, *Human chr-21* and *Drosophila melanogaster*), reads were simulated for two different Illumina platforms (HiSeq 2000 (100 bp), HiSeq 2500 (150 bp)) using ART [18].

Additionally, the three tools were evaluated on eight real Illumina datasets

Table 4.2 Real datasets used for the evaluation of graph aligner tools.

Abbr.	Organism	Reference ID	Genome size	Repeated 31-mers (%)	Cov.	Sequencing platform	Read length	Trimmed reads	Dataset ID
R1	*Bifidobacterium dentium*	Nc013714.1	2.6 Mbp	0.4	373 X	Illumina MiSeq	251 bp		SRR1151311
R2	*Escherichia coli K-12 DH10B*	NC010473	4.5 Mbp	3.2	418 X	Illumina MiSeq	150 bp		Ill. Data library
R3	*Escherichia coli K-12 MG1655*	NC000913	4.5 Mbp	0.6	612 X	Illumina GAII	100 bp		ERA000206
R4	*Salmonella enterica*	NC011083.1	4.7 Mbp	0.5	97 X	Illumina MiSeq	239 bp	✓	SRR1206093
R5	*Pseudomonas aeruginosa*	ERR330008	6.1 Mbp	0.6	169 X	Illumina MiSeq	120 bp	✓	ERR330008
R6	*Homo sapiens* Chr. 21	HG19	45.2 Mbp	4.3	29 X	Illumina HiSeq	100 bp		Ill. Data library
R7	*Caenorhabditis elegans*	WS222	97.6 Mbp	2.6	58 X	Illumina HiSeq	101 bp		SRR543736
R8	*Drosophila melanogaster*	Release 5	116.4 Mbp	1.1	52 X	Illumina HiSeq	100 bp		SRR823377

for which both a reference genome and sequencing data are publicly available (see Table 4.2). Genome sizes range from 2 Mbp (*Bifidobacterium dentium*) to 116 Mbp (*Drosophila melanogaster*), and read coverage varies from 29 X to 612 X. The data were produced on the Illumina HiSeq, MiSeq and GAII platforms. Read lengths range from 100 bp to 251 bp. Two data sets have a variable read length due to prior read trimming, while the others have fixed read lengths.

4.2.6 Evaluation metrics

For each simulated read, ART generates a corresponding error-free read that is used to perform the accuracy evaluation (see App. C.2). For real data, the ground truth is unknown. In this case, it is assumed that the correct alignment is represented by the alignment of the read to the linear reference genome using BWA. Only paired-end reads where both pairs map to the reference genome properly are extracted using SAMtools [19]. Finally, the pairwise alignment of each read is reconstructed based on the CIGAR string and MD tag using sam2pairwise [20] (see App. C.3). The performance of the aligners is measured based on their ability to align reads to the correct position in the graph. For a given read, the correct path in the graph is the path with the same sequence content as the error-free read corresponding to that read. A detailed explanation is provided in App. C.4.

4.3 Results and discussion

4.3.1 Alignment ratio

Table 4.3 shows the percentage of correctly aligned reads for the simulated data (see App. C.5.1 for the detailed information). BrownieAligner has the highest percentage of correctly aligned reads ($\geq 98.07\%$) for all data sets. BGREAT consistently performs slightly worse ($\geq 96.16\%$) than BrownieAligner while deBGA performs slightly worse on half of the data sets (S1, S3, S5), but significantly worse ($\geq 83.01\%$) on the others (S2, S4, S6). All tools perform worse on the H. *sapiens* data (S3 and S4) than on the E. *coli* data (S1, S2) and D. *melanogaster* data (S5, S6). The performance of deBGA additionally depends on the read length

Table 4.3 Accuracy comparison of graph aligner tools in terms of correct alignment of reads to the graph on simulated data.

	S1	S2	S3	S4	S5	S6
	Percentage of correctly aligned reads.(%)					
BGREAT	99.94	99.61	98.92	96.16	99.89	99.40
BrownieAligner	100.00	99.99	99.42	98.07	99.97	99.89
BrownieAlignerNoMM	99.99	99.98	99.30	97.67	99.96	99.85
deBGA	99.52	83.48	99.07	83.01	99.37	83.37

Table 4.4 Accuracy evaluation of BrownieAlignerNoMM and BrownieAligner on the subset of the simulated reads that align to a path of at least two nodes in the graph.

	S1	S2	S3	S4	S5	S6
	Percentage of correctly aligned reads. (%)					
BrownieAligner	99.34	99.05	90.72	86.07	98.21	97.12
BrownieAlignerNoMM	98.72	98.47	87.68	82.39	97.38	96.13

or coverage, since it consistently performs significantly better on the 150bp 25x coverage data than on the 100bp 50x coverage data, for all genomes. Additionally, comparing the results for BrownieAligner and BrownieAlignerNoMM reveals that the use of the Markov model in the read alignment process always improves the overall accuracy of the alignment.

We additionally investigated the accuracy of BrownieAligner on those reads that are aligned to a walk in the graph that comprises multiple nodes, i.e., the reads for which the Markov model algorithm is actually used. Table 4.4 shows the percentage of these reads that are correctly aligned by BrownieAligner (with Markov models) and BrownieAlignerNoMM. Results indicate that the use of these Markov models offers a significant improvement for the alignment of these harder to align reads.

In order to see the effect of k-mer size on the accuracy, all tools were benchmarked with different values of k on all the simulated datasets. The results indicate that for each dataset the best accuracy for BrownieAligner is always higher than the best accuracy for other tools (see App. C.5.1). The results show BrownieAligner performs better for larger k. This has two reasons. First, BrownieAligner can use maximal exact matches during the seeding phase, enabling the identification of seeds smaller than k. Hence, the sensitivity of the seed finding procedure is not negatively affected by a larger value of k. Second, with higher values of k the repeat structure in the graph is less complex, and hence BrownieAligner is

Table 4.5 Accuracy comparison of graph aligner tools in terms of correct alignment of reads to the graph on real data.

	R1	R2	R3	R4	R5	R6	R7	R8
	Percentage of correctly aligned reads. (%)							
BGREAT	94.55	94.28	91.28	84.97	96.09	92.01	94.57	80.37
BrownieAligner	99.81	99.81	99.55	99.02	99.78	96.98	96.53	89.59
BrownieAlignerNoMM	99.81	99.80	99.52	98.99	99.78	96.67	96.47	89.55
deBGA	99.67	99.30	92.36	97.31	93.63	98.42	74.72	85.42

less prone to choosing an incorrect path in the alignment phase. The accuracy of BrownieAligner on simulated data also has been evaluated based on other values of $maxOrder$, $minLikelihoodRatio$ and $minChainCov$. The results indicate that BrownieAligner performs consistently well over a wide range of parameters setting (see App. C.5.1).

Table 4.5 shows the percentage of reads that are correctly aligned by each tool for 8 real datasets (see App. C.5.2 for the detailed tables). The accuracy of BrownieAligner for the bacterial genomes (R1-R5) is very high ($G \geq 99.02\%$) and BrownieAligner outperforms the other tools, followed by deBGA ($G \geq 92.36\%$) and then BGREAT ($G \geq 84.97\%$). For the H. *sapiens* data (R6) deBGA performs remarkably well. For the other two eukaryotic genomes (R7 and R8), BrownieAligner has again the highest percentage correctly aligned reads. The comparison between BrownieAligner and BrownieAlignerNoMM again indicates that the use of the Markov models to resolve repeats improves the accuracy of read alignment. Additionally, the difference is more significant in H. *sapiens* (R6), which is known to be repeat-rich. The effect of the MM for the alignment of reads that span multiple nodes is further investigated in real data (see App. C.5.2). Results indicate that the alignment accuracy generally benefits from using the MM.

4.3.2 Time and space requirements

Fig. 4.2 and 4.3 show the memory usage and runtime of the aligners for the simulated data (see App. C.5.3 for detailed tables). For the smallest genomes, deBGA requires the most memory, while for larger genomes BrownieAligner has the highest memory requirements. Run times for S1, S2, S5 and S6 data sets are comparable for all tools. However, BrownieAligner and BGREAT take significantly longer than deBGA to align S3 and S4. Generally speaking, BGREAT is memory-efficient and deBGA is fast.

In order to capture the effect of the branch and bound pruning strategy in algorithm 1, we disabled this feature in BrownieAlignerNoBB. Fig. 4.4 compares the amount of time that the two versions of BrownieAligner take to align only

Figure 4.2 Peak memory Peak memory usage of the aligner tools for simulated datasets.

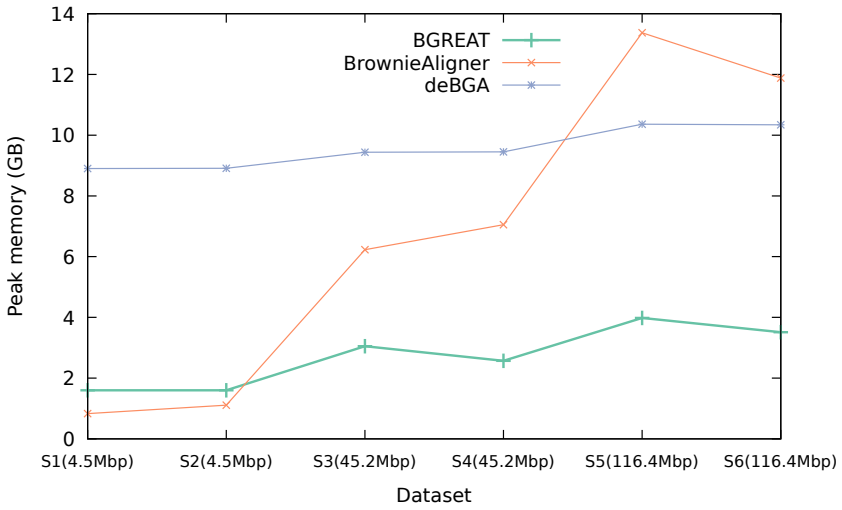
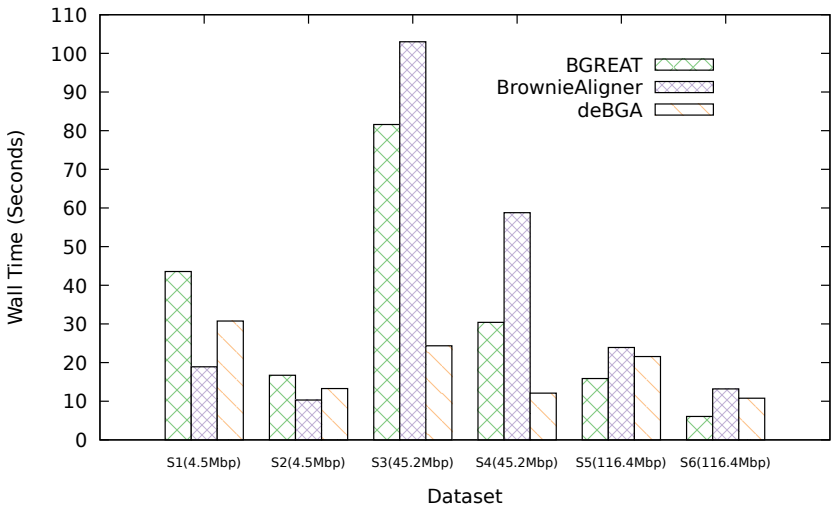


Figure 4.3 Runtime. Average runtime of tools to align 1M reads for the simulated datasets.



those reads that align to a non-trivial walk in the graph, i.e., those reads where Algorithm 1 is used. Results show that using this strategy reduces the runtime of BrownieAligner especially for more repetitive genomes (see App.C.5.3.2 for detailed tables).

Figure 4.4 Runtime. The effect of branch and bound strategy on the running time of BrownieAligner.

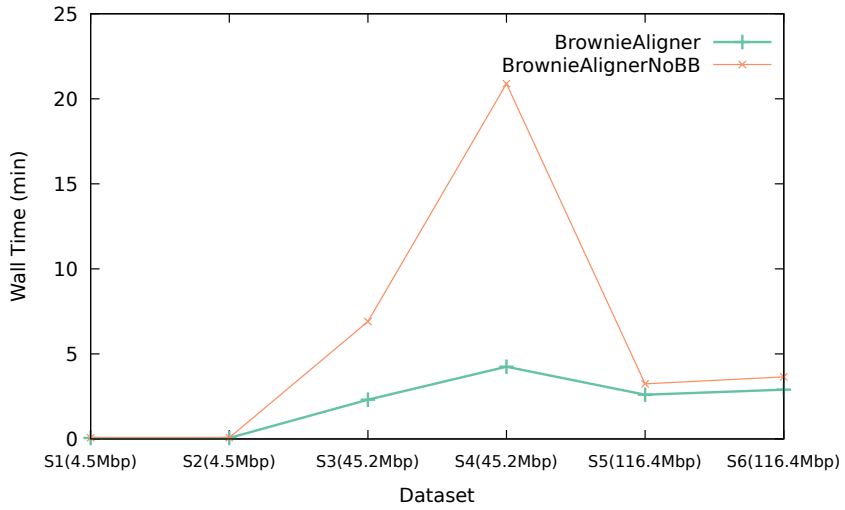


Fig. 4.5 and 4.6 show the memory usage and runtime of the aligners for the real data (see App. C.5.3 for detailed tables). Memory usage and runtime of tools in real data also follow the same pattern as the simulated data except that BrownieAligner is the slowest tool for three largest datasets.

Generally, BrownieAligner has a higher runtime to align reads to the H. *sapiens* genome (S3, S4 in simulated data sets and R6 in real data sets). This is due to the presence of more repetitive patterns in the genome making the de Bruijn graph more complex. Therefore, the DFS algorithm in BrownieAligner has to visit more nodes before it finds the optimal path in the graph.

4.4 Conclusions

BrownieAligner is proposed as a tool to align short Illumina reads to a de Bruijn graph. It uses higher-order Markov models to implicitly resolve repeats in the graph, thus avoiding reads to be aligned against paths in the de Bruijn graph that do not constitute a subsequence of the genome. Our results show that using this model always improves the accuracy of the alignment both in simulated and real

Figure 4.5 Peak memory usage. Peak memory usage of the aligner tools for real datasets.

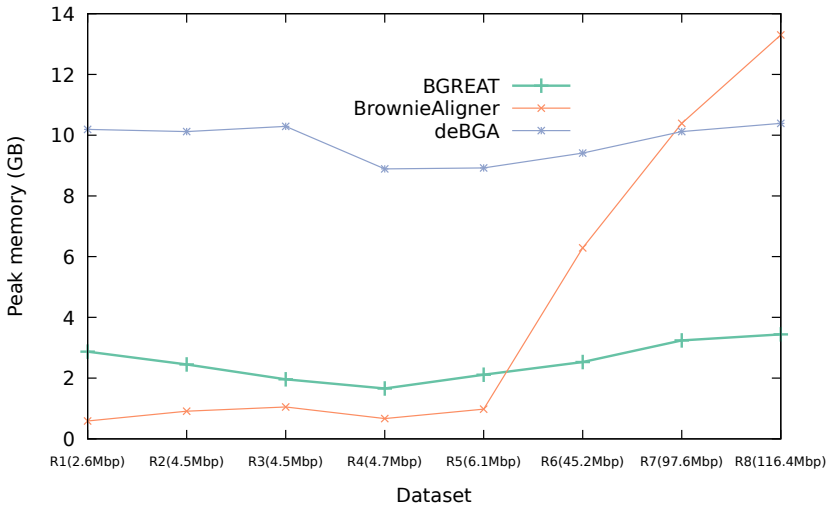
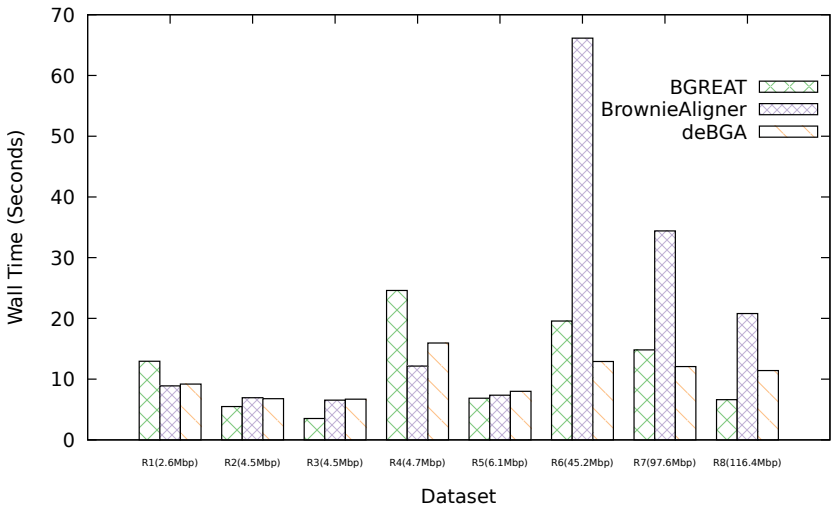


Figure 4.6 Runtime. Average runtime of tools to align 1M reads for the real datasets.



data. BrownieAligner generally outperforms other state-of-the-art tools in terms of accuracy, while demanding slightly higher runtime and memory requirements.

References

[1] *Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems.* Genome Biology, 12(11):R112, Jan. 2011. 2-2, 3-2, 4-2

[2] Phillip E C Compeau, Pavel A. Pevzner, and Glenn Tesler. *How to apply de Bruijn graphs to genome assembly*. Nature Biotechnology, 29(11):987–991, 2011. 1-16, 4-2

[3] Manfred G Grabherr, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn A Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, Zehua Chen, Evan Mauceli, Nir Hacohen, Andreas Gnirke, Nicholas Rhind, Federica di Palma, Bruce W Birren, Chad Nusbaum, Kerstin Lindblad-Toh, Nir Friedman, and Aviv Regev. Nature Biotechnology, 29(7):644–652, jul. 2011. 4-2

[4] J. Pell, A. Hintze, R. Canino-Koning, A. Howe, J. M. Tiedje, and C. T. Brown. *Scaling metagenome sequence assembly with probabilistic de Bruijn graphs*. Proceedings of the National Academy of Sciences, 109(33):13272–13277, 2012. 4-2

[5] Lorenzo Tattini, Romina D'Aurizio, and Alberto Magi. *Detection of Genomic Structural Variants from Next-Generation Sequencing Data*. Frontiers in Bioengineering and Biotechnology, 3(June):1–8, 2015. 4-2

[6] Ben Langmead and Steven L. Salzberg. *Fast gapped-read alignment with Bowtie 2*. Nature Methods, 9(4):357–359, 2012. 4-2

[7] Heng Li and Richard Durbin. *Fast and accurate short read alignment with Burrows-Wheeler transform*. Bioinformatics, 25(14):1754–1760, 2009. 1-25, 2-7, 3-5, 4-2

[8] Heng Li and Nils Homer. *A survey of sequence alignment algorithms for next-generation sequencing*. Briefings in Bioinformatics, 11(5):473–483, 2010. 4-3

[9] Andrey D. Prjibelski, Irina Vasilinetc, Anton Bankevich, Alexey Gurevich, Tatiana Krivosheeva, Sergey Nurk, Son Pham, Anton Korobeynikov, Alla Lapidus, and Pavel A. Pevzner. *ExSPAnder: A universal repeat resolver for DNA fragment assembly*. Bioinformatics, 30(12):293–301, 2014. 4-3

[10] Antoine Limasset, Bastien Cazaux, Eric Rivals, and Pierre Peterlongo. *Read mapping on de Bruijn graphs*. BMC Bioinformatics, 17(1):237, 2016. 1-28, 4-3, 4-4, 5-3

[11] Bo Liu, Hongzhe Guo, Michael Brudno, and Yadong Wang. *DeBGA: Read alignment with de Bruijn graph-based seed and extension*. Bioinformatics, 32(21):3224–3232, 2016. 1-28, 4-3, 5-3

[12] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. *STAR: Ultrafast universal RNA-seq aligner*. Bioinformatics, 29(1):15–21, 2013. 4-4

[13] Saul B Needleman and Christian D. Wunsch. *A general method applicable to the search for similarities in the amino acid sequence of two proteins*. Journal of Molecular Biology, 48(3):443–453, Mar. 1970. 1-24, 3-4, 4-4, 5-3

[14] Michaël Vyverman, Bernard De Baets, Veerle Fack, and Peter Dawyndt. *EssaMEM: Finding maximal exact matches using enhanced sparse suffix arrays*. Bioinformatics, 29(6):802–804, 2013. 4-5

[15] Wai Ki Ching and Michael K. Ng. *Markov chains: models, algorithms and applications*. Kluwer Academic Publishers, Dordrecht, 2006. 4-7

[16] Steffen Bickel, Peter Haider, and Tobias Scheffer. *Predicting sentences using N-gram language models*. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT '05, volume 2, pages 193–200, Morristown, NJ, USA, 2005. Association for Computational Linguistics. 4-7

[17] Rayan Chikhi, Antoine Limasset, and Paul Medvedev. *Compacting de Bruijn graphs from sequencing data quickly and in low memory*. Bioinformatics, 32(12):i201–i208, 2016. 4-10

[18] Weichun Huang, Leping Li, Jason R Myers, and Gabor T Marth. *ART: a next-generation sequencing read simulator.* Bioinformatics, 28(4):593–4, Feb. 2012. 2-6, 4-10

[19] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. *The Sequence Alignment/Map format and SAMtools*. Bioinformatics, 25(16):2078–2079, 2009. 4-11

[20] Matthew C. LaFave and Shawn M. Burgess. *sam2pairwise version 1.0.0*, Aug. 2014. 4-11

5 Conclusion and further directions

"Either write something worth reading or do something worth writing.[1]*"*

5.1 Discussion

Advances in sequencing technologies over the past decade were a turning point in our history. Nowadays, the sequencing cost has decreased, and we can sequence the entire human genome much faster than before. This enhancement in the technology leads to the accumulation of a massive amount of sequencing data which needs to be analyzed. Here, we:

- Studied Error Correction (EC) methods, their ability to correct sequencing errors, and their impact on *de novo* genome assembly. The focus of our study was on Illumina data, and 12 EC tools that have been released after 2012 were studied. Based on this study, we noticed that most EC tools reduce the error rate in sequencing data without introducing many new errors. However, EC tools perform poorly in certain sequencing areas such as shallow-coverage regions or regions that contain highly repetitive elements such as a poly (A/T). Reads overlapping these regions are often poorly corrected in an inconsistent way; this inevitably leads to increased assembly fragmentation. As a result, state-of-the-art assemblers like SPAdes or DISCOVAR do not benefit, and in fact, may even produce worse assemblies when using precorrected data.

[1] Benjamin Franklin

- Introduced a new EC tool, BrownieCorrector, a targeted EC tool that focuses on the correction of reads that contain highly repetitive elements such as a poly (A/T); BrownieCorrector first extracts the reads that contain a given highly repetitive k-mer. Next, it clusters the reads into compatible groups using the read sequence and paired-end read information. Reads in each group are assumed to arise from the same genomic region and are corrected consistently and independently from reads in other clusters. To correct reads in each group, BrownieCorrector first constructs the de Bruijn graph from the reads in that group. Then, it performs typical graph cleaning procedures like tip clipping and bubble detection. Tips and bubbles are often erroneous nodes that represent erroneous k-mers in the reads. Finally, BrownieCorrector aligns back the reads to the cleaned de Bruijn graph to correct existing sequencing errors. We compared the assembled contigs and scaffolds produced by SPAdes without and with performing error correction using state-of-the-art EC tools as well as BrownieCorrector. The results indicate that BrownieCorrector improves the quality of assembly and performs better in most cases even though it corrects less than 2% of reads in total.

- Introduced a new graph aligner tool, BrownieAligner, to align short reads to the de Bruijn graph. BrownieAligner is used in BrownieCorrector's pipeline when reads in each cluster need to be aligned to the cleaned de Bruijn graph. Similar to most of the short-read aligner tools, BrownieAligner uses the same seed-and-extend approach to find the optimal path in the graph to which a given read aligns. In contrast to the linear reference, the extension phase in the graph alignment is computationally intractable. To reduce the search space in the graph, we suggested a branch and bound algorithm which still promises to return the optimal solution. Furthermore, to avoid aligning reads to the paths that do not correspond to any true sub-sequences in the initial data (such as reads or the reference genome), we proposed a higher-order Markov model (MM). BrownieAligner aligns reads using a two-step approach: in the first step, reads are aligned to the graph to train the MM. In the second round, the MM guides the alignment procedure. BrownieAligner has the highest accuracy of the read alignment compared with other state-of-the-art graph aligner tools for both simulated and real data.

5.2 Future work

Both BrownieAligner and BrownieCorrector can be enhanced in terms of accuracy and efficiency. In the following two sections, we outline some potential improvements in these tools.

5.2.1 BrownieAligner

Although there are many short-read aligners available for the community, there are only a few that can map reads to a graph. For instance,BrownieAligner is compared with two other available tools: BGREAT [1] and deBGA [2]. The lack of enough competitors itself implies the potential possibility of further work in this field. BrownieAligner shows a promising efficiency compared with other state-of-the-art tools on average; yet, the accuracy of BrownieAligner drops for larger datasets particularly for more repetitive genomes like Homo sapiens. For example, BrownieAligner can accurately align 96.98% of the reads to the graph, while deBGA shows a higher accuracy with 98.42%. The main difference between BrownieAligner and deBGA is that deBGA uses the paired-end read information for the read alignment. The use of paired-end read information can be beneficial especially for aligning reads that arise from repetitive regions. While one of the reads in a pair which originate from the repetitive context can potentially align to multiple paths in the graph, the other one in the pair determines the correct path. Therefore, we believe BrownieAligner can be enhanced if it employs the paired-end read information.

BrownieAligner attempts to maximize the similarity score as used in the Needleman Wunsch algorithm [3]. Another improvement on BrownieAligner could be using the affine gap model which has a linear gap cost. Because, in real datasets, it is more likely to see two consecutive indels in a read than two independent ones.

BrownieAligner does not align a read to the graph if it can map to two or more paths with the same score. This conservative approach is useful when we want to avoid the wrong alignment as much as possible; however, a more creative idea can narrow down the number of possibilities and suggest the best option as well. For example, using the quality scores in the alignment procedure can increase the ratio of correctly aligned reads to the graph albeit at the cost of a higher runtime [4].

BrownieAligner is applied only to Illumina data; however, the algorithm can be utilized for other kinds of data with a proper modification of the settings and parameters. For example, 10x-Genomics data are a very similar type of data to Illumina with an extra tag that informs us that certain reads originate from the same molecule. The same branch and bound algorithm can be applied to align 10x-Genomics reads to the de Bruijn graph. In this case, the tag information in the reads can be used to align more precisely the reads which originate from the same molecule to the same set of nodes in the graph.

Short-read aligners like BWA often return the output for the user in a standard format like a SAM file; it is a TAB-delimited text format comprising an optional header section before an obligatory alignment section. Each alignment line has 11 required fields for essential alignment information such as mapping position, CIGAR string, etc. However, in the context of graph alignment, some of these fields are meaningless. For example, the POS field contains the alignment position

in the linear reference genome, which is not useful for graph alignments. Alternatively, one can define a new standard format for graph alignment output based on the existing SAM file. In this format, other relevant information such as the chain of nodes to which the read aligns can be stored.

5.2.2 BrownieCorrector

One of the advantages of BrownieCorrector's pipeline is its modularity where each step can be substituted by a method of choice. For example, the Louvain community detection algorithm can be replaced by an alternative clustering algorithm. Different metrics can be used to infer the similarity score between pairs of reads. New EC tools can be implemented and used to correct clusters. Because the number of reads in each cluster is limited (for example the max cluster size is set to 500 reads in BrownieCorrector), other expensive methods which need more CPU cycles can be applied to infer the consensus sequence. In the case of using the current error correction method, any improvement on BrownieAligner can also improve BrownieCorrector. We believe this flexibility allows the pipeline to evolve further in the future.

We use BrownieCorrector as a preprocessing tool to correct sequencing errors near highly repetitive k-mers. The default k-mer is a poly (A/T). However, BrownieCorrector can also be used as a postprocessing tool to correct reads that overlap the edges of the breakpoints in the assembled contigs. In this way, one can first assemble the reads with SPAdes then find the most frequent k-mer at the end or the start of assembled contigs. Based on our investigations, a poly (A/T) k-mer is often one of the most highly frequent k-mers in these regions. However, if other k-mers appear to be more frequent, they can be used instead of the default poly (A/T) in the pipeline.

Although BrownieCorrector corrects less than 2% of the whole datasets, it is among the slowest EC tools. Compared with other EC tools, only Karect [5] and ACE [6] are slightly slower. The most time-consuming parts in the pipeline are calculating the similarity score between pairs of reads, and read error correction. For the first part, we only calculate the similarity score between two reads if they share one or more mutual k-mers; otherwise, we assume the similarity score is zero. This heuristic idea already decreased the running time. However, we assume that there is still room for further improvement to run it more efficiently. For the error correction part, because clusters can be corrected separately and independently, new parallelization techniques and ideas can be applied to run the whole pipeline faster.

BrownieCorrector and BrownieAligner are both designed to align DNA sequencing data, however, a similar approach can be used for RNA-Seq data as well. For example, performing error correction of RNA sequencing data may also im-

prove the quality of *de novo* transcriptome assembly. However, dealing with RNA-Seq data, error detection is more difficult. Genes can express at a different level and hence there is more coverage variability. Biological events like RNA editing and alternative splicing also increase the complexity of the problem. The idea in BrownieCorrector to cluster the reads and then doing the error correction may be a good approach to deal with these challenges. Aligning RNA reads to a reference is a preliminary step in some RNA-Seq data analysis. For example, to quantify the gene expression levels, reads are aligned to the reference. In this case, RNA sequencing data can be aligned to an assembly graph instead of a linear reference using BrownieAligner. This could be particularly useful when the reference is incomplete, i.e., shorter contigs are often discarded after the assembly. Aligning RNA reads to an assembly graph that contains all the contigs of any size can improve the alignment rate.

Finally, BrownieCorrector and BrownieAligner are both designed and implemented to run in the Linux environment. Both tools can be modified to support the Windows operating system as well.

5.3 Current limitations and future perspective

To evaluate the EC tools, we mainly focused on their impact on *de novo* genome assembly. However, are there other contexts for which EC tools can be useful? For example, one may suggest using an EC tool to reduce the error rate prior to performing variant calling or read alignment. In variant calling pipelines, reads are first mapped to a known reference genome. The read mapper uses the entire read and information from the paired-end reads to determine the most suitable alignment position. To separate sequencing errors from true variants, read pileups or even multiple alignments of reads are used. Base quality scores are recalibrated to maximize their information content. Additional prior information is gathered from, e.g., dbSNP and local assemblies of haplotypes are created to maximize specificity and sensitivity. In contrast, Illumina EC methods have less information to make decisions (most notably: no reference genome and no dbSNP database). Therefore, EC methods should probably not be used prior to variant calling and for the same reasons for read alignment. In addition, to our knowledge, EC tools are not being advertised for any other practical purpose than *de novo* genome assembly.

There are several metrics to evaluate the quality of assembly results such as contiguity, genome coverage fraction, number of contigs. Here, in this thesis, we mainly focused on the value of NGA50. We have already discussed in the introduction and research chapters why NGA50 is privileged relative to other relevant metrics. However, another interesting approach that the future evaluation methods can take into account is to evaluate the quality of assembled contigs based on their

ability to identify coding regions and gene annotations. In this regard, one needs to set up an annotation pipeline to evaluate the differences across assemblies. We assume that the continuity of assembled contigs (higher NGA50) can expand its capacity to annotate more genes. In particular, with larger contig sizes, the relative order of genes also can be studied, which is useful for evolutionary studies [7].

In this thesis, we noticed that EC tools often reduce a high fraction of sequencing errors without improving the quality of assembly. The fact that they can reduce a high fraction of errors shows that the underlying algorithms of many of these EC are sound but, perhaps, with some enhancement on the methodology, they can improve the quality of assembly as well. For example, they can distinguish problematic regions of data and perform careful error correction in those particular regions. They can also contribute the paired-end read information in their methods. There is no other EC tool currently that uses the paired-end read information except BrownieCorrector. Another approach for a future EC tool could be hybrid error correction. In this regard, long reads from the third generation can be used to supplement the short Illumina reads in two steps. In this way, a fast EC tool that can reduce a huge fraction of errors of Illumina sequencing data can be employed in the first step. This EC tool may make some mistakes and eliminate some existing overlaps between reads in problematic regions. In the second step, longer reads from the third generation can be used to fix these mistakes to retrieve these missing overlaps.

One of the challenges for error correction or genome assembly is the existence of repeats in the genome. We showed in BrownieCorrector by using the paired-end read information we could improve the quality of error correction and later the assembly. However, what if the repeat size extends spanned by paired-end reads? In this case, using BrownieCorrector cannot help. However, using reads from different libraries or using accurate mate pairs with a low deviation in insert size can be helpful. In the case of having multiple libraries, BrownieCorrector corrects each library independently. However, a better idea could be using the information from all the libraries altogether. Especially for the clustering, for example, one can verify and correct the read clustering obtained from the reads in one library with the reads in another library.

The problem of *de novo* genome assembly was born and exists today due to the limitation in the technology. There has not been a sequencer machine that can retrieve the whole DNA sequence in one run. However, with the current advances in sequencing technology, today it is possible to generate reads even up to a few Mb in length. To assemble repeat-rich genomes using longer reads can help to resolve more repeats which can result in longer and more accurate contigs. However, it does not mean that short reads are not useful for the genome assembly, still, over 90% of the available sequencing data are Illumina short reads. Generating short reads costs less. For example, the price to sequence 1 Mb data with PacBio is

approximately 20 times more than the price to sequence the same amount of data with Illumina MiSeq. Another issue is the error rate, while long reads suffer from a higher error rate (up to 15%), short reads are very accurate(1–2% error rate). Maybe the best approach is taking advantage of both types, using both short and long reads in a hybrid fashion.

Finally, the EC tools try to detect and correct sequencing errors; however, we assume that the sequencing data are obtained from healthy tissue. For example, in cancer data, even a single difference of nucleotide could be a sign of a somatic mutation and doing error correction is equal to destroying information and evidence. Therefore, we highly suggest not to perform any kind of error correction on sequencing data that are obtained from cancer tissues.

References

[1] Antoine Limasset, Bastien Cazaux, Eric Rivals, and Pierre Peterlongo. *Read mapping on de Bruijn graphs*. BMC Bioinformatics, 17(1):237, 2016. 1-28, 4-3, 4-4, 5-3

[2] Bo Liu, Hongzhe Guo, Michael Brudno, and Yadong Wang. *DeBGA: Read alignment with de Bruijn graph-based seed and extension*. Bioinformatics, 32(21):3224–3232, 2016. 1-28, 4-3, 5-3

[3] Saul B Needleman and Christian D. Wunsch. *A general method applicable to the search for similarities in the amino acid sequence of two proteins*. Journal of Molecular Biology, 48(3):443–453, Mar. 1970. 1-24, 3-4, 4-4, 5-3

[4] Gerwin Dox. *Efficient algorithms for pairwise sequence alignment on graphs*. Master dissertation, University of Ghent, 2018. 5-3

[5] Amin Allam, Panos Kalnis, and Victor Solovyev. *Karect: accurate correction of substitution, insertion and deletion errors for next-generation sequencing data.* Bioinformatics, 31(July):3421–3428, July 2015. 1-17, 2-3, 3-3, 3-6, 5-4

[6] Siavash Sheikhizadeh and Dick de Ridder. *ACE: accurate correction of errors using K-mer tries.* Bioinformatics, 31(19):3216–8, 2015. 1-17, 2-3, 3-3, 3-6, 5-4

[7] Mahdi Heydari, Sayed-Amir Marashi, Ruzbeh Tusserkani, and Mehdi Sadeghi. *Reconstruction of phylogenetic trees of prokaryotes using maximal common intervals*. Biosystems, 124:86–94, October 2014. 5-6

A

Supplementary Data: Evaluation of the Impact of Illumina Error Correction Tools on de novo Genome Assembly

"... Yesterday I was clever, so I wanted to change the world. Today I am wise, so I am changing myself.[1] *"*

A.1 Error Correction Tool Parameter Settings

All error correction tools were executed with 32 threads. Some tools need to be provided with the approximate genome size. For those tools, the exact genome size was provided. Some tools that internally operate on k-mers allow the user to specify the value of k. For all tables and figures in the chapter 2 and in supplementary data, except Table 4 in the chapter 2, the default or recommended value of k was taken for each tool, regardless of dataset or assembly tool that was used.

To generate the results of Table 4 in the chapter 2, the optimal value of k was selected for each EC tool/dataset combination by running the EC tool multiple times with different k-mer sizes. The optimal k-mer size then corresponds to the SPAdes assembly that yields the highest scaffold N50. This way of selecting the optimal value of k would be identical to an end-user who wants to optimally assem-

[1] Rumi

ble a new genome in absence of a reference genome. Below, the actual parameters are specified for each tool individually:

A.1.1 ACE

The k-mer size for ACE is built-in and cannot be specified by the user.

```
$ size=$(stat -c\%s genome.fasta)
$ ./ace $size $inputreads aceOut/aceCorrected

```

A.1.2 BayesHammer v. 3.7.1

The k-mer size for BayesHammer is built-in and cannot be specified by the user.

```
$ ./spades.py -t 32 --careful --12 $inputreads -o
   bayesHammerOut  --only-error-correction --disable-
   gzip-output
```

A.1.3 BFC v. r181

```
$ size=$(stat -c\%s genome.fasta)
$ ./bfc -s $size -k 33 -t 32 $inputreads >bfcOut/
   bfcCorrected
```

Table A.1 shows the scaffold N50 of the SPAdes assembly from BFC-corrected reads for different values of k used in BFC.

Table A.1 Scaffold N50 of the SPAdes assembly from BFC-corrected reads

k-mer size	D1	D2	D3	D4	D5	D6	D7	D8
29	287 949	108 254	133 309	723 537	289 353	**16 529**	**7 879**	82 402
31	287 949	108 254	133 309	723 537	289 353	16 294	7 869	80 872
33	**287 949**	**108 254**	**133 309**	**723 537**	**289 353**	16 484	7 871	82 203
35	287 949	107 839	133 309	723 537	289 353	16 273	7 857	**83 364**
37	287 949	107 839	133 309	723 537	289 353	15 856	7 865	82 486

Therefore, in Table 4 of the chapter 2, k = 33 was used for datasets D1, D2, D3, D4 and D5; k = 29 was used for datasets D6 and D7; k = 35 was used for dataset D8. The default value of k = 33 was used for all other tables and figures.

A.1.4 BLESS 2 v. 1.02

```
$ ./bless -read $inputreads -prefix blessOut/
   blessCorrected -kmerlength 31 -smpthread 32
```

Table A.2 shows the scaffold N50 of the SPAdes assembly from BLESS-corrected reads for different values of k used in BLESS 2.

Table A.2 Scaffold N50 of the SPAdes assembly from BLESS 2-corrected reads

k-mer size	D1	D2	D3	D4	D5	D6	D7	D8
27	397 392	**108 357**	**133 309**	381 567	264 881	**14 672**	3 894	44 783
29	397 392	108 254	**133 309**	381 537	264 881	13 970	**3 996**	**45 633**
31	**397 392**	108 254	126 410	381 864	264 881	13 678	3 931	43 315
33	397 392	108 254	126 410	**412 097**	264 881	14 267	3 834	41 579
35	397 392	108 254	126 410	412 031	**289 353**	14 180	3 723	41 677

Therefore, in Table 4 of the chapter 2, k = 27 was used for datasets D2, D3 and D6; k = 29 was used for datasets D7 and D8; k = 33 was used for dataset D4; k = 35 was used for dataset D5. The default value of k = 31 was used for all other tables and figures.

A.1.5 Blue v. 1.1.2

```
$  size=$(stat -c\%s genome.fasta)
$  mono Tessel.exe -k 25 -g $size  -t 32 Cspor $
   inputreads
$ mono Blue.exe -r blueCorrected -t 32  -o blueOut
   Cspor_31.cbt reads
```

Table A.3 shows the scaffold N50 of the SPAdes assembly from Blue-corrected reads for different values of k used in Blue.

Table A.3 Scaffold N50 of the SPAdes assembly from Blue-corrected reads

k-mer size	D1	D2	D3	D4	D5	D6	D7	D8
21	**292 264**	107 776	133 088	723 550	**314 485**	13 572	7 454	77 807
25	287 948	107 774	**133 309**	**723 550**	289 314	13 214	**7 708**	83 277
27	287 948	**108 189**	133 309	723 366	289 314	**13 690**	7 686	84 876
29	287 948	**108 189**	133 309	723 537	289 314	13 397	7 685	85 463
31	287 948	**108 189**	133 309	723 537	289 314	13 300	7 682	**86 523**

Therefore, in Table 4 of the chapter 2, k = 21 was used for datasets D1 and D5; k = 25 was used for datasets D3, D4 and D7; k = 27 was used for dataset D6; k =

31 was used for datasets D2 and D8. The default value of k = 25 was used for all other tables and figures.

A.1.6 Fiona v. 0.2.5

The k-mer size for Fiona is built-in and cannot be specified by the user.

```
$ size=$(stat -c\%s genome.fasta)
$ ./fiona  -nt 32 -g $size  $inputreads
```

A.1.7 Karect v. 1.0

```
$ ./karect -correct -inputfile=$inputreads -matchtype=
   hamming -celltype=diploid -resultdir=karectOut -
   kmer=9 -memory=32 -threads=32
```

Table A.4 shows the scaffold N50 of the SPAdes assembly from Karect-corrected reads for different values of k used in Karect.

Table A.4 Scaffold N50 of the SPAdes assembly from Karect-corrected reads

k-mer	D1	D2	D3	D4	D5	D6	D7	D8
9	**287 949**	**108 228**	**133 309**	**725 282**	**289 353**	**17 170**	7 923	**88 533**
11	287 949	108 228	133 309	725 282	289 353	17 170	7 923	88 533
13	287 949	108 228	133 309	723 537	289 353	16 805	**7 944**	88 135
14	287 949	108 228	133 309	723 537	289 353	16 866	7 917	87 347

Therefore, in Table 4 of the chapter 2, k = 9 was used for datasets D1, D2, D3, D4, D5, D6 and D8; k = 13 was used for dataset D7; The default value of k = 9 was used for all other tables and figures. It should be noted that Karect sometimes overrides the user-specified value for k to a value it considers to be more suitable.

A.1.8 Lighter v. 1.1.0

```
$ size=$(stat -c\%s genome.fasta)
$ ./lighter  -t 32 -K 17 $size -r $inputreads -od
   lighterOut/lighterCorrected
```

Table A.5 shows the scaffold N50 of the SPAdes assembly from Lighter-corrected reads for different values of k used in Lighter.

Therefore, in Table 4 of the chapter 2, k = 13 was used for datasets D2, D5 and D7; k = 15 was used for dataset D8; k = 17 was used for datasets D1, D3 and D4;

Table A.5 Scaffold N50 of the SPAdes assembly from Lighter-corrected reads

k-mer	D1	D2	D3	D4	D5	D6	D7	D8
13	287 949	**108 254**	133 309	412 163	**289 353**	15 197	**7 946**	83 760
15	287 949	107 839	133 309	723 537	289 353	16 096	7 599	**84 674**
17	**287 949**	107 839	**133 309**	**723 537**	264 881	16 666	7 154	79 363
19	287 949	107 839	133 309	723 537	289 353	16 734	7 455	80 331
21	287 949	107 839	133 309	723 537	289 353	**16 826**	7 726	82 346

k = 21 was used for dataset D6. The default value of k = 17 was used for all other tables and figures.

A.1.9 Musket v. 1.1

```
$  ./musket -inorder -p 32 $inputreads -o musketOut/
   musketCorrected
```

Table A.6 shows the scaffold N50 of the SPAdes assembly from Musket-corrected reads for different values of k used in Musket.

Table A.6 Scaffold N50 of the SPAdes assembly from Musket-corrected reads

k-mer	D1	D2	D3	D4	D5	D6	D7	D8
17	287 949	107 839	133 309	723 537	**289 353**	16 152	5 808	58 138
21	**287 949**	107 839	**133 309**	**723 537**	264 881	**16 419**	6 334	63 519
23	287 949	107 839	125 608	723 537	**289 353**	16 414	7 190	70 009
25	287 949	**108 254**	125 608	723 537	**289 353**	16 260	7 598	69 990
27	287 949	**108 254**	133 309	723 537	**289 353**	15 816	**7 704**	**75 521**

Therefore, in Table 4 of the chapter 2, k = 21 was used for datasets D1, D3, D4 and D6; k = 27 was used for dataset D2, D5, D7 and D8. The default value of k = 21 was used for all other tables and figures.

A.1.10 RACER v. 1.0.1

The k-mer size for RACER is built-in and cannot be specified by the user.

```
$  size=$(stat -c\%s genome.fasta)
$ ./racer.exe $inputreads racerOut/Corrected $size
```

A.1.11 SGA-EC v. 0.10.14

No k-mer value has to be specified for SGA-EC.

```
$ ./sga preprocess --permute-ambiguous  --no-primer-
   check -o sgaOut/temp -p=2 -m 11 $inputreads
$ ./sga index -a ropebwt -t 32 --no-reverse sgaOut/
   temp
$ ./sga correct --learn -t 32 -o sgaOut/sgaCorrectede
   sgaOut/temp
```

A.1.12 Trowel v. 0.2.0.4

```
$ echo $inputreads> trowelOut/trowelInput
$ ./trowel.0.2.0.4.linux.64 -k 11 -t 32 -f trowelOut/
   trowelCorrected
```

Table A.7 shows the scaffold N50 of the SPAdes assembly from Trowel-corrected reads for different values of k used in Trowel.

Table A.7 Scaffold N50 of the SPAdes assembly from Trowel-corrected reads

k-mer	D1	D2	D3	D4	D5	D6	D7	D8
11	**287 948**	**108 254**	**133 309**	**723 537**	**289 353**	13 493	7 913	77 879
13	287 948	108 254	133 189	723 537	289 353	14 000	**7 938**	77 308
15	287 948	108 254	133 309	723 537	289 353	**14 042**	7 909	**79 217**

Therefore, in Table 4 of the chapter 2, k = 11 was used for datasets D1, D2, D3, D4 and D5; k = 13 was used for dataset D7; k = 15 was used for datasets D6 and D7. The default value of k = 11 was used for all other tables and figures.

A.2 Data simulation

To compare the performance of error correction tools (EC tools) on simulated data, we produced synthetic Illumina reads for the same set of organisms for which real data was used. The ART read simulator is used to generate reads, with the following command:

```
$ ./art_illumina -i genome.fasta -p -l [len] -f [cov]
  -m 300 -s 30 -o reads
```

Where `cov` and `len` correspond to the coverage and length and change according to the values in real datasets. The mean fragment size is 300 bp, and the fragment standard deviation is 30 bp.

A.3 Error Metrics

A.3.1 Alignment ratio

Reads are grouped based on the number of mismatches (m) after aligning them to the reference genome using BWA v. 0.7.12:

```
$ ./bwa mem -M -t 32 -p reference/genome.fasta reads.
  fastq >alignment/samfileName.sam
```

Table A.8 shows the percentage of reads that align to the reference genome without mismatches ($m = 0$). Table A.9 shows the percentage of reads that do not align to the reference genome with $<$10 mismatches. The 'Uncorrected' row shows the results of the raw data.

Table A.8 Percentage of reads that mapped with 0 mismatches (%).

Tool	*B. dentium*	*E. coli DH10B*	*E. coli MG1655*	*S. enterica*	*P. aeruginosa*	*H. sapiens* Chr. 21	*C. elegans*	*D. melanogaster*
Uncorrected	70.16	76.31	61.07	55.17	83.64	69.36	70.41	50.32
ACE	98.04	98.83	98.83	94.61	98.75	82.33	79.09	60.14
BayesHammer	93.40	95.77	89.37	90.15	95.95	81.46	78.01	57.75
BFC	97.41	98.86	96.65	92.35	98.48	83.09	78.57	59.82
BLESS 2	96.86	98.43	96.48	90.14	98.24	83.46	78.48	59.66
Blue	98.58	99.64	99.15	94.23	99.11	82.28	78.41	60.07
Fiona	96.26	97.93	96.61	89.49	97.64	82.16	78.11	59.23
Karect	98.04	99.29	98.84	93.94	98.51	83.17	79.07	59.97
Lighter	96.72	98.58	97.87	89.63	97.09	81.36	77.15	58.70
Musket	97.67	98.99	96.82	92.17	98.40	81.78	77.80	59.72
RACER	97.30	98.58	98.01	91.40	98.40	82.21	78.51	59.69
SGA-EC	90.14	93.76	78.88	86.35	94.97	81.51	78.06	58.06
Trowel	85.63	90.76	82.72	79.01	92.37	75.85	74.68	54.90

Table A.9 Percentage of reads that do not align with <10 mismatches.

Tool	*B. dentium*	*E. coli DH10B*	*E. coli MG1655*	*S. enterica*	*P. aeruginosa*	*H. sapiens* Chr. 21	*C. elegans*	*D. melanogaster*
Uncorrected	2.76	0.95	0.39	9.99	0.90	0.62	18.11	6.80
ACE	0.33	0.01	0.11	3.20	0.28	0.33	17.12	6.34
BayesHammer	1.27	0.15	0.02	3.33	0.40	0.22	16.69	6.47
BFC	1.45	0.19	0.30	5.20	0.66	0.54	18.05	6.67
BLESS 2	1.32	0.19	0.49	5.07	0.66	0.61	18.07	6.84
Blue	1.09	0.03	0.20	4.31	0.49	0.44	17.88	6.60
Fiona	1.48	0.11	0.17	5.40	0.51	0.41	17.78	6.60
Karect	1.11	0.02	0.17	3.88	0.43	0.39	17.73	6.58
Lighter	1.47	0.10	0.19	5.45	0.59	0.47	17.84	6.63
Musket	1.34	0.15	0.26	4.96	0.62	0.47	17.89	6.60
RACER	1.31	0.14	0.23	4.88	0.51	0.45	17.70	6.65
SGA-EC	2.54	0.76	0.35	9.14	0.75	0.55	18.06	6.70
Trowel	1.95	0.56	0.23	7.03	0.69	0.53	17.69	6.70

A.3.2 EC gain

A.3.2.1 Accuracy comparison method

Accuracy comparison of EC tools in simulated data is straightforward since the perfect read is known. Let R represent an input read. For each read R, there is a corresponding read C which is corrected by any of the EC tools. In artificial data, a perfect read P is provided together with R. Therefore, for the evaluation of tools in simulated data, bases in these three reads (R, P and C) are compared and classified as follows:

$$tp : \mathrm{R}_c \neq \mathrm{P}_c \text{ and } \mathrm{C}_c = \mathrm{P}_c, \mathrm{TP} = \sum_{c \in R} tp;$$
$$tn : \mathrm{R}_c = \mathrm{P}_c \text{ and } \mathrm{C}_c = \mathrm{P}_c, \mathrm{TN} = \sum_{c \in R} tn;$$
$$fp : \mathrm{R}_c = \mathrm{P}_c \text{ and } \mathrm{C}_c \neq \mathrm{P}_c, \mathrm{FP} = \sum_{c \in R} fp;$$
$$fn : \mathrm{R}_c \neq \mathrm{P}_c \text{ and } \mathrm{C}_c \neq \mathrm{P}_c, \mathrm{FN} = \sum_{c \in R} fn.$$

The error correction gain (EC gain) is then defined as

$$\text{EC gain} = \frac{\mathrm{TP} - \mathrm{FP}}{\mathrm{TP} + \mathrm{FN}}.$$

An EC gain of 100% means all errors were corrected and no new errors were introduced. Additionally, sensitivity and false positive rate (FP rate) are defined as follows:

$$\text{Sensitivity} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}},$$
$$\text{FP rate} = \frac{\mathrm{FP}}{\mathrm{TN} + \mathrm{FP}}.$$

A.3.2.2 Real Data

Exact numbers of TP, TN, FP and FN for real data are shown in Table A.10.

Table A.10 Detailed confusion matrices for real data.

Tool	*B. dentium*	*E. coli DH10B*	*E. coli MG1655*	*S. enterica*	*P. aeruginosa*	*H. sapiens* Chr. 21	*C. elegans*	*D. melanogaster*
				Dataset				
				True Positives (TP) – corrected errors				
ACE	3076751	8332547	22291105	3701334	3292214	5032509	10795486	25245221
BFC	2482403	7142877	18036534	3083801	2670043	4444675	9330314	18081922
Blue	3112185	8400946	22279091	3674979	3246700	4915665	10326983	22965194
Fiona	2757539	8018906	21957781	3239983	3111415	4978875	10486053	23052956
Karect	3133184	8450316	22414636	3732389	3331053	5287884	11220434	25992487
Lighter	2694538	7952212	20837800	3038404	2927161	4014919	8650557	17759054
Musket	2878390	7922398	21005281	3335539	2961995	4345534	9554662	19344147
RACER	2932759	8112897	22077431	3374675	3215041	4865921	9987137	24864167
SGA-EC	1742936	5690254	10231654	2010468	2215005	4137311	8965373	16986433
Trowel	1229337	4226181	9761215	1550337	1613647	1936362	4575638	8640337
				True Negatives (TN) – initially correct bases left untouched				
ACE	963422263	1931235385	2746256917	440139607	1039050327	1319616601	4737356765	5606869096
BFC	964033360	1931274464	2746348400	441245208	1039248109	1321102045	4748425050	5612713036
Blue	964029362	1931244564	2746290398	441235585	1039242118	1320502326	4747747773	5611271694
Fiona	964035021	1931270828	2746332983	441244937	1039242743	1320774172	4747710215	5611944730
Karect	964034920	1931277673	2746357548	441245411	1039250137	1321107050	4748328371	5612244594
Lighter	964033340	1931268196	2746328586	441242591	1039192194	1320945136	4747204149	5611259898
Musket	964034290	1931275869	2746352663	441244510	1039244912	1320929786	4747597357	5610975777
RACER	964010310	1931159567	2746106691	441215449	1039223611	1317855370	4743412687	5604761363
SGA-EC	964034237	1931275385	2746351399	441245198	1039248037	1321074542	4748393243	5612994706
Trowel	964015541	1931230735	2745348701	441227802	1039224947	1320368303	4744149598	5612170398
				False Positives (FP) – newly introduced errors				
ACE	42115	45236	110238	66473	201973	1607588	11276793	6306244
BFC	2017	5828	18069	721	2751	110312	73154	411719
Blue	7660	38812	81267	13731	9982	722472	795185	1916850
Fiona	1939	12563	37718	2501	9327	458681	867700	1224861
Karect	457	2624	8951	518	723	105756	171201	881629
Lighter	2038	12099	37887	3339	58686	267293	1294892	1865549
Musket	1087	4430	13816	1419	5951	282661	901130	2149247
RACER	20414	119807	266323	25624	27673	3439109	5209526	8553829
SGA-EC	1143	4912	15115	735	2890	138742	106266	134578
Trowel	19913	49616	1032119	18304	26030	854611	4415136	964727
				False Negatives (FN) – remaining errors				
ACE	72832	129122	189116	76350	100688	3178135	3838512	47947216
BFC	669265	1318746	4443590	704083	722347	3763717	5294722	55097487
Blue	39497	60730	201157	112984	145698	3293805	4298919	50218318
Fiona	393729	442232	522381	547490	280918	3236482	4143215	50179596
Karect	18484	11307	65491	55495	61337	2920538	3404740	47187117
Lighter	457130	509413	1642324	749481	465237	4193495	5974735	55420668
Musket	273278	539225	1474842	452345	430394	3862860	5070408	53835292
RACER	223971	352560	403619	418534	177627	3351498	4654443	48340247
SGA-EC	1408740	2771371	12248650	1777441	1177881	4071424	5660045	56205417
Trowel	1922344	4235705	12724566	2237577	1778750	6274148	10071360	64540776

A.3.2.3 Simulated Data

Table A.11 shows the EC gain, sensitivity, and specificity expressed as number of errors introduced per Mbp of read data. Table A.12 shows the exact numbers of TP, TN, FP and FN for all tools on the simulated data. BFC has the highest gain on four datasets, Karect and Fiona both have the highest gain on two datasets.

Table A.11 Accuracy comparison in terms of EC gain, percentage of corrected errors, and number of errors introduced per Mbp in simulated data.

	Dataset							
Tool	*B. dentium*	*E. coli DH10B*	*E. coli MG1655*	*S. enterica*	*P. aeruginosa*	*H. sapiens* Chr. 21	*C. elegans*	*D. melanogaster*
				Error correction gain (%)				
ACE	99.8	99.7	99.8	99.9	99.2	81.2	74.3	95.8
BFC	99.9	99.8	98.7	99.9	99.8	96.2	99.3	99.7
Blue	99.7	99.0	99.1	99.7	99.5	85.1	91.0	97.2
Fiona	99.9	99.9	99.7	99.8	99.9	90.3	96.2	98.1
Karect	99.9	99.8	99.9	99.9	99.8	92.8	98.0	99.4
Lighter	99.3	99.6	99.5	99.2	98.2	84.8	83.6	90.0
Musket	99.8	99.7	99.8	99.8	99.4	88.5	93.3	97.9
RACER	99.2	98.9	98.8	99.2	98.6	58.9	69.1	87.6
SGA-EC	99.6	99.7	15.8	99.8	99.8	89.8	96.8	97.6
Trowel	84.1	83.5	74.4	85.7	81.6	67.9	76.1	75.8
				Sensitivity – percentage of corrected errors (%)				
ACE	100.0	99.9	99.9	100.0	99.9	94.3	98.5	98.7
BFC	99.9	99.8	98.7	99.9	99.9	97.8	99.4	99.8
Blue	100.0	100.0	99.6	100.0	100.0	91.3	97.3	98.3
Fiona	99.9	99.9	99.8	99.9	99.9	95.6	98.8	98.6
Karect	100.0	100.0	100.0	100.0	100.0	93.1	99.3	99.5
Lighter	99.4	99.7	99.6	99.3	98.9	88.6	89.3	92.7
Musket	99.8	99.7	99.8	99.8	99.5	90.0	94.1	98.2
RACER	99.9	99.9	99.9	99.9	99.9	93.0	97.7	98.6
SGA-EC	99.6	99.7	15.9	99.8	99.9	92.1	97.8	97.8
Trowel	84.3	83.5	74.6	85.8	81.6	70.0	76.8	76.0
				Number of errors introduced per Mbp of read data				
ACE	12	4	10	9	12	1271	389	280
BFC	1	0	1	1	0	156	2	8
Blue	25	19	43	28	8	603	102	105
Fiona	3	2	13	5	1	520	41	49
Karect	5	3	4	5	2	35	22	13
Lighter	5	1	11	7	13	368	91	267
Musket	1	0	2	1	1	147	13	28
RACER	71	20	101	60	24	3312	460	1063
SGA-EC	1	1	6	2	1	221	16	19
Trowel	18	1	13	6	0	203	11	16

Table A.12 Detailed confusion matrices in simulated data

	Dataset							
Tool	*B. dentium*	*E. coli DH10B*	*E. coli MG1655*	*S. enterica*	*P. aeruginosa*	*H. sapiens* Chr. 21	*C. elegans*	*D. melanogaster*
				True Positives (TP) – corrected errors				
ACE	9667046	3814496	27260933	4359012	1847493	12266571	9180759	59347214
BFC	9659287	3808751	26926157	4357828	1846499	12728393	9263433	59962209
Blue	9666564	3815467	27161340	4359254	1848540	11881359	9070130	59069399
Fiona	9660940	3815149	27232134	4356236	1847967	12446652	9208110	59246494
Karect	9669223	3815493	27270690	4359797	1848298	12116343	9257419	59816346
Lighter	9608363	3806153	27164122	4328575	1828813	11529659	8318765	55726718
Musket	9650131	3805811	27216841	4352679	1839474	11713659	8768168	59020638
RACER	9676980	3815294	27244781	4362894	1847896	12114016	9116775	59266390
SGA-EC	9632731	3806414	4340616	4351854	1846414	11986144	9115211	58761640
Trowel	8151159	3188984	20336922	3740932	1508617	9109108	7155992	45677232
				True Negatives (TN) – initially correct bases left untouched				
ACE	973643656	1954926851	2812136997	469845584	1056817153	1339880180	5805025579	6198006667
BFC	973654353	1954933880	2812161687	469849372	1056829305	1341366583	5807245715	6199676406
Blue	973637810	1954901136	2812074791	469840224	1056822852	1340786550	5806685096	6199146502
Fiona	973652784	1954931510	2812129698	469847730	1056828691	1340902736	5807025162	6199437609
Karect	973650880	1954929616	2812153109	469847558	1056827837	1341529923	5807130529	6199647692
Lighter	973650027	1954932142	2812133532	469846600	1056816359	1341081880	5806729873	6198071892
Musket	973654027	1954933908	2812157389	469849138	1056829003	1341378585	5807181419	6199553951
RACER	973568663	1954896583	2811887873	469814560	1056805849	1337241860	5804649423	6193298979
SGA-EC	973653813	1954932504	2812146301	469848917	1056828854	1341281680	5807162255	6199608343
Trowel	973637311	1954932965	2812127117	469847097	1056829444	1341303372	5807194307	6199626029
				False Positives (FP) – newly introduced errors				
ACE	11677	8424	27799	4317	12971	1704723	2259541	1733689
BFC	530	784	2325	308	453	208911	9755	48551
Blue	23995	37162	120731	12952	8471	809236	590384	650265
Fiona	2641	3237	36199	2201	1114	697478	239447	301954
Karect	4853	5856	11865	2481	2273	46854	128678	80121
Lighter	4864	2539	30517	3084	13411	493732	525707	1653388
Musket	864	774	6650	547	765	196952	74116	171110
RACER	68975	39371	284440	27985	24868	4443967	2673149	6591309
SGA-EC	1169	2366	17850	811	1015	296150	95000	118356
Trowel	17589	1703	36902	2593	322	272190	61239	98964
				False Negatives (FN) – remaining errors				
ACE	3077	2683	14041	1593	1548	748107	139964	765300
BFC	10832	8426	348803	2775	2539	285163	55299	148821
Blue	3572	1715	113685	1364	503	1132600	248773	1041986
Fiona	9208	2036	42942	4406	1075	569220	112234	866386
Karect	898	1685	4283	806	740	897266	61366	294752
Lighter	61758	11026	110850	32029	20236	1483959	1000030	4384507
Musket	19988	11366	58119	7925	9564	1299912	550568	1090406
RACER	13078	2956	31690	5943	1316	916081	214034	863661
SGA-EC	37401	10771	22934352	8758	2629	1027556	203649	1349522
Trowel	1518960	628194	6938039	619671	340421	3904451	2162752	14433803

A.4 Assembly Result for Real Data

In order to see the impact of EC tools on assembly results, we used SPAdes, DISCOVAR, IDBA and Velvet to assemble both corrected and uncorrected data. Quast provides comprehensive information on the assembly quality. For the Quast analyses, scaffolds were used. The following commands were used to run the assemblers.

- SPAdes (v. 3.7.1)

```
$ spades.py -t 32 --only-assembler --12 reads.
    fastq -o outputDir
```

- DISCOVAR

```
$ DiscovarDeNovo READS=reads.fastq OUT_DIR=
    outputDir  MEMORY_CHECK=true NUM_THREADS=32
```

- Velvet (v. 1.2.10)

```
$ ./velveth asmDir 31 -fastq -shortPaired reads.
    fastq
$ ./velvetg asmDir -exp_cov auto -cov_cutoff auto
```

- IDBA(v. 1.1.1)

```
$ ./idba --no_correct -r reads.fa -o outputDir --
    num_threads 32
```

Quast results for each dataset are shown in the following subsections. Assemblies were named after the EC tool that was used to preprocess the data. The 'Uncorrected' assembly was obtained from uncorrected data. Default parameter settings are used for Quast, i.e., all statistics are based on contigs of size $\geq$ 500 bp. The Quast (v. 4.4) command line was:

```
./quast.py asmDir/contigs.fa -R genome.fasta -o
    quastReport --plots-format ps -1
    uncorrectedForwardRead.fq  -2
    uncorrectedReverseRead.fq --labels\ "toolName"
```

A.4.1 DISCOVAR

A.4.1.1 *B. dentium*

Table A.13 contains the Quast report after assembling dataset *B. dentium* with DISCOVAR.

A.4.1.2 *E. coli str. K-12 substr. DH10B*

Table A.14 contains the Quast report after assembling dataset *E. coli str. K-12 substr. DH10B* with DISCOVAR.

A.4.1.3 *E. coli str. K-12 substr. MG1655*

Table A.15 contains the Quast report after assembling dataset *E. coli str. K-12 substr. MG1655* with DISCOVAR.

A.4.1.4 *S. enterica*

Table A.16 contains the Quast report after assembling dataset *S. enterica* with DISCOVAR.

A.4.1.5 *P. aeruginosa*

Table A.17 contains the Quast report after assembling dataset *P. aeruginosa* with DISCOVAR.

A.4.1.6 *H. sapiens* Chr. 21

Table A.18 contains the Quast report after assembling dataset *H. sapiens* Chr. 21 with DISCOVAR.

A.4.1.7 *C. elegans*

Table A.19 contains the Quast report after assembling dataset *C. elegans* with DISCOVAR.

A.4.1.8 *D. melanogaster*

Table A.20 contains the Quast report after assembling dataset *D. melanogaster* with DISCOVAR.

Table A.13 Assembly quality metrics for *B. dentium*

Assembly	Uncorrected	ACE	BayesHammer	BFC	BLESS2	Blue	Karect	Lighter	Musket	RACER	SGA-EC
# contigs (≥ 0 bp)	41	40	34	34	28	40	31	34	37	31	34
# contigs (≥ 1000 bp)	19	22	18	18	17	23	18	18	20	18	18
# contigs (≥ 5000 bp)	17	18	16	16	15	18	16	16	18	15	16
# contigs (≥ 10000 bp)	15	16	14	14	13	16	14	14	16	13	14
# contigs (≥ 25000 bp)	15	15	14	14	13	14	14	14	15	13	14
# contigs (≥ 50000 bp)	11	10	10	10	10	10	10	10	11	10	10
Total length (≥ 0 bp)	2620312	2620550	2618818	2618819	2615877	2618828	2618035	2618817	2618219	**2907328**	2618819
Total length (≥ 1000 bp)	2613881	2614837	2613775	2613776	2612673	2613044	2613959	2613774	2612796	**2903513**	2613776
Total length (≥ 5000 bp)	2610478	2607949	2610372	2610373	2609270	2604556	2610556	2610371	2609393	**2896805**	2610373
Total length (≥ 10000 bp)	2599192	2596663	2599086	2599087	2597984	2593270	2599270	2599085	2598107	**2885519**	2599087
Total length (≥ 25000 bp)	2599192	2581604	2599086	2599087	2597984	2567611	2599270	2599085	2580596	**2885519**	2599087
Total length (≥ 50000 bp)	2440269	2386840	2440165	2440166	2480462	2408689	2440349	2440165	2421675	**2767997**	2440166
# contigs	20	23	19	19	18	25	19	19	21	19	19
Largest contig	610357	573238	610358	**610359**	608981	608981	610358	610358	534355	610358	**610359**
Total length	2614583	2615539	2614477	2614478	2613375	2614479	2614661	2614476	2613498	**2904215**	2614478
Reference length	2636367	2636367	2636367	2636367	2636367	2636367	2636367	2636367	2636367	2636367	2636367
GC (%)	58.51	58.51	58.51	58.51	58.50	58.51	58.51	58.51	58.51	58.63	58.51
Reference GC (%)	58.54	58.54	58.54	58.54	58.54	58.54	58.54	58.54	58.54	58.54	58.54
N50	288053	271664	288054	288054	288054	271663	288053	288054	288054	**397536**	288054
NG50	288053	271664	288054	288054	288054	271663	288053	288054	288054	**397536**	288054
N75	148905	158597	158597	158597	172541	158597	158597	158597	158597	**174159**	158597
NG75	148905	158597	158597	158597	172541	158597	158597	158597	158597	**190580**	158597
L50	4	4	4	4	4	4	4	4	4	**3**	4
LG50	4	4	4	4	4	4	4	4	4	**3**	4
L75	7	7	7	7	7	7	7	7	7	**6**	7
LG75	7	7	7	7	7	7	7	7	7	**5**	7
# misassemblies	**0**	**0**	**0**	**0**	1	1	**0**	**0**	1	1	**0**
# misassembled contigs	**0**	**0**	**0**	**0**	1	1	**0**	**0**	1	1	**0**
Misassembled contigs length	**0**	**0**	**0**	**0**	608981	608981	**0**	**0**	534355	578715	**0**
# local misassemblies	**1**	2	**1**	**1**	2	**1**	**1**	**1**	**1**	2	**1**
# unaligned mis. contigs	0	0	0	0	0	0	0	0	0	0	0
# unaligned contigs	0 + 0 part	0 + 0 part	0 + 0 part	0 + 0 part	0 + 0 part	0 + 0 part	0 + 0 part	0 + 0 part	0 + 0 part	0 + 0 part	0 + 0 part
Unaligned length	0	0	0	0	0	0	0	0	0	0	0
Genome fraction (%)	**99.116**	**99.116**	**99.116**	**99.116**	99.064	99.064	**99.116**	**99.116**	99.064	**99.116**	**99.116**
Duplication ratio	**1.001**	**1.001**	**1.001**	**1.001**	**1.001**	**1.001**	**1.001**	**1.001**	**1.001**	1.111	**1.001**
# N's per 100 kbp	**0.00**	3.82	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	3.44	**0.00**
# mismatches per 100 kbp	0.31	0.31	0.34	**0.23**	0.27	0.27	0.31	0.27	**0.23**	0.31	**0.23**
# indels per 100 kbp	0.31	0.38	0.31	0.34	0.31	**0.23**	0.27	0.38	0.31	0.27	0.34
Largest alignment	610357	572957	610358	**610359**	425802	425802	610358	610358	425802	610358	**610359**
Total aligned length	2614583	2615073	2614477	2614478	2613100	2614294	2614476	2614476	2613498	**2903840**	2614478
NA50	288053	271664	288054	288054	288054	271663	288053	288054	288054	**289208**	288054
NGA50	288053	271664	288054	288054	288054	271663	288053	288054	288054	**289208**	288054
NA75	148905	158597	158597	158597	172541	158597	158597	158597	148905	**174159**	158597
NGA75	148905	158597	158597	158597	172541	158597	158597	158597	148905	**190305**	158597
LA50	4	4	4	4	4	4	4	4	4	4	4
LGA50	4	4	4	4	4	4	4	4	4	4	4
LA75	**7**	**7**	**7**	**7**	8	8	**7**	**7**	8	**7**	**7**
LGA75	7	7	7	7	8	8	7	7	8	**6**	7

Table A.14 Assembly quality metrics for *E. coli str. K-12 substr. DH10B*

Assembly	Uncorrected	ACE	BayesHammer	BFC	BLESS2	Blue	Karect	Lighter	Musket	RACER	SGA-EC
# contigs (≥ 0 bp)	274	323	291	238	264	297	207	269	308	295	186
# contigs (≥ 1000 bp)	79	73	74	74	71	75	73	71	74	72	75
# contigs (≥ 5000 bp)	61	58	60	60	57	60	58	56	59	57	58
# contigs (≥ 10000 bp)	56	53	55	54	52	53	53	51	54	51	53
# contigs (≥ 25000 bp)	49	47	49	48	47	48	48	46	48	46	46
# contigs (≥ 50000 bp)	31	31	31	30	31	32	30	28	30	28	29
Total length (≥ 0 bp)	4592120	4872509	4554459	5034288	5148745	5301868	5248314	5234472	4721486	**5364848**	4399093
Total length (≥ 1000 bp)	4531525	4797173	4487930	4982888	5091870	5233985	5207238	5173957	4651894	**5297842**	4362699
Total length (≥ 5000 bp)	4487709	4759362	4451693	4945793	5055238	5196661	5166445	5135383	4613533	**5259841**	4314781
Total length (≥ 10000 bp)	4449112	4716994	4409556	4896653	5012927	5139964	5124088	5093028	4571240	**5209142**	4272644
Total length (≥ 25000 bp)	4337336	4624344	4317759	4804003	4935021	5062002	5046126	5015066	4478624	**5131267**	4151701
Total length (≥ 50000 bp)	3690981	4050737	3670597	4157698	4350700	**4461391**	4393097	4369294	3833349	4455347	3534867
# contigs	94	88	89	89	82	88	82	84	87	88	85
Largest contig	343551	652557	411914	1074043	421583	652365	**1172452**	652461	652557	837099	326329
Total length	4542042	4807144	4497803	4993249	5099259	5242952	5213422	5183003	4660793	**5308813**	4369755
Reference length	4686137	4686137	4686137	4686137	4686137	4686137	4686137	4686137	4686137	4686137	4686137
GC (%)	50.72	50.68	50.72	50.74	50.85	50.71	50.76	50.74	50.72	50.76	50.75
Reference GC (%)	50.78	50.78	50.78	50.78	50.78	50.78	50.78	50.78	50.78	50.78	50.78
N50	97356	133298	107848	132162	158602	163361	133298	165875	122330	**171773**	117698
NG50	95058	158602	97317	156704	163311	165947	163410	173935	122330	**177622**	107823
N75	59899	62449	59899	67455	**74835**	70471	67455	**74835**	59899	**74835**	61504
NG75	57141	70655	55107	76722	88864	87147	85684	88864	59899	**112586**	54707
L50	12	11	13	10	10	10	9	**8**	11	9	12
LG50	13	10	14	8	9	9	**7**	**7**	11	**7**	14
L75	26	24	26	23	23	24	22	**20**	25	**20**	25
LG75	28	22	29	20	19	18	17	16	25	**15**	29
# misassemblies	3	5	4	5	8	8	6	7	**2**	11	3
# misassembled contigs	3	5	4	3	8	8	4	7	**2**	8	3
Misassembled contigs length	634863	1149960	755823	1310461	1749325	1874510	1759738	2375577	774887	2222845	**401102**
# local misassemblies	14	3	6	**2**	3	8	5	**2**	3	9	9
# structural variations	0	0	0	0	0	0	0	0	0	0	0
# unaligned mis. contigs	0	0	0	0	0	0	0	0	0	0	0
# unaligned contigs	0 + 0 part	0 + 0 part	0 + 0 part	0 + 0 part	0 + 0 part	0 + 0 part	0 + 0 part	0 + 0 part	0 + 0 part	0 + 0 part	0 + 0 part
Unaligned length	0	0	0	0	0	0	0	0	0	0	0
Genome fraction (%)	92.370	92.439	**93.254**	92.496	92.422	92.584	92.419	92.452	92.424	92.655	92.391
Duplication ratio	1.049	1.110	1.029	1.152	1.177	1.208	1.204	1.196	1.076	1.223	**1.009**
# N's per 100 kbp	41.83	14.56	22.23	12.02	25.49	24.80	21.10	19.29	**10.73**	22.60	27.46
# mismatches per 100 kbp	1.85	1.02	1.21	0.62	**0.42**	1.01	1.34	0.65	1.06	0.92	1.62
# indels per 100 kbp	0.14	0.05	0.05	**0.02**	0.07	0.05	0.12	0.07	0.05	**0.02**	0.16
Largest alignment	**326329**	326328	326328	326129	326319	326232	326129	326328	326328	326322	**326329**
Total aligned length	4540136	4806444	4496802	4992649	5097903	5241579	5212322	5182003	4660277	**5307613**	4368499
NA50	95058	112586	89397	117698	117685	117698	**132162**	131963	117698	131963	107823
NGA50	92505	117698	88680	132162	133298	132162	**156703**	**156703**	117698	133298	97317
NA75	57141	59800	56402	62449	62250	62449	**67455**	**67455**	59799	62449	56567
NGA75	54607	67455	54261	76722	82808	83023	**85684**	82987	59799	82958	45709
LA50	14	13	14	13	14	14	13	13	**12**	14	13
LGA50	15	12	15	12	12	12	**11**	**11**	12	12	14
LA75	29	27	29	27	29	29	27	27	**26**	29	27
LGA75	31	25	32	24	24	23	**22**	**22**	26	23	32

Table A.15 Assembly quality metrics for *E. coli str. K-12 substr. MG1655*

Assembly	Uncorrected	ACE	BayesHammer	BFC	BLESS2	Blue	Karect	Lighter	Musket	RACER	SGA-EC
# contigs (≥ 0 bp)	136	205	1168	207	215	304	243	215	311	208	271
# contigs (≥ 1000 bp)	124	77	608	105	90	92	82	95	87	88	207
# contigs (≥ 5000 bp)	3	61	247	74	67	67	61	71	64	69	3
# contigs (≥ 10000 bp)	0	54	133	69	57	59	55	63	54	62	0
# contigs (≥ 25000 bp)	0	46	30	54	47	45	46	52	44	46	0
# contigs (≥ 50000 bp)	0	30	4	35	31	31	32	37	30	31	0
Total length (≥ 0 bp)	236904	4583890	4673678	4581495	4582720	4681774	4816991	4583866	**5001900**	4587380	398754
Total length (≥ 1000 bp)	233516	4547305	4511593	4549381	4542301	4625105	4767154	4548075	**4935390**	4554309	380114
Total length (≥ 5000 bp)	21752	4503937	3550363	4463067	4482054	4556845	4716578	4490654	**4877839**	4502287	24634
Total length (≥ 10000 bp)	0	4450193	2720691	4428272	4401991	4499387	4668891	4427303	**4798997**	4448767	0
Total length (≥ 25000 bp)	0	4316532	1134734	4194714	4221717	4276337	4525945	4245050	**4628383**	4182324	0
Total length (≥ 50000 bp)	0	3749345	256307	3471755	3672133	3780324	4057531	3703361	**4138053**	3644641	0
# contigs	124	82	624	111	102	99	96	103	99	94	210
Largest contig	9147	365228	73863	221402	238045	429293	471577	211687	418937	**480612**	9536
Total length	233516	4551290	4522598	4554385	4551550	4629907	4777246	4553997	**4944315**	4558950	382165
Reference length	4639675	4639675	4639675	4639675	4639675	4639675	4639675	4639675	4639675	4639675	4639675
GC (%)	39.56	50.73	50.70	50.74	50.74	50.69	50.72	50.72	50.67	50.73	41.64
Reference GC (%)	50.79	50.79	50.79	50.79	50.79	50.79	50.79	50.79	50.79	50.79	50.79
N50	1725	130621	13296	88143	112270	117964	112492	95324	**134676**	105110	1743
NG50	-	125346	12916	87072	106364	117964	117885	92024	164309	105110	-
N75	1423	59654	5929	57478	58977	59666	**67528**	58857	66071	**67528**	1396
NG75	-	58850	5266	49405	58105	59666	78892	57312	87166	61545	-
L50	44	12	94	18	13	12	12	17	**11**	15	74
LG50	-	13	99	19	14	12	11	18	10	15	-
L75	80	25	218	35	27	26	25	32	**23**	27	135
LG75	-	26	234	36	28	26	23	33	20	28	-
# misassemblies	**0**	9	1	9	6	13	7	5	10	16	**0**
# misassembled contigs	**0**	7	1	7	6	6	7	5	9	12	**0**
Misassembled contigs length	**0**	1135500	62521	494731	942233	1089006	1516331	510573	1791746	1387624	**0**
# local misassemblies	**0**	13	24	17	18	16	6	11	9	5	**0**
# unaligned mis. contigs	0	0	0	0	0	0	0	0	0	0	0
# unaligned contigs	0 + 0 part	0 + 0 part	0 + 0 part	0 + 0 part	0 + 0 part	0 + 0 part	0 + 0 part	0 + 0 part	0 + 0 part	0 + 0 part	0 + 0 part
Unaligned length	0	0	0	0	0	0	0	0	0	0	0
Genome fraction (%)	5.015	97.960	95.944	97.913	97.916	97.741	**98.317**	97.985	98.207	98.097	8.205
Duplication ratio	1.004	**1.001**	1.016	1.003	1.002	1.021	1.047	1.002	1.085	1.002	1.004
# N's per 100 kbp	**0.00**	68.11	159.20	103.20	105.46	88.55	31.40	59.29	50.56	48.26	**0.00**
# mismatches per 100 kbp	3.01	0.95	1.98	0.64	**0.18**	0.31	0.31	0.46	0.50	1.74	3.68
# indels per 100 kbp	**0.00**	0.35	1.10	0.59	0.57	0.37	0.13	0.31	0.15	0.24	**0.00**
Largest alignment	9147	268205	73663	221199	237845	209037	**268448**	209311	209518	209273	9536
Total aligned length	233516	4547442	4514693	4548941	4546078	4625272	4775037	4550674	**4941146**	4555709	382165
NA50	1725	110240	13241	79327	95454	95121	**112492**	91824	112450	86536	1743
NGA50	-	105571	12901	78999	95454	95121	112492	87309	132756	86536	-
NA75	1423	58208	5855	45590	57827	56387	**59588**	57312	58666	48231	1396
NGA75	-	58094	5259	43901	57311	56387	63364	56944	63620	44021	-
LA50	44	**14**	96	19	16	17	**14**	18	15	19	74
LGA50	-	15	100	20	16	17	14	19	14	19	-
LA75	80	**29**	220	38	31	34	**29**	34	30	36	135
LGA75	-	30	236	39	32	34	27	35	27	38	-

Table A.16 Assembly quality metrics for *S. enterica*

Assembly	Uncorrected	ACE	BayesHammer	BFC	BLESS2	Blue	Karect	Lighter	Musket	RACER	SGA-EC
# contigs (≥ 0 bp)	57	58	60	53	53	96	49	60	64	57	60
# contigs (≥ 1000 bp)	26	26	25	24	29	26	24	26	26	25	27
# contigs (≥ 5000 bp)	21	21	20	20	26	21	20	21	21	20	23
# contigs (≥ 10000 bp)	20	20	19	18	24	20	18	20	20	20	21
# contigs (≥ 25000 bp)	19	19	18	17	22	19	17	19	19	19	20
# contigs (≥ 50000 bp)	17	17	17	16	21	18	16	17	18	18	18
Total length (≥ 0 bp)	8163172	6749101	5393710	7876026	8740562	6725492	7983640	5267128	7609062	**8827553**	7139332
Total length (≥ 1000 bp)	8152999	6737889	5382222	7866252	8732154	6704366	7974783	5255839	7596472	**8815758**	7127974
Total length (≥ 5000 bp)	8138322	6723525	5367858	7854968	8723916	6689707	7963123	5241077	7581909	**8801195**	7116576
Total length (≥ 10000 bp)	8132122	6717325	5361658	7842409	8711266	6683507	7950564	5234877	7575709	**8801195**	7104017
Total length (≥ 25000 bp)	8121235	6706944	5351283	7832038	8669276	6673137	7939669	5224509	7565341	**8780558**	7093636
Total length (≥ 50000 bp)	8037149	6623285	5313751	7794506	8631738	6635805	7902137	5140850	7527809	**8741247**	7009977
# contigs	27	29	27	25	30	28	26	28	28	29	30
Largest contig	**2045651**	**2045651**	1022875	**2045651**	1059914	1319742	**2045651**	1022875	**2045651**	2045552	1227131
Total length	8153995	6740191	5383886	7867248	8733150	6705863	7976367	5257503	7598186	**8818658**	7130310
Reference length	4888768	4888768	4888768	4888768	4888768	4888768	4888768	4888768	4888768	4888768	4888768
GC (%)	52.02	51.95	52.18	52.08	52.10	52.18	52.06	52.08	52.00	52.12	51.99
Reference GC (%)	52.09	52.09	52.09	52.09	52.09	52.09	52.09	52.09	52.09	52.09	52.09
N50	**1128667**	564383	471517	879542	556744	597403	879542	439239	879542	890143	492948
NG50	1227131	614042	471517	1128568	899660	879165	1128568	439239	1128568	**1319841**	1060013
N75	407470	321307	298751	439351	407470	323519	439318	231546	323519	**597104**	321307
NG75	**1128667**	564383	321307	879542	763376	597403	879542	298751	879542	900602	492948
L50	**3**	4	4	**3**	6	4	**3**	4	**3**	4	4
LG50	**2**	**2**	4	**2**	3	3	**2**	4	**2**	**2**	3
L75	7	8	8	7	10	7	7	9	7	**6**	8
LG75	**3**	4	7	**3**	4	4	**3**	7	**3**	**3**	4
# misassemblies	32	26	**18**	27	46	26	31	19	29	36	27
# misassembled contigs	11	10	10	11	20	11	11	**9**	11	12	11
Misassembled contigs length	7380321	5730649	4495114	7369278	8485367	5978777	7476855	**4086405**	6776877	8144235	6117395
# local misassemblies	6	9	**3**	6	7	6	7	4	6	8	6
# unaligned mis. contigs	0	0	0	0	0	0	0	0	0	0	0
# unaligned contigs	3 + 4 part	3 + 4 part	3 + 4 part	3 + 4 part	3 + 4 part	3 + 4 part	3 + 4 part	3 + 4 part	3 + 4 part	3 + 4 part	3 + 4 part
Unaligned length	85006	84608	**60978**	66482	91260	82781	88086	82781	84807	85007	88086
Genome fraction (%)	96.588	96.583	96.589	96.590	90.498	96.560	**96.605**	96.584	96.588	95.827	96.580
Duplication ratio	1.709	1.410	1.127	1.652	1.953	1.403	1.670	**1.096**	1.591	1.864	1.492
# N's per 100 kbp	13.49	14.84	**7.43**	13.98	38.93	13.42	16.30	9.51	13.16	19.28	14.02
# mismatches per 100 kbp	7.71	7.94	8.03	7.69	**6.83**	8.09	8.91	7.98	7.54	8.03	7.48
# indels per 100 kbp	0.47	**0.36**	0.38	0.40	0.38	0.38	1.06	**0.36**	**0.36**	0.38	0.49
Largest alignment	**453219**	**453219**	**453219**	**453219**	383351	**453219**	**453219**	**453219**	**453219**	**453219**	385791
Total aligned length	8068080	6654783	5322508	7799766	8638780	6622282	7887181	5174322	7512575	**8732343**	7041424
NA50	**231546**	216238	216238	231347	175570	215208	231347	215202	**231546**	216238	215202
NGA50	**349181**	339455	231546	348982	297351	271281	339455	216238	339455	339455	339455
NA75	140342	134245	106253	**161809**	109398	117146	161610	106253	161610	140241	133818
NGA75	**235808**	215202	161809	231546	195759	180274	231546	134245	231546	235198	215202
LA50	12	10	**9**	12	17	11	12	**9**	11	14	12
LGA50	**7**	**7**	8	**7**	8	**7**	**7**	8	**7**	**7**	**7**
LA75	23	20	**17**	22	32	21	23	**17**	21	26	22
LGA75	**10**	12	14	11	13	13	11	15	11	11	12

Table A.17 Assembly quality metrics for *P. aeruginosa*

Assembly	Uncorrected	ACE	BayesHammer	BFC	BLESS2	Blue	Karect	Lighter	Musket	RACER	SGA-EC
# contigs (≥ 0 bp)	235	257	214	237	285	155	248	175	208	235	198
# contigs (≥ 1000 bp)	100	58	63	51	98	46	59	70	56	66	78
# contigs (≥ 5000 bp)	86	43	49	39	79	36	46	53	46	50	64
# contigs (≥ 10000 bp)	86	43	47	39	79	35	46	53	46	50	63
# contigs (≥ 25000 bp)	75	40	46	36	69	33	43	48	42	44	58
# contigs (≥ 50000 bp)	60	39	44	34	58	32	42	45	41	42	51
Total length (≥ 0 bp)	11625729	11634936	10503702	10420741	**12042989**	11140745	11989806	11767482	11524362	11873579	10961929
Total length (≥ 1000 bp)	11582521	11575379	10456654	10365266	**11986568**	11108207	11932686	11734935	11477897	11822085	10924088
Total length (≥ 5000 bp)	11551657	11542556	10428297	10338168	**11935032**	11084356	11903187	11700878	11453202	11787070	10891866
Total length (≥ 10000 bp)	11551657	11542556	10417440	10338168	**11935032**	11077115	11903187	11700878	11453202	11787070	10886855
Total length (≥ 25000 bp)	11383796	11509234	10406408	10304846	11770514	11054584	**11869865**	11644056	11394722	11704610	10800207
Total length (≥ 50000 bp)	10849911	11477737	10343168	10230546	11371155	11022244	**11838368**	11541634	11363225	11636822	10559422
# contigs	112	74	77	66	113	53	74	81	71	79	89
Largest contig	555743	950726	758235	1712382	1099014	**1712467**	1101178	1099225	1101178	1101178	727357
Total length	11590443	11585582	10465902	10374758	**11996711**	11113273	11942534	11742174	11487346	11830623	10931407
Reference length	6264404	6264404	6264404	6264404	6264404	6264404	6264404	6264404	6264404	6264404	6264404
GC (%)	66.55	66.58	66.53	66.58	66.58	66.57	66.55	66.59	66.61	66.55	66.56
Reference GC (%)	66.56	66.56	66.56	66.56	66.56	66.56	66.56	66.56	66.56	66.56	66.56
N50	226123	475847	336413	509581	233095	436693	457755	354411	365189	**578979**	253019
NG50	353311	651197	488647	760385	655399	**776261**	580931	580931	697483	727189	457723
N75	127449	231991	188258	232707	133440	**238833**	230657	232839	233095	221660	137433
NG75	257909	582785	364869	574308	356909	651944	526453	376883	595827	**656305**	329577
L50	18	9	11	**7**	13	**7**	10	11	9	8	13
LG50	8	4	5	**3**	4	**3**	4	4	4	4	6
L75	34	18	21	**15**	29	**15**	19	21	18	18	27
LG75	13	7	9	6	8	**5**	7	8	6	6	10
# misassemblies	88	45	51	44	83	**43**	51	54	50	54	62
# misassembled contigs	85	43	47	39	79	**35**	46	52	45	50	62
Misassembled contigs length	11373850	11542556	10417440	**10338168**	11935032	11077115	11903187	11583348	11333515	11787070	10864735
# local misassemblies	9	7	**2**	**2**	11	16	4	4	3	**2**	6
# unaligned mis. contigs	0	0	0	0	0	0	0	0	0	0	0
# unaligned contigs	0 + 1 part	0 + 1 part	0 + 1 part	0 + 1 part	0 + 1 part	0 + 1 part	0 + 1 part	0 + 1 part	0 + 1 part	0 + 1 part	**0 + 0 part**
Unaligned length	20158	20258	20258	20258	20258	20358	20258	20258	20258	20258	**0**
Genome fraction (%)	94.090	92.618	83.725	82.911	**96.082**	88.807	95.430	94.785	92.714	94.584	87.741
Duplication ratio	**1.963**	1.993	1.992	1.994	1.990	1.994	1.994	1.974	1.974	1.993	1.989
# N's per 100 kbp	74.20	39.70	45.86	38.56	78.35	**35.09**	42.70	44.28	40.91	43.95	55.80
# mismatches per 100 kbp	0.78	**0.59**	0.93	0.83	0.88	1.73	0.85	0.74	1.31	0.71	0.73
# indels per 100 kbp	0.64	0.72	0.78	**0.62**	0.83	0.88	0.84	0.66	0.77	0.74	0.80
Largest alignment	277921	475263	379167	856091	549407	**856283**	550788	549662	550788	550788	363728
Total aligned length	11561645	11560824	10440921	10350599	**11967149**	11089016	11917271	11716808	11461927	11805252	10925274
NA50	112912	237774	168256	188488	115225	188311	228927	177255	182644	**289340**	126360
NGA50	177807	325449	244174	**380043**	277543	349328	290316	290316	348592	363445	228712
NA75	61958	112823	91289	111071	63883	**115483**	110730	100361	109533	100505	68567
NGA75	130009	291243	182283	254840	178305	218396	263077	188491	297764	**328003**	164639
LA50	35	18	21	**13**	26	16	19	21	17	16	26
LGA50	15	8	10	**6**	9	**6**	8	8	8	8	12
LA75	68	36	42	**30**	61	34	39	42	36	36	54
LGA75	25	14	18	**11**	16	12	14	15	12	12	20

Table A.18 Assembly quality metrics for *H. sapiens* Chr. 21

Assembly	Uncorrected	ACE	BayesHammer	BFC	BLESS2	Blue	Karect	Lighter	Musket	RACER	SGA-EC
# contigs (≥ 0 bp)	27142	28951	27677	30461	27602	31994	29444	28813	28902	29266	28622
# contigs (≥ 1000 bp)	8724	8872	8615	8922	8659	8846	8956	8854	8903	8966	8830
# contigs (≥ 5000 bp)	291	300	300	358	295	385	313	308	311	313	301
# contigs (≥ 10000 bp)	45	51	45	72	50	85	52	54	50	55	55
# contigs (≥ 25000 bp)	2	2	3	5	4	4	3	3	3	2	3
# contigs (≥ 50000 bp)	0	0	0	0	0	0	0	0	0	0	0
Total length (≥ 0 bp)	23408496	24205535	23398741	25197489	23563324	**25978783**	24713094	24238862	24388697	24638523	24099519
Total length (≥ 1000 bp)	18534837	18873907	18350432	19502701	18544529	**19856763**	19251190	18939574	19071438	19245204	18855664
Total length (≥ 5000 bp)	2236239	2390389	2345949	2943213	2358076	**3238607**	2509559	2460569	2447363	2507608	2409899
Total length (≥ 10000 bp)	657387	772745	689726	1107992	789211	**1277133**	814402	824225	769572	846768	824665
Total length (≥ 25000 bp)	58348	59509	84445	**154711**	119213	121978	85547	85543	84420	59716	85544
Total length (≥ 50000 bp)	0	0	0	0	0	0	0	0	0	0	0
# contigs	8925	9108	8816	9142	8844	9092	9209	9079	9132	9197	9037
Largest contig	30142	31299	30202	**36962**	32650	33532	31299	31300	30195	31488	31299
Total length	18682551	19045980	18496737	19660516	18683473	**20033594**	19441662	19105744	19242695	19415777	19007561
Reference length	46709983	46709983	46709983	46709983	46709983	46709983	46709983	46709983	46709983	46709983	46709983
GC (%)	39.74	39.97	39.70	40.18	39.65	40.44	40.00	39.95	40.00	40.00	39.91
Reference GC (%)	42.22	42.22	42.22	42.22	42.22	42.22	42.22	42.22	42.22	42.22	42.22
N50	2151	2158	2166	2227	2184	**2314**	2187	2173	2179	2182	2171
N75	1533	1522	1529	1548	1537	**1576**	1538	1529	1536	1537	1532
L50	2674	2686	2610	2611	2615	**2517**	2701	2669	2689	2695	2665
L75	5266	5335	5178	5282	5186	**5171**	5374	5312	5343	5370	5299
# misassemblies	7	**5**	**5**	6	**5**	**5**	**5**	7	**5**	10	8
# misassembled contigs	7	**5**	**5**	6	**5**	**5**	**5**	7	**5**	10	8
Misassembled contigs length	11595	**7184**	7866	21301	7237	32262	16937	22039	19282	36377	35449
# local misassemblies	39	40	**33**	60	47	56	42	35	40	52	36
# structural variations	0	0	0	0	0	0	0	0	0	0	0
# unaligned mis. contigs	**0**	**0**	**0**	**0**	**0**	**0**	1	**0**	**0**	**0**	**0**
# unaligned contigs	11 + 1 part	9 + 2 part	5 + 1 part	8 + 3 part	5 + 4 part	**4 + 2 part**	9 + 4 part	11 + 3 part	7 + 2 part	7 + 4 part	7 + 2 part
Unaligned length	13200	12644	6862	12574	10467	**5235**	14528	14858	12580	12369	11621
Genome fraction (%)	39.758	40.515	39.397	41.844	39.795	**42.658**	41.365	40.658	40.951	41.310	40.459
Duplication ratio	**1.005**	1.006	**1.005**	**1.005**	**1.005**	**1.005**	**1.005**	**1.005**	**1.005**	1.006	**1.005**
# N's per 100 kbp	**11.24**	13.13	14.06	21.36	15.52	22.46	13.89	13.61	12.99	15.97	13.68
# mismatches per 100 kbp	86.72	86.79	86.70	88.20	86.67	88.13	87.61	87.28	87.45	89.43	**86.12**
# indels per 100 kbp	18.62	18.87	**18.26**	19.58	18.36	18.56	19.36	18.86	18.86	19.26	18.82
Largest alignment	29939	31199	30002	**36761**	32348	33332	31199	31200	30095	31388	31199
Total aligned length	18662994	19026207	18481892	19637168	18665168	**20017474**	19419142	19083052	19222660	19393898	18988684
NA50	2149	2156	2165	2225	2183	**2313**	2186	2172	2178	2180	2169
NGA50	-	-	-	-	-	-	-	-	-	-	-
NA75	1531	1520	1528	1544	1533	**1574**	1536	1526	1533	1533	1529
LA50	2676	2688	2612	2615	2617	**2522**	2704	2672	2692	2700	2669
LA75	5270	5339	5182	5290	5190	**5179**	5379	5319	5349	5378	5305

Table A.19 Assembly quality metrics for *C. elegans*

Assembly	Uncorrected	ACE	BayesHammer	BFC	BLESS2	Blue	Karect	Lighter	Musket	RACER	SGA-EC
# contigs (≥ 0 bp)	33879	35629	34072	34338	34851	36053	35285	34300	34770	37198	34010
# contigs (≥ 1000 bp)	27381	27222	27337	26996	27904	26458	27217	27363	27294	27887	27244
# contigs (≥ 5000 bp)	3874	3685	3893	4101	3041	4098	4012	3840	3723	3391	3989
# contigs (≥ 10000 bp)	584	588	616	689	411	768	672	595	574	497	648
# contigs (≥ 25000 bp)	23	28	25	33	17	61	31	24	24	22	28
# contigs (≥ 50000 bp)	1	1	1	3	3	6	2	1	1	1	1
Total length (≥ 0 bp)	88711959	88080769	89019134	90477027	83674565	**90945967**	90382710	88648128	87898196	87242922	89720536
Total length (≥ 1000 bp)	86109765	84805863	86321966	**87577530**	80876971	87415714	87218441	85875614	84947836	83429141	87003150
Total length (≥ 5000 bp)	30047522	28818579	30349677	32512747	23082132	**33990281**	31726003	29867219	28965455	26124208	31301586
Total length (≥ 10000 bp)	8115815	8324546	8544789	9742241	5873440	**11724956**	9576884	8347025	8049310	7028244	9090758
Total length (≥ 25000 bp)	716932	885165	797767	1060134	668624	**2171500**	1089069	780103	766392	699825	898189
Total length (≥ 50000 bp)	54399	57601	54393	164608	248925	**471404**	174936	54399	54399	57601	54399
# contigs	28936	29090	28958	28685	29587	28212	29020	29012	29008	30336	28871
Largest contig	54399	57601	54393	57890	108677	**115103**	**115103**	54399	54399	57601	54399
Total length	87209529	86131218	87468860	**88775004**	82072189	88662798	88498583	87045920	86166399	85171657	88159881
Reference length	100286070	100286070	100286070	100286070	100286070	100286070	100286070	100286070	100286070	100286070	100286070
GC (%)	38.03	38.41	38.06	38.27	38.35	38.35	38.13	38.18	38.27	38.34	38.12
Reference GC (%)	35.44	35.44	35.44	35.44	35.44	35.44	35.44	35.44	35.44	35.44	35.44
N50	3687	3628	3707	3860	3283	**3935**	3777	3668	3635	3430	3775
NG50	3192	3086	3213	3388	2669	**3429**	3310	3165	3097	2904	3289
N75	2238	2202	2241	2296	2047	**2311**	2265	2225	2204	2094	2265
NG75	1707	1641	1719	**1809**	1416	1808	1775	1695	1642	1536	1765
L50	7038	7047	7009	6811	7521	**6433**	6900	7038	7050	7400	6936
LG50	8944	9165	8866	8405	10603	**8017**	8566	8981	9159	9797	8658
L75	14674	14720	14642	14312	15484	**13839**	14510	14704	14716	15387	14523
LG75	19700	20305	19539	18540	23504	**18096**	18916	19820	20284	21714	19071
# misassemblies	**113**	141	114	130	121	141	149	117	120	915	116
# misassembled contigs	**105**	137	108	124	119	138	145	112	116	894	111
Misassembled contigs length	965372	1074547	889920	1049107	989929	1155666	1240877	912972	**882752**	2312963	958991
# local misassemblies	90	92	96	89	93	**79**	87	86	85	118	91
# unaligned mis. contigs	3	**2**	**2**	5	3	**2**	3	**2**	**2**	**2**	3
# unaligned contigs	3929 + 37 part	3966 + 36 part	3895 + 34 part	3636 + 38 part	4009 + 33 part	**3130 + 34 part**	3936 + 38 part	3935 + 37 part	3948 + 35 part	3988 + 43 part	3879 + 36 part
Unaligned length	10018419	10738295	10217534	11068366	**9903052**	11424350	10691877	10453916	10459102	10525274	10518396
Genome fraction (%)	76.482	74.634	76.543	76.947	71.452	76.381	**76.965**	75.919	75.037	73.737	76.931
Duplication ratio	**1.006**	1.007	**1.006**	1.007	1.007	1.008	1.008	**1.006**	**1.006**	1.009	**1.006**
# N's per 100 kbp	8.14	9.98	8.35	10.03	10.23	13.65	11.30	8.16	**7.89**	8.34	9.53
# mismatches per 100 kbp	**4.25**	8.39	4.38	4.60	4.87	5.48	5.24	13.21	15.35	14.20	**4.25**
# indels per 100 kbp	2.87	3.10	**2.61**	2.83	2.63	3.07	3.01	3.09	3.19	4.08	2.91
Largest alignment	**27606**	25320	25354	**27606**	25315	**27606**	**27606**	**27606**	25347	25347	**27606**
Total aligned length	77165313	75351336	77225216	77677096	72141274	77206533	**77775255**	76561502	75676690	74548243	77615090
NA50	3269	3124	3257	3254	2886	3208	3278	3203	3158	2955	**3284**
NGA50	2786	2613	2785	2830	2303	2777	**2837**	2715	2643	2433	2834
NA75	**1818**	1742	1809	1788	1681	1753	1800	1787	1766	1656	1810
NGA75	1274	633	1276	**1301**	-	1266	**1301**	1236	994	-	1298
LA50	**7909**	8188	7952	8052	8538	8097	7970	8039	8094	8586	7941
LGA50	10073	10664	10077	9946	12076	10042	**9903**	10282	10535	11408	9928
LA75	**16849**	17423	16957	17253	17870	17439	17081	17152	17226	18240	16977
LGA75	23261	24993	23256	22895	-	23267	**22838**	23799	24516	-	22887

Table A.20 Assembly quality metrics for *D. melanogaster*

Assembly	Uncorrected	ACE	BayesHammer	BFC	BLESS2	Blue	Karect	Lighter	Musket	RACER	SGA-EC
# contigs (≥ 0 bp)	27260	30244	29569	27393	28469	21402	25185	28576	24938	26720	28174
# contigs (≥ 1000 bp)	12634	14793	14325	12449	21034	10019	13299	13371	13790	15317	13061
# contigs (≥ 5000 bp)	4940	5340	5242	4864	5765	4093	4965	5093	5172	5268	5046
# contigs (≥ 10000 bp)	2986	3026	2979	2946	2574	2690	2968	3023	3022	2934	2970
# contigs (≥ 25000 bp)	1136	1030	1065	1165	515	1256	1166	1125	1099	1024	1150
# contigs (≥ 50000 bp)	364	242	286	378	54	504	323	318	273	243	328
Total length (≥ 0 bp)	122045333	122366187	122006358	**122808874**	114602943	121557134	121827634	122473174	120004005	120942507	122466576
Total length (≥ 1000 bp)	117164542	117232929	116938531	117847785	112114552	117816277	**117894949**	117394289	116463039	117195927	117387138
Total length (≥ 5000 bp)	97048788	92303485	92940657	97807043	71943394	**102159157**	95701726	95560675	93443597	90424200	96364085
Total length (≥ 10000 bp)	83408784	76258640	77232839	84501065	50210781	**92385558**	81803352	81166940	78545771	74170057	81868999
Total length (≥ 25000 bp)	53624596	44485080	46772910	55830709	18559399	**68976793**	52576273	50764105	47763708	43698411	52402012
Total length (≥ 50000 bp)	26851850	17448614	19919080	28437865	3330371	**42383546**	23590856	22668748	19399922	16853382	23901384
# contigs	13973	16228	15743	13841	21908	11072	14427	14800	14729	16459	14508
Largest contig	212522	159060	187339	225726	91562	**368590**	198380	198532	171899	198165	181461
Total length	118112708	118247885	117934833	**118830170**	112727088	118562819	118699015	118403026	117124660	118001344	118409815
Reference length	120381546	120381546	120381546	120381546	120381546	120381546	120381546	120381546	120381546	120381546	120381546
GC (%)	42.44	42.49	42.44	42.45	42.65	42.48	42.44	42.45	42.53	42.50	42.45
Reference GC (%)	42.42	42.42	42.42	42.42	42.42	42.42	42.42	42.42	42.42	42.42	42.42
N50	22091	17377	18249	22788	8075	**33082**	21345	20330	19153	16735	21287
NG50	21451	16988	17561	22303	7119	**32109**	20764	19863	18358	16199	20838
N75	7864	5805	6121	7934	3744	**12284**	7079	6997	6524	5360	7320
NG75	7184	5389	5567	7502	3214	**11308**	6643	6479	5776	5016	6759
L50	1368	1739	1642	1316	3263	**919**	1460	1500	1593	1775	1448
LG50	1420	1801	1710	1350	3769	**947**	1500	1550	1680	1847	1495
L75	3573	4664	4421	3464	8671	**2377**	3827	3935	4182	4894	3779
LG75	3800	4949	4736	3614	10327	**2493**	4011	4156	4580	5241	3989
# misassemblies	991	1017	999	1021	1092	**983**	1016	993	1011	1250	988
# misassembled contigs	903	932	914	923	1033	**867**	927	901	927	1132	900
Misassembled contigs length	20252411	15883432	15918142	19834439	**9381705**	27249040	18786258	17767341	16750517	19442661	18461236
# local misassemblies	1907	1879	1896	1840	**1763**	1835	1811	1838	1883	1821	1852
# unaligned mis. contigs	109	111	105	122	**65**	82	100	110	84	100	109
# unaligned contigs	1269 + 844 part	1303 + 859 part	1310 + 840 part	1283 + 861 part	1308 + 624 part	1290 + 691 part	1257 + 879 part	1272 + 848 part	1314 + 756 part	1393 + 869 part	**1250 + 867 part**
Unaligned length	5979319	5979218	6019433	6131439	**4567032**	5380727	6035967	6027963	5434932	5734625	6056066
Genome fraction (%)	92.002	91.859	91.758	92.304	88.732	**92.553**	92.363	92.076	91.677	91.737	92.023
Duplication ratio	**1.012**	1.015	1.013	1.014	1.013	1.016	1.013	1.014	**1.012**	1.017	1.014
# N's per 100 kbp	205.23	212.86	213.51	215.52	**170.23**	200.74	241.03	210.13	224.80	224.40	201.84
# mismatches per 100 kbp	543.91	546.01	543.28	545.31	**542.17**	545.48	547.20	545.75	551.17	548.97	544.36
# indels per 100 kbp	129.04	128.27	128.22	129.21	**123.93**	129.62	130.06	129.19	128.88	127.95	129.12
Largest alignment	175992	158169	152405	152426	84393	**278506**	181494	178460	171298	162199	154126
Total aligned length	111608713	111717063	111375299	112144165	107696193	**112649626**	112079362	111833973	111127562	111640564	111830509
NA50	19250	15446	16182	19817	7324	**27349**	18525	17833	16821	14528	18818
NGA50	18832	15025	15582	19469	6537	**26744**	18121	17358	16161	14081	18357
NA75	6427	4927	5068	6460	3427	**9819**	5724	5697	5462	4737	6010
NGA75	5852	4679	4766	6051	2881	**9058**	5285	5223	4915	4461	5485
LA50	1570	1953	1842	1510	3582	**1113**	1669	1703	1803	2064	1638
LGA50	1629	2023	1919	1550	4135	**1146**	1715	1759	1902	2147	1691
LA75	4143	5365	5096	4043	9439	**2862**	4465	4559	4816	5716	4378
LGA75	4420	5698	5469	4229	11261	**3007**	4694	4831	5290	6104	4635

A.4.2 IDBA

A.4.2.1 *B. dentium*

Table A.21 contains the Quast report after assembling dataset *B. dentium* with IDBA.

A.4.2.2 *E. coli str. K-12 substr. DH10B*

Table A.22 contains the Quast report after assembling dataset *E. coli str. K-12 substr. DH10B* with IDBA.

A.4.2.3 *E. coli str. K-12 substr. MG1655*

Table A.23 contains the Quast report after assembling dataset *E. coli str. K-12 substr. MG1655* with IDBA.

A.4.2.4 *S. enterica*

Table A.24 contains the Quast report after assembling dataset *S. enterica* with IDBA.

A.4.2.5 *P. aeruginosa*

Table A.25 contains the Quast report after assembling dataset *P. aeruginosa* with IDBA.

A.4.2.6 *H. sapiens* Chr. 21

Table A.26 contains the Quast report after assembling dataset *H. sapiens* Chr. 21 with IDBA.

A.4.2.7 *C. elegans*

Table A.27 contains the Quast report after assembling dataset *C. elegans* with IDBA.

A.4.2.8 *D. melanogaster*

Table A.28 contains the Quast report after assembling dataset *D. melanogaster* with IDBA.

Table A.21 Assembly quality metrics for *B. dentium*

Assembly	Uncorrected	ACE	BayesHammer	BFC	BLESS2	Blue	Fiona	Karect	Lighter	Musket	RACER	SGA-EC	Trowel
# contigs (≥ 0 bp)	5027	2032	4055	3632	2220	3338	3103	3945	3423	3001	2523	4845	4177
# contigs (≥ 1000 bp)	154	99	109	82	57	70	83	74	97	88	80	162	140
# contigs (≥ 5000 bp)	112	76	81	58	42	51	63	55	71	66	60	115	101
# contigs (≥ 10000 bp)	85	64	69	53	39	48	53	49	58	58	55	91	79
# contigs (≥ 25000 bp)	31	34	37	38	31	32	36	31	38	36	35	36	40
# contigs (≥ 50000 bp)	9	17	16	21	18	19	18	19	16	15	18	5	11
Total length (≥ 0 bp)	**3587351**	2819442	3496247	3480758	3117137	3444880	3414588	3476847	3453335	3417411	3367147	3568629	3518137
Total length (≥ 1000 bp)	2609222	2604126	2611819	2614744	2607722	2615696	2610144	2613192	2614549	2614679	**2615893**	2606824	2612925
Total length (≥ 5000 bp)	2514677	2553828	2553978	2567890	2580180	**2582139**	2571400	2581371	2560121	2572371	2576281	2499243	2533087
Total length (≥ 10000 bp)	2315068	2469354	2466567	2532364	2556678	**2560520**	2504335	2533329	2459554	2509701	2541682	2317366	2380329
Total length (≥ 25000 bp)	1438392	1953661	1957421	2283949	**2433293**	2273090	2233421	2217746	2110486	2125279	2211135	1418973	1758731
Total length (≥ 50000 bp)	589275	1303207	1203707	1649327	**1965790**	1810343	1562706	1780649	1310233	1354916	1577255	363499	748566
# contigs	352	117	296	268	125	252	259	260	284	276	262	357	327
Largest contig	99811	157035	119824	174678	197580	197580	210973	**211639**	147538	197572	174086	103372	149572
Total length	**2731975**	2615596	2727231	2729908	2647585	2726849	2717295	2728488	2729152	2730793	2727323	2727106	2728098
Reference length	2636367	2636367	2636367	2636367	2636367	2636367	2636367	2636367	2636367	2636367	2636367	2636367	2636367
GC (%)	58.64	58.49	58.63	58.63	58.51	58.64	58.61	58.64	58.63	58.64	58.63	58.64	58.63
Reference GC (%)	58.54	58.54	58.54	58.54	58.54	58.54	58.54	58.54	58.54	58.54	58.54	58.54	58.54
N50	30669	47845	44252	58654	**98980**	71013	55343	64662	47671	48847	62439	25330	31841
NG50	31451	47845	45395	60136	**98980**	73724	56050	64662	47845	58654	62741	25469	32380
N75	14834	24492	21701	35593	**47845**	35678	34935	34935	29210	27070	32388	15784	16984
NG75	15684	24492	23506	37464	**47845**	39437	35678	42204	29913	29824	36267	16326	19214
L50	29	18	20	16	**10**	12	15	12	18	16	15	34	27
LG50	27	18	19	15	**10**	11	14	12	17	15	14	33	26
L75	64	35	41	31	**19**	25	30	26	36	34	30	70	55
LG75	59	35	38	29	**19**	23	28	24	34	31	28	66	51
# misassemblies	0	0	0	0	0	0	0	0	0	0	0	0	0
# misassembled contigs	0	0	0	0	0	0	0	0	0	0	0	0	0
Misassembled contigs length	0	0	0	0	0	0	0	0	0	0	0	0	0
# local misassemblies	2	2	2	2	2	2	2	2	2	2	2	2	2
# unaligned mis. contigs	0	0	0	0	0	0	0	0	0	0	0	0	0
# unaligned contigs	194 + 0 part	**12 + 0 part**	189 + 0 part	188 + 0 part	66 + 0 part	186 + 0 part	174 + 0 part	186 + 0 part	187 + 0 part	190 + 0 part	185 + 0 part	192 + 0 part	188 + 0 part
Unaligned length	124776	**6960**	122785	121410	40083	118938	109568	120541	119583	122203	118530	123616	121155
Genome fraction (%)	98.738	98.815	98.701	98.844	98.847	98.848	98.812	98.845	98.852	98.834	**98.863**	98.649	98.746
Duplication ratio	1.002	**1.001**	**1.001**	**1.001**	**1.001**	**1.001**	**1.001**	**1.001**	**1.001**	**1.001**	**1.001**	**1.001**	**1.001**
# N's per 100 kbp	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
# mismatches per 100 kbp	**0.23**	**0.23**	**0.23**	**0.23**	**0.23**	0.35	0.27	**0.23**	0.27	0.31	**0.23**	**0.23**	**0.23**
# indels per 100 kbp	0.27	**0.23**	0.27	0.31	0.58	**0.23**	0.27	0.31	0.31	0.27	**0.23**	0.27	0.27
Largest alignment	99811	156760	119824	174678	197580	197580	210973	**211639**	147538	197572	174086	103372	149572
Total aligned length	2606924	2608361	2604171	2608223	2607227	2607636	2607452	2607672	**2609294**	2608315	2608518	2603215	2606668
NA50	30669	47845	44252	58654	**98980**	71013	55343	64662	47671	48847	62439	25330	31841
NGA50	31451	47845	45395	60136	**98980**	73724	56050	64662	47845	58654	62741	25469	32380
NA75	14834	24492	21701	35593	**47845**	35678	34935	34935	29210	27070	32388	15784	16984
NGA75	15684	24492	23506	37189	**47845**	39437	35678	42204	29913	29824	36267	16326	18972
LA50	29	18	20	16	**10**	12	15	12	18	16	15	34	27
LGA50	27	18	19	15	**10**	11	14	12	17	15	14	33	26
LA75	64	35	41	31	**19**	25	30	26	36	34	30	70	55
LGA75	59	35	38	29	**19**	23	28	24	34	31	28	66	51

Table A.22 Assembly quality metrics for *E. coli str. K-12 substr. DH10B*

Assembly	Uncorrected	ACE	BayesHammer	BFC	BLESS2	Blue	Fiona	Karect	Lighter	Musket	RACER	SGA-EC	Trowel
# contigs (≥ 0 bp)	2360	1011	974	641	614	752	704	845	681	674	667	1735	1154
# contigs (≥ 1000 bp)	125	121	122	110	107	115	112	110	112	112	114	167	127
# contigs (≥ 5000 bp)	97	95	97	86	85	90	89	85	88	88	89	130	101
# contigs (≥ 10000 bp)	85	79	82	76	73	77	77	74	75	74	76	106	89
# contigs (≥ 25000 bp)	63	60	61	57	56	59	58	55	57	58	56	63	64
# contigs (≥ 50000 bp)	29	32	33	33	32	31	31	32	33	31	32	25	30
Total length (≥ 0 bp)	**4515653**	4411601	4409433	4386875	4383754	4393892	4391199	4400623	4390421	4389610	4387773	4462636	4424457
Total length (≥ 1000 bp)	4329090	4327143	4328658	**4329413**	4327908	4328587	4327815	4328681	4328305	4328052	4328697	4326431	4329230
Total length (≥ 5000 bp)	4265809	4275841	4276153	4281299	**4284862**	4280221	4280926	4278797	4280020	4279741	4279445	4240734	4270897
Total length (≥ 10000 bp)	4173490	4150608	4159568	**4201688**	4190257	4176438	4185157	4189340	4176318	4166508	4175910	4060828	4178528
Total length (≥ 25000 bp)	3825722	3839346	3813633	3902949	3912933	3894188	3887474	3876118	3891890	**3917393**	3859576	3335440	3762545
Total length (≥ 50000 bp)	2633615	2854610	2791981	3072002	**3082420**	2893257	2935976	3070569	3044346	2971324	3026965	1953025	2503554
# contigs	133	129	131	119	115	124	121	118	122	121	122	176	135
Largest contig	239163	269573	206204	**269702**	**269702**	177084	269660	269637	**269702**	**269702**	269699	174732	178824
Total length	4334631	4333487	4335683	**4336423**	4334181	4335568	4334707	4335004	4336308	4335381	4334729	4332925	4335159
Reference length	4686137	4686137	4686137	4686137	4686137	4686137	4686137	4686137	4686137	4686137	4686137	4686137	4686137
GC (%)	50.75	50.74	50.75	50.75	50.75	50.74	50.75	50.75	50.75	50.74	50.75	50.75	50.75
Reference GC (%)	50.78	50.78	50.78	50.78	50.78	50.78	50.78	50.78	50.78	50.78	50.78	50.78	50.78
N50	60090	61037	62810	**77601**	**77601**	76358	76364	77573	73020	**77601**	76339	45611	58765
NG50	58765	58765	59786	66088	67274	63502	66087	**67318**	63521	66088	67228	43033	55566
N75	34689	40132	41267	**43033**	43015	41478	40710	43011	42648	41772	41772	28506	35180
NG75	30831	31552	31676	**34059**	34044	34058	32622	34058	34058	32648	31675	21579	30973
L50	21	20	23	**18**	**18**	19	19	**18**	19	**18**	19	30	24
LG50	24	23	26	21	**20**	22	21	**20**	22	21	21	34	28
L75	44	41	44	37	**36**	40	39	**36**	38	38	37	60	48
LG75	53	49	51	44	**43**	46	46	**43**	45	45	45	71	56
# misassemblies	**0**	**0**	1	1	2	1	1	2	1	**0**	1	1	**0**
# misassembled contigs	**0**	**0**	1	1	1	1	1	1	1	**0**	1	1	**0**
Misassembled contigs length	**0**	**0**	73043	50204	160773	115259	115095	160695	73020	**0**	73061	57259	**0**
# local misassemblies	**0**	1	**0**	2	2	**0**	1	1	**0**	**0**	**0**	**0**	**0**
# structural variations	0	0	0	0	0	0	0	0	0	0	0	0	0
# unaligned mis. contigs	0	0	0	0	0	0	0	0	0	0	0	0	0
# unaligned contigs	**0 + 0 part**	**0 + 0 part**	**0 + 0 part**	**0 + 0 part**	**0 + 0 part**	**0 + 0 part**	1 + 0 part	**0 + 0 part**	**0 + 0 part**	**0 + 0 part**	**0 + 0 part**	**0 + 0 part**	**0 + 0 part**
Unaligned length	**0**	**0**	**0**	**0**	**0**	**0**	777	**0**	**0**	**0**	**0**	**0**	**0**
Genome fraction (%)	92.524	92.538	92.551	92.557	92.510	92.549	92.514	92.541	**92.558**	92.536	92.527	92.493	92.539
Duplication ratio	1.000	**0.999**	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
# N's per 100 kbp	**0.00**	**0.00**	**0.00**	**0.00**	1.32	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
# mismatches per 100 kbp	**0.16**	0.88	0.67	0.92	0.95	0.95	1.01	0.85	1.01	1.18	2.01	0.83	0.48
# indels per 100 kbp	1.75	4.34	2.56	1.78	**1.41**	1.94	2.05	2.35	2.01	1.68	2.01	2.24	2.31
Largest alignment	239163	269573	206204	**269702**	**269702**	177084	269660	269637	**269702**	**269702**	269699	174732	178824
Total aligned length	4334630	4333450	4335683	**4336421**	4334122	4335559	4333913	4335004	4336306	4335379	4334723	4332925	4334971
NA50	60090	61037	61004	**77601**	**77601**	76358	76364	77573	67318	**77601**	76339	45611	58765
NGA50	58765	58765	59634	**66088**	66087	63502	66087	66063	62810	**66088**	66087	43033	55566
NA75	34689	40132	40148	**43033**	43015	41478	40710	41772	41731	41772	41267	27454	35180
NGA75	30831	31552	31552	**34059**	34044	32647	32622	34058	32818	32648	31675	21281	30973
LA50	21	20	23	**18**	**18**	19	19	**18**	19	**18**	19	30	24
LGA50	24	23	26	**21**	**21**	22	**21**	**21**	22	**21**	**21**	34	28
LA75	44	41	45	**37**	**37**	40	39	38	39	38	38	61	48
LGA75	53	49	52	**44**	**44**	47	46	**44**	46	45	45	72	56

Table A.23 Assembly quality metrics for *E. coli str. K-12 substr. MG1655*

Assembly	Uncorrected	ACE	BayesHammer	BFC	BLESS2	Blue	Fiona	Karect	Lighter	Musket	RACER	SGA-EC	Trowel
# contigs (≥ 0 bp)	2764	2174	803	1082	810	1210	1352	936	1220	1176	1332	2229	1314
# contigs (≥ 1000 bp)	116	129	99	115	110	118	117	110	114	121	120	116	116
# contigs (≥ 5000 bp)	93	100	75	88	81	90	87	84	88	92	93	91	91
# contigs (≥ 10000 bp)	80	88	70	78	73	77	77	74	77	81	85	79	78
# contigs (≥ 25000 bp)	58	61	54	56	54	57	57	54	56	59	58	58	57
# contigs (≥ 50000 bp)	33	36	33	33	33	34	33	34	34	34	34	33	34
Total length (≥ 0 bp)	**4747542**	4690218	4605109	4619712	4599723	4625294	4633230	4609966	4630594	4624123	4635398	4706232	4641206
Total length (≥ 1000 bp)	4525935	4527292	**4532359**	4531014	4531064	4529382	4530010	4529620	4530595	4530717	4529205	4526427	4531535
Total length (≥ 5000 bp)	4472536	4458330	4472750	4470195	4462677	4464111	4460408	4469372	4468887	4461488	4464995	4465235	**4477643**
Total length (≥ 10000 bp)	4373336	4373951	**4435795**	4394143	4400650	4363962	4386938	4392757	4386462	4379290	4406336	4373714	4379980
Total length (≥ 25000 bp)	4037922	3955590	**4213354**	4077735	4113882	4052339	4087233	4100329	4072451	4058027	4017375	4050236	4056754
Total length (≥ 50000 bp)	3136081	3037009	**3436733**	3234735	3355993	3217140	3218457	3382460	3272272	3148687	3147332	3159213	3235091
# contigs	129	138	110	126	119	129	128	123	124	132	132	129	127
Largest contig	209341	173958	**221477**	209350	221408	209346	209346	209344	173964	173963	173962	209302	209349
Total length	4535920	4534370	**4540813**	4539316	4537947	4537626	4538026	4539311	4538316	4538868	4538199	4535889	4540024
Reference length	4639675	4639675	4639675	4639675	4639675	4639675	4639675	4639675	4639675	4639675	4639675	4639675	4639675
GC (%)	50.74	50.73	50.74	50.74	50.74	50.73	50.73	50.74	50.73	50.74	50.74	50.73	50.74
Reference GC (%)	50.79	50.79	50.79	50.79	50.79	50.79	50.79	50.79	50.79	50.79	50.79	50.79	50.79
N50	82762	66115	**95664**	86971	95437	80468	82764	86968	82761	80468	82031	82766	82762
NG50	80467	65113	**95664**	82764	86972	78592	82764	86968	82550	78592	80378	80469	82552
N75	42704	41736	**50620**	44004	48769	43791	43792	48769	45417	43177	42696	41954	43792
NG75	41267	41031	**48769**	42697	45437	42509	42700	45437	43217	41718	41267	41019	41780
L50	19	23	**17**	18	**17**	19	19	18	19	20	20	19	19
LG50	20	24	**17**	19	18	20	19	18	20	21	21	20	20
L75	39	44	**33**	37	34	39	38	35	37	40	40	39	38
LG75	41	46	**34**	39	36	40	39	37	39	42	42	41	40
# misassemblies	**0**	**0**	1	1	**0**	**0**	**0**	1	**0**	**0**	2	**0**	**0**
# misassembled contigs	**0**	**0**	1	1	**0**	**0**	**0**	1	**0**	**0**	2	**0**	**0**
Misassembled contigs length	**0**	**0**	10212	10203	**0**	**0**	**0**	10153	**0**	**0**	104717	**0**	**0**
# local misassemblies	**1**	2	12	**1**	7	**1**	2	2	2	**1**	2	**1**	**1**
# unaligned mis. contigs	0	0	0	0	0	0	0	0	0	0	0	0	0
# unaligned contigs	3 + 0 part	2 + 0 part	4 + 0 part	3 + 0 part	**0 + 1 part**	3 + 0 part	3 + 0 part	3 + 0 part	3 + 0 part	3 + 0 part	3 + 0 part	3 + 0 part	3 + 0 part
Unaligned length	2325	1336	3165	2325	**1232**	2325	2275	2328	2321	2325	2326	2325	2325
Genome fraction (%)	97.695	97.691	97.771	97.767	97.756	97.735	97.743	97.770	97.748	97.751	97.741	97.692	**97.777**
Duplication ratio	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
# N's per 100 kbp	0.44	0.55	14.12	0.29	5.02	0.46	0.37	0.24	0.64	0.11	**0.09**	0.42	0.22
# mismatches per 100 kbp	0.22	1.08	0.73	0.73	0.97	0.53	1.68	1.41	0.60	1.01	1.17	**0.20**	0.46
# indels per 100 kbp	1.76	4.70	1.76	1.59	**1.50**	1.98	1.81	1.85	1.74	1.92	2.40	1.77	1.90
Largest alignment	209341	173958	**221392**	209350	221360	209346	209346	209344	173964	173963	173962	209302	209349
Total aligned length	4533238	4532587	4536848	4536542	4536266	4535022	4535432	4536704	4535629	4536019	4535447	4533187	**4537044**
NA50	82762	66115	**95658**	86971	95394	80468	82764	86968	82761	80468	77721	82766	82762
NGA50	80467	65113	**95658**	82764	86972	78592	82764	86968	82550	78592	77721	80469	82552
NA75	42704	41727	**50620**	44004	48769	43791	43792	48769	45342	43175	41869	41954	43792
NGA75	41267	41031	**48769**	42697	45437	42509	42700	45437	43217	41718	41267	41019	41780
LA50	19	23	**17**	18	**17**	19	19	18	19	20	21	19	19
LGA50	20	24	**17**	19	18	20	19	18	20	21	21	20	20
LA75	39	44	**33**	37	34	39	38	35	37	40	41	39	38
LGA75	41	46	**34**	39	36	40	39	37	39	42	43	41	40

Table A.24 Assembly quality metrics for *S. enterica*

Assembly	Uncorrected	ACE	BayesHammer	BFC	BLESS2	Blue	Fiona	Karect	Lighter	Musket	RACER	SGA-EC	Trowel
# contigs (≥ 0 bp)	812	823	869	491	443	558	506	479	600	686	572	565	563
# contigs (≥ 1000 bp)	79	77	70	71	67	71	73	69	75	73	71	78	79
# contigs (≥ 5000 bp)	66	65	62	63	58	62	62	60	66	63	60	66	66
# contigs (≥ 10000 bp)	61	58	59	57	54	55	57	55	60	59	55	61	61
# contigs (≥ 25000 bp)	51	50	49	48	45	48	48	47	50	50	46	50	49
# contigs (≥ 50000 bp)	40	37	39	39	35	37	38	37	38	40	36	39	40
Total length (≥ 0 bp)	**4821474**	4804536	4806987	4790253	4783105	4792819	4790169	4786252	4798745	4801080	4793989	4798672	4799077
Total length (≥ 1000 bp)	4725334	4722054	4725977	4725529	**4729204**	4724999	4725076	4726267	4723966	4723639	4726072	4726664	4727836
Total length (≥ 5000 bp)	4700100	4695972	4706517	4701537	**4708332**	4705864	4701692	4705593	4703476	4701522	4700545	4702472	4698906
Total length (≥ 10000 bp)	4661202	4644548	**4685098**	4656215	4674900	4653437	4662492	4669155	4658992	4672133	4663193	4663624	4663395
Total length (≥ 25000 bp)	4513568	4509580	4521938	4515340	4539096	**4550383**	4527714	4546144	4512460	4537287	4510828	4495372	4461605
Total length (≥ 50000 bp)	4076548	4009703	4141396	4164954	4156507	4129418	4146808	**4165139**	4040894	4157010	4129115	4085924	4115384
# contigs	89	92	83	87	79	84	87	87	85	88	87	91	89
Largest contig	238647	250669	251312	233440	**459069**	270864	271383	271408	251213	270980	251397	300827	251469
Total length	4731590	4731287	4733912	4734797	4736663	4732933	4733552	**4737085**	4730461	4733123	4735739	4734531	4734203
Reference length	4888768	4888768	4888768	4888768	4888768	4888768	4888768	4888768	4888768	4888768	4888768	4888768	4888768
GC (%)	52.12	52.12	52.12	52.12	52.12	52.12	52.12	52.12	52.12	52.12	52.12	52.12	52.12
Reference GC (%)	52.09	52.09	52.09	52.09	52.09	52.09	52.09	52.09	52.09	52.09	52.09	52.09	52.09
N50	95967	103744	103707	104432	105442	105471	104212	105364	103730	103041	**113432**	96275	96157
NG50	95967	96604	103707	103449	104432	104432	103730	104151	103730	96499	**105573**	96162	95970
N75	60821	73757	73753	76273	77767	76975	76712	**77948**	59513	73751	77944	60821	60821
NG75	58498	72340	59289	60614	**76975**	73711	60821	**76975**	58498	60821	73752	58498	58498
L50	17	16	16	15	**13**	15	15	15	16	16	14	16	16
LG50	17	17	16	16	**14**	16	16	16	16	17	15	17	17
L75	31	30	29	29	**26**	28	28	28	30	30	27	30	30
LG75	33	31	31	31	**27**	30	30	29	32	32	29	32	32
# misassemblies	12	12	12	**11**	12	12	12	12	12	12	12	12	12
# misassembled contigs	11	11	11	**10**	11	11	11	11	11	11	11	11	11
Misassembled contigs length	1033227	948436	1034172	**943010**	1263534	1032332	1031396	1032285	991683	1031666	1032946	1014811	1033117
# local misassemblies	67	57	53	63	**43**	53	62	51	68	63	53	61	56
# unaligned mis. contigs	0	0	0	0	0	0	0	0	0	0	0	0	0
# unaligned contigs	5 + 5 part	4 + 5 part	**3 + 5 part**	6 + 6 part	**3 + 5 part**	5 + 5 part	5 + 5 part	6 + 5 part	4 + 5 part	5 + 5 part	5 + 5 part	5 + 5 part	5 + 5 part
Unaligned length	61435	61416	**60729**	62794	61137	62242	61957	62807	61625	62237	62200	61715	61855
Genome fraction (%)	95.501	95.466	95.571	95.548	95.592	95.521	95.502	**95.595**	95.481	95.525	95.579	95.552	95.546
Duplication ratio	**1.000**	1.001	**1.000**	**1.000**	**1.000**	**1.000**	1.001	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
# N's per 100 kbp	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
# mismatches per 100 kbp	7.52	6.49	7.34	6.34	8.56	6.51	8.57	6.33	7.50	6.55	6.53	7.43	**6.29**
# indels per 100 kbp	0.47	0.28	0.49	0.36	0.75	0.30	0.69	0.28	0.54	**0.26**	0.30	0.49	0.28
Largest alignment	238647	250669	251312	233440	283927	270864	271383	271408	251213	270980	251397	**300827**	251469
Total aligned length	4669316	4667770	4672698	4671545	4673500	4670212	4669307	**4673820**	4668360	4670394	4673052	4671906	4671511
NA50	82209	93265	90110	90108	**96263**	93583	89837	95926	89730	89868	93347	86748	86993
NGA50	79980	93265	86993	86993	**93583**	89628	89837	**93583**	86993	86852	93347	82085	82103
NA75	55119	56400	56890	56890	**58734**	56707	56601	58724	55119	56662	56890	55119	55119
NGA75	52699	53069	53855	55119	**56890**	56354	55119	**56890**	53069	55119	55119	53069	53069
LA50	19	18	17	17	**15**	16	17	16	17	18	16	18	18
LGA50	20	18	18	18	**16**	17	17	17	18	19	**16**	19	19
LA75	36	34	34	34	**30**	33	33	32	35	35	32	35	36
LGA75	39	36	37	36	**32**	35	35	34	37	37	34	37	38

Table A.25 Assembly quality metrics for *P. aeruginosa*

Assembly	Uncorrected	ACE	BayesHammer	BFC	BLESS2	Blue	Fiona	Karect	Lighter	Musket	RACER	SGA-EC	Trowel
# contigs (≥ 0 bp)	640	1130	634	443	445	575	501	633	494	600	842	558	576
# contigs (≥ 1000 bp)	99	130	108	106	109	115	110	115	109	108	110	102	108
# contigs (≥ 5000 bp)	74	103	81	78	77	84	80	86	81	81	80	75	81
# contigs (≥ 10000 bp)	68	94	76	73	73	79	74	79	76	76	74	71	77
# contigs (≥ 25000 bp)	59	76	64	63	63	67	64	66	64	66	62	60	65
# contigs (≥ 50000 bp)	38	41	39	40	39	40	40	40	39	40	39	38	39
Total length (≥ 0 bp)	6239779	**6253244**	6234266	6224805	6223542	6231832	6227762	6231337	6227904	6233154	6245761	6232323	6233959
Total length (≥ 1000 bp)	6174563	6152014	6169579	**6175263**	6174246	6170609	6171734	6166898	6173390	6170887	6169227	6170833	6170970
Total length (≥ 5000 bp)	**6113582**	6088350	6105032	6108520	6101292	6102324	6102432	6099653	6107461	6105881	6100101	6107264	6106618
Total length (≥ 10000 bp)	6076163	6027150	6072784	6076383	6074339	6070208	6064233	6051239	6075891	6073767	6062610	**6084914**	6084289
Total length (≥ 25000 bp)	**5923485**	5709368	5869363	5909336	5908572	5858988	5898372	5805141	5870750	5907954	5857882	5902501	5878244
Total length (≥ 50000 bp)	**5112838**	4369494	4949439	5058315	5020886	4874531	5007305	4838181	4945028	4965029	5027585	5087636	4912645
# contigs	114	148	124	123	126	132	126	132	125	126	125	119	125
Largest contig	326181	250136	275768	326189	326267	326202	325991	325840	326196	**326346**	326048	326183	326193
Total length	6184663	6163897	6180295	**6186722**	6185705	6181504	6182450	6178149	6184137	6183023	6179265	6182323	6182463
Reference length	6264404	6264404	6264404	6264404	6264404	6264404	6264404	6264404	6264404	6264404	6264404	6264404	6264404
GC (%)	66.59	66.59	66.59	66.59	66.59	66.59	66.59	66.59	66.59	66.59	66.59	66.59	66.59
Reference GC (%)	66.56	66.56	66.56	66.56	66.56	66.56	66.56	66.56	66.56	66.56	66.56	66.56	66.56
N50	**146539**	94524	128842	116297	112489	112485	112476	109839	112487	112487	116297	146417	112480
NG50	**146539**	94524	128842	112478	112489	112485	111578	109839	112487	112487	116297	146417	112480
N75	**71679**	47785	63533	67237	63639	58292	63360	58397	63700	63532	63700	68180	59169
NG75	**67253**	47710	59080	63700	63533	55195	59240	55255	63318	59079	63533	67245	58397
L50	**15**	23	16	16	16	17	16	18	17	17	16	**15**	17
LG50	**15**	23	16	17	16	17	17	18	17	17	16	**15**	17
L75	**30**	47	34	33	33	36	34	37	34	35	33	31	35
LG75	**31**	48	35	34	34	37	35	38	35	36	34	32	36
# misassemblies	2	**1**	**1**	**1**	**1**	**1**	**1**	**1**	2	**1**	**1**	2	2
# misassembled contigs	2	**1**	**1**	**1**	**1**	**1**	**1**	**1**	2	**1**	**1**	2	2
Misassembled contigs length	160470	**115647**	116061	116297	116297	116297	116154	116186	160337	116297	116297	160337	160337
# local misassemblies	53	171	78	47	**44**	52	65	100	55	62	75	59	60
# unaligned mis. contigs	0	0	0	0	0	0	0	0	0	0	0	0	0
# unaligned contigs	0 + 1 part	0 + 1 part	0 + 1 part	0 + 1 part	0 + 1 part	0 + 1 part	0 + 1 part	0 + 1 part	0 + 1 part	0 + 1 part	0 + 1 part	0 + 1 part	0 + 1 part
Unaligned length	10079	10079	10079	10079	10079	10079	10079	10079	10079	10079	10079	10079	10079
Genome fraction (%)	98.548	98.209	98.477	**98.580**	98.570	98.494	98.512	98.441	98.539	98.517	98.459	98.507	98.509
Duplication ratio	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
# N's per 100 kbp	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
# mismatches per 100 kbp	1.34	1.56	1.64	**1.26**	1.39	1.56	1.39	1.30	1.43	1.51	1.39	1.33	1.59
# indels per 100 kbp	0.73	0.75	0.71	0.71	1.51	**0.70**	0.76	0.79	0.78	**0.70**	0.75	0.71	0.71
Largest alignment	326126	250136	275767	326188	326267	326157	325973	325821	326054	**326346**	326048	326128	326138
Total aligned length	6174529	6153680	6170206	**6176642**	6175542	6171363	6172352	6168051	6173914	6172792	6169085	6172189	6172328
NA50	**146539**	93281	128842	112478	111603	110973	111578	103345	111389	111395	112464	146417	110996
NGA50	**146539**	93281	128842	111591	111603	110973	109872	101701	111389	111395	112464	146417	110996
NA75	**66126**	47674	59080	63700	63533	55195	59240	55255	63318	59079	63700	**66126**	59169
NGA75	**65205**	47653	58397	63533	59240	54953	58397	54953	59240	58256	63533	63700	58397
LA50	**15**	23	16	16	16	17	16	18	17	17	16	**15**	17
LGA50	**15**	23	16	17	16	17	17	19	17	17	16	**15**	17
LA75	**31**	48	35	34	34	37	35	38	35	36	33	32	35
LGA75	**32**	49	36	35	35	38	36	39	36	37	34	33	36

Table A.26 Assembly quality metrics for *H. sapiens* Chr. 21

Assembly	Uncorrected	ACE	BayesHammer	BFC	BLESS2	Blue	Fiona	Karect	Lighter	Musket	RACER	SGA-EC	Trowel
# contigs (≥ 0 bp)	102309	138880	94922	105611	91455	134746	122964	108526	107836	112054	123551	105099	110982
# contigs (≥ 1000 bp)	4098	4449	3534	4123	4049	4664	4251	3925	4119	4178	4693	4100	4343
# contigs (≥ 5000 bp)	2034	2045	1951	2040	2005	2065	2057	2019	2034	2047	2098	2026	2054
# contigs (≥ 10000 bp)	1016	990	1056	1025	1020	945	1001	1040	1034	1026	967	1036	985
# contigs (≥ 25000 bp)	188	157	243	181	187	143	178	194	186	177	132	187	165
# contigs (≥ 50000 bp)	13	4	22	11	14	6	8	15	12	8	2	10	10
Total length (≥ 0 bp)	39562976	**41770782**	39121526	39740631	38469011	41232601	40747821	39951470	39932723	40122382	40677687	39778296	39988114
Total length (≥ 1000 bp)	31826082	31561819	**31942713**	31768142	31597457	31429606	31687664	31831542	31799582	31755892	31449067	31821983	31669381
Total length (≥ 5000 bp)	26458220	25306302	**27837942**	26343661	26237295	24743806	25962836	26857358	26371881	26231497	24731102	26427931	25757309
Total length (≥ 10000 bp)	19170911	17745297	**21343647**	19066792	19172151	16764356	18430354	19863425	19237728	18935031	16662665	19340614	18113654
Total length (≥ 25000 bp)	6517940	5250017	**8675262**	6214022	6368441	4743003	6056407	6908248	6445975	6072489	4278541	6430367	5568310
Total length (≥ 50000 bp)	761378	256413	**1300609**	633021	817430	356391	456768	891449	707687	469569	129147	604189	583509
# contigs	4726	5146	3993	4755	4593	5367	4883	4526	4742	4801	5458	4722	5043
Largest contig	69952	78911	**83462**	66569	83445	69936	66569	80638	68396	69938	69524	69950	66579
Total length	**32278872**	32061659	32275141	32224901	31991512	31940973	32141287	32264069	32248737	32202968	32001416	32272288	32172503
Reference length	46709983	46709983	46709983	46709983	46709983	46709983	46709983	46709983	46709983	46709983	46709983	46709983	46709983
GC (%)	40.61	40.57	40.60	40.59	40.52	40.53	40.58	40.60	40.60	40.60	40.55	40.60	40.57
Reference GC (%)	42.22	42.22	42.22	42.22	42.22	42.22	42.22	42.22	42.22	42.22	42.22	42.22	42.22
N50	12581	11303	**14993**	12387	12711	10627	11882	13052	12534	12237	10400	12646	11664
NG50	7039	6210	**8395**	7017	6960	5817	6564	7422	6984	6885	5803	7103	6470
N75	6478	5751	**7869**	6377	6538	5444	6124	6785	6423	6331	5409	6481	6021
L50	746	829	**632**	759	737	868	785	715	756	769	903	750	798
LG50	1513	1695	**1274**	1534	1518	1808	1604	1445	1525	1556	1843	1510	1634
L75	1640	1810	**1379**	1656	1612	1914	1722	1564	1649	1677	1959	1635	1758
# misassemblies	51	60	52	77	76	**44**	49	57	68	56	106	66	81
# misassembled contigs	48	58	49	74	73	**42**	47	55	65	54	100	63	79
Misassembled contigs length	659650	594475	800208	1074952	899972	**472393**	669011	815815	936482	707009	1290025	1022192	963237
# local misassemblies	76	67	109	73	91	68	77	78	76	72	69	76	**66**
# structural variations	0	1	0	0	0	1	0	0	1	0	0	0	0
# unaligned mis. contigs	2	3	**0**	**0**	1	**0**	**0**	2	1	**0**	**0**	2	**0**
# unaligned contigs	106 + 22 part	123 + 17 part	102 + 25 part	116 + 19 part	**84 + 28 part**	116 + 27 part	115 + 20 part	108 + 23 part	109 + 19 part	118 + 17 part	118 + 20 part	124 + 24 part	104 + 24 part
Unaligned length	143652	162260	142194	156455	**125693**	181182	145244	150318	138623	156759	154912	166104	138168
Genome fraction (%)	**68.680**	68.153	68.565	68.527	68.004	67.876	68.382	68.637	68.622	68.492	68.027	68.615	68.448
Duplication ratio	**1.002**	**1.002**	1.003	**1.002**	1.003	**1.002**	**1.002**	**1.002**	**1.002**	**1.002**	**1.002**	**1.002**	**1.002**
# N's per 100 kbp	384.89	408.14	605.02	396.11	568.92	434.75	410.71	411.02	390.94	394.41	**370.62**	397.72	395.88
# mismatches per 100 kbp	136.06	139.06	135.48	137.33	143.15	136.86	136.10	**133.75**	137.70	138.98	147.99	135.92	137.26
# indels per 100 kbp	54.77	55.70	56.20	55.33	55.66	56.15	55.37	55.33	55.54	54.97	56.07	55.09	**54.62**
Largest alignment	69952	78850	**83373**	66487	83356	69936	66486	80577	68230	69938	63517	69950	66496
Total aligned length	**32088467**	31842867	32029480	32016988	31768444	31709911	31947406	32064852	32062277	31999522	31788360	32057890	31980829
NA50	12444	11114	**14723**	12151	12322	10516	11679	12915	12315	12013	10157	12388	11431
NGA50	6948	6106	**8199**	6804	6788	5755	6478	7278	6843	6720	5635	6949	6360
NA75	6384	5651	**7677**	6250	6390	5384	6029	6662	6336	6238	5273	6370	5890
LA50	751	836	**641**	771	752	875	793	722	764	777	923	761	810
LGA50	1529	1717	**1298**	1563	1553	1823	1623	1463	1546	1577	1887	1536	1660
LA75	1657	1835	**1406**	1687	1650	1931	1743	1584	1673	1700	2005	1664	1786

Table A.27 Assembly quality metrics for *C. elegans*

Assembly	Uncorrected	ACE	BayesHammer	BFC	BLESS2	Blue	Fiona	Karect	Lighter	Musket	RACER	SGA-EC	Trowel
# contigs (≥ 0 bp)	203415	241287	191757	198118	205987	222398	219816	199468	212134	220202	242493	204134	202600
# contigs (≥ 1000 bp)	23820	27679	23229	23760	28064	23956	25503	23696	25433	26169	28621	23808	23779
# contigs (≥ 5000 bp)	5265	3822	5305	5311	3471	5287	5014	5378	5078	4789	3898	5288	5254
# contigs (≥ 10000 bp)	1637	741	1712	1656	647	1613	1296	1684	1329	1170	789	1651	1644
# contigs (≥ 25000 bp)	198	69	209	191	59	175	125	184	118	117	90	191	195
# contigs (≥ 50000 bp)	45	9	46	39	5	34	24	37	25	27	16	42	45
Total length (≥ 0 bp)	118799992	114170074	117537858	118214042	109961204	**119444609**	118285707	118206096	118387916	118072643	118233003	118714248	118615565
Total length (≥ 1000 bp)	**96127803**	85118518	95937053	96044197	82706061	95032072	93009114	96009464	93691897	92334252	87446335	96076605	95935042
Total length (≥ 5000 bp)	53843135	32228125	**55090132**	53980078	28994075	52574207	46636697	54131678	47273617	44165284	33627639	53830727	53633067
Total length (≥ 10000 bp)	28768855	11599807	**30181631**	28686276	10153864	27119766	21103125	28524948	21543242	19282752	12875429	28667755	28646245
Total length (≥ 25000 bp)	8569575	2562224	**8944599**	8082901	2063112	6905329	5001497	7424989	4707794	4885306	3436135	8148320	8344784
Total length (≥ 50000 bp)	3442593	559485	**3548962**	3017440	283191	2283060	1670084	2574568	1568823	1916567	1002804	3236938	3446748
# contigs	33262	41130	32320	33118	42334	33807	36584	32950	36368	37821	43721	33173	33241
Largest contig	158036	81042	**161446**	159367	64717	106190	139294	139293	80140	139290	82845	148754	158036
Total length	**102988100**	94836365	102540189	102842496	93026396	102180262	101021330	102718424	101620228	100768560	98311793	102877857	102814415
Reference length	100286070	100286070	100286070	100286070	100286070	100286070	100286070	100286070	100286070	100286070	100286070	100286070	100286070
GC (%)	38.56	38.56	38.52	38.55	38.52	38.54	38.49	38.52	38.38	38.57	38.75	38.56	38.57
Reference GC (%)	35.44	35.44	35.44	35.44	35.44	35.44	35.44	35.44	35.44	35.44	35.44	35.44	35.44
N50	5329	3414	**5560**	5365	3178	5208	4532	5400	4594	4285	3390	5329	5302
NG50	5534	3187	**5756**	5547	2898	5345	4581	5575	4657	4306	3304	5534	5509
N75	2461	1766	**2531**	2474	1665	2399	2169	2491	2201	2074	1685	2469	2461
NG75	2639	1534	**2690**	2650	1376	2527	2210	2662	2278	2099	1603	2640	2629
L50	4810	7512	**4580**	4817	7916	4997	5828	4844	5818	6134	7711	4825	4822
LG50	4561	8338	**4381**	4583	9111	4817	5747	4623	5674	6078	8006	4587	4588
L75	11969	17217	**11458**	11921	18097	12253	13928	11886	13827	14651	18076	11956	11958
LG75	11174	19698	**10810**	11171	21694	11675	13676	11177	13379	14477	18977	11194	11212
# misassemblies	**598**	1860	721	657	1678	757	990	675	830	1435	2774	665	628
# misassembled contigs	**594**	1787	708	654	1635	740	975	669	818	1380	2642	658	620
Misassembled contigs length	**3122101**	6343367	4041907	3343362	5160442	3652768	4258832	3495943	3923723	5540661	8112551	3344928	3242140
# local misassemblies	169	175	176	169	193	159	179	166	174	185	**120**	167	165
# unaligned mis. contigs	1	1	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	1	1
# unaligned contigs	4673 + 21 part	4271 + 42 part	4609 + 31 part	4709 + 22 part	**3856 + 42 part**	4949 + 28 part	5128 + 40 part	4783 + 29 part	5175 + 33 part	5273 + 32 part	5215 + 36 part	4690 + 23 part	4699 + 21 part
Unaligned length	15457774	13185257	15271613	15438170	**12771696**	15253161	14748271	15321756	14650849	14909845	14753767	15464643	15423697
Genome fraction (%)	**87.069**	81.060	86.857	86.948	79.860	86.459	85.771	86.946	86.467	85.306	82.599	86.948	86.934
Duplication ratio	**1.002**	1.004	**1.002**	**1.002**	**1.002**	1.003	1.003	**1.002**	1.003	1.004	1.009	**1.002**	**1.002**
# N's per 100 kbp	**0.00**	**0.00**	0.82	**0.00**	8.93	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
# mismatches per 100 kbp	**15.40**	44.37	15.81	17.49	22.64	17.76	47.72	17.15	45.84	110.51	101.37	16.20	15.46
# indels per 100 kbp	**8.60**	12.72	9.59	8.80	11.48	9.26	16.91	8.68	9.11	11.10	25.60	8.63	8.65
Largest alignment	51074	23778	**51077**	51069	22624	40495	40508	51069	36434	36039	33712	40497	51069
Total aligned length	**87429792**	81366347	87191165	87306376	80143767	86810073	86135010	87297553	86834040	85673728	83041928	87307082	87290550
NA50	3971	2519	**4112**	3992	2352	3889	3414	4039	3495	3181	2413	3975	3977
NGA50	4132	2329	**4241**	4140	2123	4008	3451	4181	3567	3202	2348	4131	4124
NA75	1382	1060	**1422**	1388	1024	1354	1258	1403	1309	1188	937	1382	1381
NGA75	1561	833	**1582**	1557	750	1474	1299	1569	1382	1212	864	1555	1545
LA50	6644	10118	**6360**	6600	10701	6699	7679	6505	7608	8167	10642	6629	6637
LGA50	6310	11243	**6090**	6285	12326	6459	7572	6210	7419	8092	11056	6309	6325
LA75	17451	24444	**16788**	17359	25515	17681	19681	17121	19304	20936	26838	17429	17429
LGA75	16071	28787	**15663**	16054	31704	16676	19249	15890	18560	20635	28482	16103	16131

Table A.28 Assembly quality metrics for *D. melanogaster*

Assembly	Uncorrected	ACE	BayesHammer	BFC	BLESS2	Blue	Fiona	Karect	Lighter	Musket	RACER	SGA-EC	Trowel
# contigs (≥ 0 bp)	223581	248508	208606	219200	159809	232478	225631	204944	227250	213768	210008	227076	216608
# contigs (≥ 1000 bp)	10040	11133	9437	9629	9261	9643	8912	8550	9507	8784	10450	9914	9747
# contigs (≥ 5000 bp)	4821	5486	4549	4709	5244	4582	4630	4330	4750	4656	5536	4800	4796
# contigs (≥ 10000 bp)	2952	3223	2830	2936	3387	2851	3031	2851	2985	3149	3487	2913	2935
# contigs (≥ 25000 bp)	1174	1065	1180	1183	1222	1169	1253	1238	1193	1305	1179	1181	1177
# contigs (≥ 50000 bp)	386	292	435	412	338	413	429	482	403	431	292	395	405
Total length (≥ 0 bp)	136848302	136462329	135724785	136392768	127932504	136892327	136300390	135379354	136339660	135162866	134486672	**137098799**	136237946
Total length (≥ 1000 bp)	114247383	112818345	114082986	114339445	113270501	113759034	114780121	**115028366**	114022280	114743642	114636046	114134174	114341269
Total length (≥ 5000 bp)	101715090	98832982	102305243	102547056	103248114	101718710	104544426	**105023741**	102447259	104854744	102687782	101822912	102455498
Total length (≥ 10000 bp)	88402381	82657559	90155500	89922855	89823586	89456079	93072964	**94439411**	89850423	93985779	88004113	88377312	89177890
Total length (≥ 25000 bp)	60421211	48860642	64093407	62278840	55230759	63079258	64942162	**68852849**	61545757	64708831	51593221	61036017	61243879
Total length (≥ 50000 bp)	32991908	22560206	38365354	35361693	24899345	36643703	36021083	**42374171**	34077796	34198243	20727594	33514313	34119167
# contigs	13390	13999	12526	12905	11272	12987	11912	11595	12438	11844	13716	13298	13078
Largest contig	444298	390345	420926	357542	390331	444337	408728	**444753**	444255	418612	197844	444372	390304
Total length	116592521	114830494	116237827	116630464	114687356	116107506	116882147	**117162385**	116068379	116889766	116922349	116504780	116670858
Reference length	120381546	120381546	120381546	120381546	120381546	120381546	120381546	120381546	120381546	120381546	120381546	120381546	120381546
GC (%)	42.60	42.60	42.59	42.59	42.59	42.61	42.60	42.59	42.56	42.61	42.60	42.59	42.60
Reference GC (%)	42.42	42.42	42.42	42.42	42.42	42.42	42.42	42.42	42.42	42.42	42.42	42.42	42.42
N50	26455	20337	29434	27934	24097	28588	29541	**33231**	27319	29410	21563	26890	27164
NG50	25129	18885	27801	26470	22588	27150	28304	**31616**	25898	28045	20684	25447	25730
N75	10391	8873	11384	11002	11346	11024	12504	**13413**	11183	12753	10116	10445	10768
NG75	9201	7545	9923	9797	9844	9590	11241	**12008**	9794	11724	9183	9215	9506
L50	1092	1444	960	1033	1309	982	1015	**880**	1059	1074	1475	1074	1066
LG50	1165	1586	1032	1102	1431	1058	1075	**930**	1140	1135	1557	1148	1136
L75	2858	3592	2552	2703	3031	2626	2548	**2282**	2721	2592	3456	2816	2774
LG75	3149	4101	2844	2973	3434	2936	2769	**2472**	3030	2807	3726	3113	3049
# misassemblies	**626**	698	628	664	748	670	658	647	640	672	967	639	670
# misassembled contigs	557	642	**553**	590	686	593	581	568	574	609	868	563	600
Misassembled contigs length	21033568	**17912825**	22874463	22837694	18568193	24054399	22543860	25176362	21502728	21660038	21288114	21894291	22474232
# local misassemblies	3655	3730	3808	3708	**1998**	3652	3435	3277	3582	2808	2638	3836	3560
# unaligned mis. contigs	31	21	25	28	**5**	30	20	16	26	16	9	38	29
# unaligned contigs	4185 + 473 part	3304 + 583 part	3948 + 540 part	4139 + 549 part	**2271 + 264 part**	4081 + 571 part	3946 + 429 part	4065 + 412 part	3633 + 499 part	3884 + 353 part	4014 + 348 part	4191 + 527 part	4139 + 470 part
Unaligned length	7951307	7080437	8027856	8020296	**4491455**	7866518	7404636	7604236	7344462	6897357	6814426	8129395	7896669
Genome fraction (%)	90.330	89.613	90.020	90.317	**91.527**	89.995	91.010	91.072	90.402	91.406	91.440	90.111	90.431
Duplication ratio	**0.999**	**0.999**	**0.999**	**0.999**	1.000	**0.999**	**0.999**	**0.999**	**0.999**	1.000	1.000	**0.999**	**0.999**
# N's per 100 kbp	8.65	7.31	18.24	7.22	49.90	7.18	6.37	5.91	7.00	5.00	**3.87**	7.67	7.76
# mismatches per 100 kbp	516.84	520.94	**513.77**	519.00	545.03	517.06	529.10	524.86	521.12	551.98	550.98	516.38	519.40
# indels per 100 kbp	130.93	131.77	**130.35**	132.28	141.13	130.74	134.60	131.52	131.03	134.06	141.09	130.55	131.89
Largest alignment	258232	202757	**259408**	258275	181773	258234	258107	258280	227789	258274	186904	258194	205327
Total aligned length	108390962	107468245	107979664	108367980	**109936320**	107999203	109234193	109329923	108475905	109736421	109757147	108127804	108527668
NA50	23503	18425	25924	24533	21970	25239	26026	**28465**	24293	25797	19085	23965	23998
NGA50	22381	17101	24534	23383	20512	23825	24990	**27415**	23152	24610	18420	22825	22928
NA75	8961	7631	9768	9426	10121	9434	10704	**11407**	9649	11103	8537	8938	9195
NGA75	7823	6313	8263	8230	8624	8044	9575	**10244**	8239	10032	7664	7801	8075
LA50	1253	1617	1124	1202	1453	1154	1170	**1037**	1220	1225	1661	1241	1233
LGA50	1335	1774	1206	1280	1587	1242	1239	**1094**	1311	1294	1753	1324	1312
LA75	3254	4034	2938	3107	3367	3035	2911	**2639**	3101	2943	3928	3224	3181
LGA75	3595	4634	3283	3426	3824	3403	3170	**2862**	3463	3192	4249	3573	3504

A.4.3 SPAdes

A.4.3.1 *B. dentium*

Table A.29 contains the Quast report after assembling dataset *B. dentium* with SPAdes.

A.4.3.2 *E. coli str. K-12 substr. DH10B*

Table A.30 contains the Quast report after assembling dataset *E. coli str. K-12 substr. DH10B* with SPAdes.

A.4.3.3 *E. coli str. K-12 substr. MG1655*

Table A.31 contains the Quast report after assembling dataset *E. coli str. K-12 substr. MG1655* with SPAdes.

A.4.3.4 *S. enterica*

Table A.32 contains the Quast report after assembling dataset *S. enterica* with SPAdes.

A.4.3.5 *P. aeruginosa*

Table A.33 contains the Quast report after assembling dataset *P. aeruginosa* with SPAdes.

A.4.3.6 *H. sapiens* Chr. 21

Table A.34 contains the Quast report after assembling dataset *H. sapiens* Chr. 21 with SPAdes.

A.4.3.7 *C. elegans*

Table A.35 contains the Quast report after assembling dataset *C. elegans* with SPAdes.

A.4.3.8 *D. melanogaster*

Table A.36 contains the Quast report after assembling dataset *D. melanogaster* with SPAdes.

Table A.29 Assembly quality metrics for *B. dentium*

Assembly	Uncorrected	ACE	BayesHammer	BFC	BLESS2	Blue	Fiona	Karect	Lighter	Musket	RACER	SGA-EC	Trowel
# contigs (≥ 0 bp)	1688	278	1653	1684	887	1664	1637	1680	1666	1682	1658	1678	1680
# contigs (≥ 1000 bp)	56	15	54	62	37	68	49	53	56	54	57	58	61
# contigs (≥ 5000 bp)	12	12	12	12	13	12	12	12	12	12	12	12	12
# contigs (≥ 10000 bp)	11	11	11	11	11	11	11	11	11	11	11	11	11
# contigs (≥ 25000 bp)	11	11	11	11	11	11	11	11	11	11	11	11	11
# contigs (≥ 50000 bp)	9	9	9	9	9	9	9	9	9	9	9	9	9
Total length (≥ 0 bp)	3518219	2733900	3503386	**3522389**	3046753	3518493	3487142	3514203	3507178	3516367	3503223	3516842	3522183
Total length (≥ 1000 bp)	2675554	2614800	2673009	2681531	2652461	**2693555**	2659416	2669659	2673430	2671485	2672971	2679544	2684632
Total length (≥ 5000 bp)	2610109	2610110	2610110	2610111	**2615377**	2610110	2610111	2610110	2610112	2610110	2610110	2610109	2610109
Total length (≥ 10000 bp)	2604505	2604506	2604506	2604507	2604505	2604506	2604507	2604506	**2604508**	2604506	2604506	2604505	2604505
Total length (≥ 25000 bp)	2604505	2604506	2604506	2604507	2604505	2604506	2604507	2604506	**2604508**	2604506	2604506	2604505	2604505
Total length (≥ 50000 bp)	2521362	2521363	2521363	2521364	2521362	2521363	2521364	2521363	**2521365**	2521363	2521363	2521362	2521362
# contigs	821	66	830	839	204	832	825	839	821	841	832	826	829
Largest contig	610215	610215	610215	610215	610215	610215	610215	610215	610215	610215	610215	610215	610215
Total length	3139081	2643783	3142820	3152040	2755449	**3153640**	3128460	3145410	3136958	3147715	3142543	3143479	3149960
Reference length	2636367	2636367	2636367	2636367	2636367	2636367	2636367	2636367	2636367	2636367	2636367	2636367	2636367
GC (%)	58.90	58.47	58.90	58.89	58.45	58.88	58.90	58.90	58.89	58.89	58.89	58.90	58.86
Reference GC (%)	58.54	58.54	58.54	58.54	58.54	58.54	58.54	58.54	58.54	58.54	58.54	58.54	58.54
N50	287948	**397392**	287949	287949	**397392**	287948	287949	287949	287949	287949	287948	287948	287948
NG50	397392	397392	397392	397392	397392	397392	397392	397392	397392	397392	397392	397392	397392
N75	158502	**177423**	158502	158502	**177423**	158502	158502	158502	158502	158502	158502	158502	158502
NG75	177423	177423	177423	177423	177423	177423	177423	177423	177423	177423	177423	177423	177423
L50	4	**3**	4	4	**3**	4	4	4	4	4	4	4	4
LG50	3	3	3	3	3	3	3	3	3	3	3	3	3
L75	8	**6**	8	8	**6**	8	8	8	8	8	8	8	8
LG75	6	6	6	6	6	6	6	6	6	6	6	6	6
# misassemblies	0	0	0	0	0	0	0	0	0	0	0	0	0
# misassembled contigs	0	0	0	0	0	0	0	0	0	0	0	0	0
Misassembled contigs length	0	0	0	0	0	0	0	0	0	0	0	0	0
# local misassemblies	3	3	3	3	3	3	3	3	3	3	3	3	3
# unaligned mis. contigs	**0**	7	1	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
# unaligned contigs	805 + 2 part	**34 + 2 part**	815 + 1 part	824 + 1 part	190 + 0 part	816 + 2 part	811 + 0 part	823 + 1 part	805 + 2 part	824 + 2 part	815 + 3 part	811 + 1 part	814 + 1 part
Unaligned length	525292	**21151**	529055	538427	141937	539856	514946	531251	523118	533413	528712	529836	536304
Genome fraction (%)	99.094	99.094	99.094	99.094	99.094	99.094	99.094	99.094	99.094	99.094	99.094	99.094	99.094
Duplication ratio	1.001	1.004	**1.000**	**1.000**	**1.000**	1.001	**1.000**	1.001	1.001	1.001	1.001	**1.000**	**1.000**
# N's per 100 kbp	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
# mismatches per 100 kbp	**0.23**	**0.23**	0.27	**0.23**	**0.23**	0.31	0.27	**0.23**	0.27	0.27	0.27	**0.23**	0.27
# indels per 100 kbp	0.27	0.31	0.31	0.27	0.27	**0.23**	0.27	0.31	0.31	0.31	**0.23**	0.27	0.27
Largest alignment	610215	610215	610215	610215	610215	610215	610215	610215	610215	610215	610215	610215	610215
Total aligned length	2613514	**2615318**	2613490	2613338	2613237	2613509	2613239	2613445	2613565	2613554	2613556	2613368	2613381
NA50	287948	**397392**	287949	287949	**397392**	287948	287949	287949	287949	287949	287948	287948	287948
NGA50	397392	397392	397392	397392	397392	397392	397392	397392	397392	397392	397392	397392	397392
NA75	158502	**177423**	158502	158502	**177423**	158502	158502	158502	158502	158502	158502	158502	158502
NGA75	177423	177423	177423	177423	177423	177423	177423	177423	177423	177423	177423	177423	177423
LA50	4	**3**	4	4	**3**	4	4	4	4	4	4	4	4
LGA50	3	3	3	3	3	3	3	3	3	3	3	3	3
LA75	8	**6**	8	8	**6**	8	8	8	8	8	8	8	8
LGA75	6	6	6	6	6	6	6	6	6	6	6	6	6

Table A.30 Assembly quality metrics for *E. coli str. K-12 substr. DH10B*

Assembly	Uncorrected	ACE	BayesHammer	BFC	BLESS2	Blue	Fiona	Karect	Lighter	Musket	RACER	SGA-EC	Trowel
# contigs (≥ 0 bp)	186	117	109	125	114	122	116	114	112	121	121	134	134
# contigs (≥ 1000 bp)	79	79	78	79	78	79	79	81	78	79	79	79	78
# contigs (≥ 5000 bp)	61	62	61	62	62	61	61	61	61	62	62	61	61
# contigs (≥ 10000 bp)	54	55	54	55	55	54	54	54	54	55	55	54	54
# contigs (≥ 25000 bp)	49	49	49	49	49	49	49	49	49	49	49	49	49
# contigs (≥ 50000 bp)	29	28	29	28	28	29	29	29	29	28	28	29	29
Total length (≥ 0 bp)	**4390611**	4373352	4372108	4375195	4372035	4372602	4372343	4372889	4372776	4373454	4373398	4376447	4377075
Total length (≥ 1000 bp)	4365674	**4366114**	4365888	4366067	4364378	4363735	4365099	4365206	4365467	4365245	4365002	4365146	4365776
Total length (≥ 5000 bp)	4325946	4328231	4327992	**4328283**	4327750	4323821	4325928	4319347	4327639	4327413	4327169	4325543	4327967
Total length (≥ 10000 bp)	4269540	4271604	4271516	**4271867**	4271647	4267470	4269394	4262831	4271107	4270997	4270753	4269127	4271551
Total length (≥ 25000 bp)	4192073	4178567	**4194081**	4178971	4178836	4190037	4191951	4185397	4193674	4178118	4177918	4191660	4194019
Total length (≥ 50000 bp)	3456316	3395882	3455789	3396435	3396228	3455430	3453075	3453758	**3459882**	3395638	3395515	3453555	3455950
# contigs	83	82	81	82	81	82	81	84	81	82	82	82	81
Largest contig	326373	326271	326372	326373	**326386**	326277	326373	326372	326372	326372	326369	326373	326373
Total length	**4368476**	4368165	4367939	4368118	4366425	4365851	4366547	4367646	4367518	4367296	4367053	4367197	4367827
Reference length	4686137	4686137	4686137	4686137	4686137	4686137	4686137	4686137	4686137	4686137	4686137	4686137	4686137
GC (%)	50.75	50.75	50.75	50.75	50.75	50.75	50.75	50.75	50.75	50.75	50.75	50.75	50.75
Reference GC (%)	50.78	50.78	50.78	50.78	50.78	50.78	50.78	50.78	50.78	50.78	50.78	50.78	50.78
N50	108254	108227	**108384**	108254	108254	107774	107813	108228	107839	107839	107839	108254	108254
NG50	**97353**	**97353**	**97353**	**97353**	**97353**	97288	**97353**	**97353**	**97353**	**97353**	**97353**	**97353**	**97353**
N75	**57964**	57170	**57964**	57170	57147	**57964**	**57964**	**57964**	**57964**	57170	57147	**57964**	**57964**
NG75	**46124**	45930	**46124**	45889	45849	45891	**46124**	45941	**46124**	45889	45889	**46124**	**46124**
L50	13	13	13	13	13	13	13	13	13	13	13	13	13
LG50	14	14	14	14	14	14	14	14	14	14	14	14	14
L75	26	26	26	26	26	26	26	26	26	26	26	26	26
LG75	31	31	31	31	31	31	31	31	31	31	31	31	31
# misassemblies	**0**	**0**	**0**	**0**	**0**	1	**0**	**0**	1	**0**	**0**	**0**	**0**
# misassembled contigs	**0**	**0**	**0**	**0**	**0**	1	**0**	**0**	1	**0**	**0**	**0**	**0**
Misassembled contigs length	**0**	**0**	**0**	**0**	**0**	188639	**0**	**0**	91840	**0**	**0**	**0**	**0**
# local misassemblies	7	8	10	8	8	**6**	7	8	7	7	7	10	10
# structural variations	0	0	0	0	0	0	0	0	0	0	0	0	0
# unaligned mis. contigs	0	0	0	0	0	0	0	0	0	0	0	0	0
# unaligned contigs	0 + 0 part	0 + 0 part	0 + 0 part	0 + 0 part	0 + 0 part	0 + 0 part	0 + 0 part	0 + 0 part	0 + 0 part	0 + 0 part	0 + 0 part	0 + 0 part	0 + 0 part
Unaligned length	0	0	0	0	0	0	0	0	0	0	0	0	0
Genome fraction (%)	**93.179**	93.161	93.172	93.174	93.139	93.136	93.149	93.163	93.168	93.161	93.157	93.157	93.167
Duplication ratio	**1.000**	1.001	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
# N's per 100 kbp	5.08	4.85	**0.23**	4.85	4.86	5.08	5.08	5.08	5.08	4.85	4.85	**0.23**	**0.23**
# mismatches per 100 kbp	**3.57**	4.44	5.02	4.21	4.63	6.74	5.93	4.86	7.86	5.75	5.82	5.22	4.42
# indels per 100 kbp	0.46	0.50	0.48	0.48	0.48	0.62	0.57	**0.39**	0.57	0.50	0.64	0.44	0.48
Largest alignment	326373	326195	326372	326373	**326386**	326277	326373	326372	326372	326372	326369	326373	326373
Total aligned length	**4368147**	4367074	4367658	4367700	4365891	4365522	4366098	4367324	4367094	4366665	4366460	4366978	4367451
NA50	108254	108227	**108381**	108254	108254	107774	107813	108228	107839	107839	107839	108254	108254
NGA50	**97353**	**97353**	**97353**	**97353**	**97353**	97288	**97353**	**97353**	**97353**	**97353**	**97353**	**97353**	**97353**
NA75	**57964**	57170	**57964**	57170	57147	57170	**57964**	**57964**	**57964**	57170	57147	**57964**	**57964**
NGA75	**46124**	45854	**46124**	45889	45849	45889	**46124**	45940	**46124**	45889	45889	**46124**	**46124**
LA50	13	13	13	13	13	13	13	13	13	13	13	13	13
LGA50	14	14	14	14	14	14	14	14	14	14	14	14	14
LA75	**26**	**26**	**26**	**26**	**26**	27	**26**	**26**	**26**	**26**	**26**	**26**	**26**
LGA75	31	31	31	31	31	31	31	31	31	31	31	31	31

Table A.31 Assembly quality metrics for *E. coli str. K-12 substr. MG1655*

Assembly	Uncorrected	ACE	BayesHammer	BFC	BLESS2	Blue	Fiona	Karect	Lighter	Musket	RACER	SGA-EC	Trowel
# contigs (≥ 0 bp)	164	158	150	159	153	159	161	155	167	160	190	155	164
# contigs (≥ 1000 bp)	81	78	79	81	83	78	82	80	79	81	82	82	82
# contigs (≥ 5000 bp)	57	55	55	56	59	55	58	56	56	56	59	57	58
# contigs (≥ 10000 bp)	53	51	51	52	55	51	54	52	52	51	55	53	54
# contigs (≥ 25000 bp)	46	44	44	45	47	44	46	45	45	45	46	46	46
# contigs (≥ 50000 bp)	32	30	30	31	33	30	32	31	31	31	30	32	32
Total length (≥ 0 bp)	4574902	4572069	4571284	4573705	4571333	4571540	4573075	4573804	4574012	4573003	**4576108**	4572122	4573804
Total length (≥ 1000 bp)	4552276	4550704	4550707	4552733	**4553419**	4549150	4551715	4552703	4550476	4551502	4550150	4552196	4551826
Total length (≥ 5000 bp)	4498097	**4498467**	4496562	4497074	4497995	4497320	4498289	4497588	4498037	4494003	4498297	4496914	4497625
Total length (≥ 10000 bp)	4468403	**4468853**	4466868	4467460	4468381	4467706	4468675	4467894	4468343	4457829	4468683	4467220	4467931
Total length (≥ 25000 bp)	4365479	**4365929**	4363941	4364536	4350852	4364932	4353033	4364945	4365419	4365575	4326060	4364293	4352292
Total length (≥ 50000 bp)	3878734	**3880534**	3877098	3878218	3878960	3879694	3867798	3878598	3880218	3880177	3756292	3877450	3865520
# contigs	93	88	92	93	91	91	95	92	91	93	94	93	94
Largest contig	264571	268093	285227	285527	264572	**285529**	264571	285527	264571	285527	264571	264571	264571
Total length	4560939	4557864	4560016	4561050	4559088	4558416	4560905	**4561354**	4559497	4559819	4558813	4560187	4560489
Reference length	4639675	4639675	4639675	4639675	4639675	4639675	4639675	4639675	4639675	4639675	4639675	4639675	4639675
GC (%)	50.74	50.75	50.74	50.74	50.75	50.75	50.74	50.75	50.74	50.74	50.74	50.74	50.74
Reference GC (%)	50.79	50.79	50.79	50.79	50.79	50.79	50.79	50.79	50.79	50.79	50.79	50.79	50.79
N50	133309	**133713**	133309	133309	126410	133309	133309	133309	133309	133309	133309	133189	133309
NG50	132876	**133713**	133309	133088	126410	133309	132876	133058	133309	133088	132876	132876	132876
N75	64785	67335	**67340**	64785	64785	67335	64692	64785	67335	64785	62691	64399	64585
NG75	60768	**64785**	64399	60768	60768	**64785**	62691	60768	**64785**	62691	59669	60768	60768
L50	13	**12**	**12**	**12**	14	**12**	13	**12**	13	**12**	13	13	13
LG50	14	**12**	**12**	13	14	**12**	14	13	13	13	14	14	14
L75	25	**23**	**23**	24	25	**23**	25	24	24	24	25	25	25
LG75	26	**24**	**24**	25	26	**24**	26	25	25	25	26	26	26
# misassemblies	0	0	0	0	0	0	0	0	0	0	0	0	0
# misassembled contigs	0	0	0	0	0	0	0	0	0	0	0	0	0
Misassembled contigs length	0	0	0	0	0	0	0	0	0	0	0	0	0
# local misassemblies	**5**	**5**	13	6	**5**	6	6	**5**	8	6	7	11	8
# unaligned mis. contigs	0	0	0	0	0	0	0	0	0	0	0	0	0
# unaligned contigs	4 + 0 part	2 + 0 part	4 + 0 part	4 + 0 part	**0 + 0 part**	4 + 0 part	5 + 0 part	4 + 0 part	4 + 0 part	4 + 0 part	4 + 0 part	4 + 0 part	4 + 0 part
Unaligned length	2946	1443	2947	2946	**0**	2946	3819	2946	2946	2946	2946	2946	2946
Genome fraction (%)	98.203	98.170	98.188	98.207	**98.219**	98.159	98.144	98.210	98.171	98.180	98.153	98.187	98.154
Duplication ratio	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	1.001	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	1.001
# N's per 100 kbp	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
# mismatches per 100 kbp	**2.28**	3.10	3.60	3.34	3.25	3.71	3.91	3.07	3.23	3.84	3.40	2.68	2.68
# indels per 100 kbp	0.50	0.44	0.55	0.46	**0.42**	0.53	0.51	0.50	0.50	0.50	0.44	0.55	0.53
Largest alignment	264571	268093	285227	285527	264572	**285529**	264571	285527	264571	285527	264571	264571	264571
Total aligned length	4557359	4555713	4556641	4557578	**4558342**	4555134	4554498	4557803	4555851	4556239	4555066	4556637	4555113
NA50	133309	**133713**	133309	133309	126372	133309	133309	133309	133309	133309	133309	133189	133309
NGA50	132876	**133713**	133309	133088	126372	133309	132876	133058	133309	133088	132876	132876	132876
NA75	64785	67335	**67340**	64785	64785	67335	64692	64785	67335	64785	62691	64399	64585
NGA75	60768	**64785**	64399	60768	60768	**64785**	62691	60768	**64785**	62691	59669	60768	60768
LA50	13	**12**	**12**	**12**	14	**12**	13	**12**	13	**12**	13	13	13
LGA50	14	**12**	**12**	13	14	**12**	14	13	13	13	14	14	14
LA75	25	**23**	**23**	24	25	**23**	25	24	24	24	25	25	25
LGA75	26	**24**	**24**	25	26	**24**	26	25	25	25	26	26	26

Table A.32 Assembly quality metrics for *S. enterica*

Assembly	Uncorrected	ACE	BayesHammer	BFC	BLESS2	Blue	Fiona	Karect	Lighter	Musket	RACER	SGA-EC	Trowel
# contigs (≥ 0 bp)	164	68	74	115	85	104	88	89	91	96	120	136	124
# contigs (≥ 1000 bp)	26	25	26	26	28	28	24	24	25	25	27	26	26
# contigs (≥ 5000 bp)	20	17	19	19	21	19	19	18	19	19	21	19	19
# contigs (≥ 10000 bp)	19	16	18	18	20	18	18	17	18	18	20	18	18
# contigs (≥ 25000 bp)	18	15	17	17	19	17	17	16	17	17	19	17	17
# contigs (≥ 50000 bp)	15	13	14	14	15	14	14	14	14	14	15	14	14
Total length (≥ 0 bp)	**4832029**	4793717	4796241	4814811	4799092	4806157	4800907	4802582	4804245	4805547	4811040	4823822	4818253
Total length (≥ 1000 bp)	4777575	**4781314**	4779647	4780180	4779682	4780281	4777060	4778776	4778844	4778551	4777452	4780362	4779898
Total length (≥ 5000 bp)	4760565	**4762438**	4761734	4761892	4761770	4760166	4761628	4762066	4761834	4761767	4760713	4762074	4761833
Total length (≥ 10000 bp)	4754365	**4756238**	4755534	4755692	4755570	4753966	4755428	4755866	4755634	4755567	4754513	4755874	4755633
Total length (≥ 25000 bp)	4744007	**4745880**	4745176	4745334	4745212	4743653	4745070	4745508	4745276	4745209	4744155	4745516	4745275
Total length (≥ 50000 bp)	4639979	**4687841**	4641148	4641306	4610505	4639921	4641042	4687469	4641248	4641181	4609483	4641488	4641247
# contigs	39	30	33	38	38	36	36	34	37	35	40	36	37
Largest contig	**1022731**	**1022731**	**1022731**	**1022731**	**1022731**	1022686	**1022731**	**1022731**	**1022731**	**1022731**	**1022731**	**1022731**	**1022731**
Total length	4787310	4785068	4785034	**4788472**	4786552	4785774	4786046	4786049	4787682	4785715	4787246	4787493	4787748
Reference length	4888768	4888768	4888768	4888768	4888768	4888768	4888768	4888768	4888768	4888768	4888768	4888768	4888768
GC (%)	52.14	52.14	52.14	52.14	52.14	52.14	52.14	52.14	52.14	52.14	52.14	52.14	52.14
Reference GC (%)	52.09	52.09	52.09	52.09	52.09	52.09	52.09	52.09	52.09	52.09	52.09	52.09	52.09
N50	412097	**725282**	723537	723537	381864	723550	723537	**725282**	723537	723537	381603	723537	723537
NG50	412097	**725282**	723537	723537	381864	723550	723537	**725282**	723537	723537	381603	723537	723537
N75	235664	**298653**	235664	235664	231409	235230	235664	235664	235664	235664	231409	235664	235664
NG75	231409	**298653**	235664	235664	231409	235230	235664	235664	235664	235664	231409	235664	235664
L50	4	**3**	**3**	**3**	4	**3**	**3**	**3**	**3**	**3**	4	**3**	**3**
LG50	4	**3**	**3**	**3**	4	**3**	**3**	**3**	**3**	**3**	4	**3**	**3**
L75	7	**6**	7	7	8	7	7	7	7	7	8	7	7
LG75	8	**6**	7	7	8	7	7	7	7	7	8	7	7
# misassemblies	**12**	14	14	14	13	14	14	14	14	14	**12**	14	14
# misassembled contigs	8	**7**	**7**	**7**	8	**7**	**7**	**7**	**7**	**7**	8	**7**	**7**
Misassembled contigs length	3676571	3960170	3677740	3677898	3647205	3677455	3677707	3724134	3677839	3677773	**3646075**	3678080	3677839
# local misassemblies	**5**	**5**	6	**5**	6	6	**5**	**5**	**5**	**5**	**5**	**5**	**5**
# unaligned mis. contigs	0	0	0	0	0	0	0	0	0	0	0	0	0
# unaligned contigs	8 + 4 part	5 + 5 part	**5 + 4 part**	10 + 4 part	8 + 4 part	8 + 4 part	8 + 4 part	8 + 4 part	9 + 4 part	7 + 4 part	9 + 4 part	8 + 4 part	10 + 4 part
Unaligned length	65378	63645	**63216**	66496	64619	65150	64560	64446	66064	63864	65748	65335	66733
Genome fraction (%)	96.545	96.552	96.557	96.557	96.551	96.535	96.551	96.557	96.553	96.558	96.545	**96.560**	96.540
Duplication ratio	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
# N's per 100 kbp	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
# mismatches per 100 kbp	7.10	7.61	7.10	**6.91**	7.08	8.14	7.80	7.73	7.46	7.01	7.08	7.33	6.99
# indels per 100 kbp	0.64	0.64	0.61	**0.38**	0.61	0.68	0.74	0.64	0.66	0.64	0.61	0.42	0.61
Largest alignment	453219	453219	453219	453219	453219	453219	453219	453219	453219	453219	453219	453219	453219
Total aligned length	4721647	4721417	4721818	4721929	4721928	4720624	4721473	4721590	4721605	4721851	4721249	**4722158**	4721015
NA50	231409	**269985**	231409	231409	231409	231409	231409	231409	231409	231409	231409	231409	231409
NGA50	231409	231409	231409	231409	231409	231409	231409	231409	231409	231409	231409	231409	231409
NA75	140342	**161737**	140342	140342	140342	140342	140342	161730	140342	140342	140342	140342	140342
NGA75	133741	**161737**	133741	133741	133741	133930	133741	161730	133741	133741	133741	133741	133741
LA50	8	**7**	8	8	8	8	8	8	8	8	8	8	8
LGA50	8	8	8	8	8	8	8	8	8	8	8	8	8
LA75	14	14	14	14	14	14	14	14	14	14	14	14	14
LGA75	15	**14**	15	15	15	15	15	**14**	15	15	15	15	15

Table A.33 Assembly quality metrics for *P. aeruginosa*

Assembly	Uncorrected	ACE	BayesHammer	BFC	BLESS2	Blue	Fiona	Karect	Lighter	Musket	RACER	SGA-EC	Trowel
# contigs (≥ 0 bp)	106	124	110	122	115	137	117	119	113	122	134	107	113
# contigs (≥ 1000 bp)	51	54	51	53	55	52	52	52	53	53	54	51	51
# contigs (≥ 5000 bp)	40	45	42	42	44	44	43	43	44	44	45	41	41
# contigs (≥ 10000 bp)	34	38	35	36	38	37	36	36	37	37	38	35	35
# contigs (≥ 25000 bp)	33	37	34	35	37	35	35	35	36	36	35	34	34
# contigs (≥ 50000 bp)	26	29	27	27	29	27	27	27	28	28	27	27	27
Total length (≥ 0 bp)	6230021	6230354	6230652	6230542	6228474	**6232837**	6230803	6230974	6230430	6231373	6231836	6229898	6230416
Total length (≥ 1000 bp)	6216016	6215931	**6217165**	6214679	6214823	6214235	6215260	6215595	6215793	6215930	6215641	6215415	6215526
Total length (≥ 5000 bp)	6197160	6202121	**6203355**	6195823	6195967	6201464	6201450	6201785	6201983	6202120	6201831	6197595	6197706
Total length (≥ 10000 bp)	6160776	6160178	**6161412**	6159439	6159583	6159614	6159507	6159896	6160040	6160177	6159888	6161211	6161322
Total length (≥ 25000 bp)	6136845	6136247	**6137481**	6135508	6135652	6122306	6135576	6135965	6136109	6136246	6096890	6137280	6137391
Total length (≥ 50000 bp)	**5871704**	5831636	5862721	5831381	5831525	5818247	5831278	5831642	5831786	5831635	5796958	5862520	5862631
# contigs	59	60	58	61	62	61	59	60	61	60	62	59	59
Largest contig	582545	582545	582545	582545	582545	582232	582232	582747	582545	582545	**742310**	582545	582545
Total length	6221236	6220118	**6221612**	6219899	6219517	6220757	6220286	6220815	6221013	6220643	6220861	6220851	6220962
Reference length	6264404	6264404	6264404	6264404	6264404	6264404	6264404	6264404	6264404	6264404	6264404	6264404	6264404
GC (%)	66.58	66.58	66.58	66.58	66.58	66.58	66.58	66.58	66.58	66.58	66.58	66.58	66.58
Reference GC (%)	66.56	66.56	66.56	66.56	66.56	66.56	66.56	66.56	66.56	66.56	66.56	66.56	66.56
N50	**289353**	264881	**289353**	**289353**	264881	289314	**289353**	**289353**	264881	264881	264881	**289353**	**289353**
NG50	**289353**	264881	**289353**	**289353**	264881	289314	**289353**	**289353**	264881	264881	264881	**289353**	**289353**
N75	**210881**	165623	165623	165623	119649	165623	165623	165623	165623	165623	165623	165623	165623
NG75	**210881**	165623	165623	165623	119649	165623	165623	165623	165623	165623	165623	165623	165623
L50	**7**	8	**7**	**7**	8	**7**	**7**	**7**	8	8	**7**	**7**	**7**
LG50	**7**	8	**7**	**7**	8	**7**	**7**	**7**	8	8	**7**	**7**	**7**
L75	**14**	15	**14**	**14**	15	**14**	**14**	**14**	15	15	**14**	**14**	**14**
LG75	**14**	15	**14**	**14**	15	**14**	**14**	**14**	15	15	**14**	**14**	**14**
# misassemblies	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	4	**3**	**3**
# misassembled contigs	3	3	3	3	3	3	3	3	3	3	3	3	3
Misassembled contigs length	565341	565352	565341	565341	565459	**565086**	565352	565341	565341	565352	920163	565341	565341
# local misassemblies	5	7	6	**4**	7	6	**4**	5	**4**	**4**	6	7	9
# unaligned mis. contigs	0	0	0	0	0	0	0	0	0	0	0	0	0
# unaligned contigs	1 + 2 part	**0 + 2 part**	1 + 3 part	1 + 2 part	**0 + 2 part**	1 + 2 part	1 + 2 part	1 + 2 part	1 + 2 part	1 + 2 part	1 + 2 part	1 + 3 part	1 + 3 part
Unaligned length	11429	**10903**	12478	11429	**10903**	11429	11429	11429	11429	11429	11429	12478	12478
Genome fraction (%)	**99.088**	99.074	99.070	99.072	99.068	99.075	99.078	99.080	99.083	99.079	99.079	99.061	99.057
Duplication ratio	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	1.001
# N's per 100 kbp	18.39	18.55	17.95	17.80	18.28	19.74	17.80	18.71	18.55	18.55	17.79	**17.49**	17.96
# mismatches per 100 kbp	**2.84**	3.35	4.03	3.06	3.56	3.93	3.16	3.35	3.09	3.16	3.58	3.72	4.03
# indels per 100 kbp	**0.84**	0.85	0.93	0.87	0.87	1.02	0.85	**0.84**	0.87	0.85	0.90	0.92	0.98
Largest alignment	582545	582545	582545	582545	582545	582232	582232	**582618**	582545	582545	582545	582545	582545
Total aligned length	**6209300**	6208656	6208592	6208168	6208014	6209026	6208555	6208955	6209077	6208733	6208835	6207866	6208060
NA50	**289353**	264881	264881	264881	264881	264881	264881	264881	264881	264881	264881	264881	264881
NGA50	**289353**	264881	264881	264881	264881	264881	264881	264881	264881	264881	264881	264881	264881
NA75	**210881**	165623	165623	165623	119649	165623	165623	165623	165623	165623	165623	165623	165623
NGA75	**210881**	119649	165623	165623	119649	165623	165623	165623	165623	165623	165623	165623	165623
LA50	**7**	8	8	8	8	8	8	8	8	8	8	8	8
LGA50	**7**	8	8	8	8	8	8	8	8	8	8	8	8
LA75	**14**	15	**14**	**14**	15	**14**	**14**	**14**	15	15	15	**14**	**14**
LGA75	**14**	16	**14**	**14**	15	**14**	**14**	**14**	15	15	15	**14**	**14**

Figure A.1 SPAdes assembly results for P. aeruginosa for both uncorrected and corrected data. Scaffolds with length NGAx or larger produce x% of the genome.

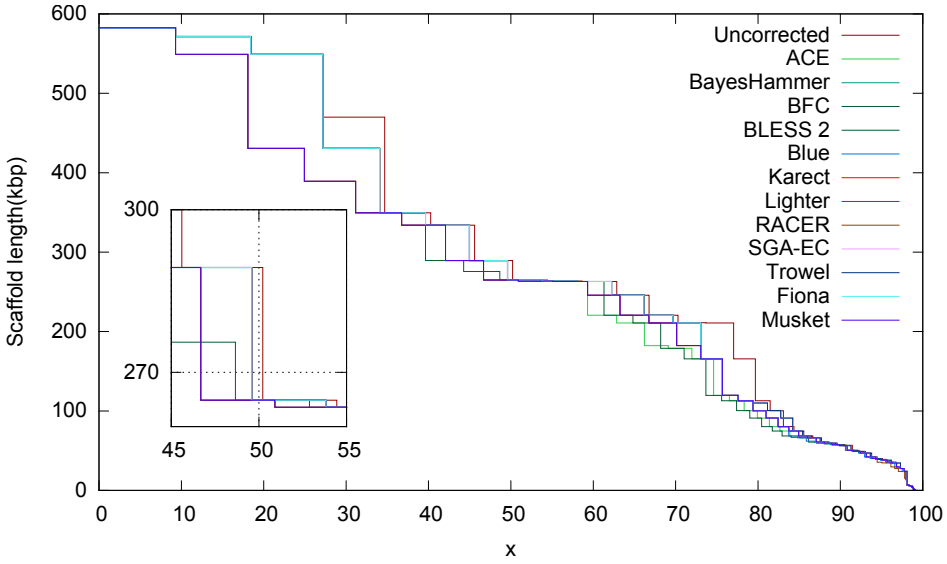


Table A.34 Assembly quality metrics for *H. sapiens* Chr. 21

Assembly	Uncorrected	ACE	BayesHammer	BFC	BLESS2	Blue	Fiona	Karect	Lighter	Musket	RACER	SGA-EC	Trowel
# contigs (≥ 0 bp)	11658	13066	10876	11247	13635	15666	11618	12646	12001	12350	13682	11803	13586
# contigs (≥ 1000 bp)	3661	3488	3376	3420	3925	3935	3368	3279	3425	3416	3879	3408	3969
# contigs (≥ 5000 bp)	1987	1961	1925	1911	2034	2085	1943	1897	1917	1929	2026	1919	2048
# contigs (≥ 10000 bp)	1145	1135	1119	1140	1098	1119	1142	1150	1137	1136	1090	1149	1103
# contigs (≥ 25000 bp)	246	250	279	282	216	210	271	295	279	274	215	270	214
# contigs (≥ 50000 bp)	24	29	27	27	12	20	32	32	23	36	17	30	17
Total length (≥ 0 bp)	34305614	34380522	34123100	34286645	34220998	**34449556**	34242236	34445755	34374980	34350021	34383404	34381972	34389916
Total length (≥ 1000 bp)	32945791	32944930	32860903	32985264	32700258	32769213	32955487	**33047117**	32997396	32974244	32838152	33005502	32846237
Total length (≥ 5000 bp)	28720164	29021061	29168621	29179374	27803086	27971993	29261315	**29545706**	29162965	29238235	27997267	29201336	27915791
Total length (≥ 10000 bp)	22672318	23065855	23340510	23624964	21088444	21094246	23475955	**24143915**	23553674	23512065	21260122	23687963	21057378
Total length (≥ 25000 bp)	8707938	9282473	10143177	10229508	7489603	7406913	10060749	**10705450**	10021930	10100507	7680611	9985753	7449787
Total length (≥ 50000 bp)	1423296	1891297	1724169	1679438	781510	1218193	2073928	2036281	1482391	**2208760**	1047851	1898371	1018260
# contigs	4303	4082	3948	4032	4584	4635	3949	3838	4038	3997	4551	4008	4702
Largest contig	74497	103923	82977	109258	82937	90065	**136866**	111589	90223	105061	85668	112665	85021
Total length	33409153	33370188	33274939	33424565	33177053	33276648	33371010	**33445850**	33437566	33390351	33323920	33437887	33379800
Reference length	46709983	46709983	46709983	46709983	46709983	46709983	46709983	46709983	46709983	46709983	46709983	46709983	46709983
GC (%)	40.79	40.82	40.76	40.80	40.75	40.79	40.81	40.82	40.81	40.81	40.79	40.81	40.78
Reference GC (%)	42.22	42.22	42.22	42.22	42.22	42.22	42.22	42.22	42.22	42.22	42.22	42.22	42.22
N50	15229	15787	16485	16484	13678	13214	15905	**17170**	16666	16419	13876	16495	13493
NG50	9391	9717	9988	10203	8262	8194	10128	**10819**	10189	10172	8448	10303	8291
N75	7932	8326	8555	8647	7180	7076	8718	**9161**	8521	8622	7219	8673	7193
L50	662	629	600	604	716	735	608	**588**	610	606	702	607	729
LG50	1216	1165	1121	1114	1347	1369	1130	**1074**	1118	1121	1318	1117	1355
L75	1412	1351	1293	1296	1546	1579	1309	**1249**	1302	1301	1529	1299	1573
# misassemblies	205	203	224	202	191	168	203	**119**	179	200	358	207	222
# misassembled contigs	196	188	207	186	183	159	179	**112**	161	184	331	196	208
Misassembled contigs length	3187244	3357505	3896326	3645633	2616542	2508072	3576383	**2217240**	3109131	3728907	5318815	3876193	3101965
# local misassemblies	148	128	153	134	**111**	121	143	134	133	146	134	141	134
# structural variations	1	0	1	1	0	1	1	0	1	1	1	1	1
# unaligned mis. contigs	2	**0**	1	1	**0**	1	**0**	1	**0**	**0**	**0**	2	**0**
# unaligned contigs	90 + 17 part	72 + 11 part	84 + 16 part	93 + 15 part	58 + 17 part	**50 + 13 part**	66 + 11 part	74 + 11 part	95 + 12 part	70 + 14 part	73 + 14 part	92 + 10 part	94 + 18 part
Unaligned length	107345	80003	115175	108173	76973	**67157**	79411	91569	108183	90024	98217	105816	111030
Genome fraction (%)	70.891	70.884	70.744	70.953	70.540	70.699	70.936	**71.080**	70.978	70.942	70.706	70.986	70.800
Duplication ratio	1.006	1.005	**1.003**	1.005	1.005	1.006	1.005	1.005	1.005	1.005	1.006	1.005	1.006
# N's per 100 kbp	12.56	13.76	11.38	12.28	28.84	14.08	**8.36**	15.66	14.59	13.19	10.92	12.46	9.72
# mismatches per 100 kbp	177.83	179.97	169.06	173.89	176.87	186.06	177.09	**168.19**	175.09	176.42	188.03	174.83	183.62
# indels per 100 kbp	37.28	37.81	37.16	**36.80**	37.01	37.70	38.14	36.96	37.49	37.43	38.00	37.48	37.17
Largest alignment	74497	103923	82445	90912	81005	90065	**136863**	111527	90220	103129	85647	90220	85007
Total aligned length	33233678	33218305	33132429	33254831	33047536	33140710	33236923	**33302749**	33270483	33245187	33141015	33274470	33199144
NA50	14406	15025	15545	15471	13158	12820	15154	**16481**	15781	15462	12579	15379	12815
NGA50	8829	9190	9443	9578	7802	7754	9491	**10302**	9618	9502	7603	9640	7863
NA75	7453	7800	8104	8018	6747	6733	8143	**8723**	8075	8070	6562	8040	6837
LA50	694	664	636	640	740	763	647	**611**	643	644	772	645	761
LGA50	1280	1231	1188	1184	1404	1428	1198	**1115**	1180	1191	1452	1188	1419
LA75	1490	1428	1371	1379	1614	1650	1388	**1297**	1375	1384	1684	1384	1648

Table A.35 Assembly quality metrics for *C. elegans*

Assembly	Uncorrected	ACE	BayesHammer	BFC	BLESS2	Blue	Fiona	Karect	Lighter	Musket	RACER	SGA-EC	Trowel
# contigs (≥ 0 bp)	64272	81082	62466	64735	87989	67408	70638	65792	67043	70931	78836	64022	65391
# contigs (≥ 1000 bp)	19715	26128	19230	19773	27860	19976	21281	19728	20822	22405	24680	19671	19746
# contigs (≥ 5000 bp)	6322	5237	6324	6345	4809	6255	6277	6345	6432	6157	5674	6320	6322
# contigs (≥ 10000 bp)	2427	1174	2456	2405	972	2373	2147	2428	2246	1919	1545	2418	2404
# contigs (≥ 25000 bp)	291	123	316	298	73	302	241	308	251	201	152	294	291
# contigs (≥ 50000 bp)	69	34	73	65	8	64	54	64	58	53	39	68	66
Total length (≥ 0 bp)	116531376	112665206	115842055	116541087	112459609	116159435	115776939	116265023	115948004	115599759	114392779	116406566	**116593128**
Total length (≥ 1000 bp)	**104888817**	96488385	104286506	104761637	93758025	104056252	102742327	104449183	103511220	102076721	98489212	104803286	104798953
Total length (≥ 5000 bp)	72682147	47647948	**73360701**	72602210	40794116	71152597	67007401	72377260	69150530	63681095	54401186	72740724	72574356
Total length (≥ 10000 bp)	45311588	20068019	**46103656**	44926759	14879662	43803433	38056990	44781786	39819043	34047844	26049004	45279671	45056640
Total length (≥ 25000 bp)	13999398	5661850	**14795023**	14054191	2637052	13691065	10590733	13714349	11128377	9601227	6786678	14041231	14181993
Total length (≥ 50000 bp)	6809697	2551613	**6959909**	6455562	459382	6022022	4471501	5853071	4791869	4687814	3002620	6744537	6897053
# contigs	26846	37484	26495	27029	40819	27393	29558	26932	28780	31318	35887	26806	26974
Largest contig	2404969	1283379	2403394	2440222	664004	2403394	1646766	2389911	2323777	2403394	1523322	2632544	**4023804**
Total length	**109983290**	104619147	109484481	109949860	103011932	109370494	108661357	109600335	109201095	108455777	106522260	109900955	109964382
Reference length	100286070	100286070	100286070	100286070	100286070	100286070	100286070	100286070	100286070	100286070	100286070	100286070	100286070
GC (%)	38.47	38.28	38.40	38.45	38.19	38.44	38.33	38.41	38.34	38.38	38.57	38.46	38.46
Reference GC (%)	35.44	35.44	35.44	35.44	35.44	35.44	35.44	35.44	35.44	35.44	35.44	35.44	35.44
N50	7948	4512	**8157**	7871	3931	7708	6873	7923	7154	6334	5139	7959	7913
NG50	8930	4738	**9140**	8876	4045	8667	7551	8861	7925	6946	5518	8926	8888
N75	3654	2270	**3736**	3636	2011	3538	3208	3633	3366	2957	2436	3661	3629
NG75	4633	2514	**4710**	4615	2142	4421	3949	4573	4191	3597	2846	4631	4606
L50	3515	6219	**3411**	3538	7235	3611	4119	3554	4000	4475	5450	3505	3522
LG50	2939	5750	**2879**	2960	6894	3055	3538	2998	3408	3859	4864	2934	2945
L75	8614	14423	**8346**	8658	16402	8846	9889	8652	9540	10757	12969	8585	8648
LG75	6849	13062	**6706**	6890	15417	7121	8124	6940	7759	8881	11193	6835	6874
# misassemblies	1377	4278	1460	**1332**	1654	1603	2032	1335	2025	2103	8293	1392	1377
# misassembled contigs	1296	3858	1371	**1254**	1593	1510	1890	1257	1874	1963	6560	1310	1297
Misassembled contigs length	9735210	17244516	10432238	9391106	**5928792**	10781051	12262942	9392527	12972582	11134702	31623064	9855510	9569483
# local misassemblies	399	326	433	392	**309**	412	473	414	394	425	346	409	401
# unaligned mis. contigs	**7**	21	9	8	13	9	9	**7**	9	12	20	**7**	10
# unaligned contigs	4743 + 91 part	**4049 + 169 part**	4899 + 88 part	4789 + 86 part	4465 + 88 part	4756 + 97 part	4798 + 108 part	4807 + 79 part	4904 + 104 part	4886 + 98 part	4976 + 177 part	4770 + 92 part	4739 + 81 part
Unaligned length	16953817	14482847	16700997	16905208	**14030377**	16726451	16200042	16687086	16399595	16371253	16133966	16937864	16931221
Genome fraction (%)	92.273	87.960	92.152	**92.298**	88.334	91.876	91.576	92.187	91.949	91.161	88.855	92.212	92.276
Duplication ratio	1.005	1.022	**1.004**	1.005	**1.004**	1.005	1.007	1.005	1.006	1.007	1.014	1.005	1.005
# N's per 100 kbp	144.78	**112.56**	148.00	144.96	115.15	148.93	144.91	144.51	144.71	143.52	136.35	146.95	146.14
# mismatches per 100 kbp	33.27	117.20	**30.55**	33.15	37.24	39.77	87.49	33.57	80.67	177.01	155.25	33.51	33.23
# indels per 100 kbp	7.22	19.86	7.82	**7.17**	8.24	8.15	22.19	7.31	8.01	10.97	21.52	7.32	7.20
Largest alignment	508136	27416	**64368**	508136	27155	508136	508137	64367	508136	41089	40668	508136	508136
Total aligned length	92814637	88413367	92662358	**92838725**	88802236	92418941	92206639	92718109	92538398	91814436	89541690	92751467	92825410
NA50	5625	2979	**5795**	5605	2928	5432	4799	5623	5037	4404	2979	5640	5593
NGA50	6472	3158	**6576**	6437	3036	6183	5385	6431	5654	4952	3266	6483	6423
NA75	1878	1124	**1939**	1874	1210	1815	1658	1901	1721	1526	1037	1880	1872
NGA75	2805	1348	**2849**	2788	1338	2644	2341	2782	2488	2143	1347	2799	2787
LA50	5079	9487	**4925**	5093	9672	5224	5940	5055	5767	6448	9342	5061	5102
LGA50	4274	8780	**4181**	4288	9214	4438	5117	4280	4931	5572	8343	4265	4293
LA75	13282	23377	**12846**	13314	23124	13702	15293	13194	14769	16627	24133	13249	13343
LGA75	10140	20740	**9931**	10174	21516	10617	12131	10183	11563	13253	20188	10133	10193

Table A.36 Assembly quality metrics for *D. melanogaster*

Assembly	Uncorrected	ACE	BayesHammer	BFC	BLESS2	Blue	Fiona	Karect	Lighter	Musket	RACER	SGA-EC	Trowel
# contigs (≥ 0 bp)	74043	78599	69271	73234	76430	74610	80861	74173	71364	80075	86186	70023	72210
# contigs (≥ 1000 bp)	5483	6789	5454	5543	6771	5503	5900	5431	5523	6199	9213	5480	5650
# contigs (≥ 5000 bp)	2533	3751	2550	2593	3730	2558	2731	2456	2640	2934	4623	2530	2643
# contigs (≥ 10000 bp)	1933	2754	1922	1951	2750	1915	2084	1844	1992	2212	3195	1907	2007
# contigs (≥ 25000 bp)	1206	1442	1193	1211	1402	1197	1271	1177	1229	1304	1412	1182	1235
# contigs (≥ 50000 bp)	704	626	701	691	615	679	711	700	711	700	498	687	692
Total length (≥ 0 bp)	130030442	128478789	129365298	129911559	127133184	129521097	**130605696**	129982716	129327660	130215917	130591397	129618485	129655323
Total length (≥ 1000 bp)	120972662	119245763	120687956	120958052	118182463	120423367	120885087	**121024055**	120546143	120456395	120306993	120966385	120866556
Total length (≥ 5000 bp)	**114719681**	112180360	114535094	114657226	111207133	114260288	114138370	114718069	114368253	113524107	109812670	114690805	114379771
Total length (≥ 10000 bp)	**110478651**	105048796	110093513	110109175	104211872	109698057	109584716	110408099	109785315	108438548	99585998	110279095	109845156
Total length (≥ 25000 bp)	98656926	83306275	98285329	97985919	82067268	97987147	96415034	**99512924**	97286836	93500096	70429725	98360112	97108585
Total length (≥ 50000 bp)	80423462	54558355	80415030	79115237	54255186	78950991	75950200	**81985969**	78612470	71818362	38126731	80250129	77361963
# contigs	8454	9707	8451	8575	9382	8580	8924	8361	8514	9443	12816	8461	8613
Largest contig	518444	344897	522527	518411	440228	482029	433735	518285	592584	367481	243181	598856	**686399**
Total length	123060751	121275348	122787001	123078343	119991674	122588231	123008213	**123082336**	122638571	122729429	122836325	123059826	122956447
Reference length	120381546	120381546	120381546	120381546	120381546	120381546	120381546	120381546	120381546	120381546	120381546	120381546	120381546
GC (%)	42.56	42.52	42.53	42.57	42.51	42.58	42.58	42.57	42.55	42.59	42.58	42.57	42.57
Reference GC (%)	42.42	42.42	42.42	42.42	42.42	42.42	42.42	42.42	42.42	42.42	42.42	42.42	42.42
N50	83804	43877	84315	82203	43315	83277	71630	**88533**	79363	63519	31144	84694	77879
NG50	86505	44167	85854	84499	43084	84749	73544	**91730**	81203	65332	31761	87595	80342
N75	34576	19862	34016	32845	19401	34311	30651	**36690**	32157	26676	13867	34383	31539
NG75	37360	20342	36806	35722	19244	36580	33263	**39101**	34418	28621	14802	37283	34705
L50	411	756	408	413	739	404	470	**393**	434	516	1091	397	436
LG50	395	746	394	396	743	391	452	**378**	420	497	1052	381	419
L75	990	1785	982	1015	1761	990	1122	**941**	1043	1248	2562	975	1061
LG75	934	1752	932	956	1776	944	1060	**888**	991	1185	2433	919	1003
# misassemblies	1100	1287	1133	1114	1304	1201	1152	**1098**	1152	1225	2236	1111	1163
# misassembled contigs	**846**	1063	861	856	1079	915	907	850	886	948	1751	847	889
Misassembled contigs length	53566639	40875670	55491362	53723282	**40614325**	57041407	51435560	54512203	55267465	50221260	49962445	55594508	55519157
# local misassemblies	1959	1875	**1417**	2007	2186	1985	1986	1846	1935	1953	1863	1903	1951
# unaligned mis. contigs	97	102	106	106	98	99	114	102	**87**	110	107	97	110
# unaligned contigs	3523 + 877 part	2780 + 902 part	3468 + 863 part	3507 + 899 part	**2298 + 788 part**	3556 + 838 part	3570 + 887 part	3507 + 866 part	3392 + 892 part	3689 + 885 part	3882 + 966 part	3518 + 889 part	3534 + 887 part
Unaligned length	8655611	7015647	8461083	8663161	**5974560**	8325736	8619025	8706997	8247034	8400074	8510536	8653846	8602555
Genome fraction (%)	94.389	94.126	**94.395**	94.385	94.007	94.283	94.339	94.381	94.364	94.288	94.094	94.390	94.340
Duplication ratio	1.007	1.008	**1.006**	1.007	1.008	1.007	1.007	1.007	1.007	1.007	1.009	1.007	1.007
# N's per 100 kbp	20.77	23.70	20.46	20.90	80.37	**19.31**	24.51	21.24	19.63	25.30	21.68	19.56	20.58
# mismatches per 100 kbp	550.95	566.10	550.06	549.94	562.32	552.77	553.75	**547.82**	551.94	574.04	572.70	549.10	548.28
# indels per 100 kbp	131.06	137.55	132.17	130.71	134.83	131.25	131.77	130.85	131.13	132.41	132.97	130.86	**130.28**
Largest alignment	446343	201781	357935	428474	285019	**446353**	406631	446139	375228	269466	179030	408292	408315
Total aligned length	113947282	113635974	113933062	113945277	113517804	113788556	113908464	113929648	113921610	113834640	113679126	**113951223**	113890263
NA50	59109	35127	57232	57455	35199	57276	53453	**60755**	54973	47234	23144	59109	54194
NGA50	60554	35392	58570	59441	34884	59109	54188	**62304**	56348	48694	23783	60636	55841
NA75	24325	15983	24327	23793	16032	23582	22424	**26073**	23625	20198	9811	24613	22828
NGA75	26503	16373	26218	25875	15841	25252	24142	**28090**	25252	21704	10521	26802	24540
LA50	570	945	582	575	930	576	640	**548**	606	700	1433	564	608
LGA50	548	932	561	552	936	557	616	**526**	586	676	1381	542	585
LA75	1374	2222	1393	1400	2202	1399	1517	**1307**	1449	1682	3435	1360	1474
LGA75	1295	2181	1321	1318	2221	1331	1432	**1232**	1379	1598	3254	1282	1392

A.4.4 Velvet

A.4.4.1 *B. dentium*

Table A.37 contains the Quast report after assembling dataset *B. dentium* with Velvet.

A.4.4.2 *E. coli str. K-12 substr. DH10B*

Table A.38 contains the Quast report after assembling dataset *E. coli str. K-12 substr. DH10B* with Velvet.

A.4.4.3 *E. coli str. K-12 substr. MG1655*

Table A.39 contains the Quast report after assembling dataset *E. coli str. K-12 substr. MG1655* with Velvet.

A.4.4.4 *S. enterica*

Table A.40 contains the Quast report after assembling dataset *S. enterica* with Velvet.

A.4.4.5 *P. aeruginosa*

Table A.41 contains the Quast report after assembling dataset *P. aeruginosa* with Velvet.

A.4.4.6 *H. sapiens* Chr. 21

Table A.42 contains the Quast report after assembling dataset *H. sapiens* Chr. 21 with Velvet.

A.4.4.7 *C. elegans*

Table A.43 contains the Quast report after assembling dataset *C. elegans* with Velvet.

A.4.4.8 *D. melanogaster*

Table A.44 contains the Quast report after assembling dataset *D. melanogaster* with Velvet.

Table A.37 Assembly quality metrics for *B. dentium*

Assembly	Uncorrected	ACE	BayesHammer	BFC	BLESS2	Blue	Fiona	Karect	Lighter	Musket	RACER	SGA-EC	Trowel
# contigs (≥ 0 bp)	164979	73360	59552	70557	44383	36024	76459	30869	76687	68649	49049	135918	108793
# contigs (≥ 1000 bp)	5	1	556	464	792	922	193	926	242	314	916	38	67
# contigs (≥ 5000 bp)	0	0	0	0	5	27	0	27	0	0	21	0	0
# contigs (≥ 10000 bp)	0	0	0	0	0	0	0	1	0	0	1	0	0
# contigs (≥ 25000 bp)	0	0	0	0	0	0	0	0	0	0	0	0	0
# contigs (≥ 50000 bp)	0	0	0	0	0	0	0	0	0	0	0	0	0
Total length (≥ 0 bp)	**13904588**	5670761	8730630	9747006	7400138	7597563	9809249	7139160	10007466	9327465	8440471	12986853	11430326
Total length (≥ 1000 bp)	5777	1010	807224	666417	1372997	1876244	254328	**1972180**	314586	422998	1835347	43440	84111
Total length (≥ 5000 bp)	0	0	0	0	30442	167860	0	**179556**	0	0	129097	0	0
Total length (≥ 10000 bp)	0	0	0	0	0	0	0	10815	0	0	**12218**	0	0
Total length (≥ 25000 bp)	0	0	0	0	0	0	0	0	0	0	0	0	0
Total length (≥ 50000 bp)	0	0	0	0	0	0	0	0	0	0	0	0	0
# contigs	508	39	2161	2003	1767	1982	1581	1930	1713	1894	1944	925	1210
Largest contig	1429	1010	3784	4501	7416	9719	3745	10815	2583	3552	**12218**	1758	3169
Total length	303382	22970	1878464	1687429	2058670	2584183	1153196	**2634272**	1284413	1470681	2521612	591102	805395
Reference length	2636367	2636367	2636367	2636367	2636367	2636367	2636367	2636367	2636367	2636367	2636367	2636367	2636367
GC (%)	59.89	49.27	58.96	58.91	58.56	58.75	58.62	58.81	58.80	58.86	58.70	58.75	58.91
Reference GC (%)	58.54	58.54	58.54	58.54	58.54	58.54	58.54	58.54	58.54	58.54	58.54	58.54	58.54
N50	576	552	908	860	1330	1660	701	**1834**	733	768	1638	603	634
NG50	-	-	676	598	1054	1616	-	1832	-	541	1576	-	-
N75	540	519	630	617	846	930	580	**993**	585	594	945	551	557
NG75	-	-	-	-	553	891	-	992	-	-	851	-	-
L50	227	**18**	695	655	496	468	588	441	625	672	468	387	487
LG50	-	-	1184	1329	739	484	-	442	-	1601	504	-	-
L75	363	**28**	1323	1243	978	987	1045	930	1119	1221	974	644	829
LG75	-	-	-	-	1612	1030	-	932	-	-	1070	-	-
# misassemblies	**0**	**0**	**0**	**0**	1	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
# misassembled contigs	**0**	**0**	**0**	**0**	1	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Misassembled contigs length	**0**	**0**	**0**	**0**	670	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
# local misassemblies	**0**	**0**	1	1	1	1	**0**	1	**0**	1	2	**0**	**0**
# unaligned mis. contigs	0	0	0	0	0	0	0	0	0	0	0	0	0
# unaligned contigs	407 + 0 part	**26 + 0 part**	495 + 0 part	444 + 0 part	73 + 0 part	442 + 0 part	393 + 0 part	478 + 0 part	412 + 0 part	437 + 0 part	433 + 0 part	429 + 0 part	432 + 0 part
Unaligned length	241295	**14816**	296562	263972	42456	264277	233261	286542	244102	259091	256233	254867	257078
Genome fraction (%)	2.354	0.309	59.850	53.862	76.212	87.487	34.830	**88.672**	39.390	45.889	85.597	12.748	20.785
Duplication ratio	**1.000**	**1.000**	1.003	1.002	1.003	1.006	1.002	1.004	1.002	1.001	1.004	**1.000**	1.001
# N's per 100 kbp	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	0.38	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
# mismatches per 100 kbp	3.22	**0.00**	1.14	0.49	0.40	0.39	**0.00**	1.41	0.67	0.41	0.44	0.30	2.55
# indels per 100 kbp	**0.00**	**0.00**	0.32	0.28	0.20	0.22	0.11	0.21	0.19	0.25	0.27	0.30	0.55
Largest alignment	1249	1010	3784	4501	7416	9719	3745	10815	2583	3552	**12218**	1442	3169
Total aligned length	62087	8154	1581896	1423454	2016200	2319904	919929	**2347650**	1040300	1211588	2265100	336235	548314
NA50	-	-	895	845	1330	1656	678	1834	715	752	1637	528	579
NGA50	-	-	638	555	1054	1611	-	1832	-	-	1575	-	-
NA75	-	-	591	579	846	918	522	985	534	549	939	-	-
NGA75	-	-	-	-	532	882	-	985	-	-	838	-	-
LA50	-	-	698	658	496	468	598	441	631	677	468	417	508
LGA50	-	-	1201	1360	739	484	-	442	-	-	504	-	-
LA75	-	-	1349	1267	978	989	1081	932	1152	1251	975	-	-
LGA75	-	-	-	-	1619	1032	-	933	-	-	1072	-	-

Table A.38 Assembly quality metrics for *E. coli str. K-12 substr. DH10B*

Assembly	ACE	BayesHammer	BFC	BLESS2	Blue	Fiona	Karect	Lighter	Musket	RACER	SGA-EC	Trowel
# contigs (≥ 0 bp)	509	491	465	468	451	472	458	502	500	448	174910	154264
# contigs (≥ 1000 bp)	90	87	85	84	81	85	80	89	85	82	0	0
# contigs (≥ 5000 bp)	73	72	69	68	69	69	68	73	72	65	0	0
# contigs (≥ 10000 bp)	66	65	63	62	63	63	63	68	65	62	0	0
# contigs (≥ 25000 bp)	53	54	54	52	51	52	54	56	53	52	0	0
# contigs (≥ 50000 bp)	31	29	30	33	31	31	32	30	28	32	0	0
Total length (≥ 0 bp)	4362719	4361014	4362674	4364353	4362388	4361867	4365893	4361911	4360726	4365116	**12078319**	11049421
Total length (≥ 1000 bp)	4324975	4325779	4329137	4330284	4329569	4327404	4332565	4325368	4324544	**4333539**	0	0
Total length (≥ 5000 bp)	4286089	4294343	4295276	4294887	4306420	4289532	**4307659**	4286673	4295327	4293367	0	0
Total length (≥ 10000 bp)	4232622	4236996	4248144	4248584	4259350	4240854	4268850	4246908	4240129	**4269045**	0	0
Total length (≥ 25000 bp)	4018923	4072492	4119676	4097964	4078566	4068100	4120739	4056582	4050242	**4120791**	0	0
Total length (≥ 50000 bp)	3239468	3169522	3232522	**3417084**	3366933	3325363	3348823	3104697	3150091	3404809	0	0
# contigs	97	92	89	89	85	92	85	97	93	87	13	23
Largest contig	299060	269346	269332	325957	326035	**326112**	269392	269442	269320	269372	709	772
Total length	4330157	4329486	4331937	4333660	4332755	4332343	4336380	4331396	4330245	**4337347**	7189	13177
Reference length	4686137	4686137	4686137	4686137	4686137	4686137	4686137	4686137	4686137	4686137	4686137	4686137
GC (%)	50.76	50.75	50.76	50.76	50.75	50.75	50.75	50.75	50.74	50.75	47.38	48.72
Reference GC (%)	50.78	50.78	50.78	50.78	50.78	50.78	50.78	50.78	50.78	50.78	50.78	50.78
N50	82393	86789	86856	85634	86758	86869	85563	82246	88175	**94801**	543	572
NG50	80642	82312	82298	80585	82465	82297	82028	77343	82274	**86919**	-	-
N75	48577	48022	49515	53169	**54670**	53109	53162	47476	45352	54645	523	527
NG75	41209	41209	41209	**48720**	42578	41209	43173	40034	40673	43982	-	-
L50	16	15	15	14	14	14	15	16	14	14	**7**	11
LG50	18	17	17	**16**	**16**	**16**	17	18	17	**16**	-	-
L75	32	31	31	30	29	30	31	33	31	30	**10**	17
LG75	38	37	37	**35**	**35**	36	36	40	37	**35**	-	-
# misassemblies	7	10	2	5	9	4	8	10	7	13	**0**	**0**
# misassembled contigs	5	6	2	5	7	4	6	8	6	9	**0**	**0**
Misassembled contigs length	477123	372804	92189	353248	456378	261461	439383	459105	481833	778996	**0**	**0**
# local misassemblies	64	39	34	30	46	38	31	49	50	42	**0**	**0**
# structural variations	0	0	0	0	0	0	0	0	0	0	0	0
# unaligned mis. contigs	0	0	0	0	0	0	0	0	0	0	0	0
# unaligned contigs	0 + 1 part	**0 + 0 part**	**0 + 0 part**	**0 + 0 part**	0 + 1 part	0 + 1 part	**0 + 0 part**	**0 + 0 part**	**0 + 0 part**	**0 + 0 part**	**0 + 0 part**	**0 + 0 part**
Unaligned length	1677	**0**	**0**	**0**	617	1037	**0**	**0**	**0**	**0**	**0**	**0**
Genome fraction (%)	92.269	92.246	92.323	92.398	92.347	92.360	**92.455**	92.281	92.281	92.420	0.153	0.281
Duplication ratio	1.001	1.002	1.001	1.001	1.001	1.001	1.001	1.002	1.001	1.001	**1.000**	**1.000**
# N's per 100 kbp	113.44	186.37	125.44	105.66	75.17	125.31	85.79	172.60	181.54	132.92	**0.00**	**0.00**
# mismatches per 100 kbp	23.59	46.89	22.40	19.86	21.77	32.12	32.20	33.69	34.20	30.25	**0.00**	**0.00**
# indels per 100 kbp	14.18	17.91	10.72	7.60	9.75	11.51	5.72	19.03	15.98	14.52	**0.00**	**0.00**
Largest alignment	298898	269196	269332	325947	325917	**326112**	269392	269344	269225	269372	709	772
Total aligned length	4323443	4322221	4326571	4329544	4327068	4328251	**4333032**	4323907	4323828	4331580	7189	13177
NA50	82293	85134	86856	82180	85556	85328	85563	82137	**86912**	82221	543	572
NGA50	75911	80601	**82198**	80193	80440	80741	80714	76130	82166	77233	-	-
NA75	44020	43628	48685	**48720**	47405	**48720**	48718	42528	42953	47742	523	527
NGA75	35956	38583	40090	40155	40137	**40443**	40154	33756	38652	39931	-	-
LA50	16	15	15	15	15	15	15	16	15	16	**7**	11
LGA50	19	**17**	**17**	**17**	**17**	**17**	**17**	19	**17**	18	-	-
LA75	34	33	31	31	31	31	32	34	32	33	**10**	17
LGA75	41	39	**37**	**37**	**37**	**37**	38	41	38	39	-	-

Table A.39 Assembly quality metrics for *E. coli str. K-12 substr. MG1655*

Assembly	Uncorrected	ACE	BayesHammer	BFC	BLESS2	Blue	Fiona	Karect	Lighter	Musket	RACER	SGA-EC	Trowel
# contigs (≥ 0 bp)	339266	626	47416	12544	556	572	605	584	621	637	566	254087	42939
# contigs (≥ 1000 bp)	0	108	0	1543	110	117	112	106	115	111	116	1	300
# contigs (≥ 5000 bp)	0	88	0	11	86	92	90	85	93	89	90	0	1
# contigs (≥ 10000 bp)	0	79	0	0	76	81	80	77	84	78	80	0	0
# contigs (≥ 25000 bp)	0	59	0	0	57	58	58	58	60	59	57	0	0
# contigs (≥ 50000 bp)	0	32	0	0	35	33	33	33	33	33	34	0	0
Total length (≥ 0 bp)	**17463920**	4568487	2369763	4817502	4566297	4565654	4564079	4562902	4566027	4566259	4567503	14472372	5409096
Total length (≥ 1000 bp)	0	4519097	0	2706314	4523572	4522624	4517645	4518802	4519150	4518503	**4524463**	1446	408072
Total length (≥ 5000 bp)	0	**4475899**	0	63787	4472985	4469468	4471178	4475747	4474711	4473760	4467668	0	5985
Total length (≥ 10000 bp)	0	4409903	0	0	4396521	4383548	4392999	**4415938**	4404793	4388487	4391238	0	0
Total length (≥ 25000 bp)	0	4089760	0	0	4089935	4022552	4051345	**4122523**	4002782	4084172	4024529	0	0
Total length (≥ 50000 bp)	0	3112482	0	0	**3283699**	3114815	3112477	3205490	3055689	3171801	3206713	0	0
# contigs	3	121	3	3067	121	128	124	118	125	124	126	67	1980
Largest contig	867	181755	840	7490	174070	179238	**232264**	208709	223529	194669	174310	1446	5985
Total length	2185	4528167	2268	3794237	4531713	4530848	4526320	4527131	4526451	4527527	**4531785**	41034	1530231
Reference length	4639675	4639675	4639675	4639675	4639675	4639675	4639675	4639675	4639675	4639675	4639675	4639675	4639675
GC (%)	44.94	50.73	45.02	50.40	50.72	50.73	50.72	50.73	50.72	50.72	50.73	38.88	47.47
Reference GC (%)	50.79	50.79	50.79	50.79	50.79	50.79	50.79	50.79	50.79	50.79	50.79	50.79	50.79
N50	808	**86050**	808	1418	80449	77816	68741	82565	67426	83207	82551	581	756
NG50	-	82691	-	1187	80449	77816	68741	82565	66935	80447	80476	-	-
N75	808	42208	620	933	**45313**	41989	42494	43670	41230	42252	42938	530	610
NG75	-	41209	-	634	42451	41057	41773	42284	39834	41187	42459	-	-
L50	**2**	20	**2**	863	20	20	20	19	20	19	20	29	716
LG50	-	21	-	1188	20	20	20	19	21	20	21	-	-
L75	**2**	39	3	1688	38	40	40	38	41	39	39	47	1283
LG75	-	41	-	2513	40	42	42	40	43	41	40	-	-
# misassemblies	-	21	-	0	5	13	23	3	13	10	9	0	0
# misassembled contigs	-	12	-	0	5	8	13	2	9	9	6	0	0
Misassembled contigs length	-	1157326	-	0	470209	704224	1125585	390367	938837	851796	543312	0	0
# local misassemblies	-	34	-	0	14	16	15	13	18	9	28	0	0
# unaligned mis. contigs	-	0	-	0	1	0	0	0	1	0	0	0	0
# unaligned contigs	3 + 0 part	0 + 1 part	3 + 0 part	4 + 0 part	**0 + 0 part**	**0 + 0 part**	**0 + 0 part**	**0 + 0 part**	0 + 1 part	**0 + 0 part**	**0 + 0 part**	4 + 0 part	4 + 0 part
Unaligned length	2185	966	2268	2876	**0**	**0**	**0**	**0**	580	**0**	**0**	2849	2885
Genome fraction (%)	-	97.314	-	81.498	97.505	97.416	97.341	97.477	97.294	97.360	97.427	0.823	32.888
Duplication ratio	-	1.003	-	1.003	1.002	1.002	1.002	1.001	1.003	1.002	1.003	1.000	1.001
# N's per 100 kbp	**0.00**	209.33	**0.00**	**0.00**	242.40	249.20	248.83	121.36	311.04	324.42	251.69	**0.00**	**0.00**
# mismatches per 100 kbp	-	28.31	-	13.38	17.00	14.25	27.50	30.23	30.46	38.63	29.07	0.00	16.38
# indels per 100 kbp	-	13.82	-	2.49	9.81	7.90	13.48	7.83	18.59	21.58	14.16	0.00	1.77
Largest alignment	-	165214	-	7490	174070	164078	173882	165316	163894	173710	173960	1446	5985
Total aligned length	-	4516869	-	3790989	4525352	4521152	4517437	4522868	4514747	4519081	4521328	38185	1527204
NA50	-	63339	-	1418	67256	67120	61390	80344	60834	60803	67089	568	755
NGA50	-	63339	-	1187	66500	63276	60805	80344	59372	60803	65967	-	-
NA75	-	39943	-	933	42223	39213	39194	43540	35870	41187	41209	513	609
NGA75	-	35877	-	633	41781	35852	38151	42197	33397	39885	39168	-	-
LA50	-	24	-	863	21	22	25	20	24	22	22	30	716
LGA50	-	24	-	1189	22	23	26	20	25	22	23	-	-
LA75	-	47	-	1688	42	45	48	39	49	44	43	49	1284
LGA75	-	50	-	2514	44	47	50	41	51	46	45	-	-

Table A.40 Assembly quality metrics for *S. enterica*

Assembly	Uncorrected	ACE	BayesHammer	BFC	BLESS2	Blue	Fiona	Karect	Lighter	Musket	RACER	SGA-EC	Trowel
# contigs (≥ 0 bp)	236914	606	610	702	553	516	784	494	928	679	490	185562	137268
# contigs (≥ 1000 bp)	0	123	125	219	119	89	265	74	337	188	74	0	2
# contigs (≥ 5000 bp)	0	88	80	131	76	62	159	57	170	116	53	0	0
# contigs (≥ 10000 bp)	0	74	69	89	64	54	103	49	110	83	45	0	0
# contigs (≥ 25000 bp)	0	50	49	37	46	41	38	42	35	45	37	0	0
# contigs (≥ 50000 bp)	0	32	31	19	25	30	21	31	19	28	25	0	0
Total length (≥ 0 bp)	**16980980**	4778207	4780385	4785569	4782707	4785804	4780122	4782938	4783429	4783406	4785994	14944644	12032854
Total length (≥ 1000 bp)	0	4725851	4728507	4722859	4734464	4738927	4715446	4740353	4696102	4726461	**4740565**	0	2320
Total length (≥ 5000 bp)	0	4640618	4619869	4486056	4627089	4672336	4442614	**4703375**	4251262	4534404	4691878	0	0
Total length (≥ 10000 bp)	0	4536709	4542588	4173914	4535891	4615276	4042426	**4642891**	3826601	4294979	4634956	0	0
Total length (≥ 25000 bp)	0	4128922	4247177	3356126	4269112	4428501	3068026	**4533152**	2717304	3707912	4513196	0	0
Total length (≥ 50000 bp)	0	3506624	3601272	2725881	3527369	4046246	2475977	**4169579**	2145483	3149161	4084512	0	0
# contigs	9	136	139	246	133	103	290	86	382	207	86	53	208
Largest contig	684	376781	383910	329670	330926	465834	297231	395892	251830	297189	**590715**	977	1164
Total length	4938	4734253	4737412	4741900	4743751	4748024	4733128	4747774	4729466	4739328	**4748477**	31165	125991
Reference length	4888768	4888768	4888768	4888768	4888768	4888768	4888768	4888768	4888768	4888768	4888768	4888768	4888768
GC (%)	51.44	52.15	52.14	52.13	52.13	52.13	52.13	52.13	52.13	52.13	52.12	48.67	50.77
Reference GC (%)	52.09	52.09	52.09	52.09	52.09	52.09	52.09	52.09	52.09	52.09	52.09	52.09	52.09
N50	543	77913	113532	79951	121679	148114	64795	159595	38266	79565	**170051**	571	597
NG50	-	75950	113466	74996	115686	116587	63079	159595	34856	77960	170051	-	-
N75	503	49652	50887	19679	49854	**72911**	15489	72527	12322	28154	72692	533	538
NG75	-	41430	46030	17932	46529	72565	13399	71471	10899	26138	72527	-	-
L50	**5**	16	14	14	12	10	20	10	25	16	8	24	91
LG50	-	17	15	15	13	11	21	10	27	17	8	-	-
L75	**7**	33	31	46	26	23	64	22	85	40	18	38	148
LG75	-	36	33	52	28	25	72	23	95	44	19	-	-
# misassemblies	**0**	17	14	16	20	18	12	19	12	15	21	**0**	**0**
# misassembled contigs	**0**	14	12	13	13	11	10	13	12	13	11	**0**	**0**
Misassembled contigs length	**0**	1661007	1372517	1655245	1842140	2131028	1196243	2286736	975040	1582981	2523514	**0**	**0**
# local misassemblies	**0**	76	75	51	61	53	47	39	32	51	59	**0**	**0**
# unaligned mis. contigs	**0**	**0**	1	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
# unaligned contigs	2 + 0 part	5 + 6 part	6 + 7 part	7 + 2 part	5 + 5 part	5 + 5 part	8 + 3 part	5 + 6 part	10 + 4 part	5 + 7 part	4 + 5 part	**0 + 0 part**	2 + 0 part
Unaligned length	1004	59327	61619	60007	59656	60395	60053	59745	58892	59686	61859	**0**	1049
Genome fraction (%)	0.080	95.526	95.561	95.648	95.687	95.799	95.483	**95.843**	95.387	95.622	95.803	0.637	2.553
Duplication ratio	**1.000**	1.001	1.001	1.001	1.001	1.001	1.001	1.001	1.002	1.001	1.001	**1.000**	1.001
# N's per 100 kbp	**0.00**	99.57	110.04	72.86	116.15	59.86	82.42	76.23	53.28	95.82	76.49	**0.00**	**0.00**
# mismatches per 100 kbp	177.94	19.79	19.16	14.20	18.98	19.30	16.82	21.45	14.60	17.11	18.58	22.46	**10.41**
# indels per 100 kbp	**0.00**	5.16	4.99	2.95	3.42	4.44	4.48	4.74	3.07	5.01	4.29	**0.00**	**0.00**
Largest alignment	684	183671	218248	215441	233900	219065	215581	233926	215296	215652	**377102**	977	1164
Total aligned length	3934	4670996	4671740	4678407	4679569	4685321	4669888	**4685535**	4667366	4675991	4684623	31165	124824
NA50	543	74294	84701	57033	106301	100709	44474	113291	35004	69369	**113670**	571	596
NGA50	-	73998	82002	52943	106301	100709	42286	113291	32911	64433	106015	-	-
NA75	502	40494	44347	18740	43571	**67823**	15059	62786	11835	27284	67822	533	536
NGA75	-	37947	40502	17163	40561	60633	13120	60793	10626	24011	62447	-	-
LA50	**5**	20	17	19	16	17	25	15	29	21	12	24	91
LGA50	-	21	18	21	16	17	27	15	31	22	13	-	-
LA75	**7**	40	36	56	34	31	73	29	92	48	26	38	148
LGA75	-	43	39	63	36	33	82	31	103	53	28	-	-

Table A.41 Assembly quality metrics for *P. aeruginosa*

Assembly	Uncorrected	ACE	BayesHammer	BFC	BLESS2	Blue	Fiona	Karect	Lighter	Musket	RACER	SGA-EC	Trowel
# contigs (≥ 0 bp)	6427	2157	1659	1464	1172	1207	1371	1117	2156	1783	1115	2128	1982
# contigs (≥ 1000 bp)	1632	563	414	419	254	291	368	239	654	506	249	730	638
# contigs (≥ 5000 bp)	112	137	140	148	124	129	136	124	142	141	128	174	154
# contigs (≥ 10000 bp)	36	74	71	76	80	75	74	81	70	70	81	69	69
# contigs (≥ 25000 bp)	13	44	49	43	47	49	48	43	43	43	47	44	34
# contigs (≥ 50000 bp)	4	25	29	23	29	28	29	29	22	26	27	22	23
Total length (≥ 0 bp)	**6346020**	6285231	6268654	6260914	6262363	6263784	6259600	6259177	6272493	6266341	6262912	6267223	6264247
Total length (≥ 1000 bp)	4750667	5952470	6034708	6062139	6116375	6114471	6080305	**6124380**	5908040	6005704	6123214	5934797	5955713
Total length (≥ 5000 bp)	1814873	4976928	5382652	5396892	5807885	5708795	5484607	**5839731**	4694888	5145507	5829866	4634435	4821830
Total length (≥ 10000 bp)	1315345	4541808	4910354	4890158	5487031	5314653	5030382	**5528596**	4208048	4660558	5491925	3914787	4262171
Total length (≥ 25000 bp)	949575	4069767	4608549	4378326	4984773	4933517	4665902	4956980	3749942	4253437	**4985637**	3558546	3731171
Total length (≥ 50000 bp)	637615	3413547	3872046	3689217	4333071	4222801	4022261	**4458330**	2969521	3619810	4248604	2767450	3372642
# contigs	2880	769	537	528	319	361	462	298	901	658	315	952	838
Largest contig	226239	328059	570411	322996	339286	**571909**	321732	571477	510496	539052	390592	359236	353318
Total length	5653892	6098246	6123822	6140602	6161700	6162690	6147408	6164910	6088141	6113380	**6168877**	6093952	6100522
Reference length	6264404	6264404	6264404	6264404	6264404	6264404	6264404	6264404	6264404	6264404	6264404	6264404	6264404
GC (%)	66.51	66.57	66.58	66.59	66.58	66.58	66.59	66.58	66.58	66.58	66.57	66.59	66.59
Reference GC (%)	66.56	66.56	66.56	66.56	66.56	66.56	66.56	66.56	66.56	66.56	66.56	66.56	66.56
N50	2619	64599	88629	109138	120313	**127693**	108906	110138	47837	78352	119291	38051	84256
NG50	2265	60742	78726	89930	**119565**	117737	108906	110138	46385	70540	118598	37313	69066
N75	1298	9231	25007	18252	41167	32431	25484	**42712**	5758	11152	38042	5176	6390
NG75	1040	7832	17828	14456	36170	31710	21018	**37686**	4988	9608	36985	4699	5662
L50	404	19	17	15	13	13	17	**11**	24	17	14	29	18
LG50	529	20	18	16	14	14	17	**11**	26	18	15	31	19
L75	1187	78	49	54	36	39	46	**33**	118	63	36	162	111
LG75	1581	92	54	60	38	41	50	**35**	143	74	38	188	131
# misassemblies	**9**	36	25	10	25	34	31	24	25	33	28	25	26
# misassembled contigs	**6**	18	13	9	17	16	20	15	17	23	20	17	16
Misassembled contigs length	**231350**	2447728	1089713	1025513	2079225	1690556	1712040	1562762	1779078	2011805	2230473	1542709	1746394
# local misassemblies	**100**	224	169	121	161	139	135	154	183	187	151	126	150
# unaligned mis. contigs	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	1	**0**	**0**	**0**	**0**
# unaligned contigs	5 + 2 part	2 + 5 part	1 + 3 part	**1 + 2 part**	1 + 3 part	2 + 4 part	1 + 4 part	2 + 3 part	2 + 7 part	2 + 6 part	1 + 6 part	1 + 5 part	1 + 7 part
Unaligned length	12588	14810	10655	**10101**	10880	15012	13650	11686	15543	15307	14447	14247	16755
Genome fraction (%)	89.452	96.792	97.408	97.657	97.964	97.931	97.733	**98.057**	96.676	97.177	98.040	96.808	96.870
Duplication ratio	1.007	1.003	**1.002**	**1.002**	**1.002**	**1.002**	**1.002**	**1.002**	1.003	**1.002**	**1.002**	1.003	1.003
# N's per 100 kbp	**118.04**	244.58	219.83	161.30	192.12	176.45	192.72	197.88	224.50	227.53	192.55	186.97	218.33
# mismatches per 100 kbp	39.63	15.06	21.86	11.38	**11.13**	13.09	16.15	23.75	20.34	17.68	15.42	26.70	34.38
# indels per 100 kbp	**12.03**	17.37	18.63	12.91	13.22	12.68	16.53	15.61	18.92	18.20	14.52	17.36	20.25
Largest alignment	226050	298012	570020	322354	339246	**571467**	321649	571031	395446	478427	311783	288241	352989
Total aligned length	5636890	6074306	6105782	6124234	6144447	6142551	6129358	6146957	6066401	6092744	**6148772**	6074595	6078072
NA50	2594	51100	73325	89839	105318	92668	88992	**110076**	45356	60272	103516	34022	52053
NGA50	2249	46410	72650	89169	92378	90062	88992	103328	40119	58932	**103516**	33437	47995
NA75	1291	8505	17828	16034	32144	29497	19133	31936	5555	10085	**33354**	5109	6104
NGA75	1023	7399	13949	13414	29880	27418	15730	29786	4829	8845	**30147**	4426	5399
LA50	411	27	20	17	16	17	19	**12**	29	20	19	34	25
LGA50	537	29	21	18	17	18	19	**13**	31	21	19	37	26
LA75	1201	99	57	59	44	48	57	**40**	132	78	46	178	129
LGA75	1598	114	63	65	47	51	62	**42**	158	90	48	205	150

Table A.42 Assembly quality metrics for *H. sapiens* Chr. 21

Assembly	Uncorrected	ACE	BayesHammer	BFC	BLESS2	Blue	Fiona	Karect	Lighter	Musket	RACER	SGA-EC	Trowel
# contigs (≥ 0 bp)	58213	58395	61700	57945	57438	57170	57769	57668	57427	57852	58297	57344	57247
# contigs (≥ 1000 bp)	6812	6402	6920	6367	6368	6277	6297	6350	6638	6471	6424	6519	6622
# contigs (≥ 5000 bp)	1664	1772	1615	1817	1811	1835	1821	1824	1704	1756	1804	1775	1730
# contigs (≥ 10000 bp)	404	496	386	477	488	535	533	516	446	471	509	470	463
# contigs (≥ 25000 bp)	11	26	16	31	28	26	32	22	19	25	21	22	20
# contigs (≥ 50000 bp)	0	0	0	0	0	0	0	0	0	0	0	0	0
Total length (≥ 0 bp)	31597631	31965979	31766176	32002529	31938926	**32041901**	32038486	31957612	31672081	31818334	32019235	31818763	31734651
Total length (≥ 1000 bp)	26827458	27507605	26550208	27599231	27535385	**27711114**	27675394	27554984	27105876	27343993	27574676	27341004	27179349
Total length (≥ 5000 bp)	14469620	16175954	13958509	16512649	16457706	**16892313**	16750500	16554128	15163420	15814857	16386813	15851231	15456691
Total length (≥ 10000 bp)	5646166	7235250	5447698	7068887	7173080	7760310	**7764712**	7437243	6381022	6825114	7312274	6776893	6569323
Total length (≥ 25000 bp)	305814	741358	449079	873737	799632	725036	**941610**	607723	524852	700429	569855	610919	568491
Total length (≥ 50000 bp)	0	0	0	0	0	0	0	0	0	0	0	0	0
# contigs	8747	8004	9061	7959	7994	7845	7855	7958	8450	8151	8059	8235	8402
Largest contig	31896	39067	39290	34106	44280	32458	**46072**	34407	32368	34493	38599	33802	38545
Total length	28232361	28675142	28101089	28756929	28718924	**28852292**	28810553	28726889	28425853	28563707	28760555	28590292	28479294
Reference length	46709983	46709983	46709983	46709983	46709983	46709983	46709983	46709983	46709983	46709983	46709983	46709983	46709983
GC (%)	39.53	39.66	39.26	39.69	39.68	39.74	39.72	39.73	39.62	39.64	39.73	39.65	39.58
Reference GC (%)	42.22	42.22	42.22	42.22	42.22	42.22	42.22	42.22	42.22	42.22	42.22	42.22	42.22
N50	5159	5885	4963	5993	6024	**6163**	6125	5985	5395	5688	5921	5646	5524
NG50	2054	2425	1914	2472	2440	**2543**	2519	2456	2194	2342	2435	2324	2217
N75	2676	3007	2563	3031	3024	**3099**	3088	3035	2812	2938	2992	2911	2815
L50	1595	1434	1634	1429	1430	**1393**	1397	1424	1522	1469	1437	1481	1499
LG50	4437	3835	4640	3772	3790	**3647**	3685	3775	4176	3953	3808	3980	4114
L75	3512	3153	3616	3122	3122	**3037**	3061	3113	3360	3219	3148	3246	3319
# misassemblies	612	620	**545**	610	732	715	653	573	668	650	644	663	682
# misassembled contigs	467	449	**403**	449	526	502	484	431	478	449	481	515	482
Misassembled contigs length	3326586	3566126	**3019786**	3568048	4202668	4253263	4091809	3584831	3642250	3563055	3822684	3936298	3766679
# local misassemblies	1010	1047	**932**	979	1023	1130	1091	1045	1063	1155	1100	1043	964
# structural variations	0	0	0	0	0	0	0	0	0	0	0	0	0
# unaligned mis. contigs	79	**44**	166	**44**	61	51	50	45	59	58	52	49	62
# unaligned contigs	324 + 573 part	173 + 432 part	547 + 925 part	178 + 385 part	168 + 395 part	168 + 435 part	**162 + 396 part**	169 + 380 part	194 + 477 part	195 + 508 part	174 + 413 part	196 + 420 part	245 + 486 part
Unaligned length	1322048	800832	2326663	748706	778800	807098	788398	**726780**	939040	983333	803632	833188	1008440
Genome fraction (%)	56.047	57.891	53.531	58.259	58.108	58.288	58.223	**58.336**	57.182	57.292	58.184	57.824	57.083
Duplication ratio	**1.028**	1.031	1.031	1.029	1.029	1.030	1.030	**1.028**	1.029	1.031	1.029	**1.028**	1.030
# N's per 100 kbp	3319.46	3550.95	4309.26	3339.67	3350.99	3317.24	3364.41	3182.06	3335.42	3523.72	**3166.64**	3175.33	3480.50
# mismatches per 100 kbp	231.08	227.19	225.64	222.01	220.04	227.97	221.99	223.26	**217.30**	226.65	235.20	230.24	224.21
# indels per 100 kbp	105.34	96.56	107.85	98.84	**95.34**	100.22	102.30	97.87	98.32	101.21	101.90	104.69	102.08
Largest alignment	29853	31527	31348	31451	31659	31447	29919	31822	31611	31801	31288	**32138**	31294
Total aligned length	26211189	27039009	25050677	27217144	27152844	27236838	27204838	**27244840**	26727737	26769049	27180503	27028047	26686854
NA50	4224	4851	3905	4988	4887	5006	4988	**5025**	4475	4661	4859	4712	4525
NGA50	1367	1756	994	1845	1786	1865	1845	**1871**	1566	1653	1807	1718	1566
NA75	2027	2350	1675	2416	2361	2396	2399	**2436**	2177	2245	2360	2295	2163
LA50	1888	1685	1957	1663	1694	1659	1649	**1647**	1808	1743	1697	1731	1785
LGA50	5584	4720	6417	4591	4699	4556	**4555**	4556	5180	4941	4689	4865	5141
LA75	4284	3814	4658	3746	3816	3743	3730	**3710**	4085	3945	3827	3908	4065

Table A.43 Assembly quality metrics for *C. elegans*

Assembly	Uncorrected	ACE	BayesHammer	BFC	BLESS2	Blue	Karect	Lighter	RACER	SGA-EC
# contigs (≥ 0 bp)	313458	247789	260292	223399	195321	207345	200640	291328	193008	228459
# contigs (≥ 1000 bp)	21506	25791	25237	26495	26687	26077	25766	23368	25455	26639
# contigs (≥ 5000 bp)	452	562	556	1374	1915	2107	2236	434	2549	1384
# contigs (≥ 10000 bp)	104	46	17	103	190	285	269	73	394	123
# contigs (≥ 25000 bp)	5	1	0	1	7	11	10	5	27	5
# contigs (≥ 50000 bp)	0	0	0	0	0	0	0	0	2	0
Total length (≥ 0 bp)	**100514891**	94677328	95949746	96223212	94442104	97069690	95356117	98710361	95898310	97728152
Total length (≥ 1000 bp)	37905692	48928957	48505197	58787718	64099874	64609003	64449307	41827510	**67692343**	59067924
Total length (≥ 5000 bp)	3832338	3819461	3531390	9485659	13824216	15914879	16622537	3441481	**19988122**	9813852
Total length (≥ 10000 bp)	1501701	613323	208156	1312103	2566271	3941268	3712693	1104325	**5832892**	1651044
Total length (≥ 25000 bp)	144182	26729	0	30830	198141	346721	302882	156627	**902471**	142834
Total length (≥ 50000 bp)	0	0	0	0	0	0	0	0	**124006**	0
# contigs	59201	55137	54557	49896	46077	46120	44909	58648	42732	50765
Largest contig	34936	26729	23739	30830	33772	38086	49047	37692	**70494**	34846
Total length	64279025	69730098	69246748	75436498	77994023	78900340	78118154	66601578	**80035157**	76221496
Reference length	100286070	100286070	100286070	100286070	100286070	100286070	100286070	100286070	100286070	100286070
GC (%)	38.74	38.50	37.36	37.78	37.65	37.84	37.60	38.67	37.64	37.96
Reference GC (%)	35.44	35.44	35.44	35.44	35.44	35.44	35.44	35.44	35.44	35.44
N50	1161	1469	1488	1925	2229	2304	2365	1247	**2625**	1897
NG50	735	966	953	1323	1610	1658	1661	804	**1917**	1329
N75	771	907	903	1078	1222	1221	1247	808	**1356**	1064
NG75	-	-	-	505	596	620	600	-	686	527
L50	16144	14108	13793	11163	9769	9315	9018	15713	**8214**	11284
LG50	35856	27027	26915	18971	15665	14807	14640	32673	**12741**	18887
L75	33287	29329	28847	24365	21616	21165	20519	32435	**18870**	24795
LG75	-	-	-	49455	40994	39514	39613	-	34527	48803
# misassemblies	**3096**	6019	5842	6981	8317	7594	8174	4317	8798	6525
# misassembled contigs	**2571**	4736	4640	5488	6301	5792	6221	3504	6669	5157
Misassembled contigs length	**3102213**	8112983	8555204	12333155	16017169	15172010	16532475	4808532	19081800	11271909
# local misassemblies	**507**	922	809	1241	1705	1479	1544	718	1429	1153
# unaligned mis. contigs	93	87	**47**	65	92	77	67	87	87	71
# unaligned contigs	5589 + 220 part	6249 + 393 part	6479 + 341 part	5193 + 420 part	4282 + 567 part	4503 + 532 part	4249 + 480 part	6133 + 347 part	**3909 + 565 part**	5144 + 398 part
Unaligned length	11242801	10009213	**7771997**	9601446	8858941	10229836	9092027	11050356	9511857	10527959
Genome fraction (%)	52.408	58.937	60.701	64.935	68.154	67.705	68.038	54.835	**69.539**	64.807
Duplication ratio	**1.009**	1.010	1.010	1.011	1.012	1.011	1.012	1.010	1.011	1.011
# N's per 100 kbp	**228.09**	487.99	461.80	435.10	527.78	490.59	484.83	348.43	484.22	434.66
# mismatches per 100 kbp	**17.40**	36.43	31.51	32.18	41.89	35.48	38.08	25.36	73.35	32.31
# indels per 100 kbp	**13.79**	28.27	26.68	29.41	34.82	32.06	33.98	19.61	38.98	28.71
Largest alignment	10782	10826	13458	16311	20974	17720	19637	7927	**21899**	14320
Total aligned length	52796255	59255569	61056499	65312826	68475519	68063868	68398255	55196310	**69843929**	65182944
NA50	865	1100	1207	1386	1589	1570	1639	929	**1755**	1336
NGA50	524	670	724	898	1075	1066	1094	565	**1205**	882
NA75	556	618	679	698	771	732	771	568	**791**	673
LA50	22902	18373	16311	14821	13371	13316	12657	21561	**12008**	15481
LGA50	49744	36229	33004	26009	21931	21611	20964	44967	**18983**	26594
LA75	46161	39575	35559	34141	31114	31843	30160	44628	**29099**	35712

Table A.44 Assembly quality metrics for *D. melanogaster*

Assembly	Uncorrected	ACE	BayesHammer	BFC	BLESS2	Blue	Fiona	Karect	Lighter	Musket	RACER	SGA-EC	Trowel
# contigs (≥ 0 bp)	100458	88911	119593	85549	79158	82740	84689	81344	93233	90739	80832	87971	89636
# contigs (≥ 1000 bp)	11374	9011	17268	8526	8241	8091	8111	8240	9556	9147	7913	8601	9091
# contigs (≥ 5000 bp)	5987	5107	6804	4877	4733	4635	4720	4722	5364	5155	4613	4931	5143
# contigs (≥ 10000 bp)	3535	3326	3069	3279	3147	3153	3151	3196	3412	3337	3117	3279	3373
# contigs (≥ 25000 bp)	1026	1250	452	1308	1323	1330	1307	1319	1216	1263	1326	1327	1279
# contigs (≥ 50000 bp)	170	326	41	370	385	430	424	399	289	343	420	352	330
Total length (≥ 0 bp)	**119882706**	118726428	119848583	118515422	117590011	118625990	118503169	117997496	118883364	118809330	118363831	119238372	118992862
Total length (≥ 1000 bp)	110945820	111997301	106901197	111963831	111698533	112321082	112161292	111864739	111467757	111860476	112364368	**112403232**	111949002
Total length (≥ 5000 bp)	97402353	102268382	81335021	102855188	102982775	103667092	103651250	103091852	100907891	101846130	**104026229**	103212029	102074935
Total length (≥ 10000 bp)	79831793	89406719	54899046	91272984	91527739	92966207	92357632	92058152	86780805	88726766	**93238860**	91212888	89264983
Total length (≥ 25000 bp)	40228323	56092233	15731892	59703455	61872079	63667742	62631987	61838260	51841237	55635456	**64062266**	59834606	55761936
Total length (≥ 50000 bp)	11289515	24250126	2493053	27166693	28988537	**32552918**	32110811	30144152	20129253	24524703	32425265	26052606	23142099
# contigs	14234	10680	23134	10165	9704	9648	9635	9757	11551	10876	9352	10330	10889
Largest contig	135981	199070	109411	221622	178719	206598	**310228**	201437	209745	192773	244683	181525	132175
Total length	112964497	113189748	111071275	113126301	112747850	113439099	113252292	112945003	112879620	113089933	113390785	**113629454**	113223579
Reference length	120381546	120381546	120381546	120381546	120381546	120381546	120381546	120381546	120381546	120381546	120381546	120381546	120381546
GC (%)	42.60	42.72	42.58	42.67	42.70	42.66	42.68	42.69	42.69	42.63	42.67	42.62	42.62
Reference GC (%)	42.42	42.42	42.42	42.42	42.42	42.42	42.42	42.42	42.42	42.42	42.42	42.42	42.42
N50	17913	24554	9867	26837	28286	29247	28679	27863	22711	24530	**29535**	26543	24455
NG50	16416	22835	8837	24772	26047	27206	26378	25842	20639	22616	**27373**	24862	22520
N75	8526	11868	4675	12862	13459	13756	13529	13415	10623	11586	**14073**	12604	11792
NG75	6914	9612	3561	10392	10598	11144	10940	10722	8757	9402	**11526**	10375	9573
L50	1795	1271	3134	1187	1117	1074	1083	1116	1409	1300	**1055**	1210	1314
LG50	2012	1422	3632	1328	1258	1197	1213	1255	1583	1455	**1179**	1342	1466
L75	4064	2910	7212	2710	2550	2481	2514	2562	3206	2974	**2430**	2744	2972
LG75	4786	3416	8915	3182	3027	2900	2953	3025	3787	3498	**2841**	3188	3478
# misassemblies	2610	2678	**2245**	2943	3004	2763	2805	2964	2837	2865	3015	2763	2752
# misassembled contigs	1702	**1591**	1642	1744	1767	1618	1613	1697	1758	1716	1708	1625	1652
Misassembled contigs length	37525466	45699234	**22889624**	50047465	51430351	51426960	51283937	51124905	44579745	47627338	53793688	48252389	46125881
# local misassemblies	4593	4126	3264	3497	3007	3631	**3003**	3514	3533	5132	3031	4441	4116
# unaligned mis. contigs	140	105	92	81	**68**	91	69	72	113	168	71	87	111
# unaligned contigs	1907 + 1478 part	1626 + 1164 part	2385 + 1185 part	1528 + 931 part	1376 + 859 part	1477 + 937 part	1458 + 809 part	1461 + 835 part	1806 + 1109 part	1672 + 1467 part	**1374 + 807 part**	1632 + 1147 part	1703 + 1267 part
Unaligned length	7337009	6218645	6918913	5229492	4762484	5188181	**4750697**	4835002	6477471	7387838	4789677	6035593	6870105
Genome fraction (%)	86.998	88.216	85.890	89.083	89.225	89.341	89.640	89.211	87.808	87.092	**89.757**	88.695	87.712
Duplication ratio	1.009	1.007	1.007	1.006	**1.005**	1.007	**1.005**	1.007	1.007	1.008	**1.005**	1.008	1.007
# N's per 100 kbp	980.75	835.91	988.73	760.40	831.13	731.83	725.02	754.48	944.90	947.85	**668.91**	817.92	925.11
# mismatches per 100 kbp	564.79	558.01	**555.27**	559.91	559.86	560.81	565.23	560.81	559.92	561.03	571.05	561.40	560.87
# indels per 100 kbp	175.12	171.61	**161.22**	169.16	163.26	167.67	171.56	167.46	173.98	175.49	172.57	171.33	174.66
Largest alignment	111277	**161258**	93672	121105	144032	139852	143266	128118	156286	111130	148210	156346	120063
Total aligned length	104800288	106279684	103443181	107269667	107402520	107600578	107947498	107465148	105741535	104875059	**108061294**	106822163	105640715
NA50	13000	17418	8022	17962	19023	**19579**	19232	18786	15721	16411	19489	18071	16834
NGA50	12003	15974	7179	16511	17330	18078	17860	17351	14419	15066	**18095**	16663	15320
NA75	5801	7874	3504	8262	8529	8942	8896	8589	7186	7128	**8961**	8189	7513
NGA75	4472	6214	2430	6613	6684	7203	7226	6872	5628	5584	**7236**	6642	5864
LA50	2444	1828	3807	1758	1647	**1608**	1631	1656	2002	1908	1622	1779	1891
LGA50	2741	2044	4420	1968	1856	**1793**	1824	1862	2251	2140	1808	1974	2114
LA75	5663	4240	8975	4058	3831	**3745**	3784	3867	4646	4488	3750	4095	4392
LGA75	6750	5011	11352	4790	4589	**4393**	4452	4590	5530	5353	4400	4780	5199

A.5 Memory and Runtime

A.5.1 Real Data

A.5.1.1 Memory

In the chapter 2 peak memory usage was measured for all EC tools. Table A.45 shows the exact numbers for reference.

Table A.45 Memory usage of EC tools (GB)

Tool	*B. dentium*	*E. coli DH10B*	*E. coli MG1655*	*S. enterica*	*P. aeruginosa*	*H. sapiens* Chr. 21	*C. elegans*	*D. melanogaster*
BayesHammer	9.97	18.87	20.68	15.06	18.48	17.00	21.78	27.38
BFC	0.95	1.06	1.47	0.96	1.18	2.43	5.16	5.22
BLESS2	3.88	3.90	3.89	2.95	3.90	3.90	3.90	3.90
Blue	1.43	2.15	2.42	1.41	1.30	3.66	8.12	8.48
Fiona	6.52	14.08	20.49	3.35	7.11	10.03	43.61	48.06
Karect	22.76	43.98	67.36	10.01	23.62	29.49	135.56	144.18
Lighter	0.35	0.36	0.36	0.36	0.37	0.55	0.82	0.90
Musket	0.78	0.66	0.98	0.74	0.51	0.76	2.40	3.04
SGA-EC	0.27	0.51	0.97	0.19	0.34	0.54	1.95	2.17
Trowel	0.49	0.57	0.90	0.44	0.50	0.52	1.79	2.14

A.5.1.2 Runtime

The runtime plot of the EC tools is shown in the chapter 2, Table A.46 shows the exact numbers.

Table A.46 Runtime of EC tools (min)

Tool	*B. dentium*	*E. coli DH10B*	*E. coli MG1655*	*S. enterica*	*P. aeruginosa*	*H. sapiens* Chr. 21	*C. elegans*	*D. melanogaster*
BayesHammer	4.66	12.15	15.30	3.54	8.94	11.55	38.50	45.64
BFC	0.78	1.55	2.67	0.43	0.84	1.17	4.44	5.54
BLESS2	0.64	1.31	2.17	0.37	0.73	1.38	4.44	6.02
Blue	1.18	2.20	4.00	0.87	1.78	2.21	10.27	12.26
Fiona	15.87	28.01	50.63	7.24	11.73	12.60	65.50	74.72
Karect	3.33	8.26	14.54	1.39	4.43	6.72	59.08	65.13
Lighter	0.28	0.70	1.26	0.17	0.49	0.78	3.01	3.87
Musket	2.99	5.83	11.10	2.61	2.93	4.11	12.36	21.94
SGA-EC	4.44	9.34	15.13	2.45	4.91	6.52	28.96	31.21
Trowel	1.00	1.79	3.21	0.54	1.13	1.74	7.98	9.13

A.5.2 Simulated Data

A.5.2.1 Memory

Table A.47 shows the peak memory usage for all tools in simulated data.

Table A.47 Peak memory usage of EC tools.

Tool	*B. dentium*	*E. coli DH10B*	*E. coli MG1655*	*S. enterica*	*P. aeruginosa*	*H. sapiens* Chr. 21	*C. elegans*	*D. melanogaster*
BFC	1.65	1.21	3.40	1.07	1.24	3.00	5.20	7.35
BLESS 2	3.90	3.90	3.91	2.95	3.90	3.89	3.90	3.90
Fiona	6.82	14.03	22.08	3.26	7.79	10.69	43.50	49.33
Karect	22.74	43.99	67.31	10.05	23.64	29.66	135.89	146.49
Lighter	0.35	0.36	0.36	0.36	0.37	0.56	0.81	0.90
Musket	1.05	0.62	1.77	0.76	0.49	1.10	1.58	7.75
SGA-EC	0.32	0.51	1.02	0.21	0.35	0.56	1.86	2.53
Trowel	0.49	0.75	0.88	0.46	0.49	0.51	1.70	1.85

A.5.2.2 Runtime

We recorded elapsed (wall clock) time to measure the runtime. Table A.48 shows the runtime of EC tools on simulated data.

Table A.48 Runtime of the EC tools.

Tool	*B. dentium*	*E. coli DH10B*	*E. coli MG1655*	*S. enterica*	*P. aeruginosa*	*H. sapiens* Chr. 21	*C. elegans*	*D. melanogaster*
BFC	0.92	1.74	2.78	0.44	0.87	1.47	4.99	6.75
BLESS 2	0.80	1.24	2.27	0.37	0.71	1.37	4.21	5.57
Fiona	27.58	22.82	53.40	6.84	11.10	12.49	64.93	70.48
Karect	3.29	7.36	16.10	1.30	3.22	9.54	27.81	41.47
Lighter	0.28	0.73	1.39	0.16	0.46	0.91	3.00	3.47
Musket	4.12	2.19	8.09	1.47	1.32	3.78	7.59	17.29
SGA-EC	5.52	8.61	14.32	2.43	4.62	7.34	27.95	34.69
Trowel	1.06	2.00	3.29	0.55	1.18	2.11	6.67	9.96

B

Supplementary Data: Illumina error correction near highly repetitive DNA regions improves de novo genome assembly

"... I learned the value of hard work by working hard. [1] *"*

B.1 Parameter settings

Tools were executed with 64 threads. BLESS2 fails to finish with 64 threads in some datasets; therefore, only 32 cores were used by this tool. For all tables and figures in the chapter 3 and in the supplementary data the parameters' default or recommended values were selected for each tool. Below, the command line parameters are specified for each tool individually:

B.1.1 ACE [2]

```
$ size=$ (stat -c\%s genome.fasta)
$ ./ace $size $inputreads aceOut/aceCorrected
```

[1] Margaret Mead

[2] https://github.com/Sheikhizadeh/ACE.git

B.1.2 BFC v. r181 [3]

```
$ size=$(stat -c\%s genome.fasta)
$ ./ace $ size $inputreads aceOut/aceCorrected
```

B.1.3 BLESS2 v. 1.02 [4]

```
$ ./bless -read $inputreads -prefix blessOut/
   blessCorrected -kmerlength 31 -smpthread 32
```

B.1.4 Browniecorrector v. 1.0 [5]

Arguments need to be provided in the following order: the first argument is the data library (note that if there are multiple libraries, each library must be corrected individually); The second argument is the *cov* which is the depth of coverage in that library; the last argument is the working directory.

```
$ ./bashScripts/runPipeLine.sh $inputreads $cov $
   workdir
```

B.1.5 Karect v. 1.0 [6]

```
$ ./karect -correct -inputfile=$inputreads -matchtype=
   hamming -celltype=diploid -resultdir=karectOut -
   kmer=9 -memory=32 -threads=32
```

B.1.6 RECKONER v. 1.1.1 [7]

```
$ size=$(stat -c\%s genome.fasta)
$ reckoner -prefix reckonerOUt -threads 64 -read $
   inputreads -genome $size
```

B.1.7 SPAdes v. 4.12[8]

In order to see the impact of EC tools on assembly results, we used SPAdes to assemble both corrected and uncorrected data. The following command was used to run the SPAdes.

```
$ spades.py -t 32 --only-assembler --12 reads.fastq -o
    outputDir
```

[3] https://github.com/lh3/bfc.git
[4] https://sourceforge.net/projects/bless-ec
[5] https://github.com/biointec/browniecorrector.git
[6] https://github.com/aminallam/karect.git
[7] https://github.com/refresh-bio/RECKONER.git
[8] http://cab.spbu.ru/software/spades/

B.1.8 Quast v. 4.6.3[9]

Quast provides comprehensive information on the assembly quality. The following command was used to run the Quast.

```
$ quast.py  asmDir/contigs.fa -R genome.fasta -o
   quastReport --plots-format pdf  --labels "toolName"
```

B.2 Data preparation

B.2.1 Illumina real data

1. R1 (*Homo sapiens* Chr. 21)
   Download[10] the row data from :Human_NA19240.7z

   ```
   $  7z e Human_NA19240.7z
   $ ./shuffleSequences_fastq.pl
      NA19240_HiSeq_100_chr21_R1_paired.fq
      NA19240_HiSeq_100_chr21_R2_paired.fq reads.
      fastq
   ```

   Download the reference genome from hs_ref_GRCh38.p12_chr21.fa.gz

2. R2 (*Homo sapiens* Chr. 14)
   Download the row data from frag_1.fastq.gz, and frag_2.fastq.gz

   ```
   $ gzip -d  frag_1.fastq.gz
   $ gzip -d  frag_2.fastq.gz
   $ ./shuffleSequences_fastq.pl frag_1.fastq frag_2.
      fastq reads.fastq
   ```

   Download the reference genome from genome.fasta

3. R3 (*Caenorhabditis elegans*)
   Download the row data from : SRR543736.sra

   ```
   $ ./fastq-dump --split-files SRR543736.sra
   $ ./shuffleSequences_fastq.pl SRR543736_1.fastq
      SRR543736_2.fastq reads.fastq
   ```

[9] http://quast.sourceforge.net/quast

[10] In case you have any difficulties of downloading the files, you can open this document with "document viewer" application, copy the link and paste it in your browser.

Download the reference genome from c_elegans.WS222.genomic.fa.gz

4. R4 (*Drosophila melanogaster*) Download the row data from SRR823377.sra

```
$ ./fastq-dump --split-files SRR823377.sra
$ ./shuffleSequences_fastq.pl SRR823377_1.fastq
   SRR823377_2.fastq reads.fastq
```

Download the reference genome via the following links: NT_033777.fna, NT_033778.fna, NT_033779.fna, NT_037436.fna, NC_004353.fna and NC_004354.fna. Then concatenate them all together:

```
$ cat NT_033777.fna NT_033779.fna NT_033778.fna
   NT_037436.fna NC_004353.fna NC_004354.fna >
   genome.fasta
```

5. R5 (*Drosophila melanogaster*) Download the row data from SRR988075.sra

```
$ ./fastq-dump --split-files SRR988075.sra
$ ./shuffleSequences_fastq.pl SRR988075_1.fastq
   SRR988075_2.fastq reads.fastq
```

The reference genoem is the same as R4.

6. R6 (*Arabidopsis thaliana*) Download the row data from SRR988075.sra

```
   $ ./fastq-dump --split-files SRR1174256.sra
   $ ./shuffleSequences_fastq.pl SRR1174256_1.
   fastq SRR1174256_2.fastq reads.fastq

```

Download the reference genome from GCF_000001735.4_TAIR10.1_genomic.fna.gz

B.2.2 Pacbio real data

1. P1 (*Drosophila melanogaster*)
   Download the row data from SRR1204466.sra

```
$ ./fastq-dump --split-files SRR1204466.sra
$ cat SRR1204466*.fastq >pacbio.reads.fastq
```

2. P2 (*Arabidopsis thaliana*)
   Download the row data from SRR1284707.sra

```
$ ./fastq-dump --split-files SRR1284707.sra
$ cat cat SRR1284707*.fastq >pacbio.reads.fastq
```

B.3 *k*-mer selection

Table B.1 represents the most frequent k-mers in each dataset. For example a poly-(A/T) 15-mer is the most frequent 15-mer in 3 datasets. We used jellyfish to count the frequency of 15-mers in the datasets as follows:

```
\texttt{ $ jellyfish count -m 15 -s 100M -t 64 -C
  reads.fastq}\\
\texttt{ $ jellyfish dump mer\_counts.jf -L $
  threshold > kmerFile.high.fasta}\\

```

The threshold values are chosen appropriately based on the dataset size to find the top-5 most frequent 15-mers.

Our experimental investigations show that most of the breakpoints in the assembly results occur in the direct vicinity of highly repetitive k-mers. For example, the top 5 most frequent 15-mers in the first or last 100 bp of the assembled contigs ($>$ 1000bp) with SPAdes in 6 different datasets are listed in table B.2:

Table B.1 The top-5 most frequent 15-mers in each dataset.

Rank	15-mer	Frequency
	R1	
1	AAAAAAAAAAAAAAA	843835
2	ACACACACACACACA	265661
3	AATGGAATGGAATGG	134236
4	AAAGTGCTGGGATTA	134174
5	CATTCCATTCCATTC	133799
	R2	
1	AAAAAAAAAAAAAAA	2199579
2	ACACACACACACACA	703103
3	GCCTGTAATCCCAGC	459711
4	AGCACTTTGGGAGGC	432208
5	AAAGTGCTGGGATTA	431474
	R3	
1	AAAAAAAAAAAAAAA	1105265
2	ATATTTTACTCTCTG	954488
3	ACTCTCTGTGGCTTC	755111
4	AAGCCACAGAGAGTA	754801
5	CCACAGAGAGTAAAA	746402
	R4	
1	AATAACATAGAATAA	5590029
2	GAATAACATAGAATA	5549407
3	ATAACATAGAATAAC	5503200
4	CATAGAATAACATAG	5201994
5	AAGAGAAGAGAAGAG	5135751
	R5	
1	AAGAGAAGAGAAGAG	16685329
2	ATAGAATAACATAGA	12988779
3	AACACAACACAACAC	12987799
4	AATAACATAGAATAA	9427421
5	ATAACATAGAATAAC	9277417
	R6	
1	ACTCCAAAACACTAA	1831402
2	CCATGAAAGCTTTGA	1804271
3	AAAAAAAAAAAAAAA	1804010
4	CTCCAAAACACTAAC	1802058
5	TATGATTGAGTATAA	1794787

Table B.2 The top-5 most frequent 15-$mers$ in the beginning or end of assembled contigs in different dataset.

Rank	15-mer	Frequency
	R1	
1	TTTTTTTTTTTTTTT	1853
2	TTTTTTTTTTTTTTG	1319
3	AAAAAAAAAAAAAAG	1164
4	AGCACTTTGGGAGGC	933
5	CCAGCACTTTGGGAG	891
	R2	
1	TTTTTTTTTTTTTTT	5609
2	TTTTTTTTTTTTTTG	4124
3	AGCACTTTGGGAGGC	3825
4	AATCCCAGCACTTTG	3664
5	AAAGTGCTGGGATTA	3658
	R3	
1	GGGGGGGGGGGGGGG	1276
2	GGGGGGGGGGGGGGA	617
3	CCCCCCCCCCCCCCA	596
4	TTTTTTTTTTTTTTT	500
5	TTCCCCCCCCCCCCC	387
	R4	
1	GGGGGGGGGGGGGGG	904
2	TTTTTTTTTTTTTTT	496
3	CCCCCCCCCCCCCCA	494
4	ATATATATATATATA	397
5	AGGGGGGGGGGGGGG	363
	R5	
1	TTTTTTTTTTTTTTT	1093
2	GGGGGGGGGGGGGGG	758
3	TTTTTTTTTTTTTTG	720
4	CCCCCCCCCCCCCCA	434
5	AAAAAAAAAAAAAAT	324
	R6	
1	TTTTTTTTTTTTTTT	1285
2	TTTTTTTTTTTTTTG	607
3	AAAAAAAAAAAAAAG	506
4	ATATATATATATATA	484
5	AAAAAAAAAAAAAAT	470

B.4 *k*-mer coverage

Assuming a genomic region and a randomly selected k-mer from this region, the average number of reads that initially cover any base of that k-mer is the initial coverage (C). The expected number of extracted reads from that region that contain that specific k-mer (C_k) is given by the following formula: $C_k = \frac{l-k+1}{l} C(1-e)^k$ where l is the read length and e is the error rate.

Proof:

It has been shown in [1], in the absence of errors the expected coverage of reads in a region of size k is $\frac{l-k+1}{l} C$. However, in the presence of errors, some of these reads fail to cover that region perfectly (i.e without any mismatch) due to the sequencing error. Let us assume the errors occur independently from each other, then the probability that all the bases of a read in an interval of size k are error-free is $(1-e)^k$. Therefore, the expected number of reads that cover a region of k is: $C_k = \frac{l-k+1}{l} C(1-e)^k$.

B.5 Results

B.5.1 Average improvement ratio of NGA50

Table 2 and 3 in the main paper show the exact values of NGA50 for contigs and scaffods after and before the error correction. Table B.3 shows the improvement rate of NGA50 for both contigs and scaffolds upon the uncorrected data for different datasets. The average improvement rate (AVG column) shows that jointly using of BrownieCorrector and Karect leads to the highest positive impact on the quality of contigs/scaffolds (+21%/+25%) whereas BrownieCorrector with (+18%/+19%), Karect with (+11%/+15%), and BFC with (+5%/+7%) are the second, third and forth best tools. On the other hand, BLESS2 (-25%/-19%), ACE (-17%/-14%), and Reckoner(-11%/-10%) deteriorate the quality of assembly on average.

B.5.2 Choice of highly repetitive *k*-mer

In order to see the performance of BrownieCorrector, we run the benchmark with two homopolymers poly-(A/T) and poly-(C/G). The results for poly-(A/T) are reported in the main paper. Table B.4 compares the number of reads in each dataset that contains specific k-mers and respectively corrected versus the total number of reads in that dataset.

Table B.5 and B.6 show the NGA50 of contigs and scaffolds when reads that contain respectively a 15-mer poly (C/G) and a 15-mer poly (AC/GT) are corrected by BrownieCorrector. Table B.5 indicates that correcting reads that contain a poly (C/G) often has a lower impact on the quality of the assembly (except D3 which

Table B.3 The improvement rate of NGA50 values for contigs and scaffolds upon the uncorrected data for different EC tools

Tools	D1	D2	D3	D4	D5	D6	D7	D8	D9	AVG
Contig NGA50 improvement rate(%)										
ACE	5	55	-51	-43	-44	-16	-31	-23	-5	-17
BFC	7	74	0	-3	-24	0	-4	-7	-1	5
BLESS2	-16	42	-53	-51	-52	-27	-24	-29	-15	-25
BrownieCorrector	23	102	0	3	8	11	3	5	10	18
Karect	15	85	0	6	-18	1	6	1	6	11
Reckoner	-16	18	-1	-17	-27	-3	-28	-16	-13	-11
BrownieCorrector+Karect	24	128	0	10	-15	11	10	3	15	21
Scaffold NGA50 improvement rate(%)										
ACE	7	52	-51	-42	-31	-5	-35	-15	-2	-14
BFC	8	71	0	-3	-9	0	-5	1	-3	7
BLESS2	-12	40	-53	-43	-39	-8	-24	-21	-12	-19
BrownieCorrector	24	104	0	1	9	12	0	8	14	19
Karect	19	82	-1	4	0	1	6	13	7	15
Reckoner	-15	15	-1	-21	-15	-3	-30	-9	-15	-10
BrownieCorrector+Karect	28	126	-1	8	5	11	8	15	23	25

yields in a higher NGA50). This is due to the fact that a poly C is less occurred than a poly A in all the datasets and the assembler can itself handle the associated complexity. Table B.6 indicates that correcting reads that contain a poly (AC/GT) has no positive impact on the quality of the assembly and sometimes the results are slightly worse. This is due to the fact that even though poly (AC/GT) is frequent, but the quality of reads that contain a poly (AC/GT) is high, and the assembler can itself handle the associated complexity. We highly suggest the user to use the poly A which is the default k-mer in the software.

B.5.3 Choice of the number of iterations

In order to find the stable cores of clusters we repeated the clustering multiple times. The default value for the number of iteration is set to 20 in the software. However, we further investigate the quality of assembly results (for D1) when it is set to 1, 5, 10 and 30 as well. Fig. B.1 shows how NGA50 of contigs and scaffolds changes for different values of iteration. This picture indicates that using the stable cores after running the clustering multiple times improves the quality of assembly. However, it also shows the accuracy of BrownieCorrector is not much affected by changing this parameter in the range of (5 to 30).

Table B.4 Two highly repetitive k-mers used in this study. The number of corrected and total number of reads in each dataset is compared.

highly repetitive k-mer	Number of corrected reads	Total number of reads
R1		
AAAAAAAAAAAAAAA	264 608 (1.96%)	13 486 136
CCCCCCCCCCCCCCC	12 180 (0.09%)	13 486 136
ACACACACACACACA	96 542 (0.71%)	13 486 136
R2		
AAAAAAAAAAAAAAA	620 500 (1.69%)	36 504 800
CCCCCCCCCCCCCCC	41 890 (0.11%)	36 504 800
ACACACACACACACA	202 770 (0.55%)	36 504 800
R3		
AAAAAAAAAAAAAAA	198 598 (0.34%)	57 721 732
CCCCCCCCCCCCCCC	112 908 (0.19%)	57 721 732
ACACACACACACACA	72 848 (0.12%)	57 721 732
R4		
AAAAAAAAAAAAAAA	576 552 (0.91%)	63 014 762
CCCCCCCCCCCCCCC	138 976 (0.22%)	63 014 762
ACACACACACACACA	477 950 (0.75%)	63 014 762
R5		
AAAAAAAAAAAAAAA	653 028 (0.85%)	75 938 276
CCCCCCCCCCCCCCC	83 506 (0.10%)	75 938 276
ACACACACACACACA	486 066 (0.64%)	75 938 276
R6		
AAAAAAAAAAAAAAA	571 806 (0.61%)	93 429 346
CCCCCCCCCCCCCCC	8 256 (0.01%)	93 429 346
ACACACACACACACA	32 320 (0.03%)	93 429 346

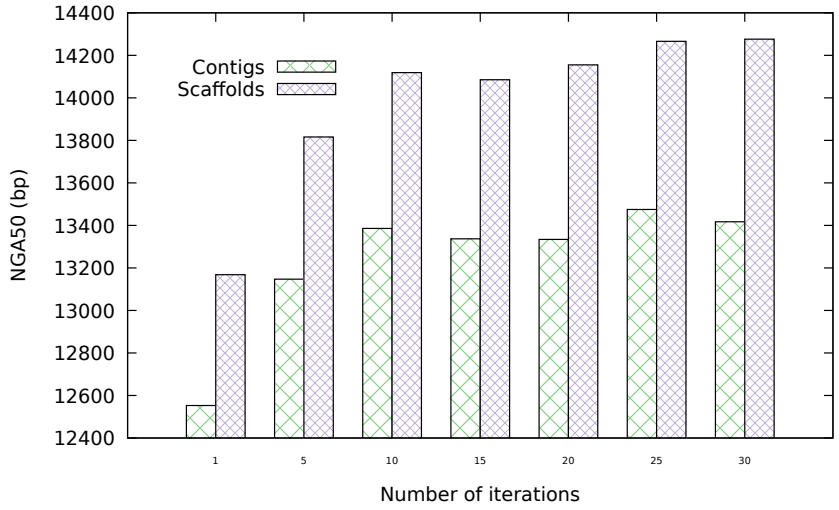
Table B.5 NGA50 of respectively contigs (top) and scaffolds (bottom) assembled by SPAdes before and after error correction. Reads that contain a 15-mer poly (C/G) are corrected by BrownieCorrector.

Tools	D1	D2	D3	D4	D5	D6	D7	D8	D9
					Contig NGA50				
Uncorrected	10 876	**5 451**	6 325	**50 833**	35 924	40 802	**80 752**	85 003	65 138
BrownieCorrector	10 876	5 449	**6 438**	50 733	**36 177**	**40 805**	79 151	85 003	**65 469**
					Scaffold NGA50				
Uncorrected	11 377	5 668	6 419	**60 714**	59 591	41 833	**96 381**	109 785	84 659
BrownieCorrector	**11 385**	5 668	**6 525**	59 130	**60 500**	**41 836**	92 852	**110 560**	84 659

Table B.6 NGA50 of respectively contigs (top) and scaffolds (bottom) assembled by SPAdes before and after error correction. Reads that contain a 15-mer poly (AC/TG) are corrected by BrownieCorrector.

Tools	D1	D2	D3	D4	D5	D6	D7	D8	D9
					Contig NGA50				
Uncorrected	**10 876**	**5 451**	**6 325**	50 833	**35 924**	**40 802**	**80 752**	**85 003**	**65 138**
BrownieCorrector	10 762	5 391	**6 325**	**50 834**	35 389	**40 802**	80 712	84 581	**65 138**
					Scaffold NGA50				
Uncorrected	**11 377**	**5 668**	**6 419**	**60 714**	59 591	41 833	**96 381**	**109 785**	**84 659**
BrownieCorrector	11 270	5 621	**6 419**	**60 714**	**59 827**	**41 836**	96 165	108 752	**84 659**

Figure B.1 The impact of changing the number of iterations in reads clustering on the quality of assembly in D1(*Homo sapiens* chr. 21).



B.5.4 Full Quast report (contigs)

This section contains the Quast evaluation report of contigs after assembling each dataset with SPAdes. Error correction by ACE, BFC, BLESS2, Brownie, Karect and Reckoner is performed prior to assembling the reads. The Uncorrected column refers to the quality of contigs without any pre-correction process. The Hybrid column shows the quality of assembly of reads which are corrected jointly by BrownieCorrector and Karect. Default parameter settings are used for Quast, therefore all statistics are based on contigs of size $\geq$ 500 bp.

B.5.4.1 D1

Table B.7 contains the Quast report after assembling dataset D1 (*Homo sapiens* Chr. 21) with SPAdes.

B.5.4.2 D2

Table B.8 contains the Quast report after assembling dataset D2 (*Homo sapiens* Chr. 14) with SPAdes.

B.5.4.3 D3

Table B.9 contains the Quast report after assembling dataset D3 (*C. elegans*) with SPAdes.

B.5.4.4 D4

Table B.10 contains the Quast report after assembling dataset D4 (*D. melanogaster*) with SPAdes.

B.5.4.5 D5

Table B.11 contains the Quast report after assembling dataset D5 (*D. melanogaster*) with SPAdes.

B.5.4.6 D6

Table B.12 contains the Quast report after assembling dataset D6 (*A. thaliana*) with SPAdes.

B.5.4.7 D7

Table B.13 contains the Quast report after a hybrid assembly of dataset D7 (*D. melanogaster*) with SPAdes. This is a hybrid assembly in which the corrected

(and uncorrected) Illumina reads (R4) are complemented with the Pacbio reads (P1).

B.5.4.8 D8

Table B.14 contains the Quast report after assembling dataset D8 *D. melanogaster* with SPAdes. This is a hybrid assembly in which the corrected (and uncorrected) Illumina reads (R5) are complemented with the Pacbio reads (P1).

B.5.4.9 D9

Table B.15 contains the Quast report after assembling dataset D9 (*A. thaliana*) with SPAdes. This is a hybrid assembly in which the corrected (and uncorrected) Illumina reads (R6) are complemented with the Pacbio reads (P2).

Table B.7 Assembly quality metrics for D1

Assembly	Uncorrected	ACE	BFC	BLESS2	Hybrid	Karect	BrownieCorrector	Reckoner
# contigs (≥ 0 bp)	15698	17387	15294	18776	16438	17247	14537	21465
# contigs (≥ 1000 bp)	3637	3501	3428	4065	3035	3269	3058	4218
# contigs (≥ 5000 bp)	1989	1953	1910	2065	1804	1898	1802	2060
# contigs (≥ 10000 bp)	1128	1115	1113	1064	1130	1126	1124	1082
# contigs (≥ 25000 bp)	239	254	271	189	314	281	321	181
# contigs (≥ 50000 bp)	24	24	28	11	46	34	43	11
Total length (≥ 0 bp)	34383421	34462802	34371704	34340929	34484783	34541537	34323631	**34746509**
Total length (≥ 1000 bp)	32610876	32607223	32648129	32311651	32677840	32667335	**32678726**	32480570
Total length (≥ 5000 bp)	28418343	28545565	28787065	27096397	**29533788**	29142408	29499002	26912697
Total length (≥ 10000 bp)	22265626	22534041	23076297	19919516	**24659630**	23596575	24593042	19861301
Total length (≥ 25000 bp)	8482791	9017654	9921318	6494787	11802375	10261128	**12015621**	6201468
Total length (≥ 50000 bp)	1439339	1504015	1767700	698377	**2815229**	2121583	2742260	674634
# contigs	4209	4018	3990	4674	3520	3776	3540	4874
Largest contig	82702	98568	**109124**	82889	88924	92853	98722	80324
Total length	33019875	32976721	**33047634**	32752359	33022607	33030548	33022956	32950991
Reference length	40988574	40988574	40988574	40988574	40988574	40988574	40988574	40988574
GC (%)	40.73	40.75	40.73	40.67	40.76	40.75	40.75	40.62
Reference GC (%)	40.93	40.93	40.93	40.93	40.93	40.93	40.93	40.93
N50	15054	15720	16310	12767	18515	17064	**18523**	12479
NG50	11384	11943	12348	9548	**13981**	12994	13969	9475
N75	7816	8068	8352	6681	**9864**	8990	9813	6589
NG75	2807	2990	2929	2365	**3519**	3269	3491	2377
L50	659	638	603	753	534	587	**532**	780
LG50	963	929	881	1123	781	854	**779**	1148
L75	1411	1361	1300	1629	**1141**	1251	1142	1679
LG75	2596	2509	2412	3101	**2093**	2291	2101	3146
# misassemblies	172	146	185	172	109	**100**	139	145
# misassembled contigs	163	134	168	163	102	**90**	130	140
Misassembled contigs length	2744756	2266007	3413239	2579081	2163845	2066353	2696928	**2019536**
# local misassemblies	117	**87**	100	93	89	90	114	93
# unaligned mis. contigs	0	0	0	0	0	0	0	0
# unaligned contigs	69 + 7 part	60 + 7 part	72 + 7 part	**47 + 6 part**	65 + 6 part	63 + 4 part	70 + 7 part	64 + 5 part
Unaligned length	75440	69319	81177	**52669**	70161	68430	76887	67237
Genome fraction (%)	80.057	79.980	80.098	79.544	**80.222**	80.191	80.200	79.862
Duplication ratio	1.004	1.004	1.004	1.003	**1.002**	1.003	**1.002**	1.005
# N's per 100 kbp	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
# mismatches per 100 kbp	166.59	163.62	161.90	162.96	**152.72**	155.00	157.31	158.16
# indels per 100 kbp	36.21	36.13	35.63	34.84	35.83	35.64	36.50	**33.08**
Largest alignment	82666	98563	96583	80972	85895	92848	**98722**	80324
Total aligned length	32887394	32846124	32913718	32658156	**32929390**	32922661	32921082	32800027
NA50	14357	15191	15483	12278	**17934**	16481	**17934**	12005
NGA50	10876	11375	11672	9183	**13526**	12507	13334	9154
NA75	7513	7758	7898	6399	**9446**	8584	9312	6292
NGA75	2672	2751	2743	2221	**3382**	3136	3344	2211
LA50	687	662	634	781	**552**	607	553	804
LGA50	1006	966	931	1168	**808**	883	812	1185
LA75	1477	1419	1375	1696	**1185**	1294	1194	1739
LGA75	2738	2629	2562	3245	**2178**	2376	2204	3280

Table B.8 Assembly quality metrics for D2

Assembly	Uncorrected	ACE	BFC	BLESS2	Hybrid	Karect	BrownieCorrector	Reckoner
# contigs (≥ 0 bp)	412807	64078	61489	65782	55134	59219	86196	83030
# contigs (≥ 1000 bp)	15216	10449	9716	11408	7780	9167	8834	13090
# contigs (≥ 5000 bp)	5457	5259	5117	5438	4672	5027	5035	5486
# contigs (≥ 10000 bp)	2048	2692	2771	2610	2797	2735	2785	2406
# contigs (≥ 25000 bp)	165	493	589	388	787	642	665	286
# contigs (≥ 50000 bp)	3	40	56	25	123	68	78	13
Total length (≥ 0 bp)	**111701232**	88494866	88685751	88477590	88259815	88449116	90604563	89660410
Total length (≥ 1000 bp)	81314165	82917065	83298061	82563788	**83582377**	83366588	83452630	82221160
Total length (≥ 5000 bp)	56548179	69077685	71125131	66852992	**75198734**	72219844	73292947	62595655
Total length (≥ 10000 bp)	32454389	50611835	54160580	46552182	**61611743**	55675021	57064546	40852265
Total length (≥ 25000 bp)	5145723	16921012	20768408	13067553	**30030862**	22904577	24067360	9419772
Total length (≥ 50000 bp)	174274	2388255	3302981	1588265	**7758002**	4107399	4809154	773905
# contigs	18340	11869	10998	13082	8702	10313	9907	15287
Largest contig	68123	85393	80374	97813	**102135**	94125	99626	81231
Total length	83591797	83953943	84236486	83787039	**84247240**	84198056	84227171	83825716
Reference length	107349540	107349540	107349540	107349540	107349540	107349540	107349540	107349540
GC (%)	40.66	40.72	40.71	40.68	40.73	40.72	40.73	40.48
Reference GC (%)	40.89	40.89	40.89	40.89	40.89	40.89	40.89	40.89
N50	7859	12805	14093	11414	**18281**	15129	15504	9663
NG50	5506	9159	10166	8182	**13099**	10719	11194	6836
N75	4000	6649	7403	5939	**9387**	7739	8062	4939
NG75	1167	1888	2116	1669	**2743**	2260	2380	1442
L50	3104	1927	1751	2173	**1355**	1633	1590	2514
LG50	4910	3012	2723	3395	**2103**	2542	2466	3958
L75	6828	4199	3815	4703	**2960**	3583	3463	5542
LG75	14473	8756	7888	9852	**6063**	7382	7039	11683
# misassemblies	119	820	640	716	353	496	**112**	689
# misassembled contigs	119	759	612	676	336	469	**110**	660
Misassembled contigs length	**985057**	10604473	9133094	8318230	6946313	7769701	2110750	6863900
# local misassemblies	44	41	49	54	40	42	**38**	44
# unaligned mis. contigs	0	0	0	0	0	0	0	0
# unaligned contigs	**13 + 6 part**	13 + 26 part	17 + 19 part	19 + 19 part	16 + 13 part	15 + 19 part	16 + 5 part	19 + 24 part
Unaligned length	16477	34851	30844	35033	26953	28976	**16274**	38607
Genome fraction (%)	77.424	77.785	78.122	77.703	**78.326**	78.150	78.312	77.521
Duplication ratio	1.006	1.005	1.004	1.004	**1.002**	1.003	**1.002**	1.007
# N's per 100 kbp	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
# mismatches per 100 kbp	112.22	126.99	119.91	125.01	105.67	114.90	**101.91**	118.25
# indels per 100 kbp	20.59	21.56	21.52	20.54	21.48	21.23	20.81	**17.42**
Largest alignment	68123	81510	80321	93310	**102135**	92098	99626	65458
Total aligned length	83425919	83704957	84047956	83625967	**84183013**	84039499	84171283	83422870
NA50	7792	11786	13177	10772	**17429**	14288	15290	9245
NGA50	5451	8475	9488	7737	**12409**	10103	11015	6440
NA75	3951	6058	6829	5537	**8948**	7315	7985	4624
NGA75	1121	1693	1929	1531	**2573**	2108	2345	1288
LA50	3124	2055	1841	2279	**1410**	1710	1610	2616
LGA50	4948	3230	2878	3569	**2192**	2675	2498	4141
LA75	6888	4522	4049	4968	**3090**	3776	3507	5822
LGA75	14685	9514	8458	10506	**6350**	7821	7138	12474

Table B.9 Assembly quality metrics for D3

Assembly	Uncorrected	ACE	BFC	BLESS2	Hybrid	Karect	BrownieCorrector	Reckoner
# contigs (≥ 0 bp)	91054	108559	91618	117149	91976	91798	91152	93938
# contigs (≥ 1000 bp)	20133	26210	20179	28117	20190	20176	20128	20188
# contigs (≥ 5000 bp)	6202	5193	6215	4644	6253	6247	6201	6189
# contigs (≥ 10000 bp)	2336	1170	2330	909	2368	2364	2337	2333
# contigs (≥ 25000 bp)	280	120	296	70	300	302	280	286
# contigs (≥ 50000 bp)	65	22	65	7	56	56	65	63
Total length (≥ 0 bp)	116361844	112509995	116388634	112348944	116081137	116062821	116365458	**116425966**
Total length (≥ 1000 bp)	**103913389**	95599801	103773800	92664921	103437334	103440805	103912841	103655165
Total length (≥ 5000 bp)	70663972	46557489	70526082	38965131	70281028	70250754	**70666015**	70314693
Total length (≥ 10000 bp)	43594203	19178514	43288714	13815325	42981410	42929158	**43606770**	43257998
Total length (≥ 25000 bp)	13566746	5009175	**13613721**	2460232	12782149	12861875	13554231	13527463
Total length (≥ 50000 bp)	6657935	1656998	6061255	396865	4955692	4959945	**6657938**	6351558
# contigs	26593	36711	26767	40384	26788	26761	26589	26789
Largest contig	**244078**	128379	240394	66400	238991	238991	**244078**	240969
Total length	108572808	103180601	108528629	101465829	108202583	108197037	**108573218**	108420265
Reference length	100286070	100286070	100286070	100286070	100286070	100286070	100286070	100286070
GC (%)	38.47	38.30	38.46	38.20	38.41	38.41	38.47	38.46
Reference GC (%)	35.44	35.44	35.44	35.44	35.44	35.44	35.44	35.44
N50	7659	4466	7625	3834	7644	7654	**7663**	7620
NG50	8517	4608	8496	3887	8464	8474	**8522**	8483
N75	3546	2270	3525	1987	3516	3524	**3547**	3515
NG75	**4365**	2428	4325	2039	4295	4296	**4365**	4306
L50	3560	6259	3588	7341	3641	3641	**3559**	3588
LG50	3047	5940	3075	7189	3148	3149	**3046**	3082
L75	8761	14393	8811	16553	8852	8849	**8759**	8818
LG75	7178	13468	7224	16114	7320	7320	**7176**	7247
# misassemblies	1232	4500	1194	1484	**1190**	1220	1233	1219
# misassembled contigs	1178	4040	1142	1435	**1132**	1160	1180	1165
Misassembled contigs length	8966387	18042488	8729971	**5420189**	8771615	8913241	8958822	8888451
# local misassemblies	257	211	253	**202**	269	271	256	273
# unaligned mis. contigs	3	7	3	**2**	**2**	**2**	3	**2**
# unaligned contigs	4921 + 67 part	**4196 + 126 part**	4969 + 53 part	4590 + 57 part	5035 + 56 part	5035 + 57 part	4922 + 67 part	4955 + 64 part
Unaligned length	16560504	14110799	16496602	**13653429**	16286086	16288035	16560915	16494872
Genome fraction (%)	91.300	86.965	**91.318**	87.248	91.229	91.231	91.302	91.222
Duplication ratio	1.005	1.021	1.005	**1.004**	1.005	1.005	1.005	1.005
# N's per 100 kbp	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
# mismatches per 100 kbp	23.69	105.56	**23.11**	27.41	23.40	23.35	23.58	24.24
# indels per 100 kbp	5.88	18.20	**5.79**	6.64	5.90	5.94	5.86	5.94
Largest alignment	54032	27212	54032	27155	**64367**	**64367**	54032	54032
Total aligned length	91822498	87395766	**91852586**	87657662	91748458	91739544	91822652	91746223
NA50	5612	2996	5598	2925	**5613**	5609	5610	5561
NGA50	6325	3116	6307	2969	6297	6295	**6328**	6281
NA75	1916	1148	1915	1229	**1938**	**1938**	1918	1904
NGA75	**2713**	1300	2702	1282	2688	2688	**2713**	2674
LA50	5036	9332	5045	9570	**5018**	5019	5035	5066
LGA50	4341	8858	4351	9370	4352	4353	**4340**	4379
LA75	13065	22879	13084	22740	**12996**	**12996**	13062	13147
LGA75	10358	21103	10388	22035	10412	10415	**10356**	10463

Table B.10 Assembly quality metrics for D4

Assembly	Uncorrected	ACE	BFC	BLESS2	Hybrid	Karect	BrownieCorrector	Reckoner
# contigs (≥ 0 bp)	94428	98953	92673	98200	92770	93552	94263	102274
# contigs (≥ 1000 bp)	5744	7518	5816	8256	5478	5570	5641	6469
# contigs (≥ 5000 bp)	2927	4348	3019	4686	2685	2771	2859	3415
# contigs (≥ 10000 bp)	2228	3088	2286	3275	2064	2129	2166	2517
# contigs (≥ 25000 bp)	1289	1412	1321	1366	1250	1280	1272	1383
# contigs (≥ 50000 bp)	687	524	674	455	687	695	690	656
Total length (≥ 0 bp)	130002600	128372226	129827421	127011744	129810748	129881140	129985751	**130506403**
Total length (≥ 1000 bp)	119818544	118251920	119812405	116972812	119873419	**119877304**	119808428	119794237
Total length (≥ 5000 bp)	113581978	110643148	113604067	108274776	**113773726**	113762300	113688120	112852066
Total length (≥ 10000 bp)	108530954	101505638	108274117	98172191	**109260361**	109119112	108675878	106389897
Total length (≥ 25000 bp)	93186659	74193597	92453290	67343513	**95998464**	95254278	94026163	88095244
Total length (≥ 50000 bp)	71303295	42715781	69206434	35471350	**75528880**	73949504	72922308	62048435
# contigs	8394	9886	8486	10424	8037	8145	8281	9166
Largest contig	481990	333642	481989	287346	517988	**517989**	481989	454387
Total length	121670559	119897165	121671950	118480841	121661557	**121679265**	121652919	121676103
Reference length	120381546	120381546	120381546	120381546	120381546	120381546	120381546	120381546
GC (%)	42.57	42.53	42.57	42.52	42.58	42.58	42.57	42.55
Reference GC (%)	42.42	42.42	42.42	42.42	42.42	42.42	42.42	42.42
N50	65546	35232	61819	30326	**75991**	71129	69567	51134
NG50	66349	35003	62540	29666	**76873**	72000	70449	52030
N75	27402	16276	26503	14275	**31163**	29995	28393	21974
NG75	28467	16087	27640	13542	**32518**	31195	29583	22982
L50	505	935	524	1071	**446**	473	485	633
LG50	496	942	514	1103	**438**	464	476	620
L75	1216	2191	1275	2496	**1082**	1136	1169	1518
LG75	1181	2214	1239	2599	**1052**	1104	1136	1475
# misassemblies	814	964	829	997	804	**799**	805	879
# misassembled contigs	635	820	650	870	**623**	628	629	705
Misassembled contigs length	43763416	31185704	43389755	**29155670**	47975869	45770362	44692094	41789688
# local misassemblies	1310	**1248**	1307	1268	1292	1298	1300	1321
# unaligned mis. contigs	28	**19**	27	22	31	31	28	21
# unaligned contigs	3858 + 378 part	2942 + 430 part	3844 + 384 part	**2440 + 373 part**	3824 + 384 part	3840 + 373 part	3852 + 378 part	3853 + 373 part
Unaligned length	8132812	6534991	8141896	**5455940**	8172885	8173915	8132689	8117761
Genome fraction (%)	93.913	93.622	93.907	93.481	93.915	**93.916**	93.910	93.890
Duplication ratio	**1.004**	1.006	**1.004**	**1.004**	**1.004**	**1.004**	**1.004**	1.005
# N's per 100 kbp	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
# mismatches per 100 kbp	551.02	561.24	551.22	560.94	**548.74**	549.34	548.84	550.02
# indels per 100 kbp	131.61	133.42	131.69	132.59	131.78	131.70	131.41	**130.95**
Largest alignment	406513	188221	280215	167357	**428481**	**428481**	406512	367383
Total aligned length	**113196867**	112830607	113186685	112666190	113178942	113185629	113187949	113182400
NA50	50135	29320	48237	25498	**55055**	53564	51322	41401
NGA50	50833	29126	49089	25133	**56046**	54106	52152	41977
NA75	21271	13471	20789	12059	**23983**	23202	22189	17553
NGA75	22149	13335	21467	11383	**24872**	23983	22983	18369
LA50	656	1110	680	1263	**600**	623	639	800
LGA50	643	1119	667	1300	**588**	611	627	785
LA75	1573	2616	1635	2929	**1429**	1479	1527	1913
LGA75	1529	2643	1589	3050	**1390**	1437	1484	1859

Table B.11 Assembly quality metrics for D5

Assembly	Uncorrected	ACE	BFC	BLESS2	Hybrid	Karect	BrownieCorrector	Reckoner
# contigs (≥ 0 bp)	170491	218463	192792	204579	188520	189526	170354	224879
# contigs (≥ 1000 bp)	7103	10411	8931	11456	8433	8613	6826	9131
# contigs (≥ 5000 bp)	3782	5247	4591	5573	4294	4399	3605	4692
# contigs (≥ 10000 bp)	2659	3014	2901	3072	2741	2794	2557	2928
# contigs (≥ 25000 bp)	1336	1059	1226	958	1196	1203	1329	1217
# contigs (≥ 50000 bp)	575	353	469	321	505	502	595	462
Total length (≥ 0 bp)	136008271	136884556	137701577	135338694	137359979	137407270	136022594	**139938723**
Total length (≥ 1000 bp)	118770197	115965322	118413811	115640342	118358084	118349944	**118780111**	118408852
Total length (≥ 5000 bp)	110806731	102620615	107643770	100398087	108148059	107936808	**111106120**	107416341
Total length (≥ 10000 bp)	102774604	86665062	95540225	82570772	97010422	96455953	**103631979**	94709945
Total length (≥ 25000 bp)	81170209	55974685	68911764	49661452	72438499	71191324	**83690560**	67424844
Total length (≥ 50000 bp)	54276219	31966506	42454776	27765715	48170727	46747732	**57723562**	41149200
# contigs	9227	12936	11267	13893	10631	10821	8939	11501
Largest contig	535439	**579123**	479301	348825	579114	459363	579114	330198
Total length	120260659	117783200	120069086	117402937	119916284	119916237	**120263583**	120090128
Reference length	120381546	120381546	120381546	120381546	120381546	120381546	120381546	120381546
GC (%)	42.42	42.52	42.43	42.53	42.44	42.43	42.42	42.42
Reference GC (%)	42.42	42.42	42.42	42.42	42.42	42.42	42.42	42.42
N50	43681	23183	31663	19964	36541	34845	**47564**	30523
NG50	43568	22418	31484	19004	36446	34551	**47534**	30359
N75	18624	9438	12639	8321	13701	13265	**19903**	12158
NG75	18613	8787	12489	7717	13517	13027	**19868**	12052
L50	700	1181	910	1364	782	820	**645**	950
LG50	702	1238	915	1441	788	827	**646**	955
L75	1753	3186	2414	3675	2135	2226	**1620**	2509
LG75	1758	3400	2433	3953	2161	2253	**1624**	2527
# misassemblies	751	802	744	757	**733**	735	742	768
# misassembled contigs	624	715	633	687	615	630	**610**	664
Misassembled contigs length	34931702	25507758	28643451	**21355751**	31273118	30251303	37027268	28868487
# local misassemblies	1112	**1011**	1066	1012	1071	1069	1119	1071
# unaligned mis. contigs	27	30	28	**18**	28	31	26	31
# unaligned contigs	3151 + 365 part	2070 + 380 part	3080 + 369 part	**1892 + 344 part**	2967 + 347 part	2973 + 349 part	3150 + 362 part	3099 + 361 part
Unaligned length	6896001	4890803	6910766	**4596029**	6739536	6742165	6889509	6879138
Genome fraction (%)	93.781	93.282	93.578	93.194	93.585	93.595	**93.789**	93.582
Duplication ratio	**1.004**	1.005	1.005	1.006	1.005	**1.004**	**1.004**	1.005
# N's per 100 kbp	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
# mismatches per 100 kbp	501.29	499.27	498.06	501.93	**495.52**	496.83	498.78	498.23
# indels per 100 kbp	119.51	117.33	118.14	**117.24**	118.01	118.21	119.22	118.08
Largest alignment	402868	313000	334917	333479	334919	334338	**402877**	314918
Total aligned length	113092386	112548843	112881644	112507933	112929042	112914657	**113119011**	112930607
NA50	35938	20690	27488	17825	30759	29547	**38675**	26330
NGA50	35924	20032	27365	17133	30557	29286	**38670**	26296
NA75	15941	8353	11070	7535	11943	11648	**16869**	10658
NGA75	15889	7768	10966	6898	11792	11491	**16840**	10579
LA50	865	1373	1086	1556	956	990	**813**	1128
LGA50	866	1436	1092	1641	964	998	**815**	1134
LA75	2115	3624	2802	4096	2516	2602	**1979**	2917
LGA75	2120	3866	2823	4406	2546	2633	**1984**	2938

Table B.12 Assembly quality metrics for D6

Assembly	Uncorrected	ACE	BFC	BLESS2	Hybrid	Karect	BrownieCorrector	Reckoner
# contigs (≥ 0 bp)	153426	169680	152423	159228	156417	157208	153325	169100
# contigs (≥ 1000 bp)	5274	5881	5208	6509	4992	5200	5007	5377
# contigs (≥ 5000 bp)	3398	3767	3381	4038	3196	3344	3196	3452
# contigs (≥ 10000 bp)	2530	2730	2516	2841	2400	2501	2389	2565
# contigs (≥ 25000 bp)	1330	1302	1311	1313	1288	1307	1286	1307
# contigs (≥ 50000 bp)	589	523	588	484	603	584	612	591
Total length (≥ 0 bp)	122790524	123570159	122779397	122265471	123023997	123085625	122779928	**123950644**
Total length (≥ 1000 bp)	108694789	108396090	**108822375**	107698581	108735045	108743759	108694535	108777289
Total length (≥ 5000 bp)	104074942	103085055	**104321365**	101443382	104314160	104125163	104260540	103974120
Total length (≥ 10000 bp)	97777795	95556603	98031267	92759620	**98505752**	97989568	98403336	97527917
Total length (≥ 25000 bp)	78164544	72474598	78262483	68171653	80366270	78482167	**80445350**	76741954
Total length (≥ 50000 bp)	51919360	44842751	52621466	39368697	55944501	52808127	**56505536**	51275596
# contigs	6731	7363	6608	8051	6416	6647	6436	6809
Largest contig	378346	360514	337032	263399	**403164**	358676	360062	292751
Total length	109681747	109400465	**109771235**	108742773	109698307	109722386	109662957	109743868
Reference length	119668634	119668634	119668634	119668634	119668634	119668634	119668634	119668634
GC (%)	35.96	35.98	35.96	36.00	35.97	35.96	35.96	35.96
Reference GC (%)	36.06	36.06	36.06	36.06	36.06	36.06	36.06	36.06
N50	46858	40291	47899	34704	51194	48131	**52341**	46233
NG50	41749	35499	42189	30758	45986	42290	**46094**	40862
N75	21989	18318	22072	16335	23637	22196	**23725**	21324
NG75	15914	13442	16240	11682	**16977**	16014	16925	15642
L50	650	743	635	844	582	626	**580**	666
LG50	762	878	745	1012	684	736	**682**	780
L75	1504	1746	1484	1972	1367	1468	**1360**	1547
LG75	1903	2236	1873	2565	1737	1859	**1733**	1953
# misassemblies	149	138	157	174	**112**	129	118	137
# misassembled contigs	148	136	154	168	**111**	123	117	132
Misassembled contigs length	6392285	5198138	8078221	4928846	**4684101**	5431837	5236199	6091039
# local misassemblies	50	**43**	44	69	50	52	46	46
# unaligned mis. contigs	2	2	2	**1**	**1**	2	2	2
# unaligned contigs	238 + 21 part	159 + 23 part	223 + 24 part	**93 + 30 part**	231 + 27 part	235 + 27 part	237 + 22 part	232 + 25 part
Unaligned length	406289	241175	405558	**184809**	394256	396474	404121	406267
Genome fraction (%)	91.207	91.079	**91.283**	90.628	91.253	91.255	91.215	91.255
Duplication ratio	**1.001**	1.002	**1.001**	**1.001**	**1.001**	**1.001**	**1.001**	**1.001**
# N's per 100 kbp	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
# mismatches per 100 kbp	20.01	25.24	19.63	22.12	18.53	19.60	**18.22**	18.97
# indels per 100 kbp	5.00	7.27	4.94	6.27	4.81	4.83	**4.75**	4.85
Largest alignment	378346	360476	322214	263398	**402910**	358628	360023	292696
Total aligned length	109234355	109062189	**109324450**	108528876	109272392	109287962	109225834	109293988
NA50	45861	39070	45958	33942	50386	46858	**51386**	45215
NGA50	40802	34273	40910	29968	**45423**	41391	45400	39605
NA75	21247	17857	21475	15900	**23153**	21665	22895	20809
NGA75	15318	13110	15574	11412	**16504**	15526	16405	15190
LA50	665	760	658	862	**591**	638	**591**	683
LGA50	781	900	772	1033	695	751	**694**	801
LA75	1544	1793	1537	2019	**1388**	1501	**1388**	1590
LGA75	1958	2296	1942	2628	**1769**	1906	1773	2007

Table B.13 Assembly quality metrics for D7

Assembly	Uncorrected	ACE	BFC	BLESS2	Hybrid	Karect	BrownieCorrector	Reckoner
# contigs (≥ 0 bp)	84104	87785	82976	85783	82832	83398	84087	92296
# contigs (≥ 1000 bp)	4269	4797	4408	4424	4195	4220	4221	5118
# contigs (≥ 5000 bp)	2128	2801	2230	2527	2012	2071	2099	2743
# contigs (≥ 10000 bp)	1600	2088	1665	1913	1486	1531	1566	2031
# contigs (≥ 25000 bp)	1013	1231	1045	1161	967	985	1004	1218
# contigs (≥ 50000 bp)	653	701	666	691	633	649	644	705
Total length (≥ 0 bp)	130506649	128784476	130252545	127655011	130240850	130280733	130480461	**130996905**
Total length (≥ 1000 bp)	**121778040**	120151535	121628181	119271937	121725554	121710656	121748117	121706993
Total length (≥ 5000 bp)	116980449	115355940	116748510	114676443	116909862	116920459	**117020950**	116233150
Total length (≥ 10000 bp)	113202715	110296487	112661638	110309610	113145023	113052834	**113234024**	111121525
Total length (≥ 25000 bp)	103664419	96387206	102536172	98158594	**104756954**	104285446	104149759	97981828
Total length (≥ 50000 bp)	90541399	77320552	88785758	81304930	**92631370**	92090502	91215132	79688544
# contigs	6250	6413	6427	5633	6104	6140	6209	7163
Largest contig	754587	545648	667406	602677	829779	829776	855558	**887190**
Total length	**123171040**	121267612	123042637	120101070	123066199	123060468	123143110	123138443
Reference length	120381546	120381546	120381546	120381546	120381546	120381546	120381546	120381546
GC (%)	42.55	42.49	42.54	42.46	42.55	42.55	42.54	42.52
Reference GC (%)	42.42	42.42	42.42	42.42	42.42	42.42	42.42	42.42
N50	120821	74655	113294	89691	**130622**	126171	124369	79327
NG50	123823	75089	114878	89541	**135170**	129017	128600	81109
N75	47031	32264	44284	37050	**50374**	49513	47556	32089
NG75	50427	33004	48111	36516	**54468**	53085	51344	34919
L50	284	431	303	369	**261**	274	276	413
LG50	273	425	291	371	**250**	263	265	396
L75	692	1041	740	896	**627**	654	668	1019
LG75	648	1021	697	902	**588**	614	626	957
# misassemblies	1048	1132	1050	1206	1014	**1001**	1029	1143
# misassembled contigs	668	809	665	791	**640**	642	661	769
Misassembled contigs length	66619419	57679056	64061457	60021320	68314711	66950147	66893425	**57223037**
# local misassemblies	1572	**1480**	1515	1548	1506	1536	1562	1655
# unaligned mis. contigs	135	127	127	**96**	118	123	120	122
# unaligned contigs	3288 + 593 part	2361 + 651 part	3294 + 629 part	**1838 + 541 part**	3296 + 640 part	3267 + 637 part	3286 + 598 part	3311 + 622 part
Unaligned length	8782559	7074733	8683828	**6035908**	8759389	8758466	8761387	8732208
Genome fraction (%)	**94.496**	94.284	94.491	94.241	94.491	94.468	94.494	94.444
Duplication ratio	1.006	1.006	**1.005**	**1.005**	**1.005**	**1.005**	1.006	1.006
# N's per 100 kbp	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
# mismatches per 100 kbp	558.01	567.44	558.79	565.53	556.90	**556.23**	556.57	558.35
# indels per 100 kbp	134.40	140.90	134.58	145.25	134.03	134.02	134.08	**133.68**
Largest alignment	486030	417035	428461	338930	**547379**	547363	447086	406192
Total aligned length	**114077574**	113784885	114050875	113749971	114018850	114005038	114071871	114037057
NA50	79151	54862	74311	61784	**86507**	82808	80392	56396
NGA50	80752	55391	77526	61609	**89065**	85226	83397	58176
NA75	32513	23228	30391	25843	**35224**	34107	33310	23753
NGA75	35171	23744	33021	25562	**37912**	36680	35880	25771
LA50	429	602	453	545	**406**	416	422	585
LGA50	412	593	435	547	**391**	400	405	561
LA75	1042	1443	1097	1291	**971**	997	1020	1417
LGA75	980	1415	1034	1299	**915**	940	961	1334

Table B.14 Assembly quality metrics for D8

Assembly	Uncorrected	ACE	BFC	BLESS2	Hybrid	Karect	BrownieCorrector	Reckoner
# contigs (≥ 0 bp)	153257	193843	171323	177917	167024	168273	153096	203849
# contigs (≥ 1000 bp)	4205	4162	4361	4367	4019	4123	3990	4569
# contigs (≥ 5000 bp)	2010	2450	2147	2596	1903	1960	1853	2326
# contigs (≥ 10000 bp)	1473	1845	1572	1909	1402	1450	1363	1701
# contigs (≥ 25000 bp)	1003	1153	1044	1206	975	992	944	1107
# contigs (≥ 50000 bp)	666	675	677	712	676	672	641	690
Total length (≥ 0 bp)	135887496	136343035	137423351	134765843	136961075	137049756	135884144	**139590516**
Total length (≥ 1000 bp)	120696663	118502898	120701123	118317880	120560238	120537455	**120706033**	120624736
Total length (≥ 5000 bp)	115654547	114306741	115613997	113904254	115684153	115538672	**115826462**	115470305
Total length (≥ 10000 bp)	111879129	109943429	111550471	109013641	112117292	111917113	**112393212**	111018421
Total length (≥ 25000 bp)	104315319	98751430	102975275	97606444	105279669	104552044	**105746772**	101450513
Total length (≥ 50000 bp)	92213430	81384118	89809178	79738854	94488203	92912592	**94775486**	86538407
# contigs	5706	5518	5821	5517	5350	5482	5466	6036
Largest contig	740414	714940	690671	676184	929948	829272	**949798**	715670
Total length	**121725596**	119424679	121699360	119105224	121471450	121469690	121719241	121630853
Reference length	120381546	120381546	120381546	120381546	120381546	120381546	120381546	120381546
GC (%)	42.38	42.47	42.38	42.47	42.39	42.39	42.38	42.38
Reference GC (%)	42.42	42.42	42.42	42.42	42.42	42.42	42.42	42.42
N50	125354	92982	112004	82650	127431	123975	**137513**	101811
NG50	126353	92254	113406	81120	129321	125108	**139117**	103109
N75	51706	38805	47534	36572	**57471**	54567	57390	41325
NG75	53948	37895	49131	35352	**59245**	55512	58973	42751
L50	270	355	297	397	261	271	**239**	333
LG50	265	360	291	405	257	267	**234**	327
L75	648	861	708	936	614	638	**577**	794
LG75	629	880	687	963	600	623	**559**	772
# misassemblies	892	919	876	**860**	870	864	894	936
# misassembled contigs	581	648	576	656	562	577	**561**	634
Misassembled contigs length	64530411	54903095	59072274	**50276098**	63420552	63011931	68069287	57826313
# local misassemblies	1487	1426	1526	1432	1451	**1422**	1493	1482
# unaligned mis. contigs	93	90	112	99	99	104	**89**	97
# unaligned contigs	2813 + 500 part	1718 + 495 part	2666 + 539 part	**1512 + 466 part**	2558 + 510 part	2563 + 512 part	2802 + 506 part	2694 + 510 part
Unaligned length	7569584	5423140	7538412	**5168064**	7336591	7325032	7588629	7445130
Genome fraction (%)	**94.400**	94.223	94.357	94.163	94.379	94.392	94.394	94.358
Duplication ratio	1.005	1.005	1.005	1.005	1.005	1.005	**1.004**	1.005
# N's per 100 kbp	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
# mismatches per 100 kbp	505.96	503.80	501.44	505.43	501.69	**501.23**	504.97	502.37
# indels per 100 kbp	120.43	119.21	**118.69**	121.82	118.72	118.84	120.22	118.87
Largest alignment	453436	453475	440801	402865	517576	**587714**	508703	453456
Total aligned length	113902949	113689117	113904771	113665384	113890437	113897046	113888974	**113906686**
NA50	83845	65926	78250	61557	86556	85122	**87799**	70604
NGA50	85003	65163	78985	60574	87822	85881	**88877**	71724
NA75	35457	29274	33797	28409	38951	37932	**39505**	30157
NGA75	36760	28366	34986	27330	40266	38829	**41434**	31616
LA50	419	507	433	529	399	403	**391**	470
LGA50	411	515	424	539	393	397	**384**	461
LA75	972	1183	1021	1234	911	936	**896**	1119
LGA75	944	1208	993	1268	890	915	**871**	1089

Table B.15 Assembly quality metrics for D9

Assembly	Uncorrected	ACE	BFC	BLESS2	Hybrid	Karect	BrownieCorrector	Reckoner
# contigs (≥ 0 bp)	142998	158391	142712	148192	146017	147021	142974	158545
# contigs (≥ 1000 bp)	4033	4220	3993	4545	3926	4013	3896	4284
# contigs (≥ 5000 bp)	2412	2542	2430	2705	2300	2377	2292	2631
# contigs (≥ 10000 bp)	1802	1898	1808	1985	1717	1777	1711	1982
# contigs (≥ 25000 bp)	1109	1099	1113	1140	1048	1081	1062	1198
# contigs (≥ 50000 bp)	638	636	633	618	621	631	620	660
Total length (≥ 0 bp)	123218064	123981967	123224625	122885449	123518075	123539458	123245160	**124331702**
Total length (≥ 1000 bp)	110256731	109999806	110353964	109476185	**110361604**	110330888	110301171	110311214
Total length (≥ 5000 bp)	106152731	105686683	**106414045**	104796912	106244053	106169772	106242186	106088913
Total length (≥ 10000 bp)	101804805	101043252	101979115	99699235	102079432	101873514	**102085247**	101431478
Total length (≥ 25000 bp)	90414664	87963335	90363257	85957292	91028806	90379967	**91415021**	88457127
Total length (≥ 50000 bp)	73498519	71549764	73040202	67808341	**75725947**	74247235	75503677	69145241
# contigs	5138	5334	5059	5669	4997	5078	4983	5346
Largest contig	513726	548881	636585	608766	638980	604382	**638981**	568204
Total length	111015037	110759642	111083449	110246718	**111096537**	111062930	111043967	111041653
Reference length	119668634	119668634	119668634	119668634	119668634	119668634	119668634	119668634
GC (%)	35.94	35.96	35.95	35.96	35.95	35.95	35.94	35.94
Reference GC (%)	36.06	36.06	36.06	36.06	36.06	36.06	36.06	36.06
N50	82512	78856	83316	71534	**94701**	86750	93967	71963
NG50	75466	70566	74799	61723	**84439**	78085	82118	64130
N75	35434	31830	34731	28802	37344	35373	**37856**	31577
NG75	25874	23118	25874	20720	26758	25818	**27539**	23731
L50	359	380	363	403	325	345	**324**	433
LG50	414	439	417	474	**373**	397	374	496
L75	870	926	877	1018	**796**	845	800	1015
LG75	1083	1174	1089	1308	**999**	1057	**999**	1252
# misassemblies	617	597	596	617	574	**570**	572	603
# misassembled contigs	463	469	454	482	437	438	**428**	452
Misassembled contigs length	27054338	25914859	27793634	**22441602**	26444291	25929567	26374921	23718330
# local misassemblies	457	**432**	453	520	445	437	459	465
# unaligned mis. contigs	9	**7**	**7**	**7**	9	8	**7**	**7**
# unaligned contigs	270 + 142 part	190 + 127 part	254 + 126 part	**142 + 142 part**	257 + 130 part	271 + 121 part	264 + 135 part	265 + 122 part
Unaligned length	677416	**488901**	635261	494795	679750	651220	685268	634358
Genome fraction (%)	91.801	91.738	**91.908**	91.321	91.861	91.856	91.817	91.843
Duplication ratio	**1.004**	**1.004**	**1.004**	**1.004**	**1.004**	**1.004**	**1.004**	1.005
# N's per 100 kbp	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
# mismatches per 100 kbp	51.77	58.12	52.17	58.54	51.78	52.60	51.13	**50.56**
# indels per 100 kbp	10.92	15.70	10.85	21.11	10.38	10.59	10.35	**10.34**
Largest alignment	513722	527818	540557	608680	635802	604382	**635803**	567176
Total aligned length	110257880	110165936	**110362885**	109666875	110338430	110333035	110274459	110310731
NA50	74168	70306	72208	64696	**84946**	77951	82261	62281
NGA50	65138	62161	64709	55639	**74620**	68873	71788	56734
NA75	30088	27435	30288	25407	32098	31063	**32833**	27426
NGA75	22319	20022	22827	17956	23545	22847	**23679**	21034
LA50	398	418	409	440	**359**	384	361	479
LGA50	461	485	472	519	**413**	443	417	552
LA75	984	1047	1003	1137	**900**	958	906	1141
LGA75	1235	1331	1251	1469	**1136**	1202	1140	1410

B.5.5 Full Quast report (scaffolds)

This section ontains the Quast evaluation report of scaffolds after assembling with each dataset with SPAdes. Error correction by ACE, BFC, BLESS2, Brownie, Karect and Reckoner are done before assembling the reads. The Uncorrected column refers to the quality of contigs without any cleaning process. The Hybrid column shows the quality of assembly of reads which are corrected jointly by BrownieCorrector and Karect. Default parameter settings are used for Quast, therefore all statistics are based on contigs of size $\geq$ 500 bp.

B.5.5.1 D1

Table B.16 contains the Quast report after assembling dataset D1 (*Homo sapiens* Chr. 21) with SPAdes. Fig. B.2 shows the corresponding NGAx plot.

B.5.5.2 D2

Table B.17 contains the Quast report after assembling dataset D2 (*Homo sapiens* Chr. 14) with SPAdes. Fig. B.3 shows the corresponding NGAx plot.

B.5.5.3 D3

Table B.18 contains the Quast report after assembling dataset D3 (*C. elegans*) with SPAdes. Fig. B.4 shows the corresponding NGAx plot.

B.5.5.4 D4

Table B.19 contains the Quast report after assembling dataset D4 (*D. melanogaster*) with SPAdes. Fig. B.5 shows the corresponding NGAx plot.

B.5.5.5 D5

Table B.20 contains the Quast report after assembling dataset D5 (*D. melanogaster*) with SPAdes. Fig. B.6 shows the corresponding NGAx plot.

B.5.5.6 D6

Table B.21 contains the Quast report after assembling dataset D6 (*A. thaliana*) with SPAdes. Fig. B.7 shows the corresponding NGAx plot.

B.5.5.7 D7

Table B.22 contains the Quast report after a hybrid assembly of dataset D7 (*D. melanogaster*) with SPAdes. This is a hybrid assembly in which the corrected

(and uncorrected) Illumina reads (R4) are complemented with the Pacbio reads (P1). Fig. B.8 shows the corresponding NGAx plot.

B.5.5.8 D8

Table B.23 contains the Quast report after assembling dataset D8 *D. melanogaster* with SPAdes. This is a hybrid assembly in which the corrected (and uncorrected) Illumina reads (R5) are complemented with the Pacbio reads (P1). Fig. B.9 shows the corresponding NGAx plot.

B.5.5.9 D9

Table B.24 contains the Quast report after assembling dataset D9 (*A. thaliana*) with SPAdes. This is a hybrid assembly in which the corrected (and uncorrected) Illumina reads (R6) are complemented with the Pacbio reads (P2). Fig. B.10 shows the corresponding NGAx plot.

Table B.16 Assembly quality metrics for D1

Assembly	Uncorrected	ACE	BFC	BLESS2	Hybrid	Karect	BrownieCorrector	Reckoner
# contigs (≥ 0 bp)	15503	17167	15110	18380	16178	17003	14341	21252
# contigs (≥ 1000 bp)	3498	3309	3284	3749	2813	3063	2907	4052
# contigs (≥ 5000 bp)	1951	1912	1876	2014	1724	1826	1748	2036
# contigs (≥ 10000 bp)	1147	1130	1131	1096	1127	1133	1126	1105
# contigs (≥ 25000 bp)	251	281	281	222	344	313	338	193
# contigs (≥ 50000 bp)	25	29	32	14	56	38	52	12
Total length (≥ 0 bp)	34398184	34481028	34385863	34366932	34506906	34561810	34339292	**34763727**
Total length (≥ 1000 bp)	32641079	32638021	32675090	32371042	**32712198**	32700154	32706081	32510652
Total length (≥ 5000 bp)	28708375	28999697	29107505	27838280	**29942475**	29527995	29774056	27339748
Total length (≥ 10000 bp)	22963404	23400423	23768650	21227019	**25630161**	24533233	25273512	20609468
Total length (≥ 25000 bp)	9022632	10086964	10372635	7729872	**13215874**	11547907	12926905	6697434
Total length (≥ 50000 bp)	1506449	1830857	1990598	877864	**3485559**	2401241	3333176	726268
# contigs	4066	3815	3837	4331	3290	3561	3388	4700
Largest contig	82702	98568	**109124**	82889	105146	92853	105053	80324
Total length	33047367	32998806	**33069170**	32792092	33051550	33057202	33049826	32976233
Reference length	40988574	40988574	40988574	40988574	40988574	40988574	40988574	40988574
GC (%)	40.73	40.75	40.73	40.68	40.76	40.75	40.75	40.62
Reference GC (%)	40.93	40.93	40.93	40.93	40.93	40.93	40.93	40.93
N50	15702	16575	17050	14101	**20504**	18195	19684	13097
NG50	11992	12721	13085	10604	**15454**	14089	14798	10037
N75	8222	8669	8917	7445	**10923**	9729	10416	6938
NG75	2972	3263	3143	2600	**3857**	3531	3671	2464
L50	633	601	581	689	**492**	548	502	746
LG50	922	874	844	1025	**715**	794	734	1094
L75	1348	1276	1241	1486	**1047**	1160	1079	1599
LG75	2470	2339	2284	2806	**1907**	2114	1975	2987
# misassemblies	173	149	186	174	111	**102**	141	146
# misassembled contigs	164	137	169	165	104	**92**	132	141
Misassembled contigs length	2890519	2407563	3535938	2710908	2289710	**2152215**	2871345	2157464
# local misassemblies	219	228	**209**	247	262	248	226	228
# unaligned mis. contigs	0	0	0	0	0	0	0	0
# unaligned contigs	79 + 8 part	59 + 8 part	76 + 8 part	**47 + 9 part**	66 + 8 part	64 + 6 part	77 + 9 part	67 + 5 part
Unaligned length	86668	69281	88409	**55386**	75372	73641	86573	73305
Genome fraction (%)	80.055	79.987	80.096	79.572	**80.221**	80.188	80.201	79.865
Duplication ratio	1.004	1.004	1.005	1.004	**1.003**	1.004	**1.003**	1.005
# N's per 100 kbp	49.01	60.56	**48.55**	83.12	72.77	67.20	51.37	57.92
# mismatches per 100 kbp	166.81	163.82	162.11	164.19	**152.47**	155.03	157.12	158.29
# indels per 100 kbp	36.34	36.37	35.73	35.69	35.94	35.79	36.58	**33.18**
Largest alignment	82666	98563	96583	80972	104946	92848	**104953**	80324
Total aligned length	32885932	32847613	32911743	32670138	**32927101**	32919824	32920529	32799489
NA50	14909	15985	16161	13519	**19844**	17626	18884	12630
NGA50	11377	12135	12294	10034	**14613**	13528	14155	9670
NA75	7740	8245	8290	7020	**10287**	9290	9838	6609
NGA75	2737	2944	2867	2411	**3622**	3345	3485	2297
LA50	661	625	613	716	**509**	567	523	769
LGA50	965	912	894	1068	**742**	823	766	1131
LA75	1416	1336	1317	1553	**1090**	1204	1130	1659
LGA75	2617	2463	2438	2959	**1999**	2206	2083	3127

Figure B.2 SPAdes assembly results for dataset D1 (*Homo sapiens* Chr. 21) for both uncorrected and corrected data. Contigs with length NGAx or larger produce x% of the genome.

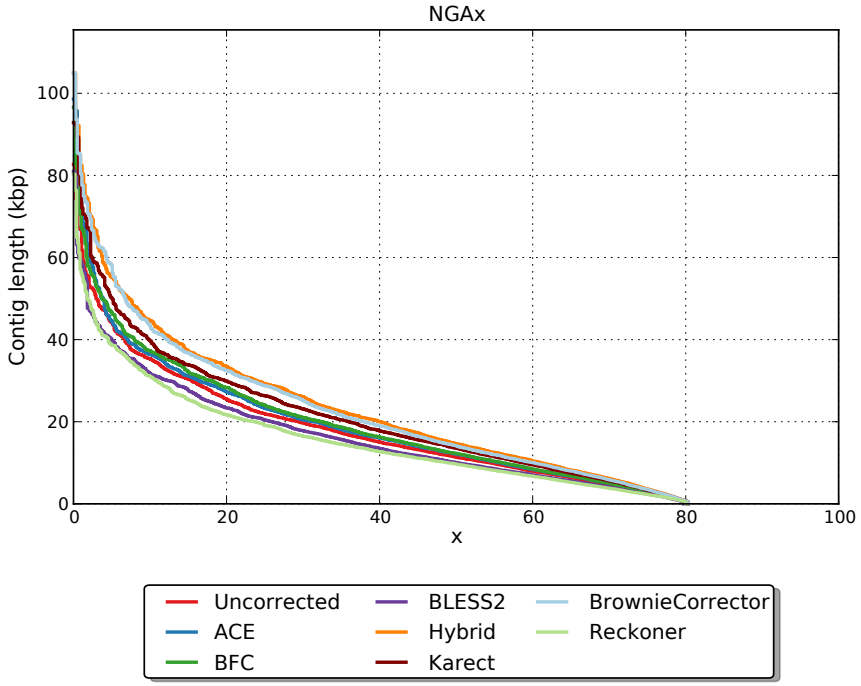


Table B.17 Assembly quality metrics for D2

Assembly	Uncorrected	ACE	BFC	BLESS2	Hybrid	Karect	BrownieCorrector	Reckoner
# contigs (≥ 0 bp)	412300	63904	61280	65498	54926	59015	85814	82873
# contigs (≥ 1000 bp)	14739	10310	9552	11193	7617	9004	8477	12975
# contigs (≥ 5000 bp)	5460	5223	5064	5399	4603	4971	4930	5473
# contigs (≥ 10000 bp)	2137	2714	2776	2629	2798	2742	2792	2424
# contigs (≥ 25000 bp)	184	503	600	408	804	655	710	294
# contigs (≥ 50000 bp)	4	42	65	27	134	76	95	14
Total length (≥ 0 bp)	**111749252**	88521825	88702085	88492512	88275916	88464779	90639963	89672578
Total length (≥ 1000 bp)	81372306	82954897	83332150	82608365	**83614800**	83395425	83495313	82247586
Total length (≥ 5000 bp)	57788314	69389231	71436004	67364079	**75477644**	72512288	74019938	62869914
Total length (≥ 10000 bp)	34244755	51349355	54876683	47470113	**62391757**	56434088	58583750	41338299
Total length (≥ 25000 bp)	5779031	17392790	21535238	13906135	**31052540**	23670662	26095142	9724112
Total length (≥ 50000 bp)	224901	2538630	3934138	1701829	**8524714**	4647418	5858991	830500
# contigs	17855	11720	10815	12831	8522	10136	9543	15158
Largest contig	68123	100666	**126928**	97813	102135	100683	100685	81231
Total length	83643950	83985139	84256551	83807676	**84267170**	84217172	84265213	83842223
Reference length	107349540	107349540	107349540	107349540	107349540	107349540	107349540	107349540
GC (%)	40.66	40.71	40.71	40.68	40.73	40.72	40.73	40.48
Reference GC (%)	40.89	40.89	40.89	40.89	40.89	40.89	40.89	40.89
N50	8250	13067	14341	11696	**18672**	15461	16361	9825
NG50	5758	9330	10372	8410	**13471**	10946	11777	6921
N75	4159	6750	7542	6069	**9688**	7877	8480	4998
NG75	1185	1918	2157	1702	**2806**	2303	2481	1455
L50	2974	1894	1708	2114	**1317**	1593	1510	2483
LG50	4693	2955	2659	3304	**2045**	2479	2341	3909
L75	6542	4125	3729	4583	**2881**	3499	3293	5476
LG75	13948	8606	7721	9619	**5908**	7222	6700	11552
# misassemblies	119	822	640	720	354	496	**112**	689
# misassembled contigs	119	760	612	679	337	469	**110**	660
Misassembled contigs length	**997983**	10744993	9204509	8534721	7092262	7896311	2186146	6895634
# local misassemblies	480	150	186	165	176	176	349	**132**
# unaligned mis. contigs	**0**	1	**0**	**0**	**0**	**0**	**0**	**0**
# unaligned contigs	**11 + 8 part**	12 + 27 part	14 + 22 part	19 + 19 part	14 + 16 part	13 + 22 part	16 + 5 part	17 + 26 part
Unaligned length	16677	35163	31344	35033	27906	29729	**16274**	38932
Genome fraction (%)	77.427	77.786	78.123	77.707	**78.327**	78.150	78.313	77.523
Duplication ratio	1.006	1.005	1.004	1.004	**1.002**	1.004	**1.002**	1.007
# N's per 100 kbp	59.38	19.79	23.88	20.83	23.75	23.18	44.74	**17.91**
# mismatches per 100 kbp	112.08	127.07	119.92	125.10	105.69	114.90	**101.65**	118.33
# indels per 100 kbp	20.59	21.61	21.57	20.78	21.54	21.28	20.80	**17.48**
Largest alignment	68123	100531	**126784**	93310	102135	100548	100547	70727
Total aligned length	83424246	83718057	84046502	83630149	**84181506**	84037475	84168916	83423730
NA50	8159	12005	13492	11050	**17885**	14695	16172	9345
NGA50	5668	8597	9698	7909	**12795**	10298	11570	6509
NA75	4106	6173	6949	5671	**9168**	7437	8356	4663
NGA75	1126	1714	1953	1553	**2629**	2138	2413	1293
LA50	2996	2021	1796	2222	**1372**	1669	1530	2584
LGA50	4738	3173	2810	3480	**2133**	2612	2374	4091
LA75	6614	4448	3962	4848	**3010**	3693	3339	5757
LGA75	14206	9372	8300	10283	**6202**	7670	6814	12352

Figure B.3 SPAdes assembly results for dataset D2 (*Homo sapiens* Chr. 14) for both uncorrected and corrected data. Contigs with length NGAx or larger produce x% of the genome.

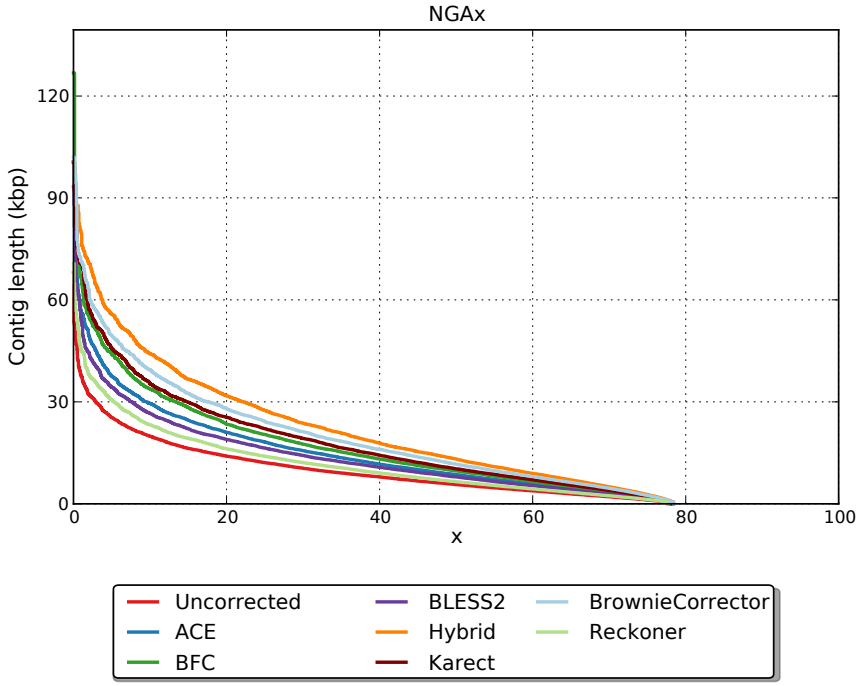


Table B.18 Assembly quality metrics for D3

Assembly	Uncorrected	ACE	BFC	BLESS2	Hybrid	Karect	BrownieCorrector	Reckoner
# contigs (≥ 0 bp)	90335	107935	90891	116183	91250	91070	90434	93218
# contigs (≥ 1000 bp)	19533	25744	19588	27553	19590	19574	19528	19596
# contigs (≥ 5000 bp)	6275	5301	6295	4769	6332	6328	6274	6262
# contigs (≥ 10000 bp)	2407	1217	2398	974	2435	2431	2408	2399
# contigs (≥ 25000 bp)	287	122	302	74	308	310	287	293
# contigs (≥ 50000 bp)	68	23	68	8	59	59	68	66
Total length (≥ 0 bp)	116427720	112565978	116454662	112409711	116147212	116129035	116431324	**116491367**
Total length (≥ 1000 bp)	**104025257**	95726723	103891921	92919344	103553564	103556644	104024612	103767981
Total length (≥ 5000 bp)	72097772	47758221	71982022	40520190	71740186	71720794	**72099815**	71747596
Total length (≥ 10000 bp)	44967590	19948617	44604367	14894822	44301236	44247396	**44980157**	44564052
Total length (≥ 25000 bp)	13938243	5143754	**13939203**	2681884	13170145	13249871	13925728	13879239
Total length (≥ 50000 bp)	6929307	1749408	6340634	478678	5235685	5239938	**6929310**	6619857
# contigs	25948	36175	26126	39647	26140	26112	25944	26150
Largest contig	**244078**	128379	240394	80553	238991	238991	**244078**	240969
Total length	108653017	103255711	108611564	101595435	108285144	108279807	**108653330**	108501463
Reference length	100286070	100286070	100286070	100286070	100286070	100286070	100286070	100286070
GC (%)	38.47	38.30	38.46	38.20	38.41	38.41	38.47	38.46
Reference GC (%)	35.44	35.44	35.44	35.44	35.44	35.44	35.44	35.44
N50	7969	4582	7916	3962	7923	7932	**7971**	7922
NG50	**8827**	4746	8788	4025	8738	8752	**8827**	8779
N75	3702	2326	3682	2037	3674	3679	**3704**	3670
NG75	**4553**	2491	4521	2100	4481	4484	**4553**	4491
L50	3458	6109	3489	7085	3541	3541	**3457**	3488
LG50	**2958**	5791	2990	6921	3061	3062	**2958**	2995
L75	8459	14033	8509	16035	8547	8544	**8457**	8514
LG75	6929	13108	6975	15560	7067	7067	**6928**	6995
# misassemblies	1250	4509	1212	1497	**1207**	1237	1251	1235
# misassembled contigs	1193	4044	1158	1445	**1147**	1174	1195	1178
Misassembled contigs length	9261752	18237042	9018821	**5602595**	9086566	9223611	9254187	9177693
# local misassemblies	402	328	393	**307**	410	412	401	411
# unaligned mis. contigs	4	9	4	3	**2**	**2**	4	3
# unaligned contigs	4505 + 68 part	**3840 + 129 part**	4553 + 54 part	4189 + 58 part	4615 + 58 part	4615 + 59 part	4506 + 68 part	4538 + 65 part
Unaligned length	16604720	14149412	16541258	**13702622**	16331635	16333588	16605131	16539717
Genome fraction (%)	91.309	86.978	**91.329**	87.304	91.240	91.241	91.311	91.232
Duplication ratio	1.005	1.022	1.005	**1.004**	1.005	1.005	1.005	1.005
# N's per 100 kbp	64.44	**58.26**	65.02	63.88	65.48	65.58	64.43	64.62
# mismatches per 100 kbp	23.77	105.67	**23.20**	27.49	23.49	23.43	23.67	24.32
# indels per 100 kbp	6.02	18.39	**5.94**	7.10	6.04	6.09	6.01	6.09
Largest alignment	54032	27212	54032	27155	**64367**	**64367**	54032	54032
Total aligned length	91830875	87408611	**91862827**	87715060	91758254	91749370	91831029	91755248
NA50	5684	3013	5674	2959	**5687**	5681	5681	5643
NGA50	6419	3143	6392	3012	6380	6377	**6420**	6354
NA75	1935	1150	1933	1238	1960	**1962**	1937	1917
NGA75	2753	1306	2739	1299	2729	2729	**2754**	2711
LA50	4987	9285	4997	9468	**4974**	4975	4986	5018
LGA50	4294	8803	4305	9249	4310	4311	**4293**	4332
LA75	12939	22780	12960	22529	**12871**	**12871**	12936	13022
LGA75	10236	20964	10265	21754	10291	10293	**10234**	10340

Figure B.4 SPAdes assembly results for dataset D3 (*C. elegans*) for both uncorrected and corrected data. Contigs with length NGAx or larger produce x% of the genome.

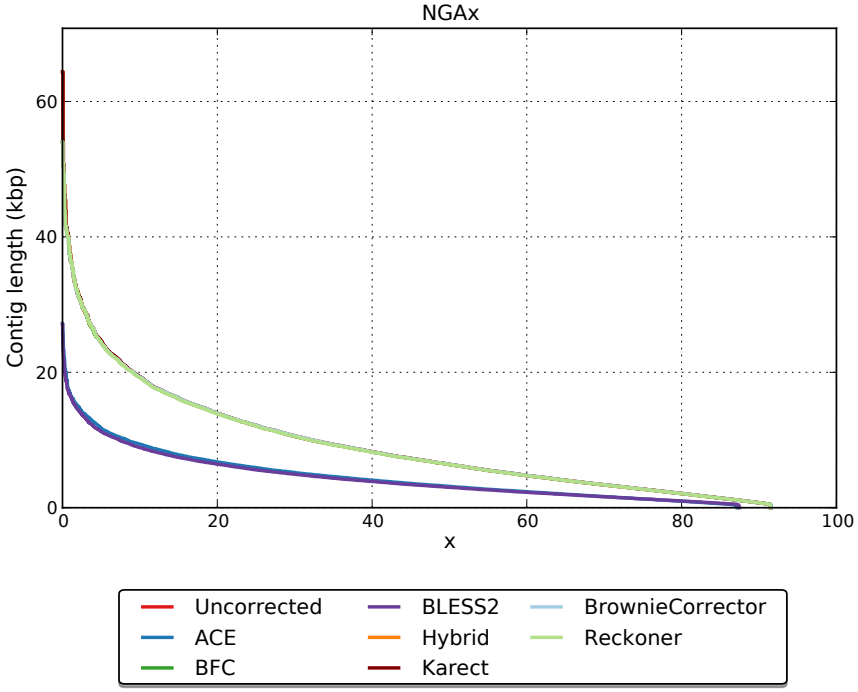


Table B.19 Assembly quality metrics for D4

Assembly	Uncorrected	ACE	BFC	BLESS2	Hybrid	Karect	BrownieCorrector	Reckoner
# contigs (≥ 0 bp)	93227	97585	91434	95653	91669	92410	93123	101162
# contigs (≥ 1000 bp)	5191	6530	5222	6405	5061	5110	5139	6011
# contigs (≥ 5000 bp)	2504	3776	2571	3704	2366	2422	2467	3104
# contigs (≥ 10000 bp)	1928	2775	1971	2737	1812	1849	1889	2305
# contigs (≥ 25000 bp)	1208	1435	1220	1402	1162	1177	1202	1331
# contigs (≥ 50000 bp)	704	624	689	608	682	692	706	692
Total length (≥ 0 bp)	130032928	128403487	129857535	127084174	129839690	129910273	130015593	**130535397**
Total length (≥ 1000 bp)	120113313	118405134	120103149	117350466	120183371	**120186079**	120097364	120089504
Total length (≥ 5000 bp)	114298074	111900442	114332150	110932876	114383380	**114416230**	114335355	113588596
Total length (≥ 10000 bp)	110173410	104680036	110024527	104013656	**110404451**	110319450	110191218	107874824
Total length (≥ 25000 bp)	98369482	82466621	97678502	82061737	**99836443**	99349015	98927274	92069860
Total length (≥ 50000 bp)	79975935	53718772	78310043	54010198	**82227505**	81465937	80907051	69202110
# contigs	7662	8866	7718	8367	7424	7491	7606	8532
Largest contig	518264	333642	518140	439905	517988	517989	518142	**557849**
Total length	121842729	120028744	121845718	118703348	121839229	**121857032**	121823758	121851136
Reference length	120381546	120381546	120381546	120381546	120381546	120381546	120381546	120381546
GC (%)	42.58	42.53	42.58	42.52	42.59	42.59	42.58	42.56
Reference GC (%)	42.42	42.42	42.42	42.42	42.42	42.42	42.42	42.42
N50	83828	43422	80542	44025	**91990**	89577	85463	60794
NG50	85463	43357	82483	43192	**93214**	90787	86717	62202
N75	35616	20093	33969	20060	**37814**	37416	35886	25734
NG75	36690	19951	35304	19140	**39194**	38652	37394	26815
L50	410	759	413	722	**373**	385	401	542
LG50	401	763	404	741	**365**	377	393	530
L75	975	1772	1005	1712	**892**	921	952	1305
LG75	945	1785	974	1777	**864**	892	923	1263
# misassemblies	832	974	844	1006	818	**814**	822	895
# misassembled contigs	622	799	628	823	**610**	616	619	695
Misassembled contigs length	50774126	**37019339**	50754826	38660485	54140750	52300325	51539311	46615389
# local misassemblies	1884	1748	1931	2074	**1732**	1784	1817	1795
# unaligned mis. contigs	29	**21**	28	23	33	32	29	23
# unaligned contigs	3774 + 386 part	2980 + 439 part	3762 + 392 part	**2371 + 381 part**	3721 + 387 part	3739 + 375 part	3772 + 386 part	3774 + 380 part
Unaligned length	8286358	6648237	8296140	**5569650**	8335405	8336511	8284422	8273018
Genome fraction (%)	93.917	93.629	93.912	93.537	93.916	**93.918**	93.914	93.893
Duplication ratio	**1.004**	1.006	**1.004**	1.005	**1.004**	**1.004**	**1.004**	1.005
# N's per 100 kbp	26.10	27.40	25.99	62.44	**24.89**	25.09	25.66	24.99
# mismatches per 100 kbp	547.86	562.17	547.37	559.26	**545.82**	546.09	546.03	546.99
# indels per 100 kbp	130.96	137.03	130.85	134.63	131.23	131.10	130.85	**130.38**
Largest alignment	**446262**	237405	428080	285020	446086	446075	446204	375337
Total aligned length	**113196227**	112863314	113184786	112744813	113175593	113182356	113188396	113182400
NA50	59641	35523	57984	35534	**64663**	62025	60667	46786
NGA50	60714	35425	59124	34856	**65857**	63400	61474	47781
NA75	25472	16277	24546	16539	**27792**	27188	26490	19823
NGA75	26604	16100	25824	15878	**29067**	28124	27588	20635
LA50	560	939	570	906	**524**	533	554	712
LGA50	548	944	557	930	**513**	521	543	697
LA75	1335	2191	1370	2133	**1243**	1265	1314	1707
LGA75	1293	2207	1326	2211	**1205**	1225	1274	1653

Figure B.5 SPAdes assembly results for dataset D4 (*D. melanogaster*) for both uncorrected and corrected data. Contigs with length NGAx or larger produce x% of the genome.

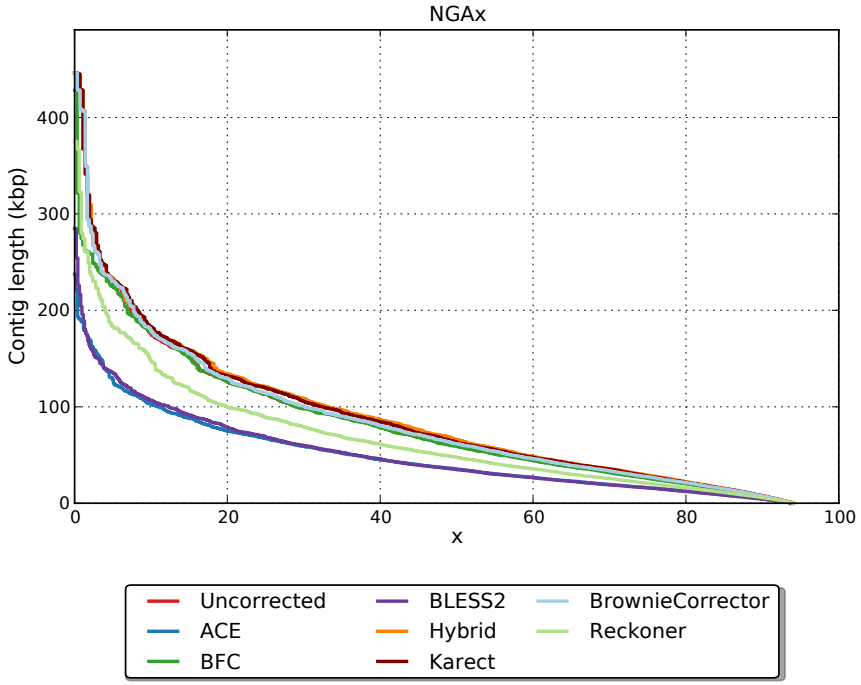


Table B.20 Assembly quality metrics for D5

Assembly	Uncorrected	ACE	BFC	BLESS2	Hybrid	Karect	BrownieCorrector	Reckoner
# contigs (≥ 0 bp)	165126	212433	185876	197317	181524	182557	164967	217940
# contigs (≥ 1000 bp)	4789	5594	5055	6310	4560	4762	4482	5243
# contigs (≥ 5000 bp)	2621	3264	2845	3636	2467	2619	2394	2970
# contigs (≥ 10000 bp)	1936	2476	2100	2698	1833	1937	1778	2193
# contigs (≥ 25000 bp)	1238	1412	1300	1401	1206	1243	1178	1334
# contigs (≥ 50000 bp)	729	660	724	620	742	737	731	724
Total length (≥ 0 bp)	136727305	137556107	138577682	136206452	138250189	138294938	136743703	**140818635**
Total length (≥ 1000 bp)	120037979	116819984	119820170	117054224	119768316	119758171	**120045889**	119817827
Total length (≥ 5000 bp)	114951585	111441304	114683614	110797759	114967414	114815029	**115192459**	114525105
Total length (≥ 10000 bp)	110058903	105831032	109406428	104056408	110451440	109964263	**110829083**	108944878
Total length (≥ 25000 bp)	98578883	88160152	96224991	82586917	100140538	98566386	**101001837**	94721840
Total length (≥ 50000 bp)	80201757	61195626	75734377	54589201	83311554	80477513	**84855860**	72897465
# contigs	7564	8407	8070	8848	7512	7720	7250	8305
Largest contig	635766	633553	593823	382830	693535	573416	**790574**	615528
Total length	122075582	118882582	122043781	118934797	121947926	121943277	**122079245**	122072919
Reference length	120381546	120381546	120381546	120381546	120381546	120381546	120381546	120381546
GC (%)	42.47	42.53	42.49	42.50	42.51	42.50	42.48	42.48
Reference GC (%)	42.42	42.42	42.42	42.42	42.42	42.42	42.42	42.42
N50	79091	52325	68696	45242	83768	79386	**90787**	65116
NG50	81525	51206	70464	44409	85018	80160	**92829**	65909
N75	34445	24360	30032	20786	38515	34907	**39744**	28474
NG75	36491	23546	31238	19878	39914	36153	**41573**	29856
L50	421	626	474	723	397	425	**375**	516
LG50	411	641	462	740	388	416	**366**	503
L75	1001	1453	1130	1692	929	1001	**882**	1216
LG75	965	1500	1089	1746	899	968	**850**	1173
# misassemblies	762	831	766	780	756	757	**754**	789
# misassembled contigs	588	698	601	671	**570**	587	575	631
Misassembled contigs length	49650199	40548248	45881007	**34458785**	50816213	49036345	54113900	45207902
# local misassemblies	**2588**	4259	3458	4442	3421	3420	**2588**	3510
# unaligned mis. contigs	27	31	28	**18**	28	31	26	31
# unaligned contigs	3627 + 370 part	2541 + 408 part	3559 + 384 part	**2410 + 375 part**	3531 + 359 part	3530 + 361 part	3629 + 365 part	3591 + 373 part
Unaligned length	8517557	**5527503**	8554387	5579538	8439808	8436001	8508422	8522280
Genome fraction (%)	93.780	93.289	93.578	93.214	93.584	93.594	**93.788**	93.581
Duplication ratio	**1.006**	1.009	1.007	1.010	1.008	1.007	**1.006**	1.008
# N's per 100 kbp	589.55	**565.60**	718.48	729.81	730.60	728.54	591.23	721.68
# mismatches per 100 kbp	500.54	500.48	497.06	502.65	**494.14**	495.59	497.83	496.90
# indels per 100 kbp	119.70	119.15	118.61	118.81	118.54	118.77	119.42	**118.52**
Largest alignment	402868	365720	385170	333479	385163	385170	**402877**	385169
Total aligned length	113080516	112542712	112866314	112513234	112914086	112899477	**113107293**	112915266
NA50	58252	41363	53544	36812	61763	58806	**63652**	50153
NGA50	59591	40860	54093	36316	62706	59526	**65174**	50834
NA75	25864	19605	23565	16968	27787	25860	**28871**	21977
NGA75	27153	18756	24911	16471	29033	26852	**30231**	23124
LA50	574	786	629	877	549	577	**527**	667
LGA50	559	805	613	897	536	564	**513**	651
LA75	1346	1814	1479	2056	1268	1345	**1218**	1579
LGA75	1298	1872	1427	2121	1226	1301	**1175**	1523

Figure B.6 SPAdes assembly results for dataset D5 (*D. melanogaster*) for both uncorrected and corrected data. Contigs with length NGAx or larger produce x% of the genome.

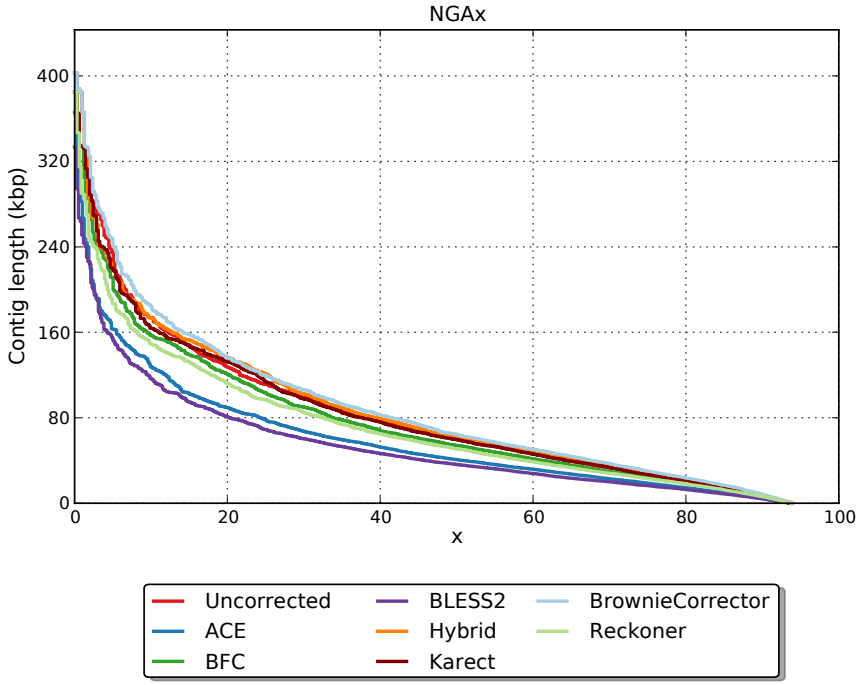


Table B.21 Assembly quality metrics for D6

Assembly	Uncorrected	ACE	BFC	BLESS2	Hybrid	Karect	BrownieCorrector	Reckoner
# contigs (≥ 0 bp)	153290	168992	152296	157691	156284	157079	153194	168975
# contigs (≥ 1000 bp)	5162	5322	5106	5472	4889	5100	4900	5279
# contigs (≥ 5000 bp)	3342	3431	3329	3432	3145	3295	3141	3402
# contigs (≥ 10000 bp)	2489	2529	2478	2497	2360	2462	2349	2529
# contigs (≥ 25000 bp)	1322	1269	1302	1282	1275	1294	1277	1300
# contigs (≥ 50000 bp)	598	575	596	565	611	593	615	599
Total length (≥ 0 bp)	122801905	123596492	122790209	122366708	123037135	123098372	122790637	**123961124**
Total length (≥ 1000 bp)	108714303	108474907	**108842601**	107970602	108759997	108767495	108713396	108797092
Total length (≥ 5000 bp)	104203084	103753669	**104453011**	102836224	104451806	104264416	104377293	104095523
Total length (≥ 10000 bp)	98015647	97195298	98261766	96043789	**98726415**	98203437	98629627	97749838
Total length (≥ 25000 bp)	79007005	76716663	78935185	76403311	80964898	79112155	**81203481**	77458191
Total length (≥ 50000 bp)	53364388	52047704	53919340	51174509	57276588	54231397	**57675590**	52534096
# contigs	6616	6774	6501	6946	6304	6539	6326	6708
Largest contig	378346	360514	337032	340077	**403164**	358676	360062	292751
Total length	109698912	109456349	**109787184**	108964952	109717059	109740727	109679469	109761199
Reference length	119668634	119668634	119668634	119668634	119668634	119668634	119668634	119668634
GC (%)	35.96	35.98	35.96	35.99	35.97	35.96	35.96	35.96
Reference GC (%)	36.06	36.06	36.06	36.06	36.06	36.06	36.06	36.06
N50	48674	47051	49223	46312	52900	49483	**53537**	47973
NG50	43039	41281	43661	40388	47318	43958	**47965**	41999
N75	22586	21015	22561	20636	24149	22647	**24323**	21691
NG75	16229	15179	16601	14482	**17410**	16410	17316	15906
L50	629	631	616	634	565	606	**561**	647
LG50	737	747	723	758	664	713	**660**	758
L75	1460	1503	1446	1517	1329	1428	**1320**	1508
LG75	1850	1934	1827	1979	1691	1811	**1685**	1906
# misassemblies	157	151	167	194	**123**	140	126	143
# misassembled contigs	156	148	163	186	**122**	134	125	138
Misassembled contigs length	7195789	6683457	8790317	7948697	**5512136**	6420888	5809898	6525648
# local misassemblies	91	95	**83**	107	90	89	86	**83**
# unaligned mis. contigs	2	2	2	**1**	**1**	2	2	2
# unaligned contigs	234 + 21 part	161 + 24 part	218 + 24 part	**104 + 34 part**	225 + 27 part	229 + 27 part	233 + 22 part	231 + 25 part
Unaligned length	409655	245951	410530	**200169**	400587	402363	407487	411574
Genome fraction (%)	91.209	91.115	**91.285**	90.734	91.256	91.258	91.219	91.257
Duplication ratio	**1.001**	1.002	**1.001**	1.002	**1.001**	**1.001**	**1.001**	**1.001**
# N's per 100 kbp	11.00	24.65	10.48	93.43	12.59	12.23	10.39	**10.08**
# mismatches per 100 kbp	20.15	31.85	19.74	22.56	18.61	19.74	**18.31**	19.02
# indels per 100 kbp	5.08	9.70	5.01	7.65	4.87	4.89	**4.84**	4.92
Largest alignment	378346	360476	322214	340077	**402910**	358628	360023	292696
Total aligned length	109238189	109114747	**109327444**	108666168	109275728	109291520	109231090	109296620
NA50	47023	45327	47171	44605	51586	48129	**52551**	46233
NGA50	41833	39895	41818	38431	46332	42256	**46678**	40779
NA75	21722	20139	21736	19707	23677	21873	**23679**	21161
NGA75	15581	14730	15864	13950	**16845**	15760	16687	15390
LA50	647	650	642	657	575	621	**573**	666
LGA50	759	770	752	786	677	731	**674**	781
LA75	1503	1556	1503	1577	1356	1467	**1351**	1554
LGA75	1909	2001	1900	2058	**1728**	1865	**1728**	1963

Figure B.7 SPAdes assembly results for dataset D6 (*D. melanogaster*) for both uncorrected and corrected data. Contigs with length NGAx or larger produce x% of the genome.

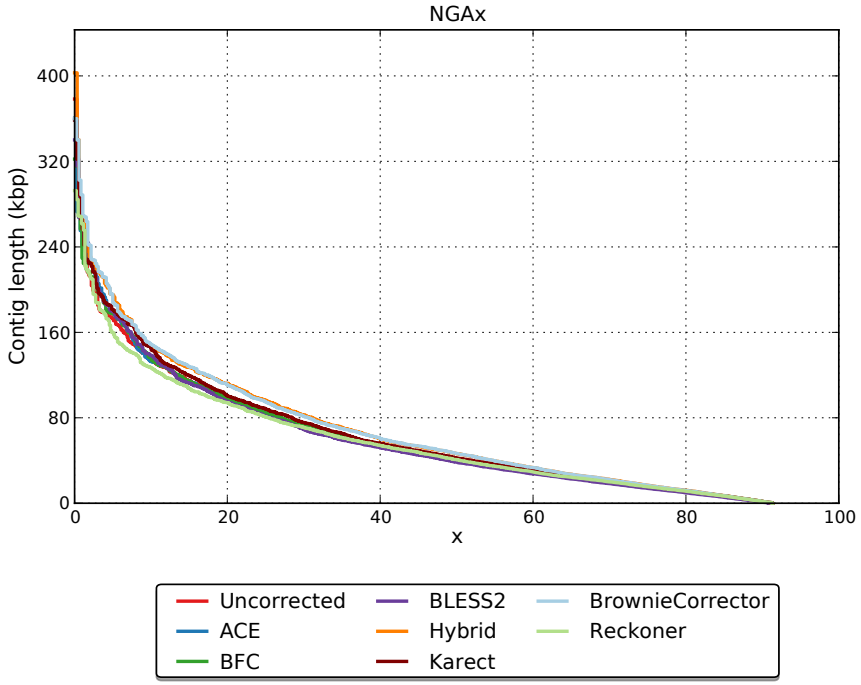


Table B.22 Assembly quality metrics for D7

Assembly	Uncorrected	ACE	BFC	BLESS2	Hybrid	Karect	BrownieCorrector	Reckoner
# contigs (≥ 0 bp)	82728	86644	81601	84454	81455	82018	82726	90888
# contigs (≥ 1000 bp)	3570	4035	3705	3493	3545	3566	3538	4403
# contigs (≥ 5000 bp)	1704	2416	1799	2032	1614	1658	1694	2334
# contigs (≥ 10000 bp)	1314	1857	1369	1571	1202	1229	1285	1751
# contigs (≥ 25000 bp)	853	1133	904	1024	798	818	854	1104
# contigs (≥ 50000 bp)	579	694	605	658	554	570	574	678
Total length (≥ 0 bp)	131187203	129316221	130842102	128194970	130823725	130846045	131147292	**131587165**
Total length (≥ 1000 bp)	**122759330**	120824487	122515719	119969315	122632419	122599419	122716522	122605532
Total length (≥ 5000 bp)	118723698	116992547	118417293	116381572	118516483	118516474	**118760696**	118026807
Total length (≥ 10000 bp)	**115939340**	113005485	115325687	113095422	115563382	115456277	115866004	113848935
Total length (≥ 25000 bp)	108413703	101200875	107777428	104320159	**108945714**	108786024	108922843	103351617
Total length (≥ 50000 bp)	98491977	85458576	96979047	91176824	**100102272**	99794613	98871234	88326978
# contigs	5317	5576	5498	4611	5199	5232	5291	6207
Largest contig	**1199604**	670790	907732	602677	1037537	908356	1016904	887190
Total length	**123986725**	121887615	123769629	120729772	123793924	123771242	123945141	123866970
Reference length	120381546	120381546	120381546	120381546	120381546	120381546	120381546	120381546
GC (%)	42.55	42.50	42.55	42.46	42.56	42.56	42.55	42.53
Reference GC (%)	42.42	42.42	42.42	42.42	42.42	42.42	42.42	42.42
N50	169610	97526	159663	125811	**184981**	174552	173854	104119
NG50	179224	99802	165472	126462	**189080**	179261	179844	108234
N75	65173	39922	58324	50582	**73720**	70434	68958	41468
NG75	74095	41210	65771	51453	**86411**	80784	79734	46078
L50	198	343	214	262	**183**	188	198	314
LG50	187	335	203	260	**174**	178	187	298
L75	483	829	529	646	**434**	453	472	779
LG75	444	801	487	641	**402**	419	436	719
# misassemblies	1093	1198	1095	1254	1058	**1049**	1073	1190
# misassembled contigs	576	751	568	703	**548**	549	570	691
Misassembled contigs length	77195891	**65434373**	74113978	72211756	79124616	77842226	77576813	65900369
# local misassemblies	1912	**1762**	1886	1911	1802	1851	1871	2032
# unaligned mis. contigs	140	127	135	**105**	125	132	129	137
# unaligned contigs	3051 + 699 part	2271 + 771 part	3086 + 724 part	**1719 + 662 part**	3061 + 730 part	3047 + 735 part	3057 + 702 part	3103 + 746 part
Unaligned length	9623179	7729395	9471622	**6671029**	9531780	9527862	9589547	9494184
Genome fraction (%)	94.484	94.275	94.474	94.248	94.478	94.456	**94.486**	94.431
Duplication ratio	**1.005**	1.006	**1.005**	**1.005**	**1.005**	**1.005**	**1.005**	1.006
# N's per 100 kbp	483.18	**408.89**	460.28	422.82	447.36	439.87	475.04	449.96
# mismatches per 100 kbp	556.84	567.59	556.70	562.67	554.80	**553.80**	555.55	556.33
# indels per 100 kbp	134.33	141.80	134.28	145.24	133.75	133.66	134.05	**133.40**
Largest alignment	570650	428847	443438	338930	570790	**570959**	570669	406192
Total aligned length	**114102548**	113780243	114027727	113763228	114012129	113990788	114097579	114030717
NA50	92022	61834	88388	73292	**99412**	97727	92238	64742
NGA50	96381	62981	91577	73377	**103872**	101753	96385	67061
NA75	37711	26304	35621	32030	**42031**	40768	38469	27161
NGA75	41779	27692	39269	32305	**45314**	44500	42580	29658
LA50	373	538	390	462	**352**	358	369	514
LGA50	354	526	371	460	**335**	341	350	487
LA75	894	1281	940	1079	**829**	848	883	1243
LGA75	826	1240	872	1071	**771**	789	817	1151

Figure B.8 SPAdes assembly results for dataset D7 (*D. melanogaster*) for both uncorrected and corrected data. Contigs with length NGAx or larger produce x% of the genome.

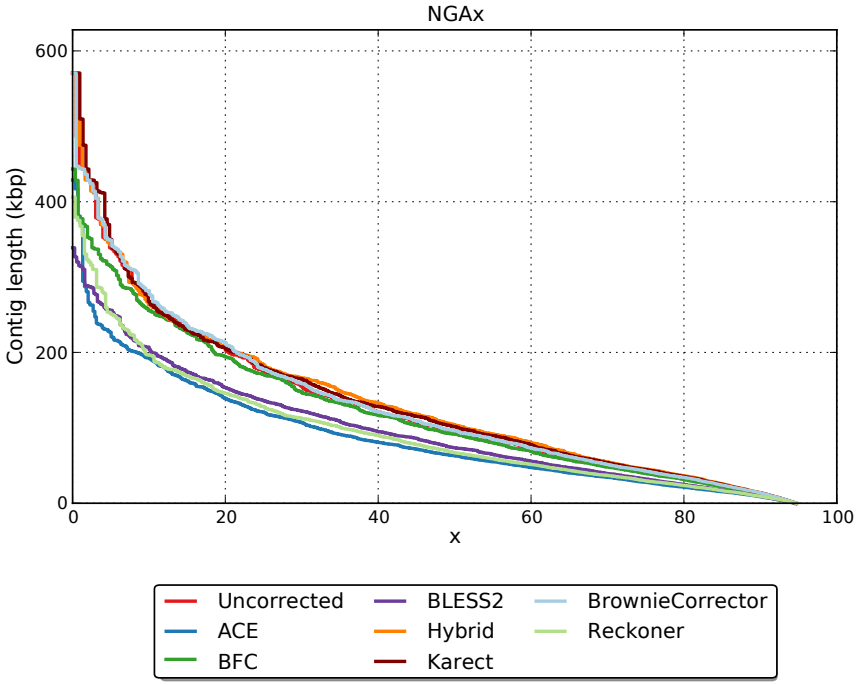


Table B.23 Assembly quality metrics for D8

Assembly	Uncorrected	ACE	BFC	BLESS2	Hybrid	Karect	BrownieCorrector	Reckoner
# contigs (≥ 0 bp)	149228	191379	166969	174905	162558	163826	149066	199481
# contigs (≥ 1000 bp)	3216	2996	3107	3393	2749	2859	3001	3290
# contigs (≥ 5000 bp)	1718	1706	1683	1903	1399	1483	1547	1855
# contigs (≥ 10000 bp)	1201	1308	1177	1441	984	1053	1068	1340
# contigs (≥ 25000 bp)	779	854	759	956	627	663	697	862
# contigs (≥ 50000 bp)	536	560	529	637	459	474	489	572
Total length (≥ 0 bp)	137006538	137208923	138629176	135737931	138182929	138240801	137010216	**140815351**
Total length (≥ 1000 bp)	122386145	119633304	**122499150**	119842489	122364837	122306783	122404017	122425002
Total length (≥ 5000 bp)	11889712	116646454	119167322	116397113	**119223919**	119086628	119025946	119068877
Total length (≥ 10000 bp)	115233784	113711648	115580814	113066925	**116307772**	116071704	115635740	115382004
Total length (≥ 25000 bp)	108540040	106412774	108957987	105232289	**110727634**	110012601	109875005	107869724
Total length (≥ 50000 bp)	99903980	95839747	100902805	93696395	**104904218**	103401295	102484228	97699745
# contigs	5319	4525	5155	4582	4748	4889	5076	5379
Largest contig	1078505	1097937	1329229	873186	**1577991**	1472989	1315292	834563
Total length	123920625	120711602	**123995028**	120697350	123832426	123797326	123920633	123949799
Reference length	120381546	120381546	120381546	120381546	120381546	120381546	120381546	120381546
GC (%)	42.44	42.48	42.44	42.45	42.46	42.46	42.44	42.44
Reference GC (%)	42.42	42.42	42.42	42.42	42.42	42.42	42.42	42.42
N50	195145	173427	209236	141236	**278523**	252831	240349	164454
NG50	202478	173448	217832	141956	**293584**	264260	250321	175352
N75	73641	66031	77292	56469	**102606**	93760	88826	64991
NG75	83516	66621	85145	56958	**113846**	104548	101313	71933
L50	170	196	156	241	**115**	126	140	196
LG50	161	195	148	240	**109**	120	133	185
L75	422	468	402	578	**294**	324	348	489
LG75	388	464	368	574	**270**	298	319	449
# misassemblies	936	999	966	**933**	952	944	934	998
# misassembled contigs	497	557	490	588	**438**	456	467	558
Misassembled contigs length	79855723	76678055	81266528	**67931475**	88449521	86859421	85282570	78755301
# local misassemblies	1646	1697	1741	1685	1646	**1619**	1658	1690
# unaligned mis. contigs	101	101	116	107	107	113	**97**	108
# unaligned contigs	3150 + 604 part	2051 + 605 part	3015 + 629 part	**1916 + 581 part**	2976 + 597 part	2979 + 602 part	3137 + 599 part	3061 + 610 part
Unaligned length	9788711	**6681807**	9833294	6696692	9705532	9659228	9818310	9743468
Genome fraction (%)	**94.396**	94.239	94.368	94.186	94.387	**94.396**	94.391	94.363
Duplication ratio	**1.004**	1.005	1.005	1.005	**1.004**	**1.004**	**1.004**	1.005
# N's per 100 kbp	883.63	**691.64**	948.10	791.18	962.03	945.40	887.88	951.17
# mismatches per 100 kbp	505.84	504.41	501.43	504.88	501.19	**500.63**	505.00	502.12
# indels per 100 kbp	120.65	119.95	118.99	122.14	**118.96**	119.05	120.48	119.18
Largest alignment	562955	488967	779743	508037	630299	**824208**	630640	562981
Total aligned length	113900519	113715614	113916864	113699300	113906488	113910162	113887441	**113924772**
NA50	106095	92928	106110	86055	**121172**	119421	113318	95924
NGA50	109785	93602	110748	86526	**126449**	124215	118192	99419
NA75	41483	39367	43182	37065	**51555**	48231	47013	37077
NGA75	45799	40003	47338	37379	**57604**	54004	52514	41418
LA50	338	362	322	388	**280**	291	307	360
LGA50	322	360	305	386	**266**	277	291	341
LA75	798	848	778	913	**663**	691	720	871
LGA75	737	842	718	907	**615**	641	667	803

Figure B.9 SPAdes assembly results for dataset D8 (*D. melanogaster*) for both uncorrected and corrected data. Contigs with length NGAx or larger produce x% of the genome.

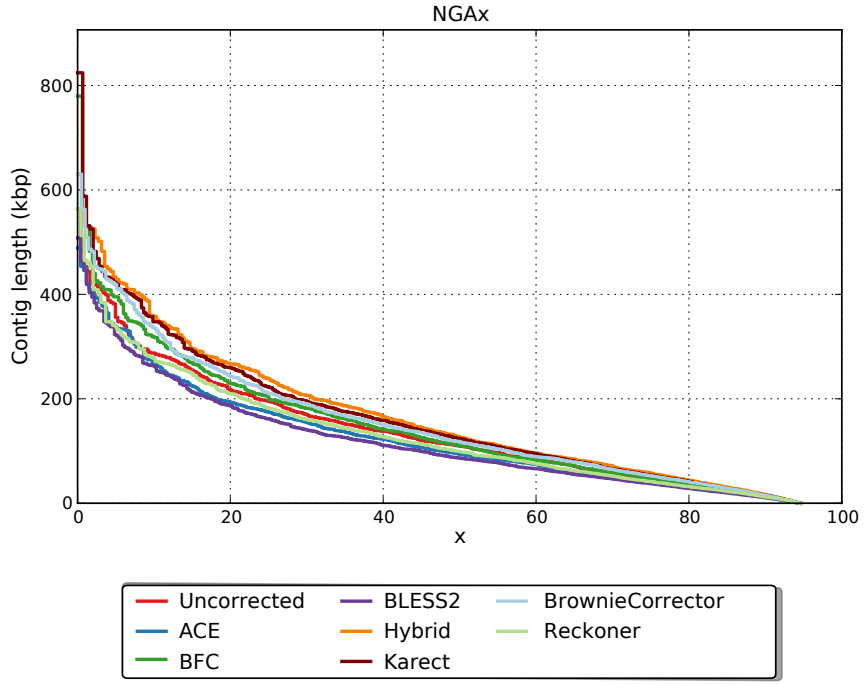
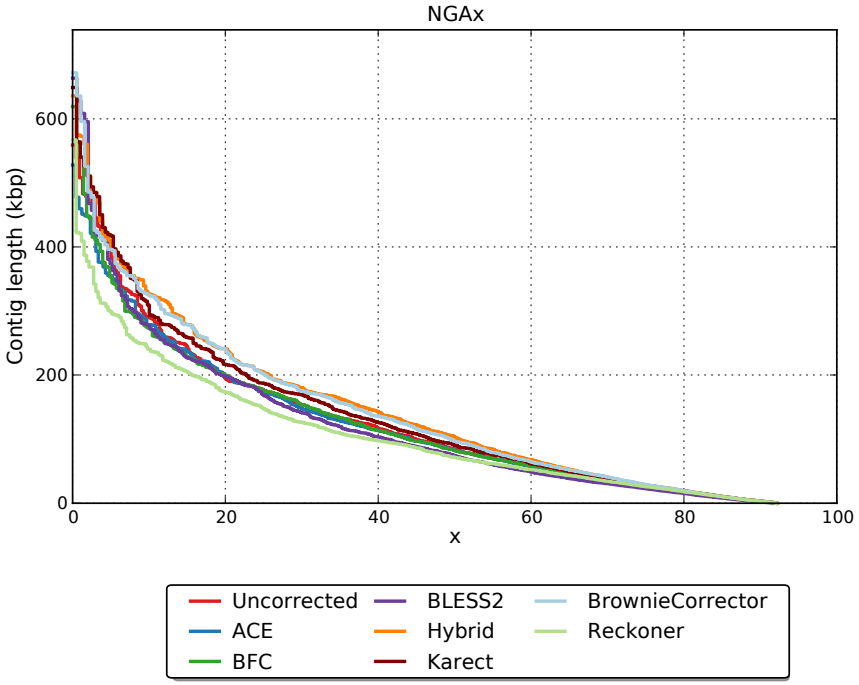


Table B.24 Assembly quality metrics for D9

Assembly	Uncorrected	ACE	BFC	BLESS2	Hybrid	Karect	BrownieCorrector	Reckoner
# contigs (≥ 0 bp)	142089	157367	141808	146982	145087	146112	142070	157568
# contigs (≥ 1000 bp)	3227	3340	3189	3629	3097	3197	3090	3412
# contigs (≥ 5000 bp)	1915	1998	1926	2157	1777	1864	1791	2077
# contigs (≥ 10000 bp)	1460	1528	1455	1629	1365	1427	1365	1612
# contigs (≥ 25000 bp)	954	947	966	1011	857	908	877	1054
# contigs (≥ 50000 bp)	615	604	613	588	559	587	565	654
Total length (≥ 0 bp)	123813843	124547852	123778880	123452408	124109646	124097329	123825918	**124894825**
Total length (≥ 1000 bp)	110920359	110650660	110973623	110165631	**111016302**	110947707	110946362	110941841
Total length (≥ 5000 bp)	107630324	107241203	**107824566**	106411218	107715793	107603069	107704107	107587644
Total length (≥ 10000 bp)	104366913	103820948	104426193	102640491	**104764365**	104467517	104638648	104241224
Total length (≥ 25000 bp)	96121787	94226910	96285383	92636453	96331162	95891626	**96638802**	95024283
Total length (≥ 50000 bp)	83940356	82224621	83679724	77761041	**85845567**	84494501	85456925	80760253
# contigs	4250	4359	4177	4656	4095	4195	4100	4398
Largest contig	692396	548881	692397	664835	706587	**799938**	673567	568204
Total length	111616529	111339614	111643887	110865743	**111696039**	111628269	111630438	111613165
Reference length	119668634	119668634	119668634	119668634	119668634	119668634	119668634	119668634
GC (%)	35.94	35.96	35.95	35.96	35.95	35.95	35.94	35.94
Reference GC (%)	36.06	36.06	36.06	36.06	36.06	36.06	36.06	36.06
N50	122106	116453	117977	101712	**150944**	130668	143649	100710
NG50	108902	107645	104580	92073	**132808**	117943	129072	93533
N75	50880	46312	49848	39472	**56026**	52705	54148	44144
NG75	39074	33291	36446	29040	39395	37369	**40588**	33255
L50	254	265	254	279	**212**	237	218	308
LG50	289	302	289	324	**240**	270	248	350
L75	611	631	615	710	**520**	572	532	717
LG75	748	791	754	904	**648**	708	661	873
# misassemblies	837	836	814	829	793	**775**	800	830
# misassembled contigs	554	582	552	581	526	528	**523**	555
Misassembled contigs length	43095471	42734679	44104032	**36877771**	45541239	43327594	45333670	38686180
# local misassemblies	644	641	648	686	**638**	**638**	640	723
# unaligned mis. contigs	13	8	10	12	8	**7**	9	8
# unaligned contigs	266 + 425 part	191 + 409 part	245 + 402 part	**147 + 388 part**	248 + 400 part	263 + 379 part	259 + 411 part	262 + 407 part
Unaligned length	1229535	1018600	1151096	**999954**	1227225	1154999	1226425	1152483
Genome fraction (%)	91.805	91.747	**91.912**	91.363	91.876	91.866	91.824	91.849
Duplication ratio	1.005	1.005	1.005	1.005	1.005	1.005	1.005	1.005
# N's per 100 kbp	511.05	490.09	475.62	493.59	505.76	**471.17**	501.25	477.68
# mismatches per 100 kbp	52.02	59.49	52.33	59.19	52.05	52.67	51.50	**51.00**
# indels per 100 kbp	11.29	16.44	11.22	21.71	10.74	10.92	**10.73**	**10.73**
Largest alignment	559314	527818	619363	663821	635802	648971	**671915**	567176
Total aligned length	110285821	110197159	**110387175**	109731628	110371573	110368712	110301676	110342848
NA50	95741	93437	94016	85059	**114187**	101700	110521	80931
NGA50	84659	83138	82101	74447	**104037**	90661	96916	71646
NA75	38622	34585	37278	31270	40532	38578	**40986**	34410
NGA75	27591	25059	27698	22761	28609	28222	**29163**	25910
LA50	315	322	322	336	**268**	292	274	376
LGA50	360	369	367	391	**304**	334	312	429
LA75	779	805	793	884	**678**	734	694	903
LGA75	965	1019	980	1131	**853**	916	868	1104

Figure B.10 SPAdes assembly results for dataset D9 (*A. thaliana*) for both uncorrected and corrected data. Contigs with length NGAx or larger produce x% of the genome.



B.5.6 Runtime and memory usage

Table B.25 and B.26 provide the detail numbers of peak memory usage and the runtime (wall time) of EC tools on datasets respectively.

Table B.25 Peak memory (GB) usage of the aligners on real data.

Tools	D1	D2	D3	D4	D5	D6
ACE	6.81	27.34	30.41	31.18	39.15	30.73
BFC	**2.43**	4.73	5.15	5.21	5.23	5.29
BLESS2	3.90	**3.90**	**3.90**	**3.90**	**3.90**	**3.89**
BrownieCorrector	2.64	20.66	17.19	4.70	7.11	7.21
Karect	29.76	86.99	136.64	145.25	171.37	188.59
Reckoner	3.75	3.94	3.95	3.94	3.97	3.96

Table B.26 Run time (min) of the aligners on real data

Tools	D1	D2	D3	D4	D5	D6
ACE	67.03	182.62	325.02	329.82	370.43	355.25
BFC	0.91	**2.71**	3.41	4.20	5.41	6.50
BLESS2	1.25	4.49	4.83	6.31	8.85	8.51
BrownieCorrector	27.72	96.40	48.85	46.23	57.95	40.11
Karect	6.96	31.44	55.03	62.78	69.68	52.35
Reckoner	**0.70**	3.44	**3.05**	**2.97**	**4.61**	**3.45**

References

[1] Mahdi Heydari, Giles Miclotte, Yves Van de Peer, and Jan Fostier. *BrownieAligner: accurate alignment of Illumina sequencing data to de Bruijn graphs*. BMC Bioinformatics, 19(1):311, sep 2018. 3-4, 3-11, B-8

C

Supplementary Data: BrownieAligner: Accurate Alignment of Illumina Sequencing Data to de Bruijn Graphs

"Don't give up on your dreams, keep on sleeping"

C.1 Parameter Settings

All tools were executed with 32 threads. For all tables and figures in the chapter 4 and in the supplementary data the default or recommended values of parameters were taken for each tool individually. Below, the command line parameters are specified for each tool individually:

C.1.1 BGREAT[1]

BCALM is used to build the de Bruijn graph from the reference genome.

```
$ ./bcalm -nb-cores 32 -in genome.fasta -kmer-size 31
   -abundance-min 1
$ ./bgreat -c -q -O -u  $inputreads -g genome.unitigs.
   fa -k 31 -t 32
```

[1] https://github.com/Malfoy/BGREAT2.git

C.1.2 BrownieAligner[2]

```
$ ./brownie index -t 32 -p $outputDir -k 31  genome.fasta
$ ./brownie align -t 32 -p $outputDir -k 31 -o $outputDir/outputFile $inputreads
```

To disable Markov Model in the alignment procedure:

```
$ ./brownie align -nMM -t 32 -p $outputDir -k 31 -o $outputDir/outputFile $inputreads
```

C.1.3 deBGA v. 0.1[3]

deBGA initially builds the graph from the reference genome. Then, it aligns reads to the graph and returns the result as a SAM file. sam2Alignment then constructs the corrected read from the reference genome based on the corresponding alignment position and the cigar string in the SAM file. deBGA sometimes reports multiple alignments for one read. In this case only the one with the lowest edit distance is considered. In order to measure the runtime and the memory usage of deBGA only two first steps (deBGA index, and deBGA aln) are taken into consideration.

```
$ ./deBGA index genome.fasta reference/ -p 32
$ ./deBGA aln reference/ $inputreads1 $inputreads2 deBGA.sam -p 32
$ ./sam2Alignment deBGA.sam genome.fasta $inputreads SMID
```

[2] https://github.com/biointec/brownie aligner
[3] https://github.com/hitbc/deBGA.git

C.2 Simulated data preparation

Synthetic Illumina reads from two different Illumina platforms and in two different read lengths and coverage (HiSeq 2000 (100 bp and 50X), HiSeq 2500 (150 bp and 25X)) are generated with ART read simulator (v. 2.6). The following commands were used:

```
$ ./art_illumina -ss HS20 -sam -i genome.fasta -p -l
   100 -f 50 -m 200 -s 10 -o reads
$ ./art_illumina -ss HS25 -sam -i genome.fasta -p -l
   150 -f 25 -m 200 -s 10 -o reads
```

The mean fragment size is 200 bp, and the fragment standard deviation is 10 bp.

C.3 Real data preparation

In the absence of ground truth for real data, it is assumed that the error-free read is represented by the segment of the reference genome to which that read aligns. Therefore, reads are initially aligned to the linear reference genome by BWA. Then paired-end reads that both pairs map to the reference genome properly are extracted and stored into mappedPairs.sam file. sam2pairwise tool uses the CIGAR and MD tag to reconstruct the pairwise alignment of each read. Finally, the python script is used to extract the mappedReads (uncorrected mapped reads), perfectReads (equivalent error free reads) from the pairwise alignment and initial real data.

```
$ bwa index reference/genome.fasta
$ bwa mem reference/genome.fasta -t 16 reads.fastq -p
   >bwa.sam
$ samtools view -S -f 0x2 -F 0x904  bwa.sam >
   mappedPairs.sam
$ sam2pairwise <mappedPairs.sam> ali.alignment
$ python extMappedFromAli.py reads.fastq ali.alignment
    perfectReads.fasta mappedReads.fastq
```

C.4 Evaluation Metric

C.4.1 Alignment ratio

Each tool reports set of reads that are aligned to the graph either explicitly (BGREAT and BrownieAligner in the output) or implicitly (deBGA: by the corresponding

flag in the SAM file). However, not all of thees reads align correctly to the graph. Generally, reads are classified into three groups as follows :

1. Aligned reads
   (a) Correctly aligned (CA)
   (b) Incorrectly aligned (IA)
2. Not aligned reads (NA)

While reads that belong to the NA class are specified by each tool, rest of the reads are needed to be further classified into CA and IA classes. The classification is straightforward when the ground truth, i.e., the perfect read, is known. Let R represent an input read. For each read R, there is a corresponding read C which is the segment of the reference genome to which that read aligns and represented by a path in the graph, and P which is the ground truth (error free read). We define a correct alignment as such P is identical to C. Then, read R is categorized into CA group if the alignment is correct, otherwise into IA group.

C.5 Results

C.5.1 Simulated Data

Detailed information about the accuracy of tools on simulated data which includes percentage of correctly aligned reads and total number of (aligned, correctly aligned, incorrectly aligned and unaligned) reads are shown in Table C.1. Default k-mer sizes are used for all tools which is 31 in BGREAT and BrownieAligner and 22 in deBGA. Additionally, BGREAT, BrownieAligner and deBGA are benchmarked against all datasets with different k in C.5.1.1, C.5.1.2 and C.5.1.3 sections respectively.

Table C.1 Accuracy evaluation of graph aligners on simulated data

	S1 (*E. coli DH10B*)	S2 (*E. coli DH10B*)	S3 (*H. sapiens* Chr. 21)	S4 (*H. sapiens* Chr. 21)	S5 (*D. melanogaster*)	S6 (*D. melanogaster*)
			Percentage of correctly aligned reads.			
BGREAT	99.94	99.61	98.92	96.16	99.89	99.40
BrownieAligner	100.00	99.99	99.42	98.07	99.97	99.89
BrownieAlignerNoMM	99.99	99.98	99.30	97.67	99.96	99.85
deBGA	99.52	83.48	99.07	83.01	99.37	83.37
			Total number of aligned reads.			
BGREAT	780 568	2 334 082	7 746 106	22 875 994	20 046 023	59 878 851
BrownieAligner	780 982	2 342 838	7 759 283	23 092 285	20 059 808	60 146 026
BrownieAlignerNoMM	780 966	2 342 698	7 750 237	23 010 993	20 057 931	60 128 263
deBGA	777 771	1 958 626	7 748 352	19 836 023	19 970 349	50 355 578
			Total number of correctly aligned reads.			
BGREAT	780547	2333855	7701160	22458474	20041869	59832087
BrownieAligner	780968	2342711	7739916	22904206	20057579	60121746
BrownieAlignerNoMM	780951	2342548	7730321	22811819	20055622	60101941
deBGA	777247	1955941	7712571	19387610	19936325	50181809
			Total number of incorrectly aligned reads.			
BGREAT	21	227	44 946	417 520	4 154	46 764
BrownieAligner	14	127	19 367	188 079	2 229	24 280
BrownieAlignerNoMM	15	150	19 916	199 174	2 309	26 322
deBGA	524	2 685	35 781	448 413	34 024	173 769
			Total number of unaligned reads.			
BGREAT	432	8 968	38 870	478 956	17 505	311 799
BrownieAligner	18	212	25 693	262 665	3 720	44 624
BrownieAlignerNoMM	34	352	34 739	343 957	5 597	62 387
deBGA	3 229	384 424	36 624	3 518 927	93 179	9 835 072

C.5.1.1 BGREAT

Table C.2 shows the accuracy of BGREAT in terms of percentage of correctly aligned reads for different values of k on simulated data. The default value is 31.

Table C.2 Accuracy evaluation of BGREAT on simulated data for different values of k.

k-mer	S1 (*E. coli DH10B*)	S2 (*E. coli DH10B*)	S3 (*H. sapiens* Chr. 21)	S4 (*H. sapiens* Chr. 21)	S5 (*D. melanogaster*)	S6 (*D. melanogaster*)
			Percentage of correctly aligned reads.			
25	99.94	**99.83**	98.53	95.23	99.86	**99.50**
31	99.94	99.61	98.92	96.16	99.89	99.40
35	99.94	99.44	99.13	96.66	99.91	99.29
41	**99.95**	99.44	99.28	97.18	99.91	99.29
51	**99.95**	99.41	**99.45**	**97.62**	**99.92**	99.25

C.5.1.2 BrownieAligner

Table C.3 shows the accuracy of BrownieAligner in terms of percentage of correctly aligned reads for different values of k on simulated data. The default value is 31.

Table C.3 Accuracy evaluation of BrownieAligner on simulated data for different values of k.

k-mer	S1 (*E. coli DH10B*)	S2 (*E. coli DH10B*)	S3 (*H. sapiens* Chr. 21)	S4 (*H. sapiens* Chr. 21)	S5 (*D. melanogaster*)	S6 (*D. melanogaster*)
			Percentage of correctly aligned reads.			
25	100.0	99.98	99.23	97.46	99.96	99.86
31	100.0	**99.99**	99.42	98.07	99.97	99.89
35	100.0	**99.99**	99.49	98.31	99.97	99.90
41	100.0	**99.99**	99.57	98.53	**99.98**	99.90
51	100.0	**99.99**	**99.66**	**98.74**	**99.98**	**99.91**

C.5.1.3 deBGA

Table C.4 shows the accuracy of deBGA for different values of k in terms of percentage of correctly aligned reads on simulated data. The default value is 22, and the accepted range is [22,28].

Table C.4 Accuracy evaluation of deBGA on simulated data for different values of k.

k-mer	S1 (*E. coli DH10B*)	S2 (*E. coli DH10B*)	S3 (*H. sapiens* Chr. 21)	S4 (*H. sapiens* Chr. 21)	S5 (*D. melanogaster*)	S6 (*D. melanogaster*)
			Percentage of correctly aligned reads.			
22	99.52	**83.48**	99.07	83.01	99.37	83.37
24	99.52	83.47	99.09	83.04	99.37	**83.38**
26	99.52	83.40	**99.12**	**85.66**	99.37	83.30
28	99.52	83.34	**99.12**	85.60	**99.38**	83.26

C.5.1.4 Choice of parameters

The results in chapter 4 are based on these parameters: the maximum MM order (maxOrder) is 10, the minLikelihoodRatio and minChainCov are set respectively to 10^5 and 10. However, we additionally investigated the accuracy of BrownieAligner on simulated data based on other values of these parameters. Table C.5 shows the accuracy of BrownieAligner in terms of percentage of correctly aligned reads for different values of maxOrder. The default value is 10.

Table C.5 Accuracy evaluation of BrownieAligner on simulated data for different values of maxOrder.

maxOrder	S1 (*E. coli DH10B*)	S2 (*E. coli DH10B*)	S3 (*H. sapiens* Chr. 21)	S4 (*H. sapiens* Chr. 21)	S5 (*D. melanogaster*)	S6 (*D. melanogaster*)
			Percentage of correctly aligned reads.			
5	100.0	99.98	99.37	97.90	99.97	99.88
10	100.0	99.99	99.42	98.07	99.97	99.89
15	100.0	99.99	99.45	98.16	99.97	99.89
20	100.0	99.99	99.46	98.21	99.97	99.89

Table C.6 shows the accuracy of BrownieAligner in terms of percentage of correctly aligned reads for different values of minLikelihoodRatio. The default value is 10^5.

Table C.6 Accuracy evaluation of BrownieAligner on simulated data for different values of minLikelihoodRatio.

minLikelihoodRatio	S1 (*E. coli DH10B*)	S2 (*E. coli DH10B*)	S3 (*H. sapiens* Chr. 21)	S4 (*H. sapiens* Chr. 21)	S5 (*D. melanogaster*)	S6 (*D. melanogaster*)
			Percentage of correctly aligned reads.			
10^2	100.0	99.99	99.43	98.08	99.97	99.89
10^5	100.0	99.99	99.42	98.07	99.97	99.89
10^{10}	99.99	99.98	99.34	98.01	99.96	99.88
10^{15}	99.99	99.98	99.34	97.78	99.96	99.86

Table C.7 shows the accuracy of BrownieAligner in terms of percentage of correctly aligned reads for different values of minChainCov. The default value is 10.

Table C.7 Accuracy evaluation of BrownieAligner on simulated data for different values of minChainCov.

minChainCov	S1 (*E. coli DH10B*)	S2 (*E. coli DH10B*)	S3 (*H. sapiens* Chr. 21)	S4 (*H. sapiens* Chr. 21)	S5 (*D. melanogaster*)	S6 (*D. melanogaster*)
			Percentage of correctly aligned reads.			
5	100.0	99.99	99.42	98.07	99.97	99.89
10	100.0	99.99	99.42	98.07	99.97	99.89
15	100.0	99.99	99.42	98.07	99.97	99.89
20	99.99	99.99	99.41	98.07	99.97	99.88

The results in this section indicates the accuracy of BrownieAligner is not affected by changing the parameters. Generally increasing the maximum order of Markov model can slightly improve the accuracy and increasing minLikelihoodRatio and minChainCov can marginally reduce the accuracy.

C.5.2 Real Data

Detailed information about the accuracy of tools on real data which includes percentage of correctly aligned reads and total number of (aligned, correctly aligned, incorrectly aligned and unaligned) reads are shown in Table C.8.

Table C.8 Accuracy comparison of graph aligners on real data

Tool	*B. dentium*	*E. coli DH10B*	*E. coli MG1655*	*S. enterica*	*P. aeruginosa*	*H. sapiens* Chr. 21	*C. elegans*	*D. melanogaster*
				Percentage of correctly aligned reads.				
BGREAT	94.55	94.28	91.28	84.97	96.09	92.01	94.57	80.37
BrownieAligner	99.81	99.81	99.55	99.02	99.78	96.98	96.53	89.59
BrownieAlignerNoMM	99.81	99.80	99.52	98.99	99.78	96.67	96.47	89.55
deBGA	99.67	99.3	92.36	97.31	93.63	98.42	74.72	85.42
				Total number of aligned reads.				
BGREAT	3 668 594	12 270 256	25 852 298	1 604 703	8 225 144	12 396 108	44 732 411	46 605 426
BrownieAligner	3 873 574	12 991 343	28 199 093	1 870 902	8 542 522	13 115 847	45 639 967	53 372 339
BrownieAlignerNoMM	3 873 499	12 990 034	28 192 797	1 870 430	8 542 199	13 067 671	45 609 800	53 329 978
deBGA	3 873 217	12 967 068	26 275 097	1 852 597	8 032 450	13 305 423	35 970 874	52 372 714
				Total number of correctly aligned reads.				
BGREAT	3668445	12268614	25837931	1604075	8224087	12254814	43710499	45520854
BrownieAligner	3872578	12987725	28176839	1869313	8539899	12916306	44616422	50741321
BrownieAlignerNoMM	3872494	12986226	28169892	1868762	8539540	12875329	44590192	50718731
deBGA	3867008	12921137	26141401	1837075	8013318	13109192	34538893	48378269
				Total number of incorrectly aligned reads.				
BGREAT	149	1 642	14 367	628	1 057	141 294	1 021 912	1 084 572
BrownieAligner	996	3 618	22 254	1 589	2 623	199 541	1 023 545	2 631 018
BrownieAlignerNoMM	1 005	3 808	22 905	1 668	2 659	192 342	1 019 608	2 611 247
deBGA	6 209	45 931	133 696	15 522	19 132	196 231	1 431 981	3 994 445
				Total number of unaligned reads.				
BGREAT	211 338	742 308	2 452 766	283 171	333 374	922 994	1 489 463	10 033 384
BrownieAligner	6 358	21 221	105 971	16 972	15 996	203 255	581 907	3 266 471
BrownieAlignerNoMM	6 433	22 530	112 267	17 444	16 319	251 431	612 074	3 308 832
deBGA	6 715	45 496	2 029 967	35 277	526 068	13 679	10 251 000	4 266 096

The accuracy of BrownieAligner on those reads that are aligned to a walk in the graph that contains multiple unitigs is shown in Table C.9 .

Table C.9 Accuracy evaluation of BrownieAlignerNoMM and BrownieAligner on the subset of the real data that are corrected by DFS Algorithm.

Tool	*B. dentium*	*E. coli DH10B*	*E. coli MG1655*	*S. enterica*	*P. aeruginosa*	*H. sapiens* Chr. 21	*C. elegans*	*D. melanogaster*
				Percentage of correctly aligned reads.				
BrownieAligner	97.83	97.15	94.94	93.75	96.08	67.95	50.70	39.09
BrownieAlignerNoMM	96.81	94.50	92.57	90.14	94.18	63.16	49.18	38.55

C.5.3 Time and space requirements

C.5.3.1 Simulated data

The memory usage and run time of the aligners are shown as plots in chapter 4, Table C.10 and Table C.11 respectively show the corresponding values. Additionally, a plot showing the effect of the branch and bound algorithm on the run time of BrownieAligner on the simulated data is shown in chapter 4, the corresponding values are provided in Table C.12.

Table C.10 Peak memory (GB) usage of the aligners on simulated data

Tools	S1 (*E. coli DH10B*)	S2 (*E. coli DH10B*)	S3 (*H. sapiens* Chr. 21)	S4 (*H. sapiens* Chr. 21)	S5 (*D. melanogaster*)	S6 (*D. melanogaster*)
BGREAT	1.60	1.60	**3.05**	**2.57**	**3.98**	**3.51**
BrownieAligner	**0.83**	**1.11**	6.23	7.05	13.37	11.88
deBGA	8.90	8.91	9.44	9.45	10.36	10.34

Table C.11 Run time (min) of the aligners on simulated data

Tools	S1 (*E. coli DH10B*)	S2 (*E. coli DH10B*)	S3 (*H. sapiens* Chr. 21)	S4 (*H. sapiens* Chr. 21)	S5 (*D. melanogaster*)	S6 (*D. melanogaster*)
BGREAT	0.57	0.65	10.59	11.84	**5.31**	**6.08**
BrownieAligner	**0.25**	**0.40**	13.36	22.88	7.99	13.23
deBGA	0.40	0.52	**3.16**	**4.70**	7.21	10.82

Table C.12 Effect of the branch and bound strategy on the run time (min) of BrownieAligner on simulated data

Tools	S1 (*E. coli DH10B*)	S2 (*E. coli DH10B*)	S3 (*H. sapiens* Chr. 21)	S4 (*H. sapiens* Chr. 21)	S5 (*D. melanogaster*)	S6 (*D. melanogaster*)
BrownieAligner	**0.06**	**0.06**	**2.31**	**4.25**	**2.60**	**2.90**
BrownieAlignerNoBB	0.07	0.07	6.90	20.89	3.24	3.65

C.5.3.2 Real data

Table C.13 shows the peak memory usage of aligners on real data. Table C.14 shows the run time (wall time) of aligners on real data.

Table C.13 Peak memory (GB) usage of the aligners on real data.

Tool	*B. dentium*	*E. coli DH10B*	*E. coli MG1655*	*S. enterica*	*P. aeruginosa*	*H. sapiens* Chr. 21	*C. elegans*	*D. melanogaster*
BGREAT	2.87	2.45	1.96	1.66	2.11	**2.53**	**3.24**	**3.44**
BrownieAligner	**0.59**	**0.91**	**1.05**	**0.67**	**0.98**	6.29	10.39	13.30
deBGA	10.19	10.12	10.29	8.89	8.92	9.41	10.12	10.39

Table C.14 Run time (min) of the aligners on real data

Tool	*B. dentium*	*E. coli DH10B*	*E. coli MG1655*	*S. enterica*	*P. aeruginosa*	*H. sapiens* Chr. 21	*C. elegans*	*D. melanogaster*
BGREAT	0.84	**1.19**	**1.66**	0.77	**0.98**	4.34	11.42	**6.26**
BrownieAligner	**0.58**	1.51	3.08	**0.38**	1.05	14.69	26.50	19.63
deBGA	0.59	1.47	3.16	0.50	1.14	**2.87**	**9.30**	10.78