COMPARISON OF DNA SEQUENCE ASSEMBLY ALGORITHMS USING MIXED DATA SOURCES

A Thesis Submitted to the
College of Graduate Studies and Research
in Partial Fulfillment of the Requirements
for the degree of Master of Science
in the Department of Computer Science
University of Saskatchewan
Saskatoon

By
Tejumoluwa Abegunde

Permission to Use

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Computer Science
176 Thorvaldson Building
110 Science Place
University of Saskatchewan
Saskatoon, Saskatchewan
Canada
S7N 5C9

Abstract

DNA sequence assembly is one of the fundamental areas of bioinformatics. It involves the correct formation of a genome sequence from its DNA fragments ("reads") by aligning and merging the fragments. There are different sequencing technologies — some support long DNA reads and the others, shorter DNA reads. There are sequence assembly programs specifically designed for these different types of raw sequencing data.

This work explores and experiments with these different types of assembly software in order to compare their performance on the type of data for which they were designed, as well as their performance on data for which they were not designed, and on mixed data. Such results are useful for establishing good procedures and tools for sequence assembly in the current genomic environment where read data of different lengths are available. This work also investigates the effect of the presence or absence of quality information on the results produced by sequence assemblers.

Five strategies were used in this research for assembling mixed data sets and the testing was done using a collection of real and artificial data sets for six bacterial organisms. The results show that there is a broad range in the ability of some DNA sequence assemblers to handle data from various sequencing technologies, especially data other than the kind they were designed for. For example, the long-read assemblers PHRAP and MIRA produced good results from assembling 454 data. The results also show the importance of having an effective methodology for assembling mixed data sets. It was found that combining contiguous sequences obtained from short-read assemblers with long DNA reads, and then assembling this combination using long-read assemblers was the most appropriate approach for assembling mixed short and long reads. It was found that the results from assembling the mixed data sets were better than the results obtained from separately assembling individual data from the different sequencing technologies. DNA sequence assemblers which do not depend on the availability of quality information were used to test the effect of the presence of quality values when assembling data. The results show that regardless of the availability of quality information, good results were produced in most of the assemblies.

In more general terms, this work shows that the approach or methodology used to assemble DNA sequences from mixed data sources makes a lot of difference in the type of results obtained, and that a good choice of methodology can help reduce the amount of effort spent on a DNA sequence assembly project.

ACKNOWLEDGEMENTS

I would like to formally thank:

My supervisor, Dr. Anthony Kusalik for providing me with this opportunity, and for his hard work, patient encouragement, and guidance throughout my studies.

My committee members, Dr. Ian Mcquillan and Dr. Barry Ziola for their guidance and support, and Dr. Andrew Sharpe for serving as my external examiner.

My fellow lab colleagues, for their friendship and support. Good luck to each of you in your future aspirations.

My parents, for their unending love and support in all my efforts and aspirations. Also to my siblings, David and Dammy for their love and support.

My best friend, Austin Ogun for your love and support. Also to Toyin Ake-Johnson for always ensuring I smile even when it seemed tough.

My friends that supported me during the course of my studies, I cannot mention all, but I appreciate you all.

I humbly dedicate this thesis to God for grace, strength, and guidance.

CONTENTS

List of Tables

List of Figures

List of Abbreviations

ABYSS	Assembly By Short Sequencing
ASCII	American Standard Code for Information Interchange
AT	*Anaerocellum thermophilum*
BAC	Bacterial Artificial Chromosome
CAF	Common Assembly Format
CAP3	Contig Assembly Program
CM	*Candidatus methano sphaerula palustris*
DNA	DeoxyriboNucleic Acid
EC	*Escherichia coli str. K12 substr. MG1655*
EDENA	Exact DE Novo Assembler
EULER-SR	Euler Short Read assembly
EXP	EXPeriment file format
FASTQ	Fasta + Qual format
HA	*Helicobacter acinoncychis*
MIRA	Mimicking Intelligent Read Assembly
MM	*Methanococcus maripaludis*
NCBI	National Centre for Bioinformatics Information
PCAP	Parallel Contig Assembly Program
PCR	Polymerase Chain Reaction
PHRAP	PHRagment Assembly Program
PHRED	PHil's Read EDitor
RAM	Random Access Memory
SCF	Standard Chromatogram Format
SS	*Streptococcus suis*
SSAKE	Short Sequence Assembly by progressive K-mer search & 3' read Extension
SPSS	Statistical Package for Social Sciences
TIGR	The Institute of Genome Research

Chapter 1

Introduction

DNA sequence assembly is one of the core areas in bioinformatics. It involves correctly aligning and joining DNA fragments to derive genome or plasmid sequences.

Currently, there are various types of sequencing technologies and some emerging ones. The Sanger dideoxy sequencing technology has been widely used since 1970; this sequencing technology produces DNA fragments of length between 400-900 bp ("long reads"). There are sequence assembly algorithms that have been designed to work with this type of data. However, since the year 2000, new sequencing technologies have emerged (and some are emerging); these are known as the *next generation sequencing technologies*. Some of these technologies include 454 sequencing, Illumina sequencing, Helicos sequencing, and Solid sequencing; and they produce DNA fragments of varying lengths from 35-400 bp ("short reads"). There are also sequence assemblers designed to work with these newer sequencing technologies.

When dealing with a sequence assembly project that involves a combination of the sequencing technologies, it is important to have a proven strategy to assemble the data sets from the mixed data sources. This research work involves experiments on various sequence assemblers to determine the most appropriate strategy to use when assembling data from mixed sources (combinations of Sanger, 454, and Illumina data) and also to determine the most appropriate sequence assembler (or assemblers) to employ.

The following points highlight the motivation for this research:

1. Sequence assemblers for Sanger sequence data ("long-read" data) existed before the development of the next generation ("short-read") sequencing technologies. Why develop new assemblers then? One of the reasons could be that earlier assemblers cannot handle the large number of reads from the next generation sequencing technologies, or can handle them, but generate poor results. Is this the case? There are many more reads involved with the short-read technology compared to the Sanger technology, and hence more overlaps. This could be more challenging, to the extent that the older assemblers may no longer be practical. For

example, the TIGR Assembler [5], which is for long-read assembly, does not work with more than 524,288 reads. However, other long-read assemblers may not have such a maximum. This work revealed how well the long-read assemblers handle short-read data.

2. One might expect the short-read assemblers to deal easily with the longer reads since the latter presents a smaller (and hence less complicated) problem. Is this the case? Or, do the short-read assemblers make assumptions that prevent them from working, or working well, with the long-read data? For example, Edena [23] (a short-read assembler) works with DNA reads of constant length, but it is uncommon for long reads to be all of the same length (with the exception of artificially-generated reads). Also, are there less obvious assumptions that could handicap the short-read assemblers when used on long-read data? Assembling the long-read data using short-read assemblers should be an "easier" problem. Does the assembly produce better results? Is it faster? This work answers these questions.

3. The short-read assemblers supposedly implement newer sequence assembly techniques, so one would expect them to have benefitted from lessons learnt from the previous generation of assemblers. Is this the case? Are the short-read assemblers just different or are they better? This work provides answers.

4. When researchers in the life sciences are involved in a sequencing project, they face the choice of which sequence assembly software to use. The problem is made worse if the read data comes from more than one sequencing technology. Which assembler and what assembly strategies should a life sciences researcher use (in a given situation)? This work serves as a partial guide.

The aim of this research was to compare some of the existing tools designed for sequence assembly. This helps contribute to establishing good methodologies for sequence assembly from DNA read data of different lengths. It also helps determine the best (or most appropriate) sequence assembler to use in various situations. Based on the results, the appropriate sequence assemblers for assembling 454 and Illumina data sets are EULER-SR and VELVET, and for Sanger, PHRAP and MIRA. When working with mixed data that involves Sanger, 454, and Illumina, the best assembly methodology based on this research is to assemble the short reads first, and merge the contigs produced from the short-read assembly with Sanger reads, and then to assemble with a long-read assembler. And when working with 454 and Illumina reads, the appropriate methodology is to assemble the 454 and Illumina reads separately, and merge the resultant contigs, and then assemble these using a long-read assembler. This document describes how these conclusions were derived.

1.1 Thesis Organization

The rest of the thesis is organized as follows. Chapter 2 describes briefly Sanger, 454, and Illumina sequencing technologies and then introduces and discusses the various sequence assemblers used in this research; it highlights the different approaches to sequence assembly implemented by each sequence assembler. Chapter 3, data and methodology, introduces the data sets (both real and artificial) and assembly approaches used for the research. Results from the assemblies and evaluations are presented in Chapter 4. The conclusions of this research and possible future work are discussed in Chapter 5.

Chapter 2

Background Information

DNA sequencing technologies have been available since the 1970s [7] and are still evolving.

2.1 DNA Sequencing

Every genome sequencing project begins with a sequencing phase (Figure 2.1). DNA sequencing is the procedure used in determining the identity and order of nucleotides in DNA fragments [9].

Earlier DNA sequencing techniques supported the production of “long” DNA fragments of length 400-800 bp (*DNA fragments* or *reads* are a series of nucleotides sequences (A, T, G, C)). These long reads are produced from the Sanger dideoxy sequencing technique [7, 41]. Newer technologies produce shorter DNA fragments. Examples are the 454 sequencing technique [22, 31] which is capable of producing DNA fragments with length of 200 - 400 bp and the Illumina sequencing technique [10] which produces DNA fragments of length 20 - 100 bp.

The two main sequencing approaches are as follows.

1. Shotgun sequencing

   This is the most widely used sequencing method. The process of shotgun sequencing begins with randomly shearing the original genome into tiny fragments. Each of the fragments is cloned or amplified (making identical copies of each fragment) and then each clone is sequenced (Figure 2.2). This can be done with any of the three previously mentioned sequencing technologies.

   - **Sanger Sequencing**

     When generating Sanger data (long reads), the fragments are attached to sequencing vectors (a *vector* is a circular genome such as a plasmid or cosmid that incorporates the fragment [6]) which serve as a mechanism for clonal amplification of each fragment (Figure 2.2) [7]. This process leads to the creation of DNA reads which serve as input to the sequence assembly phase (Figure 2.1). Cloning vectors are not used for Illumina

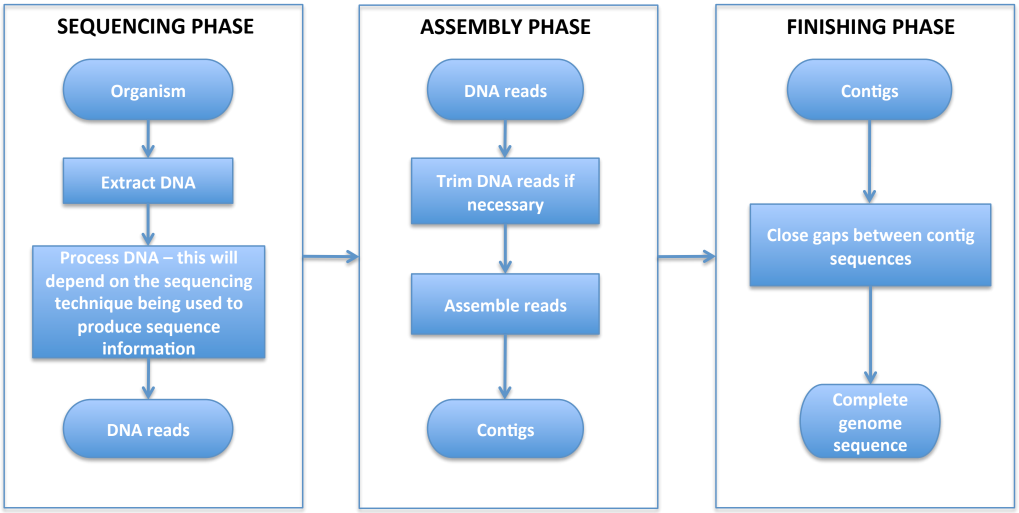


Figure 2.1: Phases involved in a shotgun genome project

and 454 data.

- **454 Sequencing**

  The 454 sequencing technology is based on pyrosequencing and uses emulsion-based clonal amplification. This involves carrying out PCR (polymerase chain reaction) in an emulsion mixture (which contains capture beads, enzyme reagents in a water mixture, synthetic oil and the DNA fragment) to generate replicates of the fragment [3, 22]. The 454 technology is faster and less expensive on a per base cost than the Sanger sequencing technology. Also, unlike the Sanger technology, unclonable regions are not skipped. This means better genome coverage.

- **Illumina Sequencing**

  The Illumina sequencing technology is based on sequencing by synthesis and uses a solid surface for bridging PCR amplification. The reaction occurs on the surface of a flow cell. A flow cell has eight lanes for simultaneous analysis. The randomly generated precursor DNA fragments are bound to the inside surface of the flow cell and then unlabelled nucleotides are added along with enzymes to begin amplification [27]. The flow cells are designed to easily present the DNA fragments in a manner that facilitates exposure to the reactants.

2. Clone-contig or Map-based sequencing

   This approach begins by cutting the genome into DNA fragments in an organized manner (Figure 2.3), such that the position of each fragment is known. The genome may be cut using a restriction enzyme. A library of overlapping DNA fragments is generated. Each of the DNA

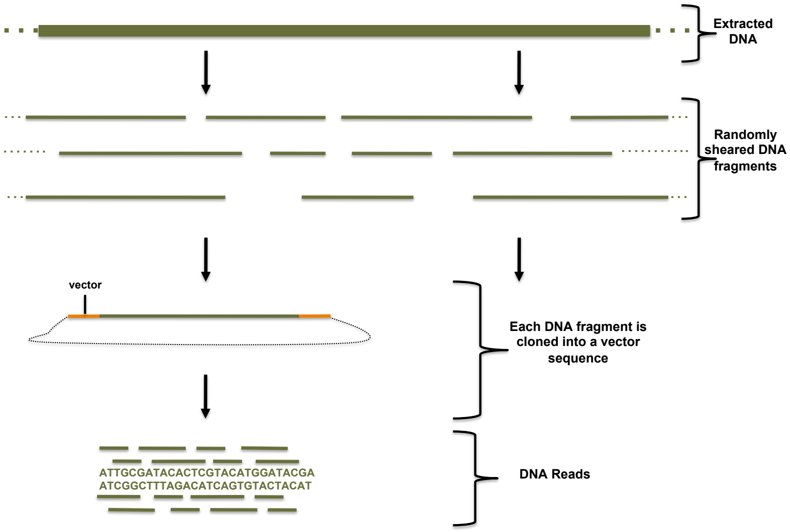


Figure 2.2: Sequencing using the shotgun approach

fragments is cloned (for example, using a BAC (Bacterial Artificial Chromosome) clone) for amplification and then sequenced [6, 9, 33]. These DNA fragments can be sequenced either using the shotgun sequencing approach described earlier or through primer walking. Primer walking involves repeatedly using the sequence of a known DNA fragment to design primers for sequencing the next contiguous fragment, terminating with the last piece of DNA [4, 33].

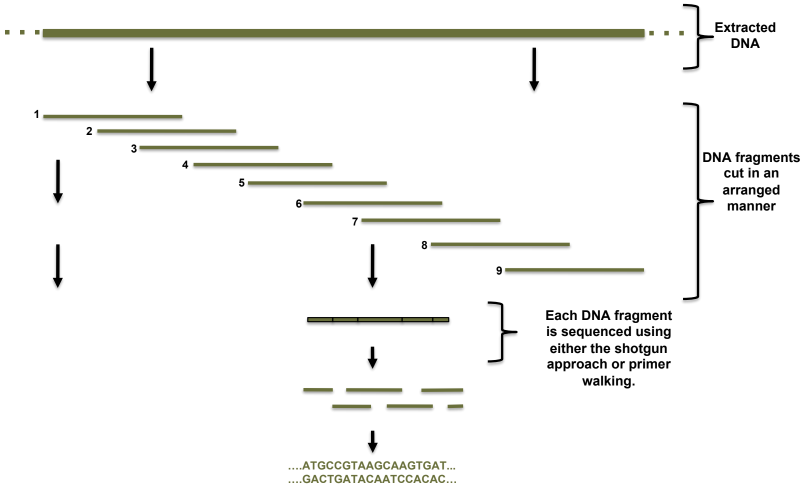


Figure 2.3: Sequencing using the map-based approach

2.2 Sequence Assembly

The shotgun sequencing approach is used in the majority of genomic sequencing projects [9, 10]. As a result of the fragmentation of DNA, the sequence data must be fit together (assembled) to derive the whole sequence of the original DNA (Figure 2.4). This process is known as *sequence assembly* (Figure 2.1).

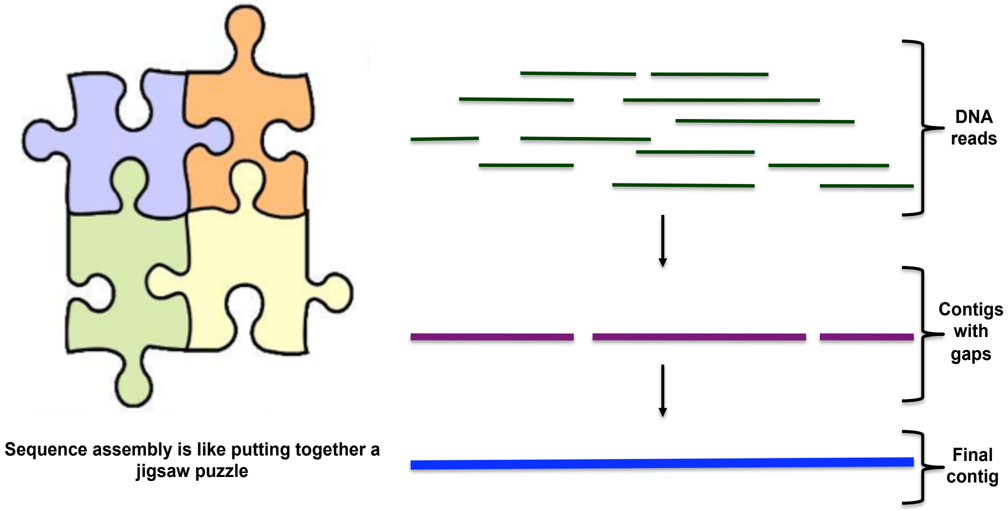


Figure 2.4: Sequence assembly involves correctly joining DNA reads

The main goal of sequence assembly is to determine the entire sequence of a target genome at a desirable level of confidence. This process involves the correct aligning and merging of DNA fragments to form long contiguous sequences (*contigs*) (Figure 2.4). There are various obstacles to the correct construction of a contiguous sequence. These obstacles include the following:

- Contaminant reads — This applies to DNA fragments which may still contain the vector sequence used for amplification during the sequencing process. This may cause errors in the assembly process.
- Chimeric reads — This occurs when DNA fragments from different regions of a genome are merged into a single DNA read which results in the production of reads with inaccurate links between them. This could result in the generation of contig(s), where portions of the contig match totally different regions of the genome (Figure 2.5) [7]. Chimerism often occurs in Sanger sequencing because clones may be accidentally merged together before sequencing.
- Bias in the reads — This occurs when not all regions of the original genome are covered equally; that is, certain regions of the genome may be cloned less often than the others during

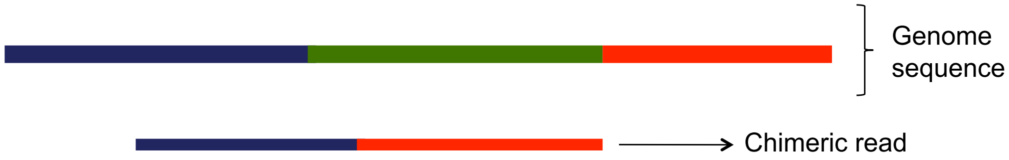


Figure 2.5: Formation of a chimeric read. The blue, green and red colours represent different regions of a genome sequence. A read is chimeric when it contains non-overlapping portions of the genome sequence. The bars represent series of sequences.

sequencing which results in lower read coverage in these regions. This is often a result of sampling or cloning biases [26].

- Insufficient genome coverage — This occurs when there are some areas of the genome not covered. This leads to gaps in the assembly.

- Repetitive reads — This is the most prominent difficulty in the assembly process. It occurs when there are sequences that appear in more than one place. The presence of repeats (repetitive portions of sequence within a DNA fragment) makes it difficult to distinguish a true overlap (which may result in merging DNA reads during contig formation) from a repeat-induced overlap (Figure 2.6). For technologies with shorter reads, the repetitive read problem is more pronounced — the shorter the read length, the larger the number of potential repeats [10].

2.3 Sequence Assemblers

There are many sequence assemblers readily available, both for the long- and short-read sequencing technologies. There are different approaches employed by these sequence assembly algorithms. Three of these approaches are described below.

1. Overlap-Layout-Consensus approach

   This is the most widely used approach in sequence assembly algorithms; it views the sequence assembly problem as a Hamiltonian path problem which involves determining a path in an undirected graph which visits all the nodes in the graph exactly once (where the nodes represent the DNA fragments and the edges between nodes represent overlapping fragments).

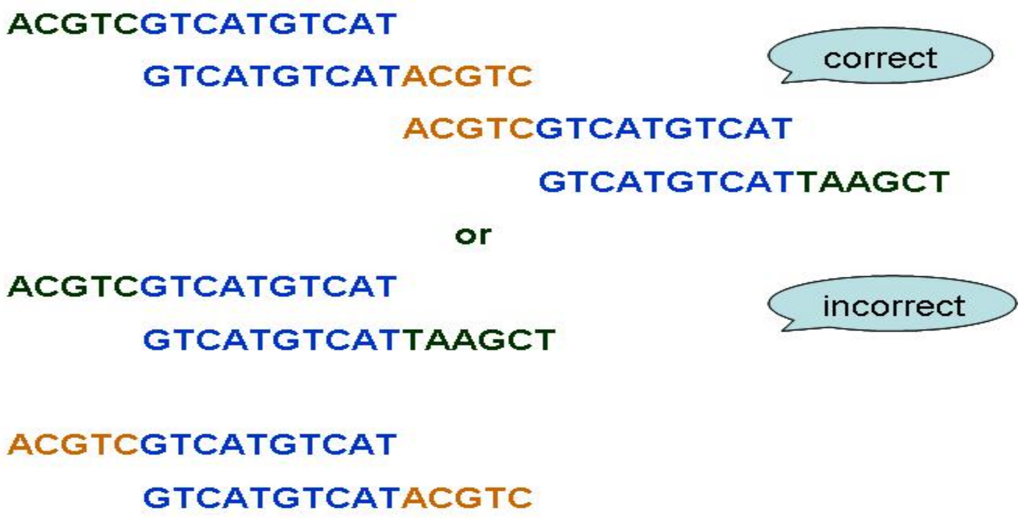


Figure 2.6: Correct layout for repetitive regions and an erroneous layout that could result in misassembly [26]. The repeats are in blue type.

- Overlap stage - This stage involves creating an overlap graph in which every node is a DNA fragment and arcs indicate an overlap (Figure 2.7). Some of the assemblers which use this approach include PHRAP (PHRagment Assembly Program) [38], CAP3 (Contig Assembly Program) [25], TIGR Assembler (The Institute of Genome Research) [5], and PCAP (Parallel Contig Assembly Program) [24].

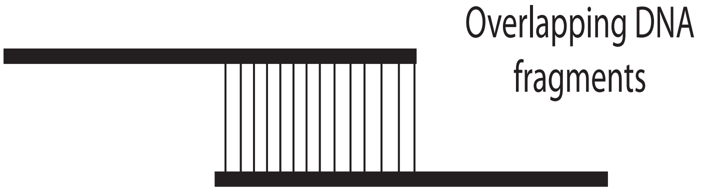


Figure 2.7: Overlap stage - find potentially overlapping fragments

- Layout stage - This stage involves determining which overlaps will be used in the final assembly. It involves finding the correct order and arrangement of the DNA fragments which will be joined into contigs (Figure 2.8). The overlap graph is simplified at this stage by removing redundant information caused by repetitive DNA fragments.

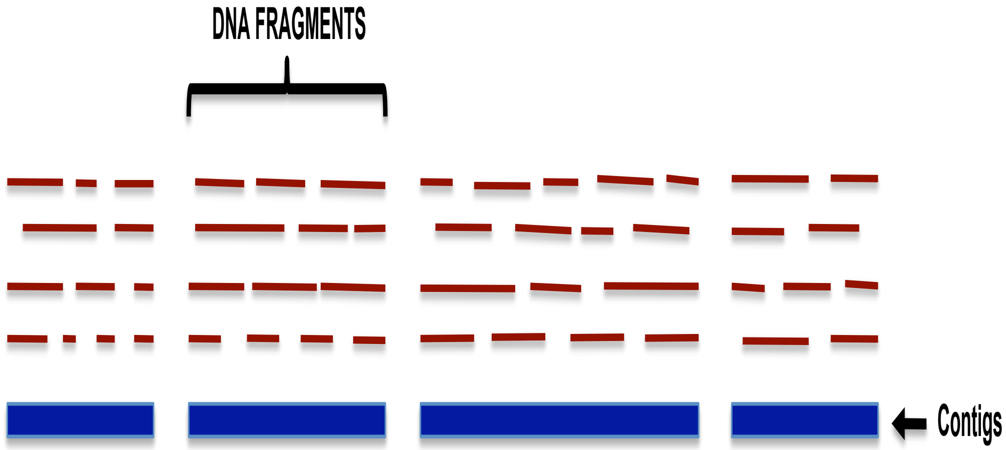


Figure 2.8: Layout stage of the overlap-layout-consensus approach.

  - Consensus stage - This phase involves building the final sequence from the layout stage and correcting DNA fragment errors.

- Eulerian path approach

  This approach involves finding a path in a de Bruijn graph (Figure 2.9) which visits every edge of the graph exactly once. A de Bruijn graph is a directed graph that represents sequences of symbols (in this context, DNA sequences), and the edges of the graph indicate where the sequence may overlap. The Eulerian path algorithm does not align pairs of DNA fragments (i.e. pairwise comparisons), rather each DNA fragment is cut into smaller pieces (k-mers) and then a de Bruijn graph is created where the nodes of the graph represent the k-mers and every path in the graph corresponds to a DNA read [30]. The fundamental structure of the de Bruijn graph is useful for representing overlaps between k-mers — this means high redundancy will naturally be handled by the de Bruijn graph without affecting the number of nodes. The de Bruijn graph presents a simpler representation of repeats than the overlap graph in the overlap-layout-consensus approach. Some of the assemblers which use this approach include Euler [35] and Euler-SR (Euler Short Read) [10].

- Align-layout-consensus approach

  There are several genomes that have already been sequenced, assembled and finished. If the genome being sequenced is closely similar to an already sequenced genome, then

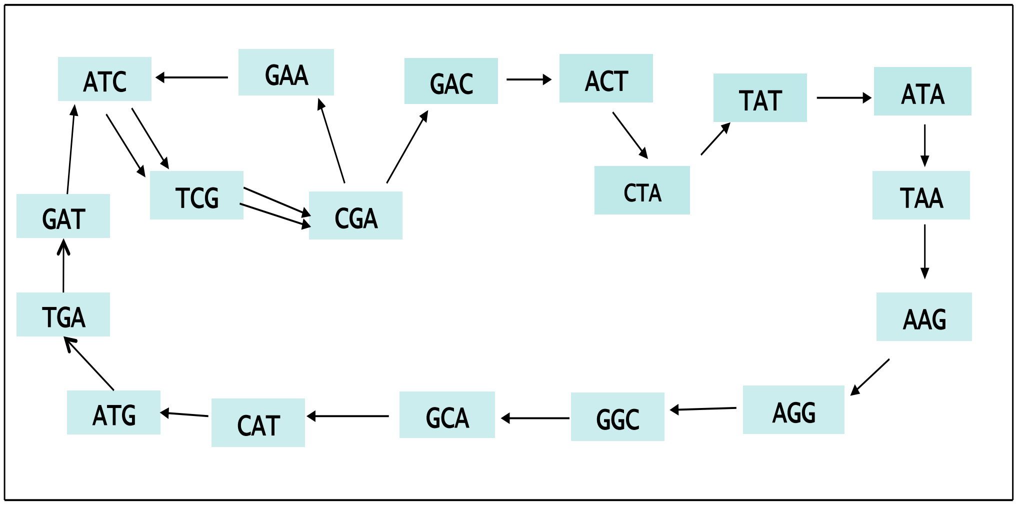


Figure 2.9: An example of a de Bruijn graph. A de Bruijn graph has nodes with exactly n incoming and n outgoing edges.

comparative sequence assembly can be used. This approach involves assembling DNA fragments by aligning the DNA fragments to the already sequenced genome [2]. An example of sequence assemblers which use this approach include AMOS (A Modular Open-Source whole genome assembler) [37] and Mosaik [48].

2.3.1 Long-read assemblers

Some of the assemblers which support long reads include PHRAP [38], CAP3 [25], TIGR Assembler [5], PCAP [24] and MIRA (Mimicking Intelligent Read Assembly) [12].

PHRAP

PHRAP is an algorithm designed to assemble shotgun sequences [38]. PHRAP was developed by Phil Green in 1999 and is widely used for sequence assembly projects. It was designed to assemble Sanger sequences and implements the overlap-layout-consensus approach. PHRAP uses the Smith-Waterman sequence alignment algorithm [46] to compare all sequences to find pairs of matching subsequences. It then generates its own quality information based on DNA read-to-read confirmation (that is, confirmed DNA matches from the alignment results). This quality information is generated regardless of the quality values that may be supplied with the input DNA reads. However the supplied quality values may improve the value of the internally generated quality information.

The process of PHRAP sequence assembly is described in the flowchart in Figure 2.10. PHRAP preassembles reads into groups before merging the groups into a contig sequence; this is useful

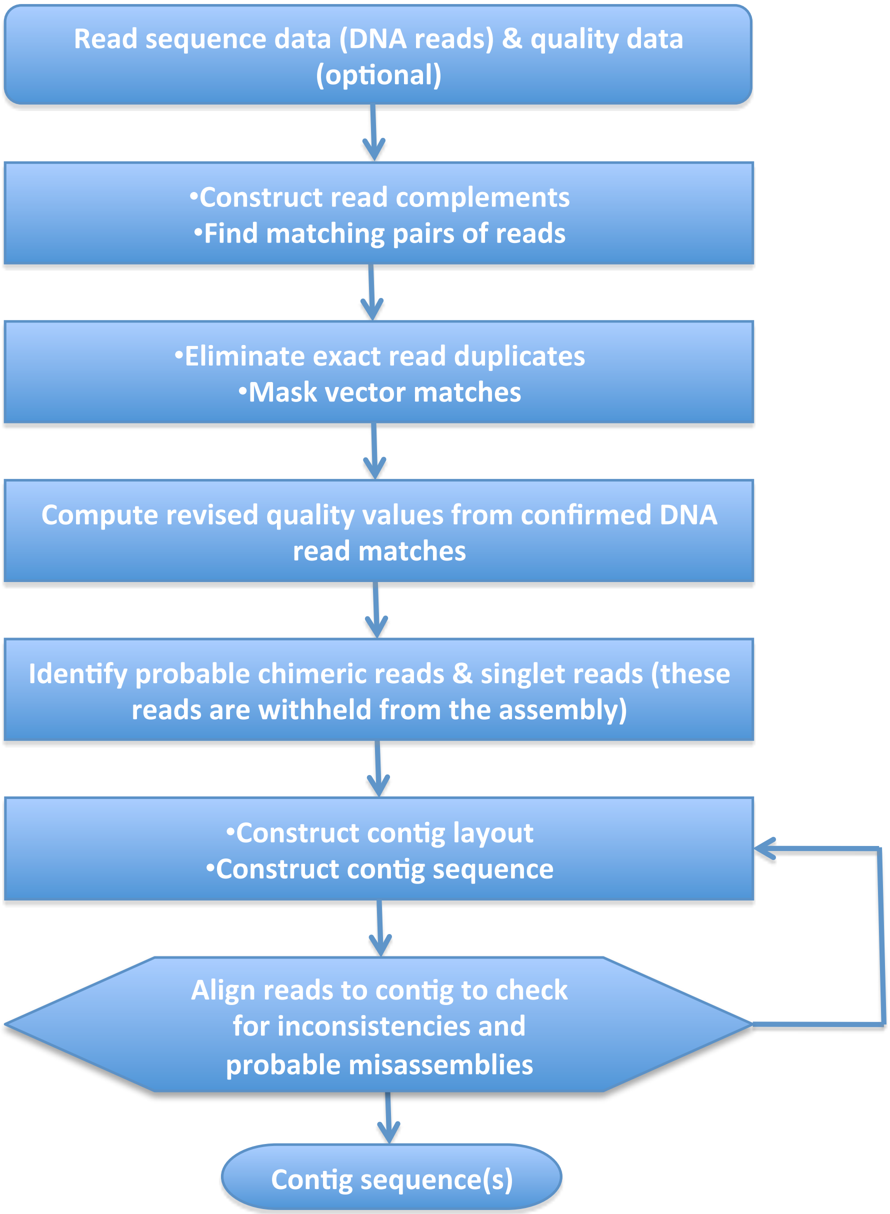


Figure 2.10: PHRAP sequence assembly process.

in reducing the risk of joining of reads incorrectly as a result of repeats (repetitive sequences). PHRAP's key features are:

- combines user-provided quality information and its internally computed quality information to improve the accuracy of contigs produced;
- constructs contig sequences from an overlap of the highest quality parts of reads. (This applies mostly when quality data is supplied at the beginning of the assembly process);
- provides exhaustive information about the assembly (singlet reads file which contains DNA

reads omitted from assembly process because they did not match any other read, quality values for contig sequences, a table displaying probable inconsistencies or mis-assemblies and other information) to help with troubleshooting;

- ability to handle large data sets; and
- portability across Unix/Linux, Mac OSX and Windows operating systems.

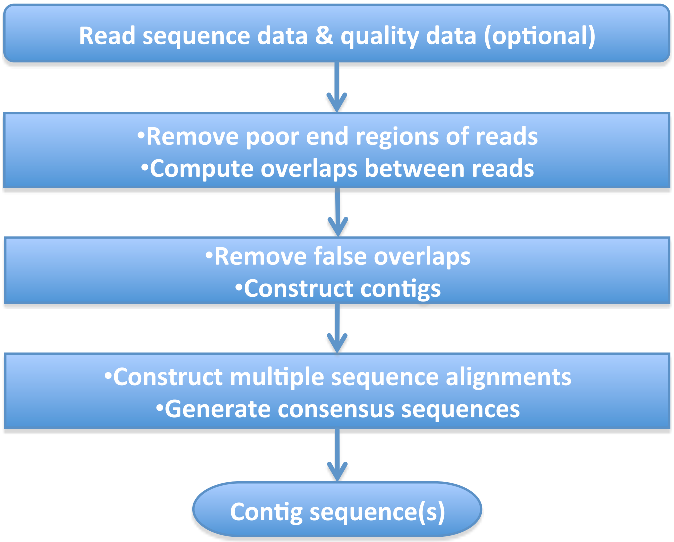
CAP3

CAP3 was developed by Huang and Madan in 1999 [25]. It is designed to work with shotgun sequences and it implements the overlap-layout-consensus approach. The operational stages within CAP3 are described in the flowchart in Figure 2.11(a). In its overlap stage, poor end regions of the DNA reads are located and removed (e.g., pieces of vector sequences). These regions are located by aligning the sequences to each other; when a long high quality region of a read is highly similar to another read, such regions are designated as good regions, otherwise they are poor regions. Once the poor ends have been removed, the Smith-Waterman sequence alignment algorithm [46] is used to determine the overlaps between the DNA reads. In the layout stage, false overlaps are removed; each overlap is measured based on length of overlap, percent identity, similarity score, and difference rate of overlaps at bases of high quality values (this applies only when quality values are supplied with the input DNA reads, if the difference rate is higher than expected, the overlap is probably false). Providing read-pair information for DNA reads improves the accuracy of contigs produced by CAP3 because CAP3 uses this information to link contigs (create scaffolds) and correct possible assembly errors. The key features of CAP3 include:

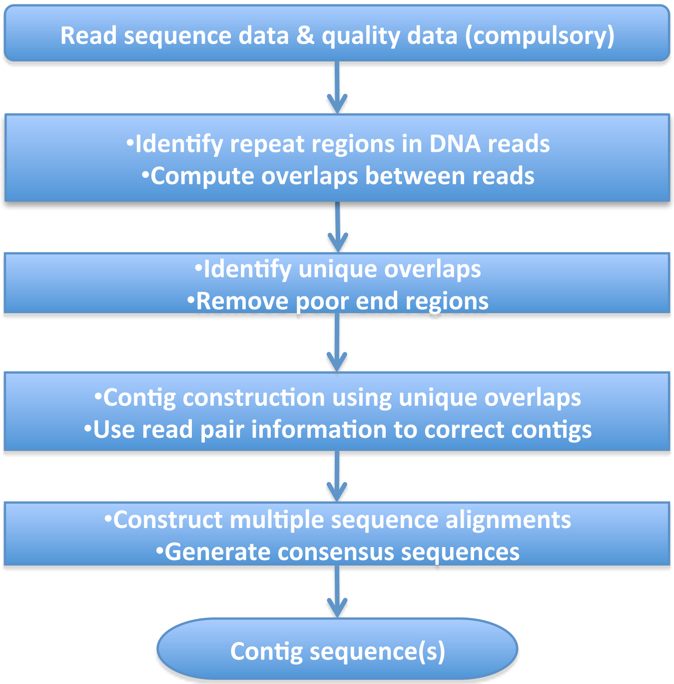
- ability to identify and remove poor end regions in the DNA fragments;
- uses quality values (provided as input) to compute overlaps between reads, for the multiple sequence alignment of the DNA fragments, and for the construction of consensus sequences;
- uses read-pair information to correct errors from the assembly process and also to link contigs (this is useful in creating scaffolds); and
- available on Unix/Linux and Mac OSX operating systems.

PCAP

PCAP was developed by Xiaoqiu Huang et al. in 2003 [24]. It is the parallel version of the CAP program (first release of the CAP3 assembler) and it is also useful for assembling large genomes. The PCAP algorithm begins by identifying and removing poor regions in the DNA reads (for example, vector contaminated regions). The next step looks for DNA reads with unique overlaps (these are considered to be potential reads to be joined). The algorithm then identifies repetitive regions (false overlaps) among the DNA reads. Contigs are constructed using the unique overlaps and if read-pair information is provided, this is used to correctly link contigs. The key feature of the PCAP algorithm (different from CAP3) is its ability to run on multiple processors. The flowchart in Figure 2.11(b) describes the PCAP algorithm.



(a) CAP3 sequence assembly process.



(b) PCAP sequence assembly process.

Figure 2.11: Flow charts for the CAP3 and PCAP sequence assembly process.

TIGR Assembler

The TIGR assembler was developed by the Institute for Genome Research in 1995 [5]. TIGR assembler was designed for shotgun sequencing projects. TIGR assembler considers all k-mer oligonucleotides (for some specific value of k) between DNA reads, and then pairs of DNA reads with a high degree of k-mer similarity are considered as potential overlaps and lists of potential overlaps are generated. The DNA reads from repeat regions are identified based on the number of potential overlaps each read has in the pair-wise comparisons [49]. When a DNA read is selected to begin assembly, TIGR uses the Smith-Waterman algorithm to evaluate the potential overlap between the current assembly and further DNA reads that would be considered and merged.

The flowchart in Figure 2.12 gives an overview of the internal processes within the TIGR assembler. Its key features are:

- identifies repeat regions before it assembles any reads,
- performs a second pair-wise comparison on DNA reads to reduce the risk of merging DNA reads incorrectly,
- uses quality information to improve accuracy of contigs generated, and
- is portable across Unix/Linux and Mac OSX operating systems.

MIRA

MIRA was developed by Bastien Chevreux in 2005 [12]. It was originally designed to work with Sanger sequences, but its current version is able to perform hybrid assemblies using Sanger/454, 454/Illumina, and Sanger/Illumina/454 mixed data sets. The MIRA assembler follows the overlap-layout-consensus approach to assemble DNA sequences. It begins with checking the DNA reads for contaminated regions (pieces of vector sequences); these regions are identified and masked. The next stage involves identifying potential read overlaps. Once this is completed, the Smith-Waterman alignment algorithm [46] is used to perform a pair-wise alignment to confirm the overlaps between the DNA reads. Once the overlaps are confirmed, the assembly (layout) graph is built. Before the contigs are generated, the MIRA algorithm checks for possible misassemblies that may occur as a result of repetitive regions or errors in DNA reads. These errors are corrected. Contig sequences are generated after the error corrections. The flowchart in Figure 2.13 describes the MIRA algorithm. Its key features are:

Figure 2.12: TIGR Assembler sequence assembly process

- depends on the provision of quality information to improve the quality of consensus sequences,
- able to perform hybrid sequence assemblies,
- uses paired-end reads to improve assembly,
- portable on Linux/Unix and Mac OS X operating systems, and
- supports different input formats: FASTA, FASTQ, CAF (Common Assembly Format), EXP (EXPeriment file format), phd (from PHRED); and different output formats understood by most finishing tools: FASTA, ace, CAF, gap4, and GenBank.

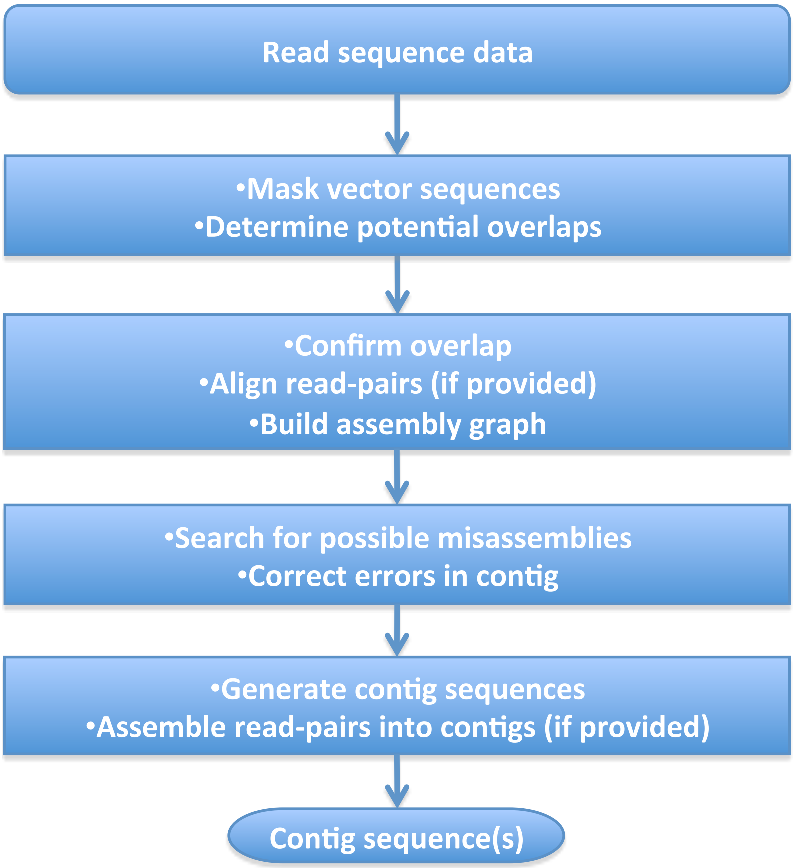


Figure 2.13: MIRA sequence assembly process

2.3.2 Short-read assemblers

Some of the available assemblers which support short reads include SSAKE (Short Sequence Assembly by k-mer search and 3' read Extension) [29], Velvet [52], EDENA (Exact DE Novo Assembler) [23], Euler-SR [10] and ABYSS (Assembly By Short Sequencing) [44].

SSAKE

SSAKE was developed by Warren et al. in 2006 [29]. It is designed to assemble Illumina sequencing data, and it provided the assembly foundation for all subsequently published short-read assemblers. The SSAKE algorithm begins by organizing the DNA reads and their reverse complements using a prefix tree. This is to enable efficient k-mer searches and determination of read overlaps. Next, the algorithm searches through all the overlaps to find the longest possible overlaps, and then begins construction of consensus sequences by joining the DNA reads from the overlaps. Its key features are:

- uses prefix trees for the k-mer searches,

- uses read-pairs to assess contig assembly quality, and
- portable on Linux/Unix operating system.

VELVET

Velvet assembler was developed by Zerbino and Birney [52]. It is designed to assemble short DNA reads. The Velvet algorithm implements a de Bruijn graph for assembling sequences. It represents DNA reads as k-mers using de Bruijn graph theory. The Velvet algorithm begins by generating k-mers from the DNA reads. For all the k-mers, a hash table is used to store the identity of the DNA read containing the k-mer and the position of its occurrence(s). The de Bruijn graph is constructed and a database is created which stores for each DNA read and the number of original k-mers that are overlapped by subsequent DNA reads. This is to keep track of all read overlaps. The next step in the velvet algorithm involves correcting errors in the graph caused by sequencing errors and repeats. After the error correction, contigs are built from the paths in the graph.

The flowchart in Figure 2.14 describes the Velvet algorithm. The key features of Velvet include:

- ablity to mix long and short DNA reads which is useful in eliminating gaps,
- use of short read-pairs to extend and link contigs,
- able to handle high redundancy of short-read sequencing data,
- able to use long reads to resolve repeats, and
- portable on Linux/Unix, Mac OS X, and Cygwin operating systems.

EDENA

EDENA was developed by Hernandez et al. in 2008 [23]. It is designed to work with Illumina sequencing data sets (data sets containing very short reads of the same length). The EDENA algorithm is based on the overlap-layout-consensus approach (Figure 2.15). It begins by processing the DNA reads to remove redundant reads and retains a single copy of each DNA read; this is done by arranging the reads in a prefix tree. A suffix tree is then used to arrange the non-redundant DNA reads; this tree structure is implemented to reveal exact read overlaps among the DNA reads, and these overlaps are used in constructing an overlap graph (based on a set overlap value, because EDENA assumes all DNA reads to be of the same length).

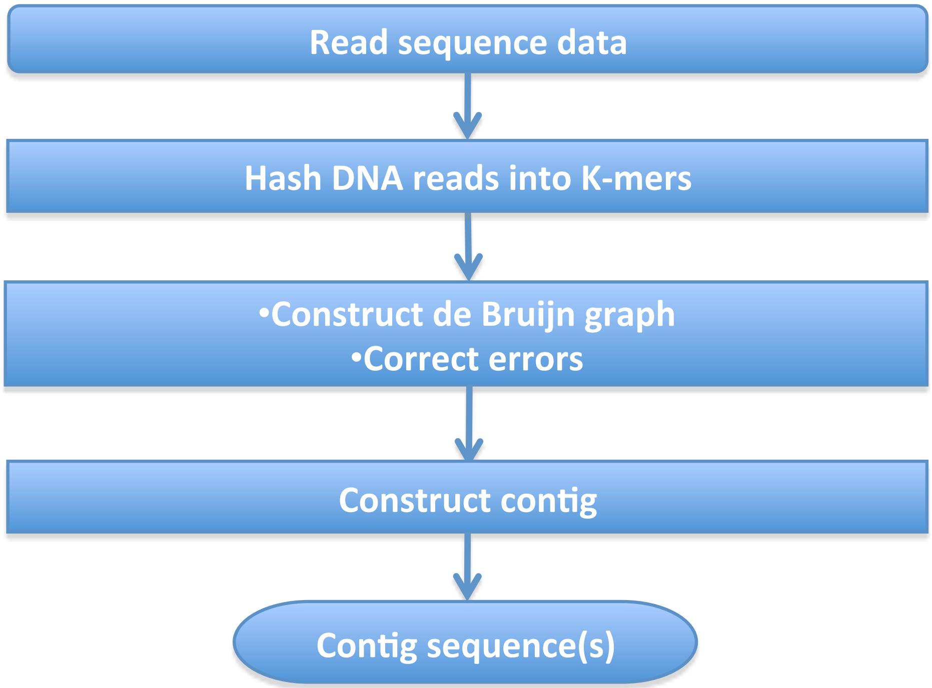


Figure 2.14: Velvet sequence assembly process

Once the overlap graph is constructed, transitive edges are removed, and then EDENA examines the branching nodes and only edges in the graph that maximize the overlap value are retained. Contigs are constructed by following the paths in the graph. Its key features are:

- relies on exact DNA read matches instead of approximate matching, this reduces false overlaps;
- it improves the assembly of DNA sequences by removing nodes from the graph resulting from base-calling errors; and
- portable on Linux/Unix operating system.

EULER-SR

Euler-SR was developed by Chaisson et al. in 2009 [10]. The Euler-SR algorithm is a modification of the Euler assembler developed by Pevzner et al. in 2001 [36, 35]. It uses the de Bruijn graph to construct an assembly. The first step in the Euler-SR algorithm involves pre-processing all the DNA reads to remove or trim erroneous reads. If quality files are provided along with the input base files, Euler-SR uses quality information of the reads to trim or remove low-quality reads. If no quality files are provided, it implements an algorithm known as the spectral alignment [35] to determine erroneous reads.

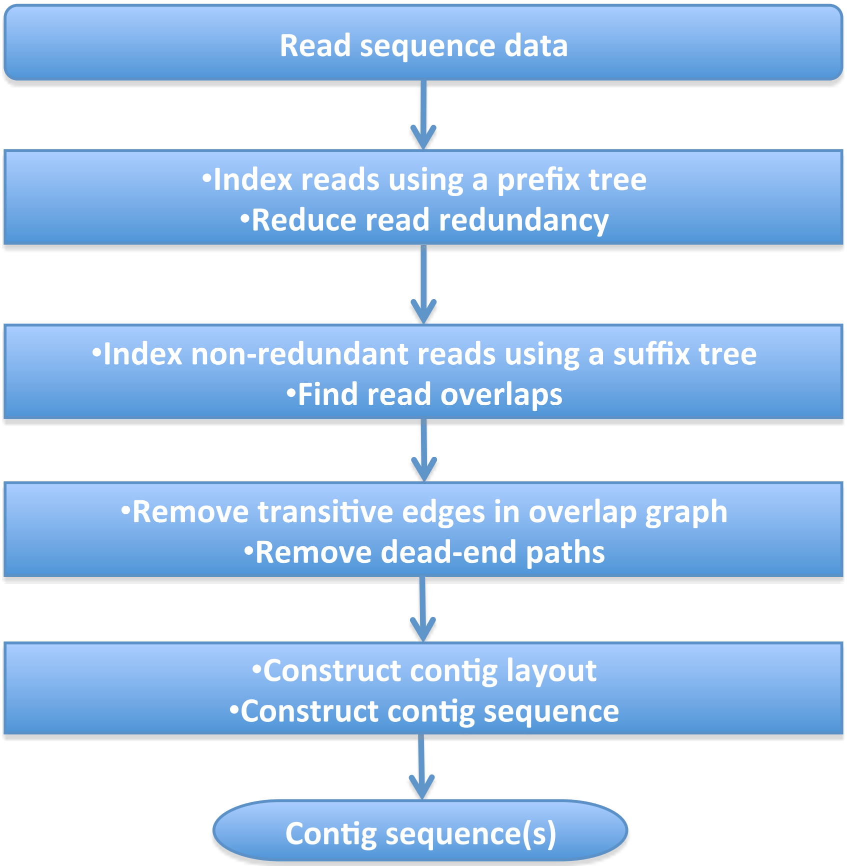


Figure 2.15: EDENA sequence assembly process

The spectral alignment algorithm selects a DNA read and a set of l-tuples called a spectrum (a l-tuple is a sequence of l DNA sequences, where l is a positive integer). The set of l-tuples is chosen by counting the frequency of all l-tuples present in all the DNA reads and selects tuples that occur with multiplicity above a certain threshold [11]. The spectral alignment algorithm determines erroneous reads by looking for the minimum number of substitutions, insertions, and deletions in each DNA read required to make every l-tuple in each DNA read belong to the spectrum. The majority of the Euler-SR execution time goes into pre-processing the DNA reads. Once the erroneous reads are removed, the de Bruijn graph is constructed, and then errors in the graph are corrected (such as, transitive edges). The last stage of the Euler-SR algorithm involves transforming the paths in the graph into assembled sequences (contigs). Its key features are:

- pre-processes DNA reads to remove errors without quality values (uses quality values if provided);
- uses mate-pairs to improve accuracy of contigs; and
- portable on Linux/Unix operating system.

ABYSS

ABYSS was developed by Simpson et al in 2009 [44]. It is designed to work with Illumina sequencing data sets. The ABYSS algorithm implements the assembly process using a distributed de Bruijn graph which allows for parallel computation of the algorithm. The algorithm begins by generating k-mers from the DNA reads; these k-mers are then loaded on a distributed de Bruijn graph (the sequences can be distributed over a cluster of computer nodes). Once the de Bruijn graph has been constructed, the adjacency information between k-mers is stored and this is independent of the actual location of the k-mers. In the next stage, erroneous vertices and edges created by sequencing errors are removed. The last stage in the ABYSS algorithm involves building contigs from the paths in the de Bruijn graph, and if read-pair information is available for the DNA reads, this information is used to link contigs and remove erroneous links caused by misaligned reads. Its key features are:

- parallel computation of sequence assemblers, which allows for short assembly run-time;
- uses mate-pair to correct errors and improve contig formation; and
- portable on Linux/Unix, Mac OS X, and Windows operating systems.

2.4 Finishing Phase

This phase involves generating an accurate complete genome sequence from contigs produced by sequence assemblers. This phase involves correcting sequence assembly errors, gap closing, and joining contigs. This phase is not within the scope of this research. However, the results of this thesis may have an impact on the finishing stage. For example, the task of finishing would be easier for the output of assemblers which produce fewer but larger contigs, and which have lower "misassembled contig" values.

2.5 Objectives of the research

Because of the large number of sequence assemblers available (some of which were discussed above), and the availability of different types of sequence reads, it is difficult to know the best sequence assembler to use for any given genomic project. The goal of this research was to compare some of the existing tools designed for sequence assembly. This will contribute to establishing good methodologies for sequence assembly from DNA reads from different sequence technologies (Sanger/454/Illumina). It will help determine the best (or most appropriate) sequence assembler to use in various situations (for example, a genome project that involves data sets from a mixture of sequencing technologies). This was achieved by answering the following questions:

1. What are the trade-offs in the following cases:

- assembly of long reads by using short-read sequence assemblers versus assembly of short reads using short-read assemblers, and
- assembly of short reads using long-read assemblers versus assembly of long reads using long-read assemblers?

2. How do long- and short-read assemblers compare in terms of:
    - accuracy of results (in terms of genome coverage, number of contigs, number of missassemblies, and N50 size of contigs),
    - execution time,
    - need for manual intervention, and
    - memory usage?
3. What is the effect of the presence (or absence) of quality values on the accuracy of contigs generated by the various sequence assemblers?

The next chapter presents the data and discusses the methodology used to answer these questions.

CHAPTER 3

DATA AND METHODOLOGY

This work involved experiments on data obtained with the Sanger, 454 and Illumina sequencing techniques. The methodology focused on the sequence assemblers and on how well each sequence assembler (both short-read and long-read types) could handle data from the three sequencing technologies or a combination of data from these technologies. These experiments were useful in determining the appropriate strategy when dealing with different kinds of data, and especially the appropriate assembler or assemblers to be used. This work also involved determining the effect of the presence or absence of quality information when assembling DNA reads.

The sequence assemblers used in this work were chosen because they are freely available to academic users and there have been publications describing them. All the organisms for this research were chosen based on the availability of DNA sequencing data and genomic sequence for them.

The following sections describe the data sets used in this research and the methodology followed.

3.1 Data

All the organisms chosen for this work are bacteria (prokaryotes) which have already-sequenced genomes. Two kinds of data sets were used for this work — real and artificial data sets. The real data sets for the organisms are the original DNA reads which were assembled to derive genome sequences in public databases for the organisms. The artificial data sets were created from the whole-genome sequences to supplement some of the real data sets; for example, if an organism had real 454 and Illumina data but no Sanger data, then an artificial Sanger data set was generated for the organism. The data sets for this work were from six organisms (the organisms are described in Table 3.4). Some of the organisms have 454 data, some have Illumina data, and some have Sanger data.

3.1.1 Real Sequencing Data

The data sets representing the Illumina sequencing technology were from the following organisms: *Escherichia coli str. K12 substr. MG1655* (denoted “EC”), *Helicobacter acinoncychis* (“HA”) [16], and *Streptococcus suis* (“SS”) [52]. The characteristics of these data sets are given in Table 3.1. The data set for HA was downloaded from the SHARCGS webpage[1] on December 19, 2008. The data set for EC was downloaded from the short read archive on the NCBI (National Centre for Bioinformatics Information) webpage[2] (accession number SRX000429) on January 14, 2009. The data set for SS was downloaded from the Sanger Centre webpage[3] for this organism’s genome on December 19, 2008.

Table 3.1: Characteristics of the datasets of real Illumina data.

Organism	# of reads	# of bases	Read length	Genome size (bp)	Quality info.
EC	20,816,448	749,392,128	36	4,639,675	yes
HA	12,288,791	442,396,476	36	1,553,927	yes
SS	2,726,374	98,149,464	36	2,007,491	yes

The data sets representing the Sanger sequencing technology were from the following organisms: *Anaerocellum thermophilum* (denoted “AT”), *Candidatus methano sphaerula palustris* (“CM”), *Methanococcus maripaludis* (“MM”) and SS. The data sets for AT, CM, and MM were downloaded from the trace archive at NCBI[4] in March, 2009. The data set for SS was downloaded from the Sanger Centre webpage[3] on December 19, 2008. The characteristics of these data sets are given in Table 3.2.

The data sets representing the 454 sequencing technology were from the organisms: AT, CM, EC, and MM. The data sets for AT, CM, and MM organisms were downloaded from the short-read archive at NCBI[2] (accession numbers SRP000444, SRP000558 AND SRP000035) in March, 2009. The data set for EC was downloaded from the CLC bio webpage[5] on March 17, 2009. The characteristics of the data sets are given in Table 3.3.

All the 454 and Illumina data sets have quality information, as do all the Sanger-read data sets

---

[1] http://sharcgs.molgen.mpg.de/download.shtml

[2] http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?

[3] http://www.sanger.ac.uk/Projects/S_suis/

[4] http://www.ncbi.nlm.nih.gov/Traces/home/

[3] http://www.sanger.ac.uk/Projects/S_suis/

[5] http://www.clcbio.com/index.php?=1290

Table 3.2: Characteristics of the datasets of real Sanger data.

Organism	# of reads	# of bases	Genome Size (bp)	Quality info.
AT	38,533	37,945,069	2,919,718	yes
CM	26,401	17,939,166	2,922,917	yes
MM	22,289	19,669,550	1,772,694	yes
SS	30,560	15,824,844	2,007,491	no

except for SS. Experiments were done on the real data sets with quality information (these are confidence values assigned to DNA bases) to determine the effect of the presence or absence of quality values when assembling reads.

Table 3.3: Characteristics of the datasets of real 454 data.

Organism	# of reads	# of bases	Genome size	Quality info.
AT	848,629	122,202,101	2,919,718	yes
CM	465,960	100,283,321	2,922,917	yes
EC	436,142	102,475,824	4,639,675	yes
MM	346,848	395,600,001	1,772,694	yes

Pre-assembly processing of real sequencing data

The Sanger data sets for the organisms AT, CM, and MM were downloaded as chromatogram files (also known as trace files) in the SCF (Standard Chromatogram Format) format [47]. These trace files are binary files that contain trace data and accuracy values generated from sequencing instruments. The trace file for each organism was processed by the base-calling program PHRED (PHil's Read EDitor) [18, 17], yielding the DNA sequences and their corresponding quality values in two separate files for each organism (for example, see Figure 3.2). The vector sequence removal programs Figaro and Lucy [13, 51] were used to trim the vector sequences from the DNA reads and quality values. The trimmed reads and quality values were the data sets used as input for the sequence assemblers (Figure 3.1). There was no pre-assembly processing of the Sanger data for SS because there was no quality information provided for the organism (only the DNA read sequences were provided).

The 454 and Illumina data sets were downloaded in FASTQ (Fasta + Qual) format, a format where both the base sequences and quality information are contained in the same file. An example of a sequence in FASTQ format is in Figure 3.2. The base sequence header line begins with '@'

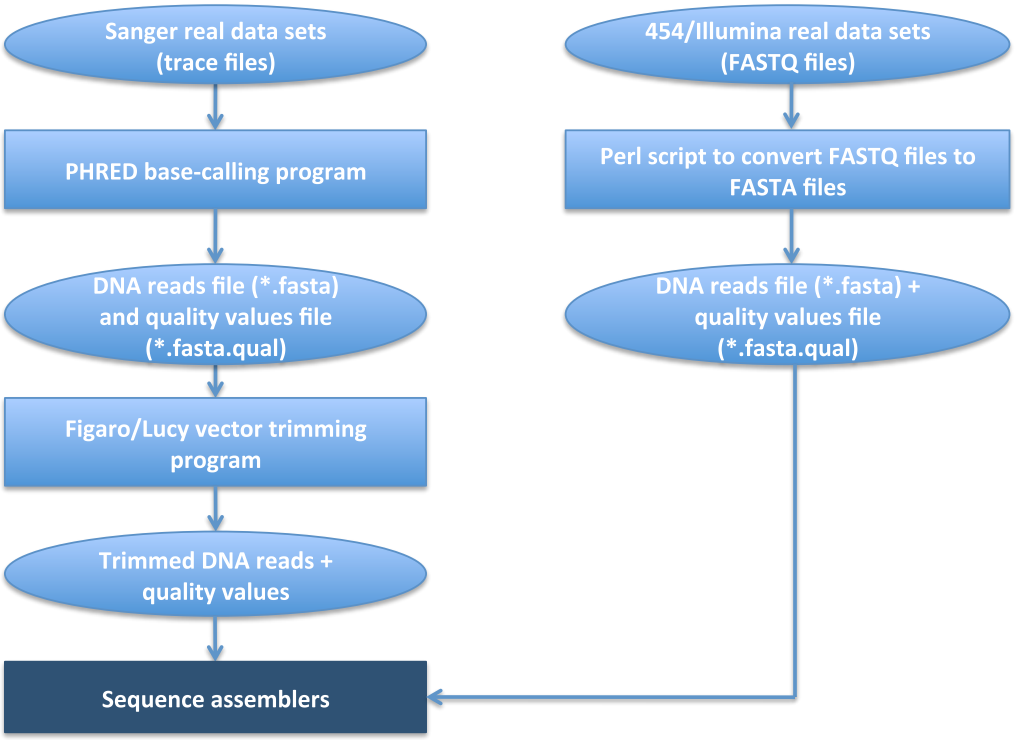


Figure 3.1: Pre-assembly processing of the real data sets.

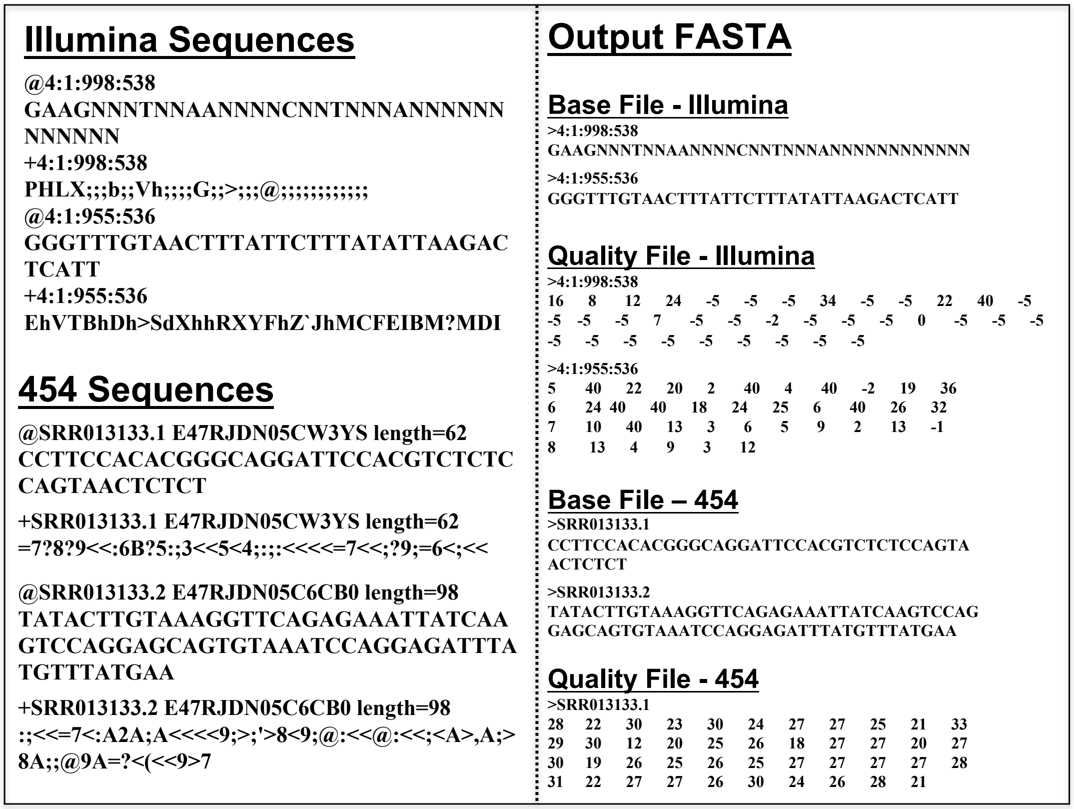
character and the quality information header line begins with '+' character. All the information in the FASTQ format is encoded in ASCII. Each quality value is encoded in one character (one byte). There are three types of FASTQ encoding [1].

- Sanger FASTQ format — encodes a PHRED quality score from 0 to 60 using ASCII 33 to 93. To convert to quality score from ASCII, subtract 33 from the ASCII value.

- Solexa/Illumina FASTQ format — encodes a Solexa/Illumina quality value from -5 to 40 using ASCII 59 to 104. To convert to quality score from ASCII, subtract 64 from the ASCII value.

- Illumina 1.3 FASTQ format - encodes a PHRED quality score from 0 to 40 using ASCII 64 to 104. To convert to quality score from ASCII, subtract 64 from the ASCII value.

A Perl script was written to convert the sequences in FASTQ format (in any of the FASTQ formats described above) to FASTA files (Figure 3.1). The output from the Perl script is a DNA reads file (containing the 454 or Illumina reads) and a corresponding quality file (after converting the ASCII values).



Illumina Sequences

```
@4:1:998:538
GAAGNNNTNNAANNNNCNNTNNNANNNNNN
NNNNNN
+4:1:998:538
PHLX;;;b;;Vh;;;;G;;>;;;@;;;;;;;;;;;;
@4:1:955:536
GGGTTTGTAACTTTATTCTTTATATTAAGAC
TCATT
+4:1:955:536
EhVTBhDh>SdXhhRXYFhZ`JhMCFEIBM?MDI
```

454 Sequences

```
@SRR013133.1 E47RJDN05CW3YS length=62
CCTTCCACACGGGCAGGATTCCACGTCTCTC
CAGTAACTCTCT

+SRR013133.1 E47RJDN05CW3YS length=62
=7?8?9<<:6B?5:;3<<5<4;:;:<<<<=7<<;?9;=6<;<<

@SRR013133.2 E47RJDN05C6CB0 length=98
TATACTTGTAAAGGTTCAGAGAAATTATCAA
GTCCAGGAGCAGTGTAAATCCAGGAGATTTA
TGTTTATGAA

+SRR013133.2 E47RJDN05C6CB0 length=98
:;<<=7<:A2A;A<<<<9;>;'>8<9;@:<<@:<<;<A>,A;>
8A;;@9A=?<(<<9>7
```

Output FASTA

Base File - Illumina

```
>4:1:998:538
GAAGNNNTNNAANNNNCNNTNNNANNNNNNNNNNNN

>4:1:955:536
GGGTTTGTAACTTTATTCTTTATATTAAGACTCATT
```

Quality File - Illumina

```
>4:1:998:538
16   8   12   24   -5   -5   -5   34   -5   -5   22   40   -5
-5  -5   -5   7    -5   -5   -2   -5   -5   -5   0    -5   -5   -5
-5   -5   -5   -5   -5   -5   -5   -5   -5

>4:1:955:536
5    40   22   20   2    40   4    40   -2   19   36
6    24 40    40   18   24   25   6    40   26   32
7    10   40   13   3    6    5    9    2    13   -1
8     13   4    9    3    12
```

Base File – 454

```
>SRR013133.1
CCTTCCACACGGGCAGGATTCCACGTCTCTCCAGTA
ACTCTCT

>SRR013133.2
TATACTTGTAAAGGTTCAGAGAAATTATCAAGTCCAG
GAGCAGTGTAAATCCAGGAGATTTATGTTTATGAA
```

Quality File - 454

```
>SRR013133.1
28   22   30   23   30   24   27   27   25   21   33
29   30   12   20   25   26   18   27   27   20   27
30   19   26   25   26   25   27   27   27   27   28
31   22   27   27   26   30   24   26   28   21
```

Figure 3.2: Examples of the FASTQ format (left) and an output sequence file in FASTA format (right) as well as a quality file for the sequence.

3.1.2 Artificial Sequencing Data

The genome sequences for the six organisms used in this study were downloaded from the NCBI webpage[5] (see Table 3.4). These genomic sequences were used as the bases for generating artificial data sets.

Artificial data sets were generated for all the organisms listed earlier. The parameters used for generating the artificial reads were derived from characteristics of the real data sets (Table 3.1, 3.2, and 3.3). A Perl script was written to collect all the read-lengths and create a frequency table of read lengths for each of the real data sets (that is, frequency tables for the Sanger, 454, and Illumina real data); the same was done for the quality values. The read-length distributions were useful in determining the average read-lengths for each of the artificial data sets or the lengths of high quality and poor quality regions. It was also useful in determining the number of reads to be

---

[5] http://www.ncbi.nlm.nih.gov/genomes/

generated for a given genome coverage. The number of sequences to be generated was determined using the formula

$$N = \frac{LEN}{ARL} * GC \tag{3.1}$$

where N is the number of DNA reads, LEN is the length of the entire genome sequence, ARL is the average read length of the DNA reads, and GC is the genome coverage. The characteristics of the artificial Sanger data sets are in Table 3.5.

For the real Sanger data sets, three frequency tables were generated for each of the parameters (read-length and quality value, one each for the start, middle, and tail portions of reads.). This was as a result of an observed variation in the quality of different regions of the Sanger reads. For example, the start portion of a Sanger read (e.g., the first 59 bases) could be of a poor quality (Figure 3.3) compared to the middle portion of the read which is of higher quality (Figure 3.4). The tail portion of the read could also be of a poor quality (see Figure 3.5). For this reason, each Sanger read was split into three to get a frequency distribution of the lengths and quality values for the start, middle, and tail portions of the reads (see Figure 3.6). The start, middle, and tail lengths for the artificial Sanger reads were determined by manual inspection of the output from the CodonCode Aligner Trace viewer [14] for sixty DNA reads taken arbitrarily from organisms AA, CM, and MM. Three files were used to collect the read lengths for the start, middle, and tail portions of the Sanger sequences, and another three files were used to collect the corresponding quality values. Each artificial Sanger read length was a sum of lengths randomly selected from each of the three length files.

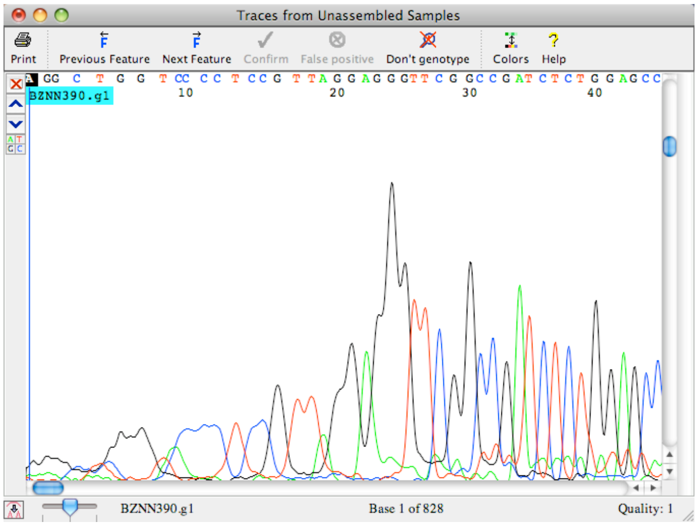


Figure 3.3: The output from CodonCode Aligner Trace viewer for the quality of the start region of a Sanger read. The y-axis displays the base peaks.

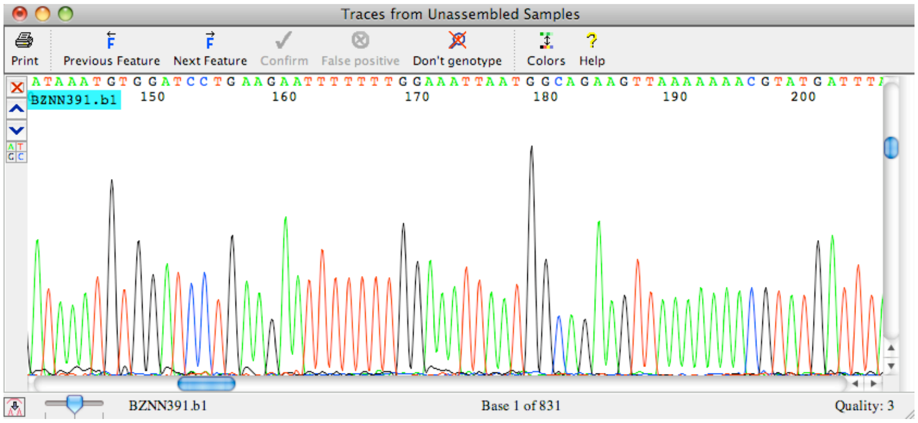


Figure 3.4: The output from CodonCode Aligner Trace viewer for the quality of the middle region of a Sanger read. It shows good quality of sequences. The y-axis shows the base peaks.

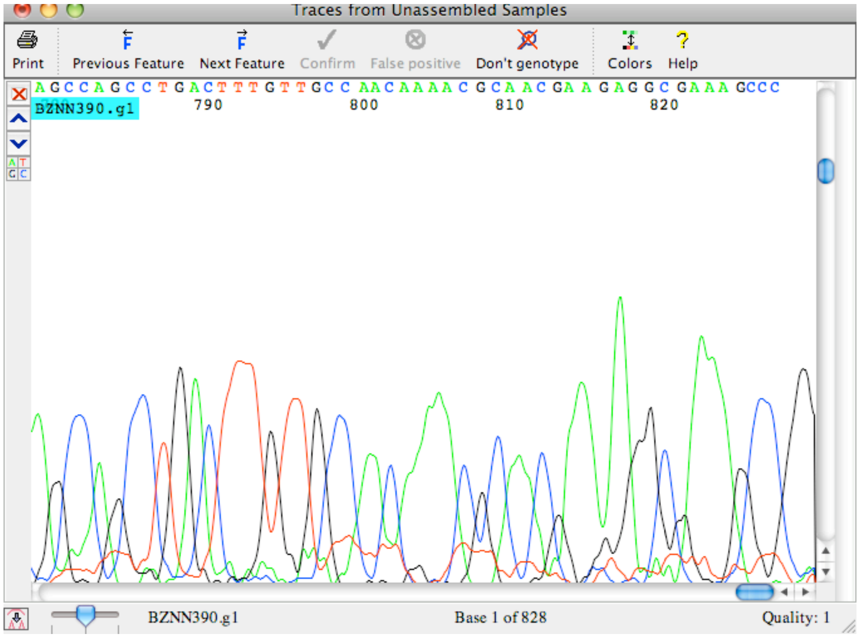


Figure 3.5: The output from CodonCode Aligner Trace viewer for the quality of the tail region of a Sanger read. It shows poor quality. The y-axis displays the base peaks.

For the Illumina data sets, there were two frequency tables generated for each of the parameters (read-length and quality value). This was because it was observed that the quality of the tail portion of the Illumina reads (e.g., the last 6 bases) was poor in comparison to the remainder of the reads. For this reason, each Illumina read was split into two (yielding two portions of length 30 and 6) to get separate frequency tables for the start and tail regions of the read. There was only one frequency table for 454 read-lengths and another table for the quality values. This was because, in general, the quality for real 454 reads was consistent for the entire length of the read.

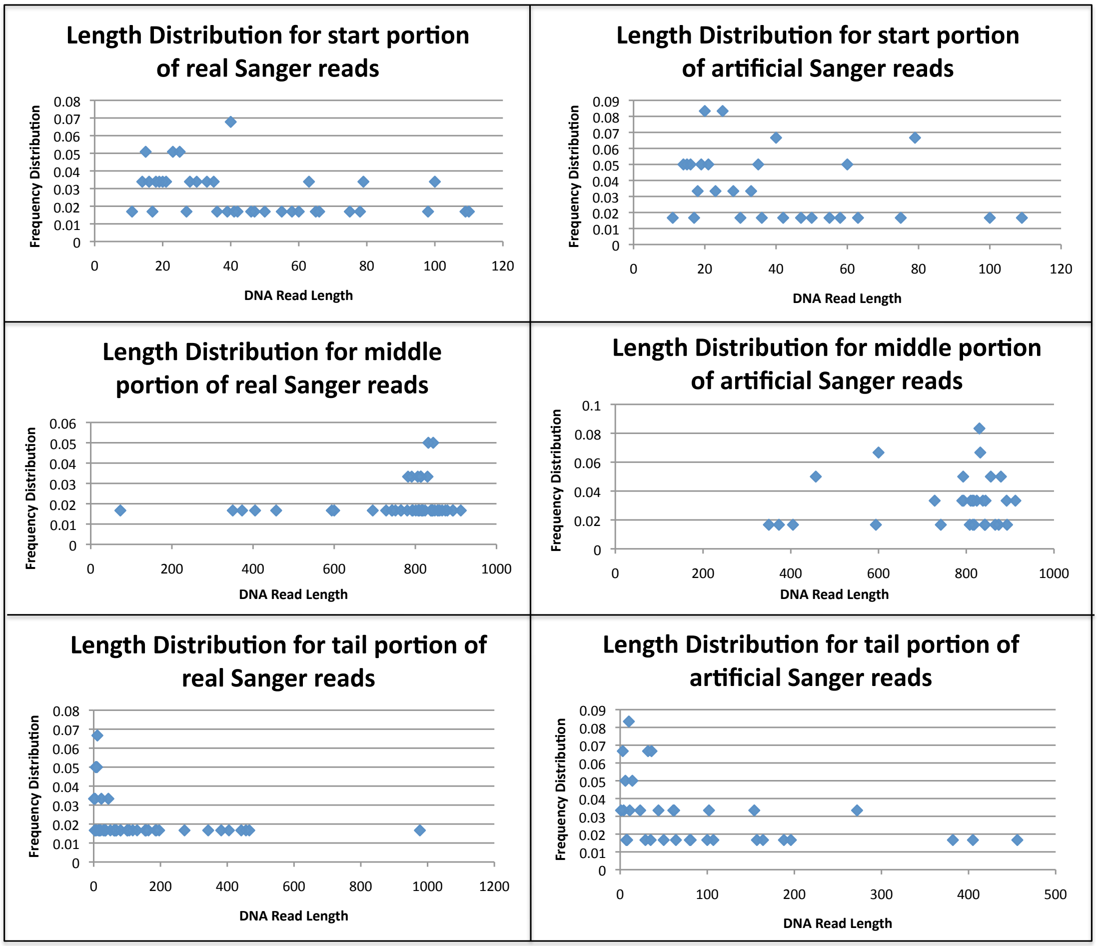


Figure 3.6: Plots for the length distribution of Sanger reads.

A Perl script was written to generate artificial reads along with a corresponding artificial quality file using the determined parameters. The Perl script also introduced errors into the artificial reads based on the artificial quality values.

The error rates for the Sanger sequences were based on the PHRED quality base-calling approach [18, 17];

$$Q_{phred} = -10 \log_{10} P \tag{3.2}$$

where Q_{phred} is the quality score and P is the estimated error probability for that base-call. For example, if Q_{phred} is 10, then there is a probability that 1 in 10 bases is wrong and if Q_{phred} is 40, then there is a probability that 1 in 10,000 bases is wrong. A Monte Carlo method was used to introduce insertion and mismatch errors into the artificially generated reads based on such probabilities. The method was done for each base in the artificial read sequence. For each base, a random number R between 0 and 1 was generated and then a test between R and the probability of error P (corresponding to the artificial quality value Q_{phred} for the base) was used to determine if a change would be made at the base-call. If P was greater than R, a mismatch error was introduced by changing the base; if P was less than R, no error was introduced. For example, suppose that the quality value associated with a particular base is 20. This corresponds to P=0.01. Suppose the random number R produced was 0.53. Since $P < R$, no errors would be introduced. If $R = 0.011$ and $P = 0.12$, an error would be introduced to the read involved.

Table 3.4: Characteristics of the genome sequences for the organisms.

Organism	Accession number	Genome size
AT	CP001393	2,919,718
CM	CP001338	2,922,917
EC	U00096	4,639,675
HA	AM260522	1,553,927
MM	CP000609	1,772,694
SS	AM946016	2,007,491

Table 3.5: Characteristics of the datasets of artificial Sanger data.

Organism	# of reads	Total # of bases
AT	47,970	43,921,464
CM	48,030	44,000,460
EC	76,230	69,763,991
HA	25,545	23,380,509
MM	29,130	26,672,830
SS	32,985	30,226,268

The Illumina and 454 real data sets were downloaded in the FASTQ format. The error rates for introducing errors into the artificial reads for the Solexa/Illumina and sequences were based on the quality base-calling approach below:

$$Q_{solexa} = -10 \log_{10} \frac{P}{1-P} \tag{3.3}$$

where Q_{solexa} is the quality score and P is the estimated error probability for that base-call. As in the case of the Sanger real data sets, frequency distributions were determined from the quality data of the real Illumina and 454 reads and used to produce artificial 454 and Illumina reads using a Monte Carlo method.

Table 3.6: Characteristics of the datasets of artificial 454 data.

Organism	# of reads	Total # of bases
AT	851,585	97,105,639
CM	852,530	97,234,516
EC	1,353,240	154,338,016
HA	453,250	51,700,348
MM	517,055	58,973,487
SS	585,550	66,783,483

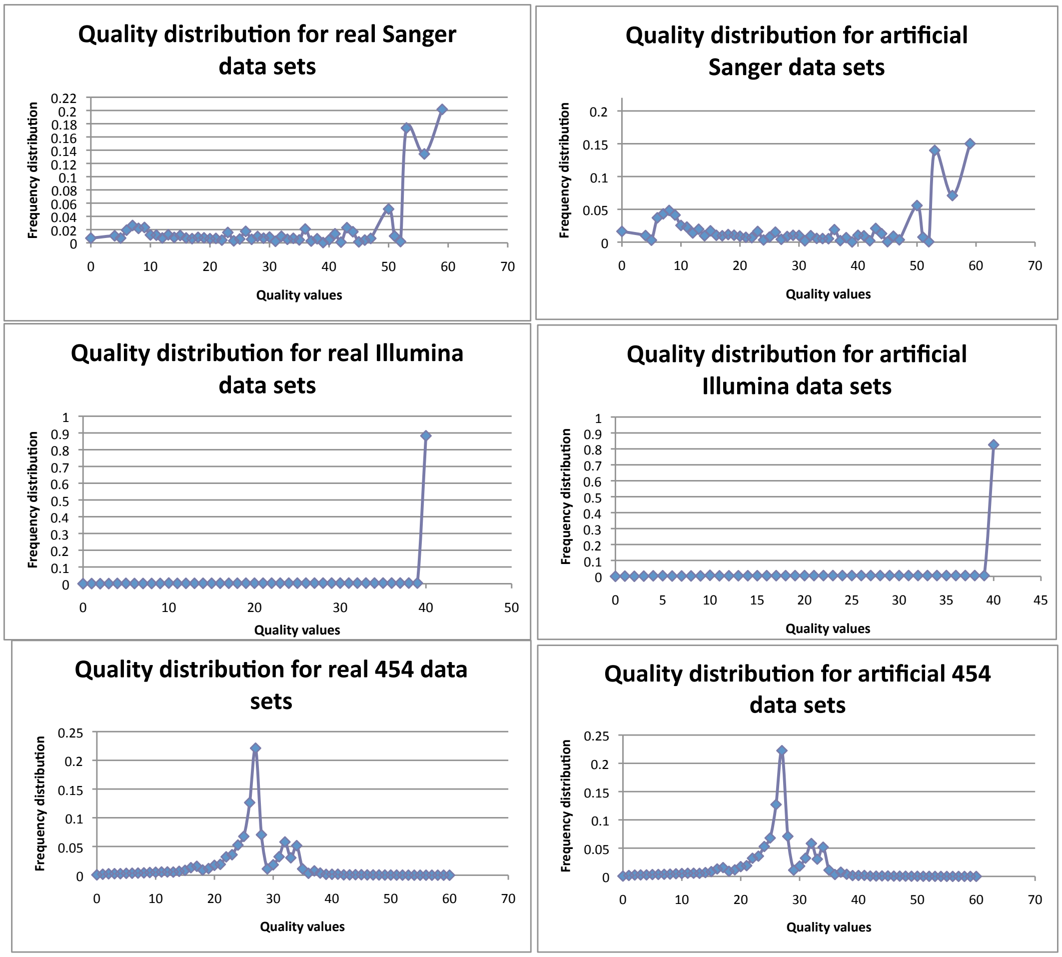
Table 3.7: Characteristics of the datasets of artificial Illumina data.

Organism	Read-length	# of reads	Total # of bases
AT	36	3,244,160	116,789,760
CM	36	3,247,720	116,917,920
EC	36	5,155,200	185,587,200
HA	36	1,726,600	62,157,600
MM	36	1,969,680	70,908,480
SS	36	2,230,560	80,800,160

Artificial Sanger, 454, and Illumina reads were generated for all the organisms regardless of whether the organism had a real data set available (see Table 3.5, Table 3.6, and Table 3.7 for characteristics of the Sanger, 454, and Illumina artificial data, respectively). This was useful in testing whether the artificial data sets were comparable to the real data sets. To test that the artificially generated reads were similar to the real data sets, the quality distributions for both (real and artificial) sets of data from the Sanger, 454 and Illumina technologies were plotted (see Figure 3.7). The results show that the real and artificial DNA reads are similar.

3.1.3 Summary of Data Sets

A summary of all the data sets used for this research is presented in Table 3.8. Each of six organisms chosen for this research had data available for 454, Illumina, and Sanger sequences. Preferentially, real data was used. For the organisms without real data sets, the artificially generated data sets were used.


Quality distribution for real Sanger data sets

Quality distribution for artificial Sanger data sets

Quality distribution for real Illumina data sets

Quality distribution for artificial Illumina data sets

Quality distribution for real 454 data sets

Quality distribution for artificial 454 data sets

Figure 3.7: Plots for the quality distribution of the real and artificial data sets for all the organisms.

Table 3.8: Summary of all the data sets used in this work.

Organism	Sequence Type		
	Sanger	454	Illumina
AT	real	real	artificial
CM	real	real	artificial
EC	artificial	real	real
HA	artificial	artificial	real
MM	real	real	artificial
SS	real	artificial	real

3.2 Methodology

The purpose of this research was to compare some of the existing sequence assemblers in order to determine the appropriate assembler (or assemblers) to use when working with data from different sequencing technologies. This was determined based on the performance of these assemblers in terms of speed, accuracy of results (number of contigs produced, genome sequence coverage, number of misassembled contigs, and the size of contigs.) and types of resources required to run these assemblers (execution time and memory usage). These are standard metrics used to determine the performance of assemblers [29, 34].

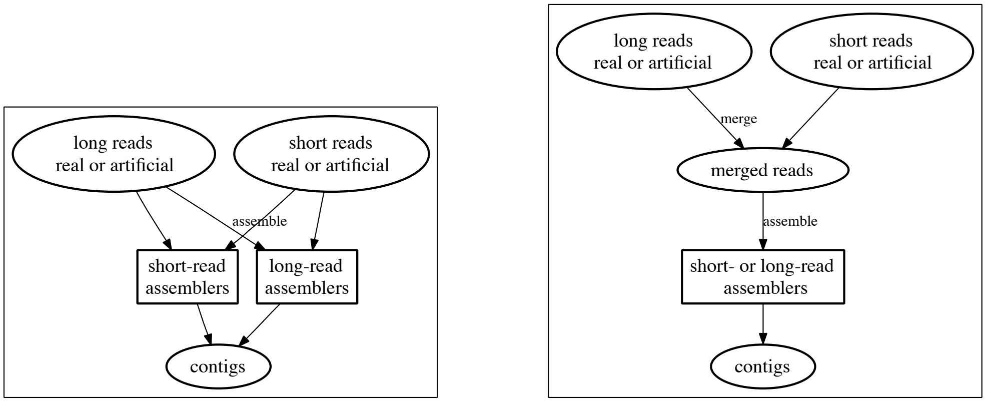
There were different strategies used for assembling the data sets (Figure 3.8). They include:

1. assemble long reads or short reads (exclusively) on long-read assemblers (Figure 3.8(a));
2. assemble long reads or short reads (exclusively) on short-read assemblers (Figure 3.8(a));
3. assemble both long and short reads combined from an organism using long- or short-read assemblers (Figure 3.8(b));
4. assemble using long-read assemblers, contigs from short reads of an organism plus long reads from the same organism (Figure 3.8(c)); and
5. assemble using long-read assemblers, contigs from both short reads and long reads (Figure 3.8(d)).

For each of these strategies, the performance of each of the sequence assemblers was determined using the metrics discussed in the following subsections for each available data set (real or artificial).

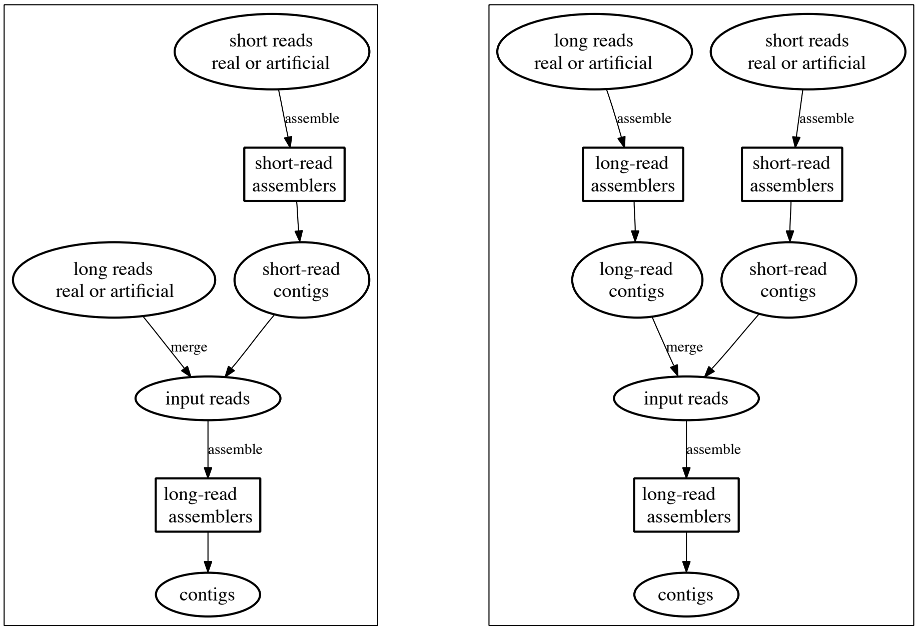
3.2.1 Accuracy of results

This is one of the most important measures in determining the performance of a sequence assembler. Since one goal of a sequence assembly project is to get highly accurate results, it is important to establish the accuracy of contigs produced from the sequence assemblers. Accuracy is measured by the degree of genome coverage, the number of misassembled contigs, the number of contigs produced, and the size of contigs. The higher the genome coverage, the better the performance of the sequence assembler. The desired result would be contigs that maximize genome coverage with minimum misassembly. Misassembly of contigs occurs when one part of a sequence is improperly joined to another part of the sequence during the assembly process; hence, fewer misassembled contigs indicate better assembly quality.



(a) Assemble long or short reads.

(b) Assemble merged short and long reads.

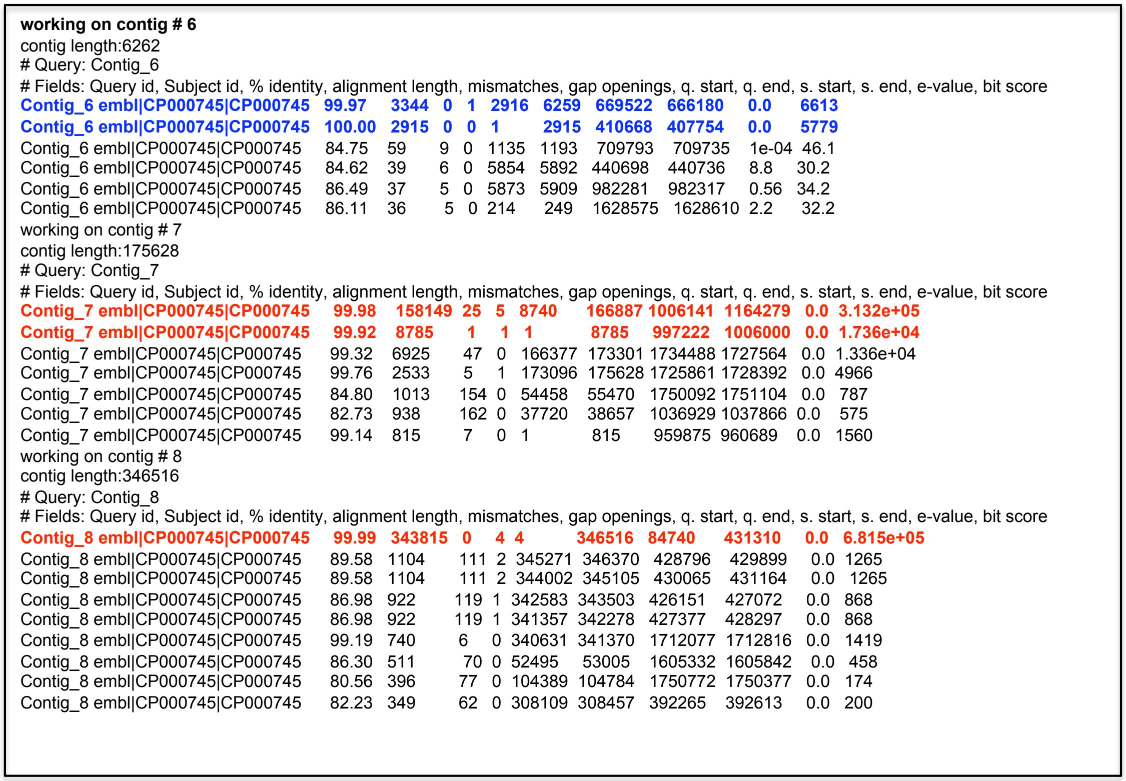


(c) Assemble long reads and short read contigs.

(d) Assemble merged contigs from short and long reads.

Figure 3.8: Flow chart displaying the approaches with which DNA reads will be submitted to the assemblers.

The genome coverages and misassemblies were calculated using the local alignment algorithm MegaBLAST and two custom scripts (programs). An awk script was used to pipeline the EMBOSS *seqret* program (which reads the contig sequences and passes the sequences one by one to the MegaBLAST algorithm), the EMBOSS *infoseq* program (which produces information about sequences such as the length, and type of sequence), and the local alignment algorithm MegaBLAST (used to compare a set of contigs to a genome sequence). A Perl script was used to analyse the output file produced by the awk script; this involved determining misassemblies and avoiding incorrect duplications when calculating the genome coverage (Figure 3.9). All good alignments (not indicative of misassemblies) were output and the numbers from the **s.start** (sequence start) column and **s.end** (sequence end) column (as shown in Figure 3.9) were used to mask the involved regions of the genome sequence. For example, if **s.start** = 5678, and **s.end** = 179081, then the sequences from positions 5678 to 178081 in the reference genome sequence would be masked (the bases were changed to 'F'). The masked sequences made it easy to count the coverage area of the genome.



```
working on contig # 6
contig length:6262
# Query: Contig_6
# Fields: Query id, Subject id, % identity, alignment length, mismatches, gap openings, q. start, q. end, s. start, s. end, e-value, bit score
Contig_6 embl|CP000745|CP000745  99.97   3344  0  1  2916  6259  669522  666180   0.0    6613
Contig_6 embl|CP000745|CP000745  100.00  2915  0  0  1     2915  410668  407754   0.0    5779
Contig_6 embl|CP000745|CP000745    84.75   59      9  0  1135  1193   709793   709735   1e-04  46.1
Contig_6 embl|CP000745|CP000745    84.62   39      6  0  5854  5892  440698   440736   8.8    30.2
Contig_6 embl|CP000745|CP000745    86.49   37      5  0  5873  5909  982281   982317   0.56   34.2
Contig_6 embl|CP000745|CP000745    86.11   36      5  0  214    249   1628575  1628610 2.2    32.2
working on contig # 7
contig length:175628
# Query: Contig_7
# Fields: Query id, Subject id, % identity, alignment length, mismatches, gap openings, q. start, q. end, s. start, s. end, e-value, bit score
Contig_7 embl|CP000745|CP000745    99.98   158149  25   5  8740     166887 1006141  1164279  0.0  3.132e+05
Contig_7 embl|CP000745|CP000745    99.92   8785     1    1  1        8785    997222   1006000  0.0  1.736e+04
Contig_7 embl|CP000745|CP000745     99.32   6925     47   0   166377  173301 1734488 1727564   0.0  1.336e+04
Contig_7 embl|CP000745|CP000745     99.76   2533     5    1   173096  175628 1725861 1728392   0.0  4966
Contig_7 embl|CP000745|CP000745     84.80   1013     154  0   54458    55470   1750092 1751104   0.0   787
Contig_7 embl|CP000745|CP000745     82.73   938      162  0   37720    38657   1036929 1037866  0.0   575
Contig_7 embl|CP000745|CP000745     99.14   815      7    0   1         815      959875  960689    0.0   1560
working on contig # 8
contig length:346516
# Query: Contig_8
# Fields: Query id, Subject id, % identity, alignment length, mismatches, gap openings, q. start, q. end, s. start, s. end, e-value, bit score
Contig_8 embl|CP000745|CP000745    99.99   343815   0    4  4        346516  84740    431310    0.0  6.815e+05
Contig_8 embl|CP000745|CP000745     89.58   1104     111  2  345271  346370  428796   429899    0.0  1265
Contig_8 embl|CP000745|CP000745     89.58   1104     111  2  344002  345105  430065   431164    0.0   1265
Contig_8 embl|CP000745|CP000745     86.98   922     119  1  342583  343503  426151   427072    0.0   868
Contig_8 embl|CP000745|CP000745     86.98   922     119  1  341357  342278  427377   428297    0.0   868
Contig_8 embl|CP000745|CP000745     99.19   740     6     0  340631  341370  1712077  1712816  0.0   1419
Contig_8 embl|CP000745|CP000745     86.30   511      70  0  52495    53005    1605332  1605842   0.0   458
Contig_8 embl|CP000745|CP000745     80.56   396     77   0  104389  104784  1750772  1750377  0.0   174
Contig_8 embl|CP000745|CP000745     82.23   349     62   0  308109  308457  392265   392613   0.0   200
```

Figure 3.9: An example of the MegaBLAST output. Both the lines in red and the lines in blue are considered to be good alignments. However, the lines in blue are also examples of misassemblies (this is because positions 1-2915 of the contig match position 407754-410668 of the reference genome, while 2916-6259 match 666180-669522). The other results in black are discarded (all are duplications or repeats).

The genome coverage was calculated as a percentage; that is, 100 times the number of bases in "good" contigs divided by the total number of bases in the genome. The percentage of misassembled contigs were calculated as the percentage ratio of the number of misassembled contigs to the total number of contigs. Another vital factor is the number of contigs produced. It is better to have fewer contigs produced because it makes scaffolding (linking and joining contigs) easier and gap closing less complicated. For the set of contigs produced, a Perl script derived the N50 size and the largest contig size. The N50 size is useful in rating the size of a contig scaffold. Given a set of contigs (arranged in the order of increasing lengths where the sum of the contig lengths is N and $M = N\ /\ 2$, the N50 size is the maximal contig length such that the cumulative sum of the contig lengths is less or equal to M. For example, if 350000, 400000, 440000, 510000, 550000, and 600000 are contig lengths, then $N = 2850000$, $M = 1425000$, and the N50 size is 440000.

3.2.2 Execution time

Although it is vital to work with a sequence assembler that generates accurate results, a sequence assembler that performs with shorter execution time would be preferred over one that takes hours or even days to generate results.

Execution time was determined by taking the total (system + user) CPU time (run time) required to finish the assembly of a complete input data set. This was determined using the standard LINUX/UNIX *time* command. All of the assembly runs were performed on one computer (see Section 3.5). The goal of using one machine was to reduce any time anomalies.

3.2.3 Memory usage

Memory usage is the amount of memory used by a program to complete its task within the execution time. This was determined by writing a script which used the process id (*pid*) of an assembler to obtain the process maximum virtual memory size from file */proc/pid/status* (where *pid* is a variable standing for the *pid* of the process running the assembler, e.g., 54663) every ten seconds on the computer running the assembly program. The memory used was calculated as the average of the numbers collected. The amount of memory data collected depended on the period the sequence assembler ran.

3.2.4 System Dependencies

This was determined based on program documentation and experiences from working with the programs. For example, a program recompiled in 64-bit may perform and produce better results than when it is in the original 32-bit form.

3.2.5 Restrictions and constraints

Some of the sequence assemblers come with restrictions and constraints. For example, the EDENA sequence assembler must have DNA reads of the same length before it can begin assembly and the PCAP assembler will not assemble DNA if there is no quality information for the reads.

3.3 Effect of Quality values on Accuracy of Contigs

Quality values (confidence values) are numbers assigned to DNA bases which indicate the level of trust for each base. A low quality value indicates low confidence in base calls and a high quality value indicates a high level of trust. Some assemblers use the quality values to remove or trim certain DNA reads to improve the quality of contigs produced. Some sequence assemblers require quality information to assembly the DNA data (for example, PCAP and MIRA). If there is no quality information provided, some assemblers use a default quality value (for example, PHRAP and CAP3).

It is expected that the quality values (when available) produce a notable improvement in the assembly process. To test this hypothesis, the real data sets for the Sanger, 454, and Illumina were assembled first, using their quality information and a second time, without their quality information. The goal for assembling the real data sets with and without their quality information was to establish the effect of the presence (or absence) of quality values on the accuracy of contigs.

3.4 Statistical Analysis

Statistical analyses were performed on the results (number of contigs, genome coverage, time, and others described in Section 3.2) to compare the group means (that is, the mean for a metric, such as genome covergae across data sets) of the assemblers, which was useful in determining if they are equal (represent similar results) or not. Due to the nature of assembly results in this research, the statistical methods used to compare group means (from the assemblers) were the t-test (independent samples t-test) and analysis of variance (one-way ANOVA). The t-test was used to check differences in means between two sequence assemblers, while the ANOVA was used to compare means among three or more sequence assemblers. These tests were performed using the SPSS software (PASW Statistics 18). The t-test assumes that the variances of both groups being tested is approximately equal, so Levene's test for equality of variances [43] was used to determine whether this assumption was met. The Levene's test value should be greater than the confidence (α) level if the assumption is met; if not, the null hypothesis should be rejected (see Table 3.9). If the Levene's test value is greater than α, then the value from the row labeled "Equal variance assumed" is used, otherwise the other “Equal variance not assumed” is used. For the statistical tests using one-way ANOVA, if a group of means are found to be significantly different, post-hoc tests were performed to find the assembler or assemblers responsible for the difference. The post-hoc tests used in this work were the Tukey honestly significant difference (HSD) test [19, 45] which assumes equal variance and equal sample size, and the Games-Howell test [19] which is a modification of the Tukey HSD test which assumes unequal variance and sample size. The Games-Howell test was useful when dealing with a table of results that contains missing values. These post-hoc tests were useful in comparing every group mean with every other group mean, and the results helped determine which assembler (or assemblers) caused the significant difference. All the tests were performed using a confidence interval of 95% which corresponds to a significance level (α) of 0.05. The presentation of the statistical results are given in Table 3.9, Table 3.10, and Table 3.11. The results in Table 3.10 and Table 3.11 indicate that there is no statistical significance between the genome coverages because the significance value (probability value) is greater than 0.05 (confidence interval of 95%).

3.5 Computer Resources

All the experiments for this research were conducted on a computer server with the Red Hat Linux operating system (version 5.3, architecture x86_64), eight dual-core AMD 3GHz Opteron(tm) 8222 SE processors, 1024 KB of cache, and 132 GB of main memory.

Table 3.9: An example of the SPSS output for the independent samples test.

Independent Samples Test						
		Levene's Test for Equality of Variances		t-test for Equality of Means		
Metrics	Equal variances	F	Sig.	t	df	Sig. (2-tailed)
NC	assumed	3.300	0.099	1.919	10	0.084
	not assumed			1.919	6.934	0.097
GC	assumed	0.411	0.536	0.214	10	0.835
	not assumed			0.214	9.138	0.836
MC	assumed	1.187	0.301	-5.718	10	<0.001
	not assumed			-5.718	7.596	0.001
N50	assumed	11.906	0.006	-1.153	10	0.276
	not assumed			-1.153	5.601	0.296
Time	assumed	2.730	0.129	-0.738	10	0.478
	not assumed			-0.738	7.257	0.484
Mem.	assumed	4.197	0.068	0.772	10	0.458
	not assumed			0.772	5.179	0.474

Sig. - test for equal variances, (the values are expected to be higher than 0.05), t- t-statistic, df- degrees of freedom, Sig. (2-tailed) - p-value, NC - number of contigs, GC - genome coverage, MC - misassembled contigs, Mem. - amount of RAM used. This result is also in Table E.13 of Appendix E.

Table 3.10: An example of one-way ANOVA output from SPSS when comparing the means for genome coverage between five assemblers.

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	49.216	4	12.304	1.192	0.340
Within Groups	247.639	24	10.318		
Total	296.855	28			

Table 3.11: An example of Games-Howell post-hoc output from SPSS when comparing the means for genome coverage between five assemblers (one-way ANOVA test). The results are an extract from Table E.17 of Appendix E.

Assemblers (I)	Assemblers (J)	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
EULERSR	MIRA	2.32333	2.91281	0.920	-9.8992	14.5459
	PHRAP	2.57167	1.46969	0.460	-2.5431	7.6864
	SSAKE	-0.37500	0.82426	0.989	-3.2089	2.4589
	VELVET	-0.25333	1.04634	0.999	-3.7017	3.1950
MIRA	EULERSR	-2.32333	2.91281	0.920	-14.5459	9.8992
	PHRAP	0.24833	3.10524	1.000	-11.6503	12.1470
	SSAKE	-2.69833	2.85691	0.867	-15.1198	9.7232
	VELVET	-2.57667	2.92872	0.892	-14.7530	9.5996
PHRAP	EULERSR	-2.57167	1.46969	0.460	-7.6864	2.5431
	MIRA	-0.24833	3.10524	1.000	-12.1470	11.6503
	SSAKE	-2.94667	1.35552	0.301	-8.0138	2.1205
	VELVET	-2.82500	1.50096	0.395	-7.9837	2.3337
SSAKE	EULERSR	0.37500	0.82426	0.989	-2.4589	3.2089
	MIRA	2.69833	2.85691	0.867	-9.7232	15.1198
	PHRAP	2.94667	1.35552	0.301	-2.1205	8.0138
	VELVET	0.12167	0.87880	1.000	-2.9378	3.1812
VELVET	EULERSR	0.25333	1.04634	0.999	-3.1950	3.7017
	MIRA	2.57667	2.92872	0.892	-9.5996	14.7530
	PHRAP	2.82500	1.50096	0.395	-2.3337	7.9837
	SSAKE	-0.12167	0.87880	1.000	-3.1812	2.9378

Chapter 4

Results

The goal of this research was to determine the most appropriate sequence assembly strategy to implement when dealing with data sets from mixed sequencing technologies as well as the best sequence assembler (or assemblers) to use. The effect of quality values on the contigs produced from sequence assemblers was also investigated. Experiments were performed on various data sets from the Sanger, 454 and Illumina sequencing technologies using short- and long-read assemblers. All the experiments were based on the methodology presented in Section 3.2 and Section 3.3. The performance of these assemblers was determined based on the accuracy of contigs produced, the execution time required to complete an assembly, the amount of memory used and a few other metrics. The results from these experiments are presented in the following sections.

4.1 Assembling short reads using long-read assemblers

The goal of assembling the short-read data sets (both 454 and Illumina) using the long-read assemblers (which are designed to work with Sanger data sets) was to determine how well these long-read assemblers handle such data, and also to find out which long-read assemblers cannot be used for assembling short reads. For the long-read assemblers which could assemble the short reads, it was also important to observe the accuracy of results (contigs) produced. The results of these investigations are presented in this subsection. More detail is provided in Tables B.1, B.2, B.3, and B.4 of Appendix B. The results observed were useful in determining the trade-offs between long-read assemblers and short-read assemblers, and also the rationality of using a long-read assembler for short-read genome projects.

Table 4.1 displays the overall ability of the long-read assemblers to assemble short reads. This was determined based on whether each long-read assembler ran to completion without returning errors on most of the data sets. The MIRA and PHRAP assemblers were able to assemble all the short-read data from 454 sequencing and some from Illumina sequencing (see Table 4.2 and Table 4.3), while the other long-read assemblers were not. The CAP3 assembler returned the error "ran out of memory" when given the short-read data as input. If a read has overlaps with many

Table 4.1: Results to show which long-read assemblers can handle short reads. An "×" indicates that the assembler was not able to successfully work with this type of data, while a check mark indicates that it could.

Assemblers	454 reads	Illumina reads
CAP3	×	×
MIRA	✓	✓
PCAP	×	×
PHRAP	✓	✓
TIGR Assembler	×	×

other DNA reads, CAP3 takes this to mean that there are many short-word matches between these reads. This poses a problem to the CAP3 assembler because it expects overlaps between only a small number of reads as is common with long-read data. There is a command line option to CAP3 that is intended to rectify or avoid the problem described above. This option increases or decreases the maximum number of word matches between DNA reads. However, specifying a large number (which would be required when assembling short reads) resulted in the CAP3 assembler running out of memory, irrespective of the amount of RAM available on the computer system. The TIGR Assembler has a limitation that the maximum number of DNA reads to be assembled must not exceed 524,000 and this means it was not able to assemble some of the short-read data (though some of the 454 data sets are less than 524,000 sequences). The TIGR assembler was re-compiled in 64-bit mode to alleviate the limitation. However, it still did not assemble any of the short-read data sets successfully. It did not return an error; it just produced empty output files. The PHRAP assembler also has a limitation that the maximum number of input sequences must not exceed 64,000. However, after re-compiling the PHRAP as a 64-bit executable program, it was able to assemble the short reads. The PCAP assembler could not assemble any of the short-read data sets; it produced no results.

For the 454 data, only MIRA and PHRAP could handle the data sets. Table 4.2 presents the results for PHRAP and MIRA assembling the 454 data sets (along with other results). More detail is provided in Graphs C.1, C.6, C.11, C.16, D.3, and D.8 of Appendix C and D. The performance of these assemblers was rated based on the metrics described in Section 3.2 and Section 3.4. Statistical analysis was done for all the results in Table 4.2. This made comparing the results from the long-read assemblers with those from the short-read assemblers easier. A one-way ANOVA test was conducted for each metric described in Section 3.2 to assess whether there was any significant difference across the various assemblers. Details of these results are provided in Table E.1 and Table E.2 of Appendix E. There was a statistically significant difference at the $p < 0.05$ level for

Table 4.2: Results from assembling 454 data using short- and long-read assemblers. The results from the long-read assemblers are shown in bold font.

Organism	Metrics	EULER-SR	**MIRA**	**PHRAP**	SSAKE	VELVET
AT	NC	808	**193**	**1232**	40830	32814
	GC (%)	93.96	**89.89**	**97.11**	95.92	93.86
	MC (%)	0.99	**12.44**	**12.74**	1.02	0.73
	N50	2196	**22381**	**9449**	295	119
	LC	14509	**104206**	**54130**	2995	1292
	ET (in seconds)	369	**-**	**22783.46**	14804.79	262.54
	Memory	115.4 Mb	**-**	**13.45 Gb**	1.11 Gb	667.84 Mb
CM	NC	1257	**337**	**116**	32166	13797
	GC (%)	98.37	**99.98**	**92.27**	97.83	98.89
	MC (%)	0.16	**10.39**	**28.45**	0.36	0.38
	N50	16659	**90096**	**133845**	485	493
	LC	78363	**209592**	**407769**	3143	4050
	ET (in seconds)	235.2	**12006.75**	**9045.55**	6684.61	201.69
	Memory	113.04 Mb	**3.18 Gb**	**4.27 Gb**	867.99 Mb	472.51 Mb
EC	NC	885	**358**	**159**	2398	3778
	GC (%)	97.79	**99.41**	**98.71**	98.35	98.03
	MC (%)	1.02	**13.69**	**31.45**	0.25	1.88
	N50	57777	**129463**	**126839**	7075	3441
	LC	165903	**349282**	**529849**	27137	16401
	ET (in seconds)	1956.23	**7452.51**	**2656.11**	2556.04	146.66
	Memory	733.76 Mb	**2.86 Gb**	**2.39 Gb**	1.54 Gb	381.67 Mb
HA	NC	251	**6435**	**91**	7525	12241
	GC (%)	98.04	**86.91**	**96.67**	98.57	98.16
	MC (%)	0	**0.2**	**15.39**	0.27	0
	N50	46591	**257**	**27204**	522	301
	ET (in seconds)	603.91	**7807.18**	**2207.08**	3260.32	112.73
	Memory	266.88 Mb	**1.38 Gb**	**2.49 Gb**	597.28 Mb	509.09 Mb
MM	NC	618	**1042**	**265**	9362	14814
	GC (%)	98.69	**99.53**	**90.56**	97.35	98.68
	MC (%)	0.16	**2.21**	**23.02**	0.14	0.84
	N50	7222	**24470**	**12319**	368	210
	LC	22488	**64854**	**58567**	3594	1996
	ET (in seconds)	105	**4549.23**	**925.63**	3885.29	69.82
	Memory	82.47 Mb	**1.48 Gb**	**1.25 Gb**	554.54 Mb	326.54 Mb
SS	NC	537	**8464**	**181**	9956	16719
	GC (%)	97.57	**-**	**93.67**	98.65	98.32
	MC (%)	1.68	**-**	**6.08**	0.07	0.08
	N50	27613	**160**	**18076**	476	281
	LC	66367	**1981**	**57886**	8855	2449
	ET (in seconds)	765.85	**7414.32**	**1983.65**	4752	146.01
	Memory	333.52 Mb	**1.49 Gb**	**3.32 Gb**	741.26 Mb	549.67 Mb

NC - Number of Contigs, GC - Genome Coverage, MC -Misassembled Contigs, LC - Largest Contig, ET - Execution Time, “-” - no result, and “**” - unreasonable results in the sense of either too many contigs or very small N50 size.

the number of contigs produced: F(4,25) = 6.008, p = 0.002 (where p is the probability value, F is the F-statistic, and (4,25) are the degrees of freedom (df)). The post-hoc comparisons indicated that there was a significant difference (a greater amount) between the number of contigs produced from SSAKE (mean = 17000) and VELVET (mean = 15700) versus the other sequence assemblers (see Table E.18 and Figure D.8). There was no significant difference for the genome coverage results (F(4,24) = 1.192, p = 0.340) and the N50 size results (F(4,25) = 2.638, p = 0.580) among the assemblers. For the misassembled contigs, there was a statistically significant difference (F(4,24) = 26.510, $p < 0.001$), and the post-hoc comparisons indicated that there was a significant difference (a greater amount) between the misassembled contigs results from PHRAP (mean = 22.21%) and the other sequence assemblers (see Table E.18 and Figure C.6). For the execution time results, there was a statistically significant difference (F(4,24) = 3.583, p = 0.020) across the evaluated assemblers, and post-hoc comparisons indicated that there was a significant difference (a lesser amount) in the results obtained for MIRA (mean = 7845.99 seconds) and the other sequence assemblers (see Figure C.16). For the memory usage, there was a statistically significant difference (F(4,24) = 4.212, p = 0.010), and post-hoc comparisons indicated that there was a significant difference (a lesser amount) between the results obtained from PHRAP (mean = 4528 MB) and the other sequence assemblers (see Figure D.3).

For the Illumina data sets, PHRAP assembled some of them. However, MIRA cannot assemble Illumina data without a reference sequence or sequences available. This means that the approach shown in Figure 3.8(a) does not apply to MIRA for Illumina reads. However, MIRA could assemble the Illumina reads by performing an hybrid assembly which involved the Illumina DNA reads and MIRA contigs from either 454 data or Sanger data (in this case, the Sanger contigs were used). Table 4.3 presents the results for PHRAP and MIRA assembling the Illumina data sets (along with results from short-read assemblers). More detail is provided in Graphs C.2, C.7, C.12, C.17, D.4, and D.9 of Appendices C and D. For the Illumina data sets for the organisms AT, CM, and MM, PHRAP assembled and ran to completion but produced no reasonable results (the N50 size and the length of the largest contig were 36 bp, which is the same as the length of the Illumina reads). PHRAP did not run to completion when assembling the EC Illumina data set. MIRA assembled all the Illumina data except the data set for HA. A one-way ANOVA test was conducted for each metric described in Section 3.2 to assess whether there was any significant difference across the various assemblers (see Table E.3 and Table E.4). There was no statistically significant difference at the $p < 0.05$ level for the number of contigs produced (F(6,33) = 1.965, p = 0.099), the genome coverage results (F(6,28) = 1.316, p = 0.283), and the misassembled contigs results (F(6,28) = 1.321, p = 0.281). For the N50 size results, there was no statistically significant difference at the $p < 0.05$ level with F(6,33) = 2.293, p = 0.05. However, the post-hoc comparisons for the N50 size indicated there was a significant difference between PHRAP (mean = 46.20) versus the other

Table 4.3: Results from assembling Illumina data using short- and long-read assemblers. The results from the long-read assemblers are in bold font.

Org.	Metrics	ABYSS	EDENA	EULER-SR	MIRA	PHRAP	SSAKE	VELVET
AT	NC	2794	772	1630	**1037**	**1339**	2611	1549
	GC (%)	95.06	93.16	94.78	**92.51**	**	96.11	94.88
	MC (%)	0	0	0	**2.51**	**	0.23	0
	N50	19918	20160	24537	**16925**	**36**	23916	11479
	LC	107384	88392	111057	**107934**	**36**	113609	54159
	ET (in sec.)	289.31	393.95	706.48	**210486.32**	**489.52**	1975.97	112.63
	Memory	236.06 Mb	521.21 Mb	315.26 Mb	**6.01 Gb**	**2.41 Gb**	2.11 Gb	569.84 Mb
CM	NC	3778	1325	1860	**311**	**791**	1826	1916
	GC (%)	98.13	97.2	98.16	**98.33**	**	97.75	98.23
	MC (%)	0.08	0.08	0	**2.25**	**	0.38	0
	N50	12106	11692	19867	**34874**	**36**	21237	7238
	LC	60771	42181	91992	**122652**	**36**	21237	7238
	ET (in sec.)	287.78	419.59	821.68	**69753.9**	**361.38**	1891.41	112.73
	Memory	280.83 Mb	526.45 Mb	616.87 Mb	**5.52 Gb**	**2.30 Gb**	2.23 Gb	568.13 Mb
EC	NC	1734	680	1276	**6141**	-	109125	666
	GC (%)	98.8	98.32	98.58	**87.83**	-	**	98.02
	MC (%)	0.17	0.15	0.31	**0.7**	-	**	0
	N50	17311	16430	24087	**127612**	-	72	48723
	LC	67095	67082	103365	**349300**	-	16468	126542
	ET (in sec.)	955.81	2721.5	9742.2	**219097.44**	-	-	776.83
	Memory	671.62 Mb	1.63 Gb	156.71 Mb	**35.22**	-	-	2.68 Gb
HA	NC	858	453	467	-	**41534**	237367	474
	GC (%)	98.29	97.98	98.98	-	**98.98**	**	98.01
	MC (%)	4.2	0.22	1.5	-	**0.62**	**	1.48
	N50	6774	7746	14405	-	**1622**	11645	28127
	ET (in sec.)	1130.68	2370.9	781.8	-	**68180**	28220.2	1159.91
	Memory	1.23 Gb	1.45 Gb	63.82 Mb	-	**31.72 Gb**	6.09 Gb	2.2 Gb
MM	NC	908	574	563	**69**	**370**	588	724
	GC (%)	98.66	98.16	98.71	**99.89**	**	98.53	98.68
	MC (%)	0	0	0	**6.45**	**	0.17	0
	N50	23396	13925	29882	**410376**	**36**	23570	9320
	LC	99232	72821	175259	**513483**	**36**	79160	35105
	ET (in sec.)	160.47	225.17	432.06	**10073.43**	**271.19**	1180.43	67.44
	Memory	160.47 Mb	321 Mb	380.39 Mb	**3.46 Gb**	**1.42 Gb**	1.39 Gb	372.76 Mb
SS	NC	2718	1278	2281	**6243**	**88131**	15632	1160
	GC (%)	98.84	95.82	95.34	**2.73**	**97.56**	98.21	98.55
	MC (%)	11.77	0	0.044	**0**	**0.02**	0.16	1.12
	N50	3554	2606	1709	**36**	**48**	768	4594
	LC	17113	13883	10963	**2012**	**1104**	6475	17831
	ET (in sec.)	301.73	604.13	152.4	**1046.07**	**57246.09**	28220.2	120.65
	Memory	314.93 Mb	446.81 Mb	74.26 Mb	**3.37 Gb**	**22.51 Gb**	6.09 Gb	579.24 Mb

Org. - Organism, NC - Number of Contigs, GC - Genome Coverage, MC -Misassembled Contigs, LC - Largest Contig, ET - Execution Time, "-" - no result, and "**" - unreasonable results in the sense of either too many contigs or very small N50 size.

sequence assemblers (see Table E.20 and Figure C.12). For the execution time results, there was a statistically significant difference at the $p < 0.05$ level with $F(6,32) = 4.551$, $p = 0.002$, and the post-hoc comparisons indicated that there was a significant difference (a greater amount) in the results obtained from MIRA (mean = 102090 seconds) versus the other sequence assemblers (see Table E.19 and Figure C.17). For the memory results, there was a statistically significant difference ($F(6,32) = 2.781$, $p = 0.027$), and the post-hoc comparisons indicated that there was a significant difference (a greater amount) in the memory usage of PHRAP (mean = 12072 MB) and MIRA (mean = 10716 MB) versus the other sequence assemblers (see Figure D.4). Although there was no statistically significant difference in the results for the number of contigs produced, the genome coverages, and the misassembled contigs, some of the results obtained from PHRAP for the number of contigs produced are quite high in comparison to the results from short-read assemblers and MIRA (for example, 88131 contigs for SS and 41534 contigs for HA). Other results also suggest that PHRAP may not be suitable for assembling Illumina reads, like the results for the amount of time and the amount of memory used (see Figure 4.1 and Figure 4.2, respectively). Furthermore, the performance of PHRAP in assembling Illumina reads from AT, CM and MM showed that PHRAP does not produce quality results; for example, the N50 size and the size of the largest contig produced for these organisms was 36 bp and this is the same as the size (36 bp) of the DNA reads submitted to the sequence assemblers. For the organism EC, due to the large number of reads for it, PHRAP returned an error "out of memory".

Among the long-read assemblers tested, PHRAP and MIRA are the best at handling short-read data. Even still, they appear to have difficulty in handling such data.

4.2 Assembling long reads on short-read assemblers

The goal for assembling long reads (Sanger data) using the short-read assemblers was to discover how these assemblers deal with Sanger reads and what kind of results these assemblers produce given such data as input. These results (Table B.5 and Table B.6) were useful in identifying the trade-offs between the long- and short-read assemblers.

Table 4.4 shows the overall ability of the short-read assemblers to assemble long reads; this was based on the assemblers that ran to completion with no errors.

Table 4.4: Results to show which short-read assemblers can handle long reads. An "✓" indicates that the assembler successfully works, while a "×" indicates otherwise.

Assemblers	Sanger reads
ABYSS	×
EDENA	×
EULER-SR	✓
SSAKE	✓
VELVET	✓

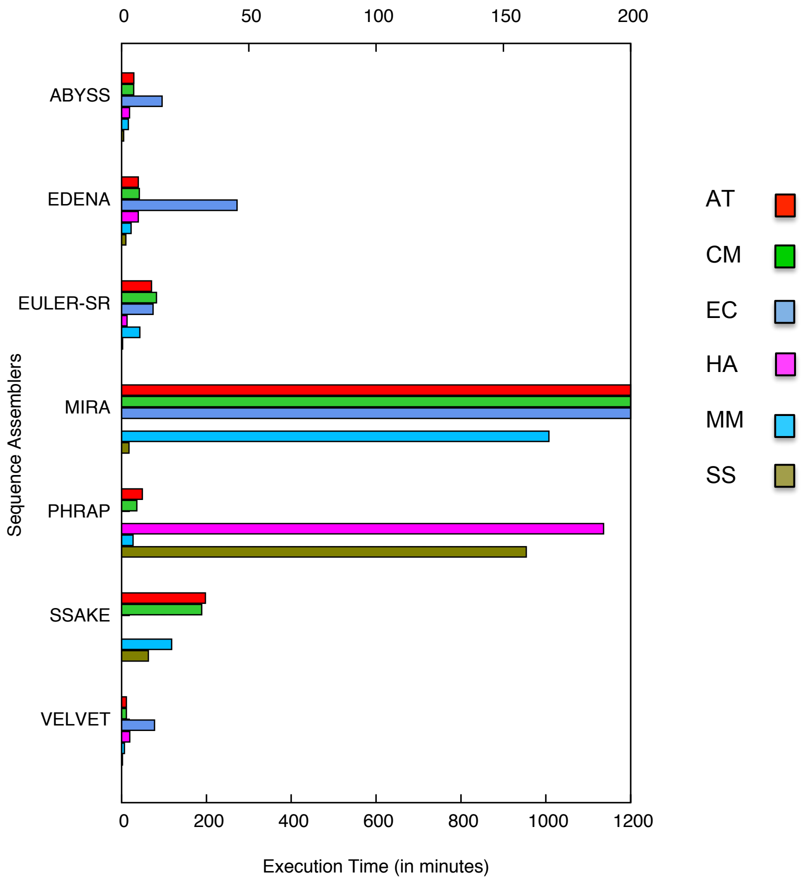


Figure 4.1: Execution time results for Illumina data. These results show that PHRAP and MIRA can take longer times to complete an assembly. A second x-axis was used because of the large difference in the time results for AT, CM, and EC (top x-axis) versus HA, MM, and SS (bottom x-axis).

The ABYSS and EDENA assemblers work with reads of the same length, and since Sanger reads are not typically of the same length, these can not assembled using the ABYSS or EDENA assemblers. However, artificial Sanger reads of the same length (three different sets of reads with lengths 900, 500 and 200) were generated to test if these assemblers could handle the Sanger reads. On this data, the ABYSS assembler ran to completion generating empty output files, while the EDENA assembler returned the Sanger reads as "erroneous" reads. The source files for these two assemblers are not publicly available, so the reasons for the incomplete assembly of the artificial same-length Sanger reads could not be identified or overcome (for example, changing a set constant in the source code and re-compiling). This experience indicates that ABYSS and EDENA assemblers were designed strictly for short reads.

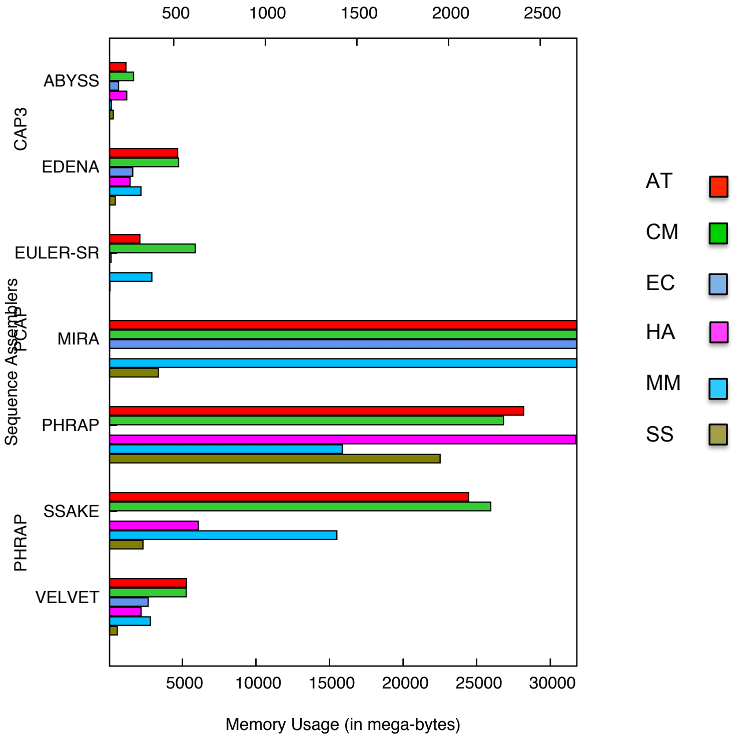


Figure 4.2: Amount of memory utilized for assembling Illumina data. These results show that PHRAP and MIRA require more RAM to complete an assembly. A second x-axis was used because of the large difference in the memory results for EC and HA (bottom x-axis) versus AT, CM, MM and SS (top x-axis).

Table 4.5: Results for running short- and long-read assemblers with Sanger data. The results from the short-read assemblers are in bold.

Organism	Metrics	CAP3	**EULER-SR**	MIRA	PCAP	PHRAP	**SSAKE**	TIGR	**VELVET**
AT	NC	472	**833**	993	701	564	**7598**	3339	**10931**
	GC (%)	92.42	**76.19**	92.51	92.62	91.77	**68.96**	94.44	**90.82**
	MC (%)	6.14	**0.36**	4.43	3.85	0.53	**0.21**	4.07	**0.66**
	N50	17409	**7819**	16908	17331	19956	**887**	5148	**583**
	LC	88604	**38063**	107897	76179	90885	**11099**	41969	**4749**
	ET(sec.)	3360.13	**55.2**	1579.77	1287.25	1168.21	**1235.69**	375.25	**47.94**
	Memory	784.05 Mb	**140.83 Mb**	480.41 Mb	341.47 Mb	503.53 Mb	**584.75 Mb**	1.53 Gb	**337.57 Mb**
CM	NC	210	**729**	320	232	100	**12778**	3291	**19502**
	GC (%)	99.01	**81.46**	97.88	98.85	98.57	**60.96**	99.11	**96.75**
	MC (%)	4.76	**0.14**	3.75	2.16	0	**0.15**	2.1	**0.17**
	N50	30047	**4539**	34847	23733	46373	**172**	11050	**290**
	LC	85321	**26025**	122630	108537	122762	**3464**	59795	**3413**
	ET(sec.)	1768.87	**30.6**	670.57	989.53	202.89	**2362.17**	1059.16	**37.58**
	Memory	575.12 Mb	**116.89 Mb**	384.541 Mb	339.02 Mb	325.09 Mb	**559.76 Mb**	797.84 Mb	**410.61 Mb**
EC	NC	17233	**-**	-	11314	463	**-**	75961	**371894**
	GC (%)	99.99	**-**	-	98.36	96.68	**-**	99.97	**-**
	MC (%)	0.69	**-**	-	0.53	18.14	**-**	0.14	**-**
	N50	2791	**-**	-	1518	43381		920	**130**
	LC	18986	**-**	-	4217	150428	**-**	2562	**833**
	ET(sec.)	12484.93	**-**	-	1629.51	1966.23	**-**	32461.52	**399.66**
	Memory	2.18 Gb	**-**	-	576.37 Mb	1.98 Gb	**-**	3.57 Gb	**2.36 Gb**
HA	NC	5698	**-**	87	2523	97	**-**	25471	**122453**
	GC (%)	99.99	**-**	99.8	93.71	98.34	**-**	99.86	**-**
	MC (%)	0.65	**-**	12.64	2.26	11.34	**-**	0.09	**-**
	N50	2910	**-**	848106	1639	86858	**-**	919	**129**
	LC	13628	**-**	848106	4236	221754	**-**	2135	**734**
	ET(sec.)	3869.75	**-**	4068.7	494.23	470.11	**-**	4771.64	**133.47**
	Memory	742.44 Mb	**-**	911.97 Mb	340.12 Mb	13.25 Gb	**-**	1.25 Gb	**666.37 Mb**
MM	NC	15	**194**	72	18	21	**4432**	623	**3701**
	GC (%)	99.63	**97.57**	99.88	99.81	99.69	**93.43**	99.8	**99.56**
	MC (%)	20	**0**	6.94	22.22	4.76	**0.09**	7.06	**0.19**
	N50	364018	**22022**	410341	308649	576887	**1390**	71954	**1521**
	LC	561631	**96652**	513463	405653	604106	**7106**	314538	**8533**
	ET(sec.)	1411.26	**15**	595.51	1328.58	181.81	**664.51**	608.81	**28.17**
	Memory	521.46 Mb	**103.89 Mb**	378.15 Mb	339.17 Mb	391.78 Mb	**456.98 Mb**	739.72 Mb	**246.74 Mb**
SS	NC	-	**479**	-	3187	484	**5400**	3920	**9241**
	GC (%)	-	**71.1**	-	93.88	87.42	**60.64**	89.79	**85.42**
	MC (%)	-	**1.88**	-	0.38	8.26	**0.2**	5.79	**0.23**
	N50	-	**6385**	-	1641	9442	**824**	1949	**513**
	LC	-	**31844**	-	4216	36205	**13503**	36204	**4732**
	ET(sec.)	-	**39**	-	656.3	172.29	**859.57**	1698.53	**31.67**
	Memory	-	**66.08 Mb**	-	341.07 Mb	322.14 Mb	**512.87 Mb**	771.63 Mb	**198 Mb**

NC - Number of Contigs, GC - Genome Coverage, MC -Misassembled Contigs, LC - Largest Contig, ET - Execution Time, and “-” - no result.

Table 4.5 presents the results from assembling the Sanger reads on both short- and long-read assemblers with the focus on the performance of the short-read assemblers (EULER-SR, SSAKE and VELVET results in bold font). A one-way ANOVA test was conducted for each metric described in Section 3.2 to assess whether there was any significant difference across the various assemblers (see Table E.5 and Table E.6). There was no statistically significant difference at the $p < 0.05$ level for the number of contigs produced ($F(7,32) = 1.568$, $p = 0.181$), the misassembled contigs results ($F(7,30) = 1.457$, $p = 0.220$), the N50 size results ($F(7,32) = 1.916$, $p = 0.100$), the execution time results ($F(7,32) = 1.156$, $p = 0.355$), and the memory results ($F(7,32) = 0.889$, $p = 0.526$). For the genome coverages, there was a statistically significant difference at the $p < 0.05$ level with $F(7,30) = 7.828$, $p < 0.001$, and the post-hoc comparisons indicated that there was significant difference in the genome coverages from the EULER-SR (mean = 81.58%) and SSAKE (mean = 70.99%) versus the other assemblers (see Table E.21 and Figure C.3). In general, EULER-SR and SSAKE have lower genome coverages than the other assemblers.

4.3 Assembling mixed data sets

The goal of this section was to determine the most appropriate approach for assembling DNA reads when working with combined data sets from different sequencing technologies. Different methodologies were used to assemble the mixed data sets, as described in Section 3.2. The results from these methodologies are presented in the following subsections.

4.3.1 Assembling 454 reads merged with Sanger reads

The only assemblers that ran to completion with output produced when the 454 reads and Sanger reads were merged (see Figure 3.8(b)) were the EULER-SR and VELVET assemblers. More details of the results are in Table B.7 and Table B.8 of Appendix B. When the PHRAP and MIRA assemblers were used in assembling these data sets, these assemblers ran for about three days, producing no outputs and just accumulating memory resources (PHRAP— up to 37 Gb). Table 4.6 presents the results from assembling these data sets using EULER-SR and VELVET. A t-test was conducted for each metric described in Section 3.2 to assess whether there was any significant difference between the two assemblers. The statistical results are provided in Table E.7 and Table E.8 of Appendix E. There was a significant difference at the $p < 0.05$ level between EULER-SR and VELVET for the following metrics: number of contigs (EULER-SR mean = 946.83, VELVET mean = 114000), genome coverage (EULER-SR mean = 97.47%, VELVET mean = 81.27%), N50 size (EULER-SR mean = 35300, VELVET mean =160), and memory used (EULER-SR mean = 128 MB, VELVET mean = 749 MB). There was no statistical significant difference in the misassembled contigs ($p = 0.152$) and the execution time ($p = 0.304$). The overall performance of the assemblers based on the metrics assessed suggest that the EULER-SR assembler performs better than VELVET on merged 454 and Sanger reads.

Table 4.6: Results for assembling 454 reads merged with Sanger reads.

Organism	Metrics	EULER-SR	VELVET
AT	NC	2289	30449
	GC (%)	93.83	68.19
	MC (%)	1.87	1.84
	N50	8682	85
	LC	51526	375
	ET (in seconds)	493.2	300.98
	Memory	176.52 Mb	915.33 Mb
CM	NC	1659	32398
	GC (%)	98.66	90.2
	MC (%)	0.24	5.03
	N50	23648	249
	LC	73702	4749
	ET (in seconds)	274.8	208.46
	Memory	131.75 Mb	698.47 Mb
EC	NC	788	328617
	GC (%)	98.04	**
	MC (%)	2.79	**
	N50	58660	133
	LC	165819	833
	ET (in seconds)	719.44	570.09
	Memory	162.24 Mb	843.47 Mb
HA	NC	177	115773
	GC (%)	98.03	**
	MC (%)	0	**
	N50	47893	301
	LC	110407	635
	ET (in seconds)	551.3	268.26
	Memory	96.74 Mb	736.74 Mb
MM	NC	278	27633
	GC (%)	99.21	85.43
	MC (%)	1.08	2.38
	N50	39976	245
	LC	172919	4749
	ET (in seconds)	129	92.89
	Memory	118.76 Mb	443.49 Mb
SS	NC	490	151594
	GC (%)	97.02	**
	MC (%)	2.96	**
	N50	32905	124
	LC	120800	635
	ET (in seconds)	201	180.12
	Memory	86.64 Mb	857.15 Mb

NC - Number of Contigs, GC - Genome Coverage, MC -Misassembled Contigs, LC - Largest Contig, ET - Execution Time, “**” - unreasonable results in the sense of either too many contigs or very small N50 size.

4.3.2 Assembling Illumina reads merged with Sanger reads

These data sets were assembled using the EULER-SR and VELVET assemblers based on the methodology described in Figure 3.8(b). Table 4.7 presents the results from the assemblies of these data sets. More details of the results are provided in Table B.9 and Table B.10 of Appendix B. A t-test was conducted for each metric described in Section 3.2 to assess whether there was any significant difference between the two assemblers. The statistical results are in Table E.9 and Table E.10 of Appendix E. There was a significant difference at the $p < 0.05$ level between EULER-SR and VELVET for the following metrics: number of contigs (EULER-SR mean = 10300, VELVET mean = 116000), genome coverage (EULER-SR mean = 97.72%, VELVET mean = 88.36%), N50 size (EULER-SR mean = 4040, VELVET mean = 684.50), and execution time (EULER-SR mean = 3660 seconds, VELVET mean = 476 seconds). There was no statistical significant difference in the misassembled contigs (p = 0.580) and the memory (p = 0.737). The overall performance of the assemblers based on the genome coverages indicate that the EULER-SR assembler performs better than VELVET.

Table 4.7: Results for assembling Illumina reads merged with Sanger reads.

Organism	Metrics	EULER-SR	VELVET
AT	NC	1241	9400
	GC (%)	95.05	94.9
	MC (%)	0.81	1.04
	N50	47133	1136
	LC	173182	12107
	ET (in seconds)	969.87	189.08
	Memory	729.39 Mb	675.39 Mb
CM	NC	1659	19061
	GC (%)	98.42	98.24
	MC (%)	0.66	0.17
	N50	36562	410
	LC	154600	4579
	ET (in seconds)	1061.96	177.87
	Memory	682.72 Mb	645.5 Mb
EC	NC	17840	351386
	GC (%)	98.35	**
	MC (%)	0.45	**
	N50	1378	109
	LC	11859	602
	ET (in seconds)	10347.79	1280.37
	Memory	3.56 Gb	3.27 Gb
HA	NC	30448	117367
	GC (%)	97.9	**
	MC (%)	0	**
	N50	234	107
	LC	3362	1359
	ET (in seconds)	5415.64	565.56
	Memory	1.86 Gb	1.29 Gb
MM	NC	423	4018
	GC (%)	98.89	98.62
	MC (%)	2.13	0.35
	N50	116741	2031
	LC	292143	16109
	ET (in seconds)	515.35	94.84
	Memory	459.59 Mb	427.2 Mb
SS	NC	-	195191
	GC (%)	-	61.66
	MC (%)	-	3.51
	N50	-	314
	LC	-	4749
	ET (in seconds)	-	548.4
	Memory	-	973.1 Mb

NC - Number of Contigs, GC - Genome Coverage, MC -Misassembled Contigs, LC - Largest Contig, ET - Execution Time, “-” - no result, “**” - unreasonable results in the sense of either too many contigs or very small N50 size.

4.3.3 Assembling Illumina reads merged with 454 reads

These data sets were assembled using the EULER-SR and VELVET assemblers based on the methodology described in Figure 3.8(a). Table 4.8 presents the results from the assemblies of these data sets. More details of the results are provided in Table B.11 and Table B.12 of Appendix B. A t-test was conducted for each metric described in Section 3.2 to assess whether there was any significant difference between the two assemblers. The statistical results are provided in Table E.11 and Table E.12 of Appendix E. There was a significant difference at the $p < 0.05$ level between EULER-SR and VELVET for the following metrics: number of contigs (EULER-SR mean = 2531.50, VELVET mean = 18600), misassembled contigs (EULER-SR mean = 4.37%, VELVET mean = 0.23%), N50 size (EULER-SR mean = 23400, VELVET mean = 667.50), and execution time (EULER-SR mean = 2670 seconds, VELVET mean = 521 seconds). There was no statistical significant difference in the genome coverages ($p = 0.637$) and the memory ($p = 0.651$). The differences between the two assemblers indicate that both assemblers are useful when assembling merged Illumina and 454 reads which is expected since both assemblers were designed for short reads.

4.3.4 Assembling Illumina contigs merged with 454 contigs

All the Illumina contigs produced by the EULER-SR assembler were merged with the 454 contigs produced by the EULER-SR, and the same Illumina contigs were merged with 454 contigs produced by PHRAP (Figure 3.8(d)). These merged contigs were assembled using PHRAP. PHRAP was chosen among other long-read assemblers because for all the assembly runs, it had no error output and it always produced results. The CAP3 and TIGR sequence assemblers had a problem with the variations in the lengths of the merged contigs (sometimes very short), and the MIRA and PCAP assemblers depend on supplied quality files to run. Table 4.9 presents the results generated from these assemblies. More details of the result are given in Table B.13 and Table B.14 of Appendix B. What is different in this assembly is that in one case contigs from just EULER-SR (Set A) were used, while in the other case contigs from both EULER-SR and PHRAP (Set B) were used. A t-test was conducted for each metric described in Section 3.2 to assess whether there was any significant difference between the two merged contig sets. The statistical results are provided in Table E.13 and Table E.14 of Appendix E. For the misassembled contig results, there was statistical significant difference between Set A (mean = 1.95%) and Set B (mean = 14.18%) with $p < 0.001$, and for the N50 results between Set A (mean = 57500) and Set B (10100) with Levene's test p-value = 0.006. There was no statistical significance observed from the other metrics; number of contigs ($p = 0.084$), genome coverages ($p = 0.835$), execution time ($p = 0.478$), and memory ($p = 0.458$). Based on the results obtained, both assemblers would be useful.

Table 4.8: Results for assembling Illumina reads merged with 454 reads.

Organism	Metrics	EULER-SR	VELVET
AT	NC	5824	37851
	GC (%)	93.25	90.87
	MC (%)	5.24	0.75
	N50	6817	102
	LC	42900	1043
	ET (in seconds)	2029.79	573.96
	Memory	1.03 Gb	2.40 Gb
CM	NC	1855	15826
	GC (%)	97.89	97.28
	MC (%)	1.02	1.09
	N50	31903	430
	LC	143473	3079
	ET (in seconds)	1505.13	397.62
	Memory	938.08 Mb	1.21 Gb
EC	NC	605	4512
	GC (%)	96.2	97.91
	MC (%)	19.67	1.31
	N50	60445	2849
	LC	294663	18945
	ET (in seconds)	5609.01	1048.25
	Memory	3.71 Gb	3.30 Gb
HA	NC	4679	15386
	GC (%)	91.68	98.11
	MC (%)	0	0.07
	N50	1382	229
	LC	8308	1981
	ET (in seconds)	4642.01	647.15
	Memory	1.99 Gb	2.16 Gb
MM	NC	761	19220
	GC (%)	99.06	97.54
	MC (%)	0.53	0.06
	N50	27296	151
	LC	123840	1464
	ET (in seconds)	675.84	168.91
	Memory	525.79 Mb	625.77 Mb
SS	NC	1465	19091
	GC (%)	97.04	98.21
	MC (%)	0.96	0.02
	N50	12753	244
	LC	65025	2463
	ET (in seconds)	1567.31	292.57
	Memory	718.81 Mb	1.01 Gb

NC - Number of Contigs, GC - Genome Coverage, MC -Misassembled Contigs, LC - Largest Contig, ET - Execution Time.

4.3.5 Assembling 454 contigs merged with Sanger reads

The contigs produced from each of the assemblers as presented in Table 4.2 were merged with Sanger reads and assembled using the long-read assemblers— PHRAP, CAP3 and PCAP (as described in Figure 3.8(c)). The results for these assemblies are given in Table B.17 and Table B.18 of Appendix B due to the large volume of results. The results indicate that the sequence assemblers produced good results; in terms of fewer number of contigs (see Figure D.11), good genome coverage results (see Figure C.5), and N50 sizes (see Figure C.15) in comparison to when the 454 reads were assembled alone. Furthermore, the results also suggest that the PHRAP assembler would be the most suitable because for all the assemblies performed, it ran to completion with good results to show. In some of the assemblies, the CAP3 and PCAP assemblers produced no results. Other than the rate of completion, all the sequence assemblers produced good results.

Table 4.9: Results for assembling Illumina contigs merged with 454 contigs using PHRAP.

Organism	Metrics	EULER-SR contigs	EULER-SR + PHRAP contigs
AT	NC	410	182
	GC (%)	96.01	97.13
	MC (%)	0.98	14.84
	N50	47193	55022
	LC	127224	146121
	ET (in seconds)	81.52	160.4
	Memory	139.4 Mb	162.6 Mb
CM	NC	254	48
	GC (%)	98.38	99.01
	MC (%)	1.18	12.5
	N50	55178	247589
	LC	200558	627649
	ET (in seconds)	235.83	680.66
	Memory	154.14 Mb	165.97 Mb
EC	NC	247	61
	GC (%)	98.082	99.22
	MC (%)	0.4	16.39
	N50	80429	175813
	LC	180280	529850
	ET (in seconds)	816.07	1558.63
	Memory	192.48 Mb	199.12 Mb
HA	NC	85	74
	GC (%)	97.80	97.42
	MC (%)	1.18	14.87
	N50	47892	45634
	LC	89053	86112
	ET (in seconds)	127.36	86.54
	Memory	127.63 Mb	117.17 Mb
MM	NC	94	94
	GC (%)	99.22	97.88
	MC (%)	1.06	20.21
	N50	86217	58567
	LC	292800	175354
	ET (in seconds)	70.76	78.33
	Memory	116.37 Mb	122.73 Mb
SS	NC	159	159
	GC (%)	97.23	95.09
	MC (%)	6.92	6.29
	N50	28024	22953
	LC	66369	58127
	ET (in seconds)	56.78	33.22
	Memory	110.64 Mb	95.81 Mb

NC - Number of Contigs, GC - Genome Coverage, MC -Misassembled Contigs, LC - Largest Contig, and ET - Execution Time.

4.3.6 Assembling Illumina contigs merged with Sanger reads

The contigs produced for the Illumina reads by each of the short-read assemblers as presented in Table 4.4 were merged with Sanger reads and assembled using the long-read assemblers— PHRAP and CAP3 (as described in Figure 3.8(c)). Due to the large volume of results generated, the results for these assemblies are given in Table B.19 and Table B.20 of Appendix B. These results indicate good output from both sequence assemblers with good genome coverages (see Figure C.4). With the exception of the misassembled contig results from the PHRAP assembler (see Figure C.9), the PHRAP assembler would be the more appropriate assembler. It seems to be more robust because it ran to completion in all of the assemblies with good results. In some of the assemblies, CAP3 produced no results.

4.3.7 Assembling merged 454, Illumina and Sanger reads

The contigs produced from assembling all the different types of data sets merged (Table 3.8(b)) are presented in Table 4.10 (also see Table B.15 and Table B.16). A t-test was conducted for each metric described in Section 3.2 to assess whether there was any significant difference between the EULER-SR and VELVET assemblers. The statistical results are provided in Table E.15 and Table E.16 of Appendix E. There was a significant difference at the $p < 0.05$ level between EULER-SR and VELVET for the following metrics: number of contigs (EULER-SR mean = 14500, VELVET mean = 92300), N50 size (EULER-SR mean = 17600, VELVET mean = 146), and execution time (EULER-SR mean = 4250 seconds, VELVET mean = 476 seconds). There was no statistical significant difference for the genome coverages ($p = 0.458$), misassembled contigs ($p = 0.337$), and memory usage ($p = 0.746$). The differences between the two assemblers indicate both assemblers work well except for the number of contigs and N50 sizes for which EULER-SR performed better than VELVET.

4.4 The effect of quality values when assembling DNA reads

The goal for assembling the real data sets for Sanger, 454, and Illumina with and without their quality information was to establish the effect of the presence (or absence) of quality values on the accuracy of contigs, execution time and memory usage; as described in Section 3.3. Tables 4.11, 4.12, and 4.13 present the results for determining the effect of quality values. T-tests were conducted to evaluate the effect of assembling data with and without quality values, and the evaluations were based on the metrics described in Section 3.2.

4.4.1 Statistical results for Sanger data with and without quality data

Three assemblers were used: CAP3, PHRAP, and TIGR Assembler. The t-tests were conducted to compare the results obtained from each of the assemblers when quality values were included against when no quality data was provided (see Table E.22, Table E.23, and Figure D.15). There was no statistically significant difference for the results obtained from the CAP3 and PHRAP assemblers. For TIGR Assembler, there was no statistically significant difference at the $p < 0.05$ level in any of the results obtained except for the N50 size results (mean when quality values were used = 29400, without quality values = 6258) with Levene's test p-value = 0.036.

Table 4.10: Results for assembling merged 454, Illumina and Sanger reads.

Organism	Metrics	EULER-SR	VELVET
AT	NC	5240	37762
	GC (%)	94.45	94.59
	MC (%)	5.59	1.21
	N50	10903	107
	LC	114287	909
	ET (in seconds)	2747.38	402.89
	Memory	1.15 Gb	1.07 Gb
CM	NC	2037	27782
	GC (%)	96.73	99.07
	MC (%)	0.64	0.14
	N50	30514	180
	LC	121176	2648
	ET (in seconds)	1860.28	295.71
	Memory	1.00 Gb	913.11 Mb
EC	NC	32288	330851
	GC (%)	98.78	**
	MC (%)	0.55	**
	N50	884	132
	LC	9006	833
	ET (in seconds)	12300.29	1174.7
	Memory	3.90 Gb	3.00 Gb
HA	NC	45342	117612
	GC (%)	98.96	**
	MC (%)	0.57	**
	N50	123	123
	LC	2962	1450
	ET (in seconds)	6546.06	584.56
	Memory	2.07 Gb	1.83 Gb
MM	NC	633	16986
	GC (%)	99.25	98.78
	MC (%)	0.16	0.21
	N50	56345	179
	LC	275961	1567
	ET (in seconds)	878.28	139.59
	Memory	606.5 Mb	501.59 Mb
SS	NC	1299	22902
	GC (%)	79.6	99.1
	MC (%)	1.85	0.32
	N50	7055	155
	LC	32350	2449
	ET (in seconds)	1188.75	260.65
	Memory	773.16 Mb	927.54 Mb

NC - Number of Contigs, GC - Genome Coverage, MC -Misassembled Contigs, LC - Largest Contig, ET - Execution Time, and "**" - unreasonable results in the sense of either too many contigs or very small N50 size.

4.4.2 Statistical results for Illumina data with and without quality data

Three assemblers were used: EDENA, EULER-SR, and VELVET. These assemblers were chosen because they accept FASTQ formats as input, unlike the other assemblers which just need a FASTA file for input (and require no quality files). The t-tests were conducted to compare the results obtained from each of the assemblers when quality values were included against when no quality data was provided (see Table E.24, Table E.25, and Figure D.14). There was no statistically significant difference for the results obtained from the EDENA and VELVET assemblers. For EULER-SR, there was no statistically significant difference at the $p < 0.05$ level in any of the results obtained except for the genome coverages (mean when quality values were used = 92.42%,

Table 4.11: Results for assembling real Sanger reads to test the effect of quality values.

Organism	Metrics	CAP3	PHRAP	TIGR
AT with quality values	NC	472	564	3339
	GC (%)	92.42	91.77	94.44
	MC (%)	6.14	0.53	4.07
	N50	17409	19956	5148
	LC	88604	90885	41969
	ET (in seconds)	3360.13	1168.21	375.25
	Memory	784.05 Mb	503.53 Mb	1.53 Gb
AT without quality values	NC	470	458	4874
	GC (%)	92.09	91.85	94.64
	MC (%)	3.83	2.4	2.32
	N50	17707	22309	2420
	LC	64144	90789	40121
	ET (in seconds)	3135.11	1106.92	3882.79
	Memory	803.31 Mb	484.73 Mb	1.50 Gb
CM with quality values	NC	210	100	3291
	GC (%)	99.01	98.57	99.11
	MC (%)	4.76	0	2.1
	N50	30047	46373	11050
	LC	85321	122762	59795
	ET (in seconds)	1768.87	202.89	1059.16
	Memory	575.12 Mb	325.09 Mb	797.84 Mb
CM without quality values	NC	216	128	6375
	GC (%)	98.78	98.25	99.27
	MC (%)	3.7	0	1.28
	N50	28319	46389	1986
	LC	73441	122874	31204
	ET (in seconds)	1436.98	212.9	1294.08
	Memory	574.82 Mb	318.35 Mb	922.09 Mb
MM with quality values	NC	15	21	623
	GC (%)	99.63	99.69	99.8
	MC (%)	20	4.76	7.06
	N50	364018	576887	71954
	LC	561631	604106	314538
	ET (in seconds)	1411.26	181.81	608.81
	Memory	521.46 Mb	391.78 Mb	739.72 Mb
MM without quality values	NC	12	13	1308
	GC (%)	99.72	99.46	99.82
	MC (%)	8.33	7.69	3.29
	N50	405649	570536	14368
	LC	595046	576889	57120
	ET (in seconds)	1406.5	174.13	717.27
	Memory	522.14 Mb	384.67 Mb	863.41 Mb

NC - Number of Contigs, GC - Genome Coverage, MC -Misassembled Contigs, LC - Largest Contig, and ET - Execution Time.

without quality values = 97.63%) with Levene's test p-value = 0.041.

4.4.3 Statistical results for 454 data with and without quality data

Two assemblers were used: EULER-SR and VELVET. The t-tests were conducted to compare the results obtained from each of the assemblers when quality values were included against when no quality data was provided (see Table E.26, Table E.27, and Figure D.13). For EULER-SR, there was no statistically significant difference at the $p < 0.05$ level in any of the results obtained except for the genome coverage (mean when quality values were used = 72.08%, without quality values = 97.20%) with Levene's test p-value = 0.029. There was no statistically significant difference for any of the results obtained from the VELVET assembler.

Table 4.12: Results for assembling real Illumina reads to test the effect of quality values.

Organism	Metrics	EDENA	EULER-SR	VELVET
EC with quality values	NC	1153	971	666
	GC (%)	97.24	97.45	98.02
	MC (%)	0.09	2.99	0
	N50	8827	18352	48723
	LC	31785	71243	126542
	ET (in seconds)	2089.87	3610.76	973.2
	Memory	1.73 Gb	2.93 Gb	2.66 Gb
EC without quality values	NC	680	1276	666
	GC (%)	98.32	98.58	98.02
	MC (%)	0.15	0.31	0
	N50	16430	24087	48723
	LC	67082	103365	126542
	ET (in seconds)	2721.5	9742.2	776.83
	Memory	1.63 Gb	156.71 Mb	2.68 Gb
HA with quality values	NC	1849	486	474
	GC (%)	94.59	97.13	98.01
	MC (%)	0	0	1.48
	N50	1275	12865	9244
	LC	5771	31069	28127
	ET (in seconds)	1975.29	1726.46	415.26
	Memory	1.68 Gb	1.51 Gb	1.12 Gb
HA without quality values	NC	453	467	474
	GC (%)	97.98	98.98	98.01
	MC (%)	0.22	1.5	1.48
	N50	7746	14405	9244
	LC	41362	62017	28127
	ET (in seconds)	2370.9	781.8	1159.91
	Memory	1.45 Gb	63.82 Mb	2.2 Gb
SS with quality values	NC	1668	4887	2758
	GC (%)	94.06	82.72	97.36
	MC (%)	0.06	0.23	0.07
	N50	1965	492	1262
	LC	13883	5388	8005
	ET (in seconds)	482.08	388.94	90.8
	Memory	501.92 Mb	397.9 Mb	521.54 Mb
SS without quality values	NC	1278	2281	1160
	GC (%)	95.82	95.34	98.55
	MC (%)	0	0.044	1.12
	N50	2606	1709	4594
	LC	13883	10963	17831
	ET (in seconds)	604.13	152.4	120.65
	Memory	446.81 Mb	74.26 Mb	579.24 Mb

NC - Number of Contigs, GC - Genome Coverage, MC -Misassembled Contigs, LC - Largest Contig, and ET - Execution Time.

For the 454 and Sanger reads, the results indicate that, in general, sequence assemblers can handle DNA reads and produce good contigs regardless of the availability of quality information (see Figure D.13 and Figure D.15). The same conclusion can be drawn for Illumina reads (see Figure D.14) with the exception of the EULER-SR assembler. When assembling without the quality values in some of the organisms, the EULER-SR assembler required less memory, and produced contigs with a significant amount of difference in genome coverages.

Table 4.13: Results for assembling real 454 reads to test the effect of quality values.

Organism	Metrics	EULER-SR	VELVET
AT with quality values	NC	8139	32814
	GC (%)	93.34	93.86
	MC (%)	10.01	0.73
	N50	3030	119
	LC	23262	1292
	ET (in seconds)	1213.43	229.74
	Memory	487.88 Mb	745.04 Mb
AT without quality values	NC	808	32418
	GC (%)	93.96	93.86
	MC (%)	0.99	0.73
	N50	2196	119
	LC	14509	1292
	ET (in seconds)	369	262.54
	Memory	115.4 Mb	667.84 Mb
EC with quality values	NC	802	3778
	GC (%)	97.13	98.03
	MC (%)	3.42	1.88
	N50	58858	3441
	LC	165818	16401
	ET (in seconds)	538.54	138.92
	Memory	450.37 Mb	386.34 Mb
EC without quality values	NC	885	3778
	GC (%)	97.79	98.03
	MC (%)	1.02	1.88
	N50	57777	3441
	LC	165903	16401
	ET (in seconds)	1956.23	146.66
	Memory	733.76 Mb	381.67 Mb
CM with quality values	NC	3385	13797
	GC (%)	97.84	98.89
	MC (%)	0.68	0.38
	N50	13677	493
	LC	76445	4050
	ET (in seconds)	635.97	173.47
	Memory	394.82 Mb	747 Mb
CM without quality values	NC	1257	13797
	GC (%)	98.37	98.89
	MC (%)	0.16	0.38
	N50	16659	493
	LC	78363	4050
	ET (in seconds)	235.2	201.69
	Memory	113.04 Mb	472.51 Mb
MM with quality values	NC	817	14814
	GC (%)	**	98.68
	MC (%)	0	0.84
	N50	134	210
	LC	551	1996
	ET (in seconds)	18.48	67.5
	Memory	36.33 Mb	383.4 Mb
MM without quality values	NC	618	14814
	GC (%)	98.69	98.68
	MC (%)	0.16	0.84
	N50	7222	210
	LC	22488	1996
	ET (in seconds)	105	69.82
	Memory	82.47 Mb	326.54 Mb

NC - Number of Contigs, GC - Genome Coverage, MC -Misassembled Contigs, LC - Largest Contig, and ET - Execution Time.

Chapter 5

Discussion

This research explored different sequence assemblers to determine their performances when working with mixed data sources (DNA reads from different sequencing technologies). These experiments were performed in order to determine a suitable methodology to use when working with these mixed data sets, and also to determine an appropriate assembler (or assemblers) for such assembly projects. This work also involved investigating the effect of quality values on the performance of the sequence assemblers.

The goal of this chapter is to summarize and interpret the results presented in Chapter 4. This chapter presents contributions, conclusions drawn from the different assembly experiments performed for this research. Some related work and possible future work are also presented.

5.1 Conclusions and Recommendations

Before the advent of the next generation sequencing techniques (for example Solexa/Illumina, 454, SOLiD), the Sanger sequencing technology was the common approach used for sequencing genomes. There were sequence assemblers designed to handle the Sanger reads. However, with the current (the newer) sequencing technologies, there are other sequence assemblers that have been developed to handle the short-read data. A part of this research involved determining the trade-off between these types of assemblers. One question of interest was; can long-read assemblers handle short reads?

Amongst the long-read assemblers used for this research, only PHRAP and MIRA could assemble the 454 and Illumina data sets. However, these long-read assemblers generally required a lot more memory and time to complete an assembly in comparison to the short-read assemblers (see Section 4.1). MIRA cannot handle the Illumina reads alone. It assembles Illumina reads as an hybrid; that is, alongside MIRA-454 or -Sanger contigs (these are 454 and Sanger contigs already produced by the MIRA assembler) which serve as a reference during the assembly process. In some of the Illumina assemblies, MIRA produced better results than those from the short-read assemblers. For the

Illumina reads, the quality of results produced by PHRAP is not as good as those from the short-read assemblers, especially in terms of the quantity of contigs produced. For example, for the SS organism, PHRAP produced 88131 contigs, compared to EULER-SR which produced 474 contigs. For the 454 reads, both assemblers produced good results (genome coverage, reasonable number of contigs, and large sized contigs). PHRAP and MIRA can therefore be used for 454 sequencing projects. The long-read assemblers would not be appropriate for Illumina sequence assembly. MIRA may be used only when there are either 454 or Sanger reads also available for the sequence assembly.

Another part of this research involved determining how the short-read assemblers handle the Sanger reads. Amongst the short-read assemblers used for this research, only EULER-SR, SSAKE, and VELVET produced results. The results from the assemblies indicate that using the short-read assemblers (with the exception of SSAKE) is much faster and requires less memory to complete an assembly in comparison to the long-read assemblers. However, the short-read assemblers (with the exception of VELVET) in some of the cases do not produce good genome coverage results in comparison to the long-read assemblers (for example, EULER-SR produced a genome coverage of just 76.29% for the AT organism compared to 91.77% from PHRAP). VELVET may not be appropriate for assembling Sanger reads because of the large quantity of contigs it produced. EULER-SR may be used for assembling Sanger reads. Generally, while the short-read assemblers required less time and memory, the quality of results produced by the long-read assemblers is still better.

Based on the assemblies conducted for the individual sequencing technologies, either EULER-SR or VELVET would be appropriate for assembling the Illumina reads. These assemblers consistently produced quality results in terms of genome coverage, number of contigs, large contigs, and small numbers of misassemblies. VELVET would be appropriate when assembling data sets in which the read-pair information is available (for example, the Illumina data set for the EC organism included the read-pair information, greatly improving VELVET's assembly performance with zero misassemblies). For the 454 reads, the EULER-SR, MIRA, and PHRAP assemblers are appropriate for sequence assembly. For long-read assemblies, PHRAP would be very appropriate, especially because of its consistency in producing good results: it always ran to completion with results to show. Another preference for the long-read assemblies would be the MIRA assembler, although its assembly process depends on the availability of quality information.

When dealing with a combination of data from different sequencing technologies (for example, 454+Sanger, Illumina+Sanger, 454+Illumina, and other combinations), it would be appropriate to have a working strategy on how to assembly these data sets. In the case of a sequencing project which involves a short-read data set and long-read data set (454+Sanger, Illumina+Sanger), different assembly strategies were tested in this research (see results in Tables B.7 - B.20). The best results were obtained from assembling the short reads first using an appropriate sequence assembler, then merging the contigs produced with the Sanger reads (long reads), and finally assembling the merged data using an appropriate long-read assembler (Figure 5.1).

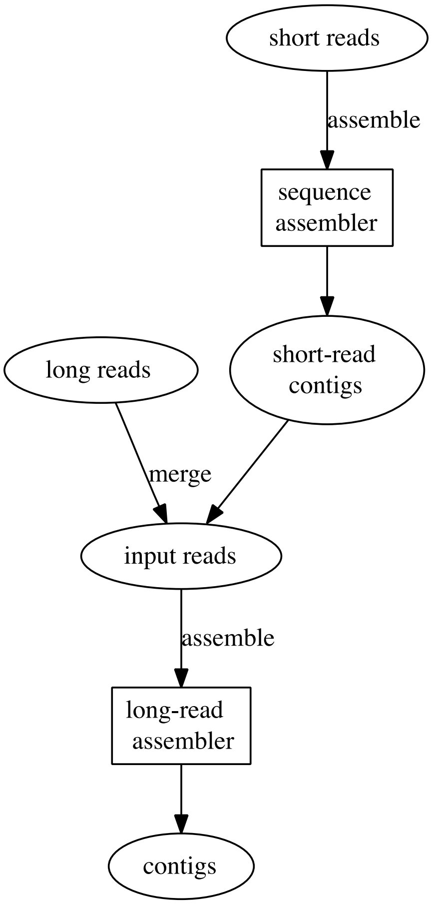


Figure 5.1: Best approach to handle an assembly that involves a combination of short and long reads

When working with a sequence assembly project that involves both Illumina and 454 reads, the results from the assemblies performed indicate that the most appropriate methodology to use would be to assemble each of the data sets separately, then merge the produced contigs, and then assembly the merged contigs on a long-read assembler (Figure 5.2). This approach produces better results in terms of the number of contigs produced, very good genome coverages, large contigs, and fast execution (see results B.11 - B.14 in Appendix B.).

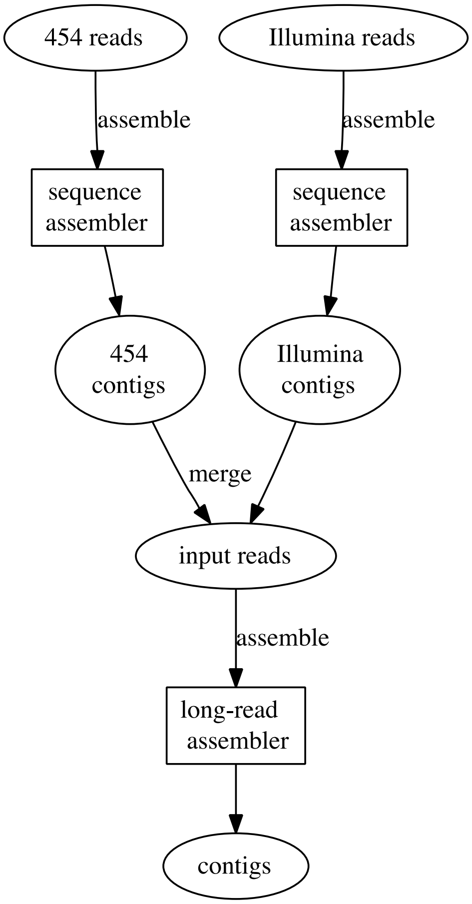


Figure 5.2: Best approach to handle an assembly that involves a combination of 454 and Illumina reads

When assembling the 454 data with and without quality values, the results obtained indicated that the short-read assemblers (EULER-SR and VELVET) produced very similar results (as shown in Figure D.15). The same applied to the Sanger data when assembled using CAP3, PCAP, and PHRAP (Figure D.13). This means that the sequence assembly algorithms are robust enough to produce good results even in the absence of quality information. The same can be concluded for the Illumina sequence assemblers (EDENA and VELVET shown in Figure D.14), with the exception of EULER-SR which showed disparity in some of its results (for example, in the amount of memory and genome coverage percentages) when quality information was unavailable.

5.2 Related Work

Recent publications [15, 20, 50] have described hybrid sequence assembly strategies for handling data from different sequencing technologies. Their results have shown the feasibility of assembling a mix of data from different sequencing technologies, and also that the resulting sequences are improvements on the final sequence of the targeted genome. In general, the results are consistent with those described in this thesis.

A Sanger / 454 hybrid approach was applied to assembling microbial genomes by Goldberg et al [20]. Their study found that the 454 sequencing platform is better used as a complement to the Sanger sequencing method rather than being used alone. They found that the best contigs (high genome coverage and small number of gaps) were obtained by first pre-assembling the 454 DNA reads using the Newbler assembler (designed by 454 Sequencing). The contigs from the pre-assembly were then "shredded" into overlapping "pseudo-reads" of length 600 to emulate the size of Sanger reads. These "pseudo-reads" were merged with the Sanger reads, and assembled using the Celera assembler [28]. The methodology used and their results are similar to the methodology (see Figure 5.1) and results presented in this thesis. However, the second-stage sequence assemblers used here (e.g. PHRAP) do not require the shredding of contigs. This prevents the problem of misassembly that may result from breaking the contigs into smaller bits.

Another study for assembling eukaryotes by integrating Sanger, 454, and Illumina sequencing data showed that eukaryotes can be accurately assembled by effectively combining these data sources [15]. Different assembly methods, from single assemblies to combined (mixed) assemblies, were assessed. The study showed that the highest quality assemblies resulted from integrating Sanger reads, 454 reads, and Illumina contigs (from a pre-assembly process using VELVET) in a single round of assembly using Forge Assembler [39]. This methodology was not duplicated in this thesis. However, the results support the importance of having an effective methodology when handling mixed data sets in any sequence assembly project.

Finally, Wall et al performed a study to determine the optimal mixture of sequencing methods to achieve the most complete and cost effective transcriptome sequencing [50]. Their study assessed the addition of next generation sequences to traditional (Sanger) sequences to establish the appropriate combinations of these technologies for transcriptome sequencing. It was found that the addition of the Solexa and SOLiD sequences to Sanger and 454 sequences increased the mean unigene length (produced larger unigenes), the transcriptome coverage, and the percentages of genes 100% covered. This particular study did not go into much detail about the assembly process or methodologies to combine the different sequencing technologies.

5.3 Future Work

One limitation encountered at the beginning of this research was availability of data; for example, getting data sets publicly available for different kinds of organisms (bacteria, eukaryotes, and archaea). All organisms used in this work were bacteria, based on the availability of data. It might be useful to look into a broader range of organisms. It would also be useful to investigate the assembly of genomes with more complexity in terms of repeats and duplication.

The next generation sequencing technologies are rapidly evolving. This research worked with data from two of these next generation technologies — 454 sequencing and Illumina sequencing. Some of the other next generation technologies include Helicos sequencing technology which uses a single molecule approach to sequencing; it directly measures single molecules of DNA. This approach improves accuracy of DNA sequences and reduces sequencing costs, since no PCR amplification is required [8, 32]. Another next generation sequencing technology is SOLiD (Sequencing by Oligonucleotide Ligation and Detection) sequencing, which is similar to the Solexa/Illumina approach. It produces DNA reads of lengths of 30-50 bp [32]. It would be worthwhile to explore these other next generation technologies, and the performances of various assemblers on the data they produce.

Sequence assembly is a vital area in bioinformatics, hence it is crucial to know the right tools to use for genome sequencing projects. This research explored some of the tools available for sequence assembly in combination with genome sequencing technologies. It would be of good use to explore finishing tools; that is, tools used for scaffolding contigs and gap closing. This would help in examining the quality of genomes or plasmids produced by sequence assembler + finisher combinations. Some of the available finishing tools include CONSED [21] and AMOS validation [42] (a part of the AMOS sequence assembler software package).

In the course of this research, it was found that VELVET produced better results (with zero misassembly) when mate pair information was provided with the DNA input reads. Mate pair sequencing involves sequencing both ends of a DNA fragment, thus producing two sequences which are oriented in opposite directions and are a certain length apart [40]. When an assembler is provided with this information, it is useful in reconstructing the sequence of the original fragment. Mate pairs make it easier for sequence assemblers to link contigs. It would be of good use to explore the performance of sequence assemblers when provided with the mate pair information against the absence of the mate pair information.

Another limitation encountered in this research was the inability of some of the long-read assemblers (CAP3, PCAP, TIGR Assembler) to successfully assemble short reads. These limitations are described in Section 4.1. It would be very useful to further investigate the sources of these problems and offer possible solutions such as modifying the source code of the sequence assemblers.

References

[1] FASTQ format. http://en.wikipedia.org/wiki/FASTQ_format. Accessed May 2009.

[2] Genome sequence assembly primer. http://www.cbcb.umd.edu/research/assembly_primer.shtml. Accessed August 2009.

[3] How is genome sequencing done? 454 Life Sciences Corporation. http://www.454.com/downloads/news-events/how-genome-sequencing-is-done_FINAL.pdf. Accessed May 2009.

[4] Primer (molecular biology). http://en.wikipedia.org/wiki/Primer_(molecular_biology). Accessed February 2009.

[5] TIGR assembler documentation. http://www.tigr.org/software/assembler/helpfile.html. Accessed December 24, 2007.

[6] S. Batzoglou. DNA sequencing. http://ai.stanford.edu/s̃erafim/CS262_2007. Accessed February 2009.

[7] A. D. Baxevanis and F. B. Ouellette. Chapter 13 of bioinformatics: A practical guide to the analysis of genes and proteins. *John Wiley Sons Inc.*, 3rd Ed., 2005.

[8] I. Braslavsky et al. Sequence information can be obtained from single DNA molecules. *PNAS*, 100 No.7:3960–3964, 2003.

[9] T.A. Brown. Chapter 4 of genomes. *John Wiley Sons Inc.*, 1999.

[10] M. Chaisson, P. Pevzner, and H. Tang. Fragment assembly with short reads. *Bioinformatics*, 20 No. 13:2067–2074, 2004.

[11] M. J. Chaisson and P. A. Pevzner. Short read fragment assembly of bacterial genomes. *Genome Research*, 18 No. 10:324–330, 2007.

[12] B. Chevreux. MIRA: An automated genome and EST assembler. Ph.D thesis, German Cancer Research Center Heidelberg, Duisburg. 2005.

[13] H. Chou and M. H. Holmes. DNA sequence quality trimming and vector removal. *Bioinformatics*, 17 No.12:1093–1104, 2001.

[14] CodonCode Corporation. Codoncode aligner - DNA sequence assembly and alignment on Windows and Mac OS X. http://www.codoncode.com/aligner. Accessed May 2008.

[15] S. DiGuistini et al. De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. *Genome Biology*, 10 No. 9, 2009.

[16] J. Dohm, C. Lottaz, T. Borodina, and H. Himmelbauer. Sharcgs, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. *Genome Research*, 17:1697–1706, 2007.

[17] B. Ewing et al. Base-calling of automated sequencer using phred. i. accuracy assessment. *Genome Research*, 8 No.10:175–185, 1998.

[18] B. Ewing and P. Green. Base-calling of automated sequencer using phred. ii. error probabilities. *Genome Research*, 8 No.10:186–194, 1998.

[19] D. Garson. ANOVA. http://computerwranglers.com/Com531/Handouts/ANOVA.doc. Accessed January 2010.

[20] S. Goldberg et al. A Sanger / pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. *PNAS*, 103 No. 30:11240–11245, 2006.

[21] D. Gordon, C. Abajian, and P. Green. Consed: A graphical tool for sequence finishing. *Genome Research*, 8 No.10:195–202, 1998.

[22] N. Hall. Advanced sequencing technologies and their wider impact in microbiology. *Experimental Biology*, 209:1518–1525, 2007.

[23] D. Hernandez, P. Francois, L. Farinelli, M. Osteras, and J. Schrenzel. De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Research*, 18:802–809, 2008.

[24] X. Huang et al. PCAP: A whole-genome assembly program. *Genome Research*, 13:2164–2170, 2002.

[25] X. Huang and A. Madan. CAP3 sequence assembly program. *Genome Research*, 9:868–877, 1999.

[26] L. Hunter. Sequence assembly. Computational Bioscience Program, University of Colorado. http://compbio.uchsc.edu/hunter/bioi7711/sequence-assembly.ppt. Accessed December 2007.

[27] Illumina Systems & Software. DNA sequencing with Solexa technology. http://corelabs.cgrb.oregonstate.edu/files/illumina/SS_DNAsequencing.pdf. Accessed May 2009.

[28] S. Istrail et al. Whole-genome shotgun assembly and comparison of human genome assemblies. *PNAS*, 101 No. 7:1916–1921, 2009.

[29] W. Jeck et al. Extending assembly of short DNA sequences to handle error. *Bioinformatics*, 23 No. 21:2942–2944, 2007.

[30] J. Kaptcianos. Graph theory aiding DNA fragment assembly. http://academics.smcvt.edu/jellis-monghan/Student%20Research/Kaptcianos.ppt. August 2009.

[31] S. Kim, H. Tang, and E. R. Mardis. Chapter 2 of genome sequencing technology and algorithms. *Artech House*, 2008.

[32] S. Kim, H. Tang, and E. R. Mardis. Chapter 8 of genome sequencing technology and algorithms. *Artech House*, 2008.

[33] A. Kusalik. Whole genome sequencing. BINF 200 notes, University of Saskatchewan. March 23, 2009.

[34] P. Pevzner, H. Tang, and M. Waterman. An Eulerian path approach to DNA fragment assembly. *PNAS*, 98 No. 17:9748–9753, 2001.

[35] P. A. Pevzner, H. Tang, and M. S. Waterman. An eulerian path approach to DNA fragment assembly. *Proc. Natl. Acad. Sci.*, 98:9748–9753, 2001.

[36] P.A. Pevzner, H. Tang, and G. Tesler. De novo repeat classification and fragment assembly. *Genome Research*, 14:1786–1796, 2004.

[37] P. M. Phillippy, A. L. Delcher, and S. L. Salzberg. Comparative genome assembly. *Briefings in bioinformatics*, 5 No. 3:237–248, 2004.

[38] Phrap and Cross_match Documentation. http://www.phrap.org/phredphrap/phrap.html. Accessed December 24, 2007.

[39] D. Platt and D. J. Evers. Forge: A parallel genome assembler combining Sanger and next generation sequence data. http://sourceforge.net/projects/forge. Accessed February 2010.

[40] J. C. Roach et al. Pairwise end sequencing: A unified approach to genomic mapping and sequencing. *Genomics*, 26 No. 2:345–353, 1995.

[41] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci.*, 74 No. 12:5463–5467, 1977.

[42] M. Schatz. AMOS assembly validation and visualization. Center for Bioinformatics and Computational Biology, University of Maryland. 2006.

[43] B. B. Schultz. Levene's Test for Relative Variation. *Systematic Zoology*, 34 No. 4:449–456, 1985.

[44] J. T. Simpson et al. ABYSS: A parallel assembler for short read sequence data. *Genome Research*, 19 No. 6:1117–1123, 2009.

[45] R. A. Smith. The effect of unequal group size on Tukey's HSD procedure. *Psychometrika*, 36 No. 1:31–34, 2006.

[46] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *J. Mol. Biol.*, 147:195–197, 1981.

[47] R Staden, K. F. Beal, and J. K. Bonfield. The staden package. *Methods in Molecular Biology*, 132:115–130, 1998.

[48] M. Stomberg. The marthlab: Mosaik. http://bioinformatics.bc.edu/marthlab/Mosaik. Accessed August 2008.

[49] G. G. Sutton et al. TIGR Assembler: A new tool for assembling large shotgun sequencing projects. *Genome science & technology*, 1 No. 1:9–19, 1995.

[50] K. P. Wall et al. Comparison of next generation sequencing technologies for transcriptome characterization. *BMC Genomics*, 10 No. 347, 2009.

[51] J. R. White et al. Figaro: a novel statistical method for vector sequence removal. *Bioinformatics*, 24 No. 4:462–467, 2008.

[52] D. R. Zerbino and E. Birney. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18:821–829, 2008.

APPENDIX A
TABLES OF DATA SETS

Table A.1: Characteristics of all the data sets used in this research.

Type of Data	Organism	Real/Artificial	# of reads
Illumina	AT	artificial	3,244,160
	CM	artificial	3,247,720
	EC	real	20,816,448
	HA	real	12,288,791
	MM	artificial	1,969,680
	SS	real	2,726,374
454	AT	real	848,629
	CM	real	465,960
	EC	real	436,142
	HA	artificial	453,250
	MM	real	346,848
	SS	artificial	585,550
Sanger	AT	real	38,533
	CM	real	26,401
	EC	artificial	76,230
	HA	artificial	25,545
	MM	real	22,289
	SS	real	30,560

Table A.2: Characteristics of the genome sequences for the data sets.

Organism	Genome size (bps)
AT	2,919,718
CM	2,922,917
EC	4,639,675
HA	1,553,927
MM	1,772,694
SS	2,007,491

APPENDIX B

TABLES OF RESULTS

The acronyms, abbreviations, and symbols used in the tables include: Org. - Organism, R/A - real or artificial, Q.F. - use of quality file, RTC - runs to completion, GC - Genome Coverage, MC -Misassembled Contigs, LC - Largest Contig, "-" - no result, and "**" - unreasonable results in the sense of either too many contigs or very small N50 size.

Table B.1: Genome coverage results for the Illumina data sets.

Org.	R/A	Assembler	Q.F.	RTC	GC(%)	# Contigs	MC(%)	N50	L.C.
AT	artificial	ABYSS	no	yes	95.06	2794	0	19918	107384
	artificial	EDENA	no	yes	93.16	772	0	20160	88392
	artificial	EULER-SR	no	yes	94.78	1630	0	24537	111057
	artificial	MIRA	yes	yes	92.51	1037	2.51	16925	107934
	artificial	PHRAP	no	yes	**	1339	**	36	36
	artificial	SSAKE	no	yes	96.11	2611	0.23	23916	113609
	artificial	VELVET	no	yes	94.88	1549	0	11479	54159
CM	artificial	ABYSS	no	yes	98.13	3778	0.08	12106	60771
	artificial	EDENA	no	yes	97.2	1325	0.08	11692	42181
	artificial	EULER-SR	no	yes	98.16	1860	0	19867	91992
	artificial	MIRA	yes	yes	98.33	311	2.25	34874	122652
	artificial	PHRAP	no	yes	**	791	**	36	36
	artificial	SSAKE	no	yes	97.75	1826	0.38	21237	74391
	artificial	VELVET	no	yes	98.23	1916	0	7238	35362
EC	real	ABYSS	no	yes	98.8	1734	0.17	17311	67095
	real	EDENA	no	yes	98.32	680	0.15	16430	67082
	real	EDENA	yes	yes	97.24	1153	0.09	8827	31785
	real	EULER-SR	no	yes	98.58	1276	0.31	24087	103365
	real	EULER-SR	yes	yes	97.45	971	2.99	18352	71243
	real	MIRA	yes	yes	87.83	6141	0.7	127612	349300
	real	PHRAP	no	no	-	-	-	-	-
	real	SSAKE	no	yes	**	109125	**	72	16468
	real	VELVET	no	yes	98.02	666	0	48723	126542
	real	VELVET	yes	yes	98.02	666	0	48723	126542
HA	real	ABYSS	no	yes	98.29	858	4.2	6774	36634
	real	EDENA	no	yes	97.98	453	0.22	7746	41362
	real	EDENA	yes	yes	94.59	1849	0	1275	5771
	real	EULER-SR	no	yes	98.98	467	1.5	14405	62017
	real	EULER-SR	yes	yes	97.13	486	0	12865	31069
	real	MIRA	yes	no	-	-	-	-	-
	real	PHRAP	no	yes	98.98	41534	0.62	75	1622
	real	SSAKE	no	yes	**	237367	**	48	11645
	real	VELVET	no	yes	98.01	474	1.48	9244	28127
	real	VELVET	yes	yes	98.01	474	1.48	9244	28127
MM	artificial	ABYSS	no	yes	98.66	908	0	23396	99232
	artificial	EDENA	no	yes	98.16	574	0	13925	72821
	artificial	EULER-SR	no	yes	98.71	563	0	29882	175259
	artificial	MIRA	yes	yes	99.89	69	6.45	410376	513483
	artificial	PHRAP	no	yes	**	370	**	36	36
	artificial	SSAKE	no	yes	98.53	588	0.17	23570	79160
	artificial	VELVET	no	yes	98.68	724	0	9320	35105
SS	real	ABYSS	no	yes	98.84	2718	11.77	3554	17113

Continued on next page

TableB.1 – continued from previous page

Org.	R/A	Assembler	Q.F.	RTC	GC(%)	# Contigs	MC(%)	N50	L.C.
SS	real	EDENA	no	yes	95.82	1278	0	2606	13883
	real	EDENA	yes	yes	94.06	1668	0.06	1965	8829
	real	EULER-SR	no	yes	95.34	2281	0.044	1709	10963
	real	EULER-SR	yes	yes	82.72	4887	0.23	492	5388
	real	MIRA	yes	yes	2.73	6243	0	36	2012
	real	PHRAP	no	yes	97.56	88131	0.02	48	1104
	real	SSAKE	no	yes	98.21	15632	0.16	768	6475
	real	VELVET	no	yes	98.55	1160	1.12	4594	17831
	real	VELVET	yes	yes	97.36	2758	0.07	1262	8005

Table B.2: Execution time and Amount of memory used on results from assembling the Illumina data sets.

Org.	R/A	Assembler	Q.F.	RTC	Time(sec.)	Memory
AT	artificial	ABYSS	no	yes	289.31	236.06 Mb
	artificial	EDENA	no	yes	393.95	521.51 Mb
	artificial	EULER-SR	no	yes	706.48	315.26 Mb
	artificial	MIRA	yes	yes	210486.32	6.01 Gb
	artificial	PHRAP	no	yes	489.52	2.41 Gb
	artificial	SSAKE	no	yes	1975.97	2.11 Gb
	artificial	VELVET	no	yes	112.63	569.84 Mb
CM	artificial	ABYSS	no	yes	287.78	280.83 Mb
	artificial	EDENA	no	yes	419.59	526.45 Mb
	artificial	EULER-SR	no	yes	821.68	616.87 Mb
	artificial	MIRA	yes	yes	69753.9	5.52 Gb
	artificial	PHRAP	no	yes	361.38	2.30 Gb
	artificial	SSAKE	no	yes	1891.41	2.23 Gb
	artificial	VELVET	no	yes	112.73	568.13 Mb
EC	real	ABYSS	no	yes	955.81	671.62 Mb
	real	EDENA	no	yes	2721.5	1.63 Gb
	real	EDENA	yes	yes	2089.87	1.73 Gb
	real	EULER-SR	no	yes	9742.2	156.71 Mb
	real	EULER-SR	yes	yes	3610.76	2.93 Gb
	real	MIRA	yes	yes	219097.44	35.22 Gb
	real	PHRAP	no	no	-	-
	real	VELVET	no	yes	776.83	2.68 Gb
	real	VELVET	yes	yes	973.2	2.66 Gb
HA	real	ABYSS	no	yes	1130.68	1.23 Gb
	real	EDENA	no	yes	2370.9	1.45 Gb
	real	EDENA	yes	yes	1975.29	1.68 Gb
	real	EULER-SR	no	yes	781.8	63.82 Mb
	real	EULER-SR	yes	yes	1726.46	1.51 Gb
	real	MIRA	yes	no	-	-
	real	PHRAP	no	yes	68180	31.72 Gb
	real	SSAKE	no	yes	28220.2	6.09 Gb
	real	VELVET	no	yes	1159.91	2.2 Gb
	real	VELVET	yes	yes	415.26	1.12 Gb
MM	artificial	ABYSS	no	yes	160.47	160.47 Mb
	artificial	EDENA	no	yes	225.17	321 Mb
	artificial	EULER-SR	no	yes	432.06	380.39 Mb
	artificial	MIRA	yes	yes	10073.43	3.46 Gb
	artificial	PHRAP	no	yes	271.19	1.42 Gb
	artificial	SSAKE	no	yes	1180.43	1.39 Gb
	artificial	VELVET	no	yes	67.44	372.76 Mb
SS	real	ABYSS	no	yes	301.73	314.93 Mb
	real	EDENA	no	yes	604.13	446.81 Mb
	real	EDENA	yes	yes	482.08	501.92 Mb
	real	EULER-SR	no	yes	152.4	74.26 Mb
	real	EULER-SR	yes	yes	388.94	397.9 Mb
	real	MIRA	yes	yes	1046.07	3.37 Gb
	real	PHRAP	no	yes	57246.09	22.51 Gb
	real	SSAKE	no	yes	3785.24	2.33 Gb
	real	VELVET	no	yes	120.65	579.24 Mb
	real	VELVET	yes	yes	90.8	521.54 Mb

Table B.3: Genome coverage results for the 454 data sets.

Org.	R/A	Assembler	Q.F.	RTC	GC (%)	# Contigs	MC (%)	N50	L.C.
AT	real	EULER-SR	no	yes	93.96	808	0.99	2196	14509
	real	EULER-SR	yes	yes	93.34	8139	10.01	3030	23262
	real	MIRA	no	yes	89.57	193	12.44	22381	104206
	real	PHRAP	no	yes	97.11	1232	12.74	9449	54130
	real	SSAKE	no	yes	95.92	40830	1.02	295	2995
	real	VELVET	no	yes	93.86	32814	0.73	119	1292
	real	VELVET	yes	yes	93.86	32814	0.73	119	1292
CM	real	EULER-SR	no	yes	98.37	1257	0.16	16659	78363
	real	EULER-SR	yes	yes	97.84	3385	0.68	13677	76445
	real	MIRA	no	yes	99.98	337	10.39	90096	209592
	real	PHRAP	no	yes	92.27	116	28.45	133845	407769
	real	SSAKE	no	yes	97.83	32166	0.36	485	3143
	real	VELVET	no	yes	98.89	13797	0.38	493	4050
	real	VELVET	yes	yes	98.89	13797	0.38	493	4050
EC	real	EULER-SR	no	yes	97.79	885	1.02	57777	165903
	real	EULER-SR	yes	yes	97.13	802	3.49	58858	165818
	real	MIRA	no	yes	99.41	358	13.69	129463	349282
	real	PHRAP	no	yes	98.71	159	31.45	126839	529849
	real	SSAKE	no	yes	98.35	2398	0.25	7075	27137
	real	VELVET	no	yes	98.03	3778	1.88	3441	16401
	real	VELVET	yes	yes	98.03	3778	1.88	3441	16401
HA	artificial	EULER-SR	no	yes	98.04	251	0	46591	89048
	artificial	MIRA	no	yes	86.91	6435	0.2	257	5239
	artificial	PHRAP	no	yes	96.67	91	15.39	27204	80696
	artificial	SSAKE	no	yes	98.57	7525	0.27	522	7498
	artificial	VELVET	no	yes	98.16	12241	0	301	2130
MM	real	EULER-SR	no	yes	98.69	618	0.16	7222	22488
	real	EULER-SR	yes	yes	**	817	0	134	551
	real	MIRA	no	yes	99.53	1042	2.21	24470	64854
	real	PHRAP	no	yes	90.56	265	23.02	12319	58567
	real	SSAKE	no	yes	97.35	9362	0.14	368	3594
	real	VELVET	no	yes	98.68	14814	0.84	210	1996
	real	VELVET	yes	yes	98.68	14814	0.84	210	1996
SS	artificial	EULER-SR	no	yes	97.57	537	1.68	27613	66367
	artificial	MIRA	no	yes	62.27	8464	0.32	160	1981
	artificial	PHRAP	no	yes	93.67	181	6.08	18076	57886
	artificial	SSAKE	no	yes	98.65	9956	0.07	476	8855
	artificial	VELVET	no	yes	98.32	16719	0.08	281	2449

Table B.4: Execution time and Amount of memory used on results from assembling the 454 data sets.

Org.	R/A	Assembler	Q.F.	RTC	Time(sec.)	Memory
AT	real	EULER-SR	no	yes	369	115.4 Mb
	real	EULER-SR	yes	yes	1213.43	487.88 Mb
	real	MIRA	no	yes	80052.16	8.98 Gb
	real	PHRAP	no	yes	22783.46	13.45 Gb
	real	SSAKE	no	yes	14804.79	1.11 Gb
	real	VELVET	no	yes	262.54	667.84 Mb
	real	VELVET	yes	yes	229.74	745.04 Mb
CM	real	EULER-SR	no	yes	235.2	113.04 Mb
	real	EULER-SR	yes	yes	635.97	394.82 Mb
	real	MIRA	no	yes	12006.75	3.18 Gb
	real	PHRAP	no	yes	9045.55	4.27 Gb
	real	SSAKE	no	yes	6684.61	867.99 Mb
	real	VELVET	no	yes	201.69	472.51 Mb
	real	VELVET	yes	yes	173.47	747 Mb
EC	real	EULER-SR	no	yes	1956.23	733.76 Mb
	real	EULER-SR	yes	yes	538.54	450.37 Mb
	real	MIRA	no	yes	7452.51	2.86 Gb
	real	PHRAP	no	yes	2656.11	2.39 Gb
	real	SSAKE	no	yes	2556.04	1.54 Gb
	real	VELVET	no	yes	146.66	381.67 Mb
	real	VELVET	yes	yes	138.92	386.34 Mb
HA	artificial	EULER-SR	no	yes	603.91	266.88 Mb
	real	MIRA	no	yes	7807.18	1.38 Gb
	artificial	PHRAP	no	yes	2207.08	2.49 Gb
	artificial	SSAKE	no	yes	3260.32	597.28 Mb
	artificial	VELVET	no	yes	112.73	509.09 Mb
MM	real	EULER-SR	no	yes	105	82.47 Mb
	real	EULER-SR	yes	yes	18.48	36.33 Mb
	real	MIRA	no	yes	4549.23	1.48 Gb
	real	PHRAP	no	yes	925.63	1.25 Gb
	real	SSAKE	no	yes	3885.29	554.54 Mb
	real	VELVET	no	yes	69.82	326.54 Mb
	real	VELVET	yes	yes	67.5	383.4 Mb
SS	artificial	EULER-SR	no	yes	765.85	333.52 Mb
	real	MIRA	no	yes	7414.32	1.49 Gb
	artificial	PHRAP	no	yes	1983.65	3.32 Gb
	artificial	SSAKE	no	yes	4752	741.26 Mb
	artificial	VELVET	no	yes	146.01	549.67 Mb

Table B.5: Genome coverage results for the Sanger data sets.

Org.	R/A	Assembler	Q.F.	RTC	GC (%)	# Contigs	MC (%)	N50	L.C.
AT	real	CAP3	yes	yes	92.42	472	6.14	17409	88604
	real	CAP3	no	yes	92.09	470	3.83	17707	64144
	real	EULER-SR	no	yes	76.19	833	0.36	7819	38063
	real	MIRA	yes	yes	92.51	993	4.43	16908	107897
	real	PCAP	yes	yes	92.62	701	3.85	17331	76179
	real	PHRAP	yes	yes	91.77	564	0.53	19956	90885
	real	PHRAP	no	yes	91.85	458	2.4	22309	90789
	real	SSAKE	no	yes	68.96	7598	0.21	887	11099
	real	TIGR	yes	yes	94.44	3339	4.07	5148	41969
	real	TIGR	no	yes	94.64	4874	2.32	2420	40121
	real	VELVET	no	yes	90.82	10931	0.66	583	4749
CM	real	CAP3	yes	yes	99.01	210	4.76	30047	85321
	real	CAP3	no	yes	98.78	216	3.7	28319	73441
	real	EULER-SR	no	yes	81.46	729	0.14	4539	26025
	real	MIRA	yes	yes	97.88	320	3.75	34847	122630
	real	PCAP	yes	yes	98.85	232	2.16	23733	108537
	real	PHRAP	yes	yes	98.57	100	0	46373	122762
	real	PHRAP	no	yes	98.25	128	0	46389	122874
	real	SSAKE	no	yes	60.96	12778	0.15	172	3464
	real	TIGR	yes	yes	99.11	3291	2.1	11050	59795
	real	TIGR	no	yes	99.27	6375	1.28	1986	31204
	real	VELVET	no	yes	96.75	19502	0.17	290	3413
EC	artificial	CAP3	yes	yes	99.99	17233	0.69	2791	18986
	artificial	EULER-SR	no	no	-	-	-	-	-
	artificial	MIRA	no	no	-	-	-	-	-
	artificial	PCAP	yes	yes	98.36	11314	0.53	1518	4217
	artificial	PHRAP	yes	yes	96.68	463	18.14	43381	150428
	artificial	SSAKE	no	no	-	-	-	-	-
	artificial	TIGR	no	yes	99.97	75961	0.14	920	2562
	artificial	VELVET	no	yes	**	371894	**	130	833
HA	artificial	CAP3	yes	yes	99.99	5698	0.65	2910	13628
	artificial	EULER-SR	no	no	-	-	-	-	-
	artificial	MIRA	yes	yes	99.89	87	12.64	848106	848106
	artificial	PCAP	yes	yes	93.71	2523	2.26	1639	4236
	artificial	PHRAP	yes	yes	98.34	97	11.34	86858	221754
	artificial	SSAKE	no	no	-	-	-	-	-
	artificial	TIGR	yes	yes	99.86	25471	0.09	919	2135
	artificial	VELVET	no	yes	**	122453	**	129	734
MM	real	CAP3	yes	yes	99.63	15	20	364018	561631
	real	CAP3	no	yes	99.72	12	8.33	405649	595046
	real	EULER-SR	no	yes	97.57	194	0	22022	96652
	real	MIRA	yes	yes	99.88	72	6.94	410341	513463
	real	PCAP	yes	yes	99.81	18	22.22	308649	405653
	real	PHRAP	yes	yes	99.69	21	4.76	576887	604106
	real	PHRAP	no	yes	99.46	13	7.69	570536	576889
	real	SSAKE	no	yes	93.43	4432	0.09	1390	7106
	real	TIGR	yes	yes	99.8	623	7.06	71954	314538
	real	TIGR	no	yes	99.82	1308	3.29	14368	57120
	real	VELVET	no	yes	98.56	3701	0.19	1521	8533
SS	real	CAP3	no	no	-	-	-	-	-
	real	EULER-SR	no	yes	71.1	479	1.88	6385	31844
	real	MIRA	no	no	-	-	-	-	-
	artificial	PCAP	yes	yes	93.88	3187	0.38	1641	4216
	real	PHRAP	no	yes	87.42	484	8.26	9442	36205
	real	SSAKE	no	yes	60.64	5400	0.2	824	13503
	real	TIGR	no	yes	89.79	3920	5.79	1949	36204
	real	VELVET	no	yes	85.42	9241	0.23	513	4732

Table B.6: Execution time and Amount of memory used on results from assembling the Sanger data sets.

Org.	R/A	Assembler	Q.F.	RTC	Time(sec.)	Memory
AT	real	CAP3	yes	yes	3360.13	784.05 Mb
	real	CAP3	no	yes	3135.11	803.31 Mb
	real	EULER-SR	no	yes	55.2	140.83 Mb
	real	MIRA	yes	yes	1579.77	480.41 Mb
	real	PCAP	yes	yes	1287.25	341.47 Mb
	real	PHRAP	yes	yes	1168.21	503.53 Mb
	real	PHRAP	no	yes	1106.92	484.73 Mb
	real	SSAKE	no	yes	1235.69	584.75 Mb
	real	TIGR	yes	yes	375.25	1.53 Gb
	real	TIGR	no	yes	3882.79	1.50 Gb
	real	VELVET	no	yes	47.94	337.57 Mb
CM	real	CAP3	yes	yes	1768.87	575.12 Mb
	real	CAP3	no	yes	1436.98	574.82 Mb
	real	EULER-SR	no	yes	30.6	116.89 Mb
	real	MIRA	yes	yes	670.57	384.54 Mb
	real	PCAP	yes	yes	989.53	339.02 Mb
	real	PHRAP	yes	yes	202.89	325.09 Mb
	real	PHRAP	no	yes	212.9	318.35 Mb
	real	SSAKE	no	yes	2362.17	559.76 Mb
	real	TIGR	yes	yes	1059.16	797.84 Mb
	real	TIGR	no	yes	1294.08	922.09 Mb
	real	VELVET	no	yes	37.58	410.61 Mb
EC	artificial	CAP3	yes	yes	12484.93	2.18 Gb
	artificial	EULER-SR	no	no	-	-
	artificial	MIRA	no	no	-	-
	artificial	PCAP	no	no	1629.51	576.37 Mb
	artificial	PHRAP	yes	yes	1966.23	1.98 Gb
	artificial	SSAKE	no	no	-	-
	artificial	TIGR	no	yes	32641.52	3.57 Gb
	artificial	VELVET	no	yes	399.66	2.36 Gb
HA	artificial	CAP3	yes	yes	3869.75	742.44 Mb
	artificial	EULER-SR	no	no	-	-
	artificial	MIRA	yes	yes	4068.7	911.97 Mb
	artificial	PCAP	yes	yes	494.23	340.12 Mb
	artificial	PHRAP	yes	yes	470.11	13.25 Gb
	artificial	SSAKE	no	no	-	-
	artificial	TIGR	yes	yes	4771.64	1.25 Gb
	artificial	VELVET	no	yes	133.47	666.37 Mb
MM	real	CAP3	yes	yes	1411.26	521.46 Mb
	real	CAP3	no	yes	1406.5	522.14Mb
	real	EULER-SR	no	yes	15	103.89 Mb
	real	MIRA	yes	yes	595.51	378.15 Mb
	real	PCAP	yes	yes	1328.58	339.17 Mb
	real	PHRAP	yes	yes	181.81	391.78 Mb
	real	PHRAP	no	yes	174.13	384.67 Mb
	real	SSAKE	no	yes	644.51	456.98 Mb
	real	TIGR	yes	yes	608.81	739.72 Mb
	real	TIGR	no	yes	717.27	863.41 Mb
	real	VELVET	no	yes	28.17	246.74 Mb
SS	real	CAP3	no	no	-	-
	real	EULER-SR	no	yes	39	66.08 Mb
	real	MIRA	no	no	-	-
	artificial	PCAP	yes	yes	656.3	341.07 Mb
	real	PHRAP	no	yes	172.29	322.14 Mb
	real	SSAKE	no	yes	859.57	512.87 Mb
	real	TIGR	no	yes	1698.53	771.63 Mb
	real	VELVET	no	yes	31.67	198 Mb

Table B.7: Genome coverage results from assembling 454 reads merged with Sanger reads.

Org.	R/A	Assembler	Q.F.	RTC	GC(%)	# Contigs	MC(%)	N50	L.C.
AT	real	EULER-SR	no	yes	93.83	2289	1.87	8682	51526
	real	VELVET	no	yes	68.19	30449	1.84	85	375
CM	real	EULER-SR	no	yes	98.66	1659	0.24	23648	73702
	real	VELVET	no	yes	90.2	32398	5.03	249	4749
EC	artificial	EULER-SR	no	yes	98.04	788	2.79	58660	165819
	artificial	VELVET	no	yes	**	328617	**	133	833
HA	artificial	EULER-SR	no	yes	98.03	177	0	47893	110407
	artificial	VELVET	no	yes	**	115773	**	124	635
MM	real	EULER-SR	no	yes	99.21	278	1.08	39976	172919
	real	VELVET	no	yes	85.43	27633	2.38	245	4749
SS	real	EULER-SR	no	yes	97.02	490	2.96	32905	120800
	real	VELVET	no	yes	**	151594	**	124	635

Table B.8: Execution time and Amount of memory used on results from assembling 454 reads merged with Sanger reads.

Org.	R/A	Assembler	Q.F.	RTC	Time(sec.)	Memory
AT	real	EULER-SR	no	yes	493.2	176.52 Mb
	real	VELVET	no	yes	300.98	915.33 Mb
CM	real	EULER-SR	no	yes	274.8	131.75 Mb
	real	VELVET	no	yes	208.46	698.47 Mb
EC	artificial	EULER-SR	no	yes	719.44	162.24 Mb
	artificial	VELVET	no	yes	570.09	843.47 Mb
HA	artificial	EULER-SR	no	yes	551.3	96.74 Mb
	artificial	VELVET	no	yes	268.26	736.74 Mb
MM	real	EULER-SR	no	yes	129	118.76 Mb
	real	VELVET	no	yes	92.89	443.49 Mb
SS	real	EULER-SR	no	yes	201	86.64 Mb
	real	VELVET	no	yes	180.12	857.15Mb

Table B.9: Genome coverage results from assembling Illumina reads merged with Sanger reads.

Org.	R/A	Assembler	Q.F.	RTC	GC(%)	# Contigs	MC(%)	N50	L.C.
AT	real	EULER-SR	no	yes	95.05	1241	0.81	47133	173182
	real	VELVET	no	yes	94.9	9400	1.04	1136	12107
CM	real	EULER-SR	no	yes	98.42	1659	0.66	36562	154600
	real	VELVET	no	yes	98.24	19061	0.17	410	4579
EC	artificial	EULER-SR	no	yes	98.35	17840	0.45	1378	11859
	artificial	VELVET	no	yes	**	351386	**	109	602
HA	artificial	EULER-SR	no	yes	97.9	30448	0	234	3362
	artificial	VELVET	no	yes	**	117367	**	107	1359
MM	real	EULER-SR	no	yes	98.89	423	2.13	116741	292143
	real	VELVET	no	yes	98.62	4018	0.35	2031	16109
SS	real	EULER-SR	no	no	-	-	-	-	-
	real	VELVET	no	yes	61.66	19591	3.51	314	4749

Table B.10: Execution time and Amount of memory used on results from assembling Illumina reads merged with Sanger reads.

Org.	R/A	Assembler	Q.F.	RTC	Time(sec.)	Memory
AT	real	EULER-SR	no	yes	969.87	729.39 Mb
	real	VELVET	no	yes	189.08	675.39 Mb
CM	real	EULER-SR	no	yes	1061.96	682.72 Mb
	real	VELVET	no	yes	177.87	645.5 Mb
EC	artificial	EULER-SR	no	yes	10347.79	3.56 Gb
	artificial	VELVET	no	yes	1280.37	3.27 Gb
HA	artificial	EULER-SR	no	yes	5415.64	1.86 Gb
	artificial	VELVET	no	yes	565.56	1.29 Gb
MM	real	EULER-SR	no	yes	515.35	459.59 Mb
	real	VELVET	no	yes	94.84	427.2 Mb
SS	real	EULER-SR	no	yes	-	-
	real	VELVET	no	yes	548.4	973.1 Mb

Table B.11: Genome coverage results from assembling Illumina reads merged with 454 reads.

Org.	R/A	Assembler	Q.F.	RTC	GC(%)	# Contigs	MC(%)	N50	L.C.
AT	artificial/real	EULER-SR	no	yes	93.25	5824	5.24	6817	42900
	artificial/real	VELVET	no	yes	90.87	37851	0.75	102	1043
CM	artificial/real	EULER-SR	no	yes	97.89	1855	1.02	31903	143473
	artificial/real	VELVET	no	yes	97.28	15826	1.09	430	3079
EC	artificial	EULER-SR	no	yes	96.2	605	19.67	60455	294663
	artificial	VELVET	no	yes	97.91	4512	1.31	2849	18945
HA	real/artificial	EULER-SR	no	yes	91.68	4679	0	1382	8308
	real/artificial	VELVET	no	yes	98.11	15386	0.07	229	1981
MM	artificial/real	EULER-SR	no	yes	99.06	761	0.53	27296	123840
	artificial/real	VELVET	no	yes	97.54	19220	0.06	151	1464
SS	real	EULER-SR	no	yes	97.04	1465	0.96	12753	65025
	real	VELVET	no	yes	98.21	19091	0.02	244	2463

Table B.12: Execution time and Amount of memory used on results from assembling Illumina reads merged with 454 reads.

Org.	R/A	Assembler	Q.F.	RTC	Time(sec.)	Memory
AT	real	EULER-SR	no	yes	2029.79	1.03 Gb
	real	VELVET	no	yes	573.96	2.40 Gb
CM	real	EULER-SR	no	yes	1505.13	938.08 Mb
	real	VELVET	no	yes	397.62	1.21 Gb
EC	artificial	EULER-SR	no	yes	5609.01	3.71 Gb
	artificial	VELVET	no	yes	1048.25	3.30 Gb
HA	artificial	EULER-SR	no	yes	4642.01	1.99 Gb
	artificial	VELVET	no	yes	647.15	2.16 Gb
MM	real	EULER-SR	no	yes	675.84	525.79 Mb
	real	VELVET	no	yes	168.91	625.77 Mb
SS	real	EULER-SR	no	yes	1567.31	718.81 Mb
	real	VELVET	no	yes	292.57	1.01 Gb

Table B.13: Genome coverage results from assembling Illumina contigs merged with 454 contigs.

Org.	R/A	Assembler	Q.F.	RTC	GC(%)	# Contigs	MC(%)	N50	L.C.
AT	artificial/real	EULER-SR	no	yes	96.01	410	0.98	47193	127224
	artificial/real	EULER-SR/PHRAP	no	yes	97.13	182	14.84	55022	146121
CM	artificial/real	EULER-SR	no	yes	98.38	254	1.18	55178	200558
	artificial/real	EULER-SR/PHRAP	no	yes	99.01	48	12.5	247589	627649
EC	real	EULER-SR	no	yes	98.08	247	0.4	80429	180280
	real	EULER-SR/PHRAP	no	yes	99.22	61	16.39	175813	529850
	real	VELVET/PHRAP	no	yes	99.21	53	16.98	178577	529849
HA	real/artificial	EULER-SR	no	yes	97.80	85	1.18	47892	89053
	real/artificial	EULER-SR/PHRAP	no	yes	97.42	74	14.87	45634	86112
MM	artificial/real	EULER-SR	no	yes	99.22	94	1.06	86217	292800
	artificial/real	EULER-SR/PHRAP	no	yes	97.88	94	20.21	58567	175354
SS	real/artificial	EULER-SR	no	yes	97.23	159	6.92	28024	66369
	real/artificial	EULER-SR/PHRAP	no	yes	95.09	159	6.29	22953	58127

Table B.14: Execution time and Amount of memory used on results from assembling Illumina contigs merged with 454 contigs.

Org.	R/A	Assembler	Q.F.	RTC	Time(sec.)	Memory
AT	artificial/real	EULER-SR	no	yes	81.52	139.46 Mb
	artificial/real	EULER-SR/PHRAP	no	yes	160.4	162.6 Mb
CM	artificial/real	EULER-SR	no	yes	235.83	9154.1 Mb
	artificial/real	EULER-SR/PHRAP	no	yes	680.66	165.97 Mb
EC	real	EULER-SR	no	yes	816.07	192.48 Mb
	real	EULER-SR/PHRAP	no	yes	1558.63	199.12 Mb
	real	VELVET/PHRAP	no	yes	2047.23	197.21 Mb
HA	real/artificial	EULER-SR	no	yes	127.36	127.63 Mb
	real/artificial	EULER-SR/PHRAP	no	yes	86.54	117.17 Mb
MM	artificial/real	EULER-SR	no	yes	70.76	116.37 Mb
	artificial/real	EULER-SR/PHRAP	no	yes	78.33	122.73 Mb
SS	real/artificial	EULER-SR	no	yes	56.78	110.64 Mb
	real/artificial	EULER-SR/PHRAP	no	yes	33.22	95.81 Mb

Table B.15: Genome coverage results from assembling merged 454, Illumina, and Sanger reads.

Org.	R/A	Assembler	Q.F.	RTC	GC(%)	# Contigs	MC(%)	N50	L.C.
AT	real/art./real	EULER-SR	no	yes	94.45	5240	5.59	10903	114287
	real/art./real	VELVET	no	yes	94.59	37762	1.21	107	909
CM	real/art./real	EULER-SR	no	yes	96.73	2037	0.64	30514	121176
	real/art./real	VELVET	no	yes	99.07	27782	0.14	180	2648
EC	real/real/art.	EULER-SR	no	yes	98.78	32288	0.55	884	9006
	real/real/art.	VELVET	no	yes	**	330851	**	132	833
HA	art./real/art.	EULER-SR	no	yes	98.96	45342	0.57	123	2962
	art./real/art.	VELVET	no	yes	**	117612	**	123	1450
MM	real/art./real	EULER-SR	no	yes	99.25	633	0.16	56345	275961
	real/art./real	VELVET	no	yes	99.78	16986	0.21	179	1567
SS	art./real/real	EULER-SR	no	yes	79.6	1299	1.85	7055	32350
	art./real/real	VELVET	no	yes	99.1	22902	0.32	155	2449

Table B.16: Execution time and Amount of memory used on results from assembling merged 454, Illumina, and Sanger reads.

Org.	R/A	Assembler	Q.F.	RTC	Time(sec.)	Memory
AT	real/art./real	EULER-SR	no	yes	2747.13	1.15 Gb
	real/art./real	VELVET	no	yes	402.89	1.07 Gb
CM	real/art./real	EULER-SR	no	yes	1860.28	1.00 Gb
	real/art./real	VELVET	no	yes	295.71	913.11 Mb
EC	real/real/art.	EULER-SR	no	yes	12300.29	3.9 Gb
	real/real/art.	VELVET	no	yes	1174.7	3.00 Gb
HA	art./real/art.	EULER-SR	no	yes	6546.06	2.07 Gb
	art./real/art.	VELVET	no	yes	584.56	1.83 Gb
MM	real/art./real	EULER-SR	no	yes	878.28	606.5 Mb
	real/art./real	VELVET	no	yes	139.59	501.59 Mb
SS	art./real/real	EULER-SR	no	yes	1188.75	773.16 Mb
	art./real/real	VELVET	no	yes	260.65	927.54 Mb

Table B.17: Genome coverage results from assembling 454 contigs merged with Sanger reads.

Org.	R/A	Assembler	Q.F.	RTC	GC (%)	# Contigs	MC (%)	N50	L.C.
AT	real	EULER-SR/PHRAP	no	yes	98.19	370	5.95	42555	185881
	real	MIRA/PHRAP	no	yes	99.76	309	17.48	70986	164364
	real	PHRAP/PHRAP	no	yes	98.07	397	22.92	31865	117484
	real	SSAKE/PHRAP	no	yes	98.86	410	12.2	33635	92421
	real	VELVET/PHRAP	no	yes	95.87	697	4.73	21079	92485
	real	EULER-SR/CAP3	no	yes	96.59	276	9.06	34636	163249
	real	MIRA/CAP3	no	yes	95.95	291	9.62	68895	278945
	real	PHRAP/CAP3	no	yes	97.56	281	21.71	42082	125419
	real	SSAKE/CAP3	no	yes	98.24	341	7.92	28777	119629
	real	VELVET/CAP3	no	yes	93.27	590	6.1	18435	64146
	real	EULER-SR/PCAP	yes	yes	97.59	535	3.93	28971	169285
	real	PHRAP/PCAP	yes	no	-	-	-	-	-
	real	SSAKE/PCAP	yes	yes	98.11	637	3.3	22478	137471
	real	VELVET/PCAP	yes	yes	94.47	855	2.69	18389	76417
CM	real	EULER-SR/PHRAP	no	yes	99.98	54	5.56	167145	442702
	real	MIRA/PHRAP	no	yes	99.41	61	24.59	181809	344716
	real	PHRAP/PHRAP	no	yes	99.99	108	12.04	113671	407691
	real	SSAKE/PHRAP	no	yes	98.6	53	30.19	263311	383498
	real	VELVET/PHRAP	no	yes	99.77	81	23.46	130353	365691
	real	EULER-SR/CAP3	no	yes	99.45	107	7.48	80593	193843
	real	MIRA/CAP3	no	yes	-	-	-	-	-
	real	PHRAP/CAP3	no	no	-	-	-	-	-
	real	SSAKE/CAP3	no	yes	99.95	43	16.28	159106	504162
	real	VELVET/CAP3	no	yes	99.68	127	8.66	69401	153779
	real	EULER-SR/PCAP	yes	no	-	-	-	-	-
	real	PHRAP/PCAP	yes	no	-	-	-	-	-
	real	SSAKE/PCAP	yes	yes	99.75	59	6.78	150914	383323
	real	VELVET/PCAP	yes	yes	99.68	152	2.63	58742	160497
EC	artificial	EULER-SR/PHRAP	no	yes	99.56	67	16.42	501610	841056
	artificial	MIRA/PHRAP	no	yes	99.76	58	29.31	384937	1113703
	artificial	PHRAP/PHRAP	no	yes	99.77	42	33.33	401641	805407
	artificial	SSAKE/PHRAP	no	yes	99.7	29	41.38	501331	751191
	artificial	VELVET/PHRAP	no	yes	99.62	21	23.81	554086	632735
	artificial	EULER-SR/CAP3	no	no	-	-	-	-	-
	artificial	MIRA/CAP3	no	no	-	-	-	-	-
	artificial	PHRAP/CAP3	no	no	-	-	-	-	-
	artificial	SSAKE/CAP3	no	yes	97.79	371	4.04	28724	156733
	artificial	VELVET/CAP3	no	yes	99.64	672	3.72	15529	69354
	artificial	EULER-SR/PCAP	yes	no	-	-	-	-	-
	artificial	PHRAP/PCAP	yes	no	-	-	-	-	-
	artificial	SSAKE/PCAP	yes	yes	99.3	2416	2.77	5179	27610
	artificial	VELVET/PCAP	yes	yes	99.34	4575	0.72	1966	17078
HA	artificial	EULER-SR/PHRAP	no	yes	99.89	13	30.77	215738	436524
	artificial	MIRA/PHRAP	no	yes	99.89	18	22.22	336254	524300
	artificial	PHRAP/PHRAP	no	yes	98.72	41	29.27	90832	157297
	artificial	SSAKE/PHRAP	no	yes	99.83	10	30	215920	333811
	artificial	VELVET/PHRAP	no	yes	99.71	9	33.33	422983	534508
	artificial	EULER-SR/CAP3	no	yes	99.78	426	1.41	6767	64784
	artificial	MIRA/CAP3	no	yes	99.27	623	0.32	5359	31771
	artificial	PHRAP/CAP3	no	yes	99.72	374	1.87	8859	49821
	artificial	SSAKE/CAP3	no	yes	99.7	372	2.15	8751	40261
	artificial	VELVET/CAP3	no	yes	99.95	609	0.82	5472	31771
	artificial	EULER-SR/PCAP	yes	no	-	-	-	-	-
	artificial	PHRAP/PCAP	yes	no	-	-	-	-	-
	artificial	SSAKE/PCAP	yes	yes	99.54	2766	1.41	1576	7770
	artificial	VELVET/PCAP	yes	yes	99.9	6787	0.72	1289	3873
MM	real	EULER-SR/PHRAP	no	yes	99.7	14	14.29	1133397	1133397
	real	MIRA/PHRAP	no	yes	99.71	28	25	375950	644592
	real	PHRAP/PHRAP	no	yes	99.77	78	52.56	163777	346516
	real	SSAKE/PHRAP	no	yes	99.68	11	18.18	1158304	1158304
	real	VELVET/PHRAP	no	yes	99.77	16	12.5	950556	950556
	real	EULER-SR/CAP3	no	yes	99.66	6	33.33	602217	742116

Continued on next page

Table B.17 – continued from previous page

Org.	R/A	Assembler	Q.F.	RTC	GC (%)	# Contigs	MC (%)	N50	L.C.
MM	real	MIRA/CAP3	no	yes	99.74	16	31.25	433493	795623
	real	PHRAP/CAP3	no	no	-	-	-	-	-
	real	SSAKE/CAP3	no	yes	99.86	9	44.44	423962	562435
	real	VELVET/CAP3	no	yes	99.68	25	8	426329	686989
	real	EULER-SR/PCAP	yes	yes	99.86	14	28.43	1001316	1001316
	real	PHRAP/PCAP	yes	yes	99.86	29	34.48	178541	590696
	real	SSAKE/PCAP	yes	yes	99.86	16	25	923003	923003
	real	VELVET/PCAP	yes	yes	99.84	47	10.64	420105	562465
SS	artificial	EULER-SR/PHRAP	no	yes	99.32	34	35.29	163970	556464
	artificial	MIRA/PHRAP	no	yes	91.6	563	8.35	9450	36357
	artificial	PHRAP/PHRAP	no	yes	95.22	81	19.75	92355	269282
	artificial	SSAKE/PHRAP	no	yes	99.56	11	18.18	1025456	1025456
	artificial	VELVET/PHRAP	no	yes	99.36	8	37.5	554712	640581
	artificial	EULER-SR/CAP3	no	yes	99.12	420	1.19	10664	56838
	artificial	MIRA/CAP3	no	no	-	-	-	-	-
	artificial	PHRAP/CAP3	no	yes	98.89	354	3.39	17207	62350
	artificial	SSAKE/CAP3	no	yes	99.63	545	1.65	8301	33074
	artificial	VELVET/CAP3	no	yes	99.68	803	0.87	5561	19354
	artificial	EULER-SR/PCAP	yes	no	-	-	-	-	-
	artificial	PHRAP/PCAP	yes	no	-	-	-	-	-
	artificial	SSAKE/PCAP	yes	yes	99.25	3771	2.44	1549	9184
	artificial	VELVET/PCAP	yes	yes	99.5	8941	2.33	1277	3483

Table B.18: Execution time and Amount of memory used on results from assembling 454 contigs merged with Sanger reads.

Org.	R/A	Assembler	Q.F.	RTC	Time(sec.)	Memory
AT	real	EULER-SR/PHRAP	no	yes	1484.3	518.43 Mb
	real	MIRA/PHRAP	no	yes	1488.9	528.62 Mb
	real	PHRAP/PHRAP	no	yes	1704.56	517 Mb
	real	SSAKE/PHRAP	no	yes	1928.23	778.96 Mb
	real	VELVET/PHRAP	no	yes	1594.66	570.89 Mb
	real	EULER-SR/CAP3	no	yes	5436.31	902.85 Mb
	real	MIRA/CAP3	no	yes	22288.95	3.45 Gb
	real	PHRAP/CAP3	no	yes	19131.66	2.27 Gb
	real	SSAKE/CAP3	no	yes	5356.39	1.24 Gb
	real	VELVET/CAP3	no	yes	3695.52	943.14 Mb
	real	EULER-SR/PCAP	yes	yes	3197.85	342.05 Mb
	real	PHRAP/PCAP	yes	no	3844.86	554.37 Mb
	real	SSAKE/PCAP	yes	yes	2134.44	349.57 Mb
	real	VELVET/PCAP	yes	yes	1495.92	343.29 Mb
CM	real	EULER-SR/PHRAP	no	yes	322.67	399.74 Mb
	real	MIRA/PHRAP	no	yes	507.73	377.01 Mb
	real	PHRAP/PHRAP	no	yes	645.6	379.77 Mb
	real	SSAKE/PHRAP	no	yes	475.7	652.3 Mb
	real	VELVET/PHRAP	no	yes	272.82	406.24 Mb
	real	EULER-SR/CAP3	no	yes	11419.42	1.73 Gb
	real	MIRA/CAP3	no	no	-	-
	real	PHRAP/CAP3	no	no	-	-
	real	SSAKE/CAP3	no	yes	3992.3	946 Mb
	real	VELVET/CAP3	no	yes	1964.11	685.33 Mb
	real	EULER-SR/PCAP	yes	no	15990.12	549.09 Mb
	real	PHRAP/PCAP	yes	no	1223.89	420.63 Mb
	real	SSAKE/PCAP	yes	yes	2336.93	343.39 Mb
	real	VELVET/PCAP	yes	yes	1314.17	339.83 Mb
EC	artificial	EULER-SR/PHRAP	no	yes	2248.96	4.35 Gb
	artificial	MIRA/PHRAP	no	yes	2456.89	1.44 Gb
	artificial	PHRAP/PHRAP	no	yes	2315.88	1.40 Gb
	artificial	SSAKE/PHRAP	no	yes	1967.7	2.03 Gb
	artificial	VELVET/PHRAP	no	yes	1952.92	5.78 Gb
	artificial	EULER-SR/CAP3	no	no	-	-
	artificial	MIRA/CAP3	no	no	-	-
	artificial	PHRAP/CAP3	no	no	-	-
	artificial	SSAKE/CAP3	no	yes	41927.29	2.89 Gb
	artificial	VELVET/CAP3	no	yes	19683.36	2.39 Gb
	artificial	EULER-SR/PCAP	yes	no	2696.89	611.51 Mb
	artificial	PHRAP/PCAP	yes	no	4089.83	653.45 Mb
	artificial	SSAKE/PCAP	yes	yes	26197.65	609.16 Mb
	artificial	VELVET/PCAP	yes	yes	13465.41	606.16 Mb
HA	artificial	EULER-SR/PHRAP	no	yes	2739.07	18.66 Gb
	artificial	MIRA/PHRAP	no	yes	498.48	9.09 Gb
	artificial	PHRAP/PHRAP	no	yes	312.4	961.81 Mb
	artificial	SSAKE/PHRAP	no	yes	483.21	7.87 Gb
	artificial	VELVET/PHRAP	no	yes	1156.03	36.22 Gb
	artificial	EULER-SR/CAP3	no	yes	12249.83	1.17 Gb
	artificial	MIRA/CAP3	no	yes	2843.31	812.99 Mb
	artificial	PHRAP/CAP3	no	yes	16357.12	3.39 Gb
	artificial	SSAKE/CAP3	no	yes	4285.55	864.16 Mb
	artificial	VELVET/CAP3	no	yes	2983.62	837.59 Mb
	artificial	EULER-SR/PCAP	yes	no	2907.8	364.68 Mb
	artificial	PHRAP/PCAP	yes	no	6227.01	542.38 Mb

Continued on next page

Table B.18 – continued from previous page

Org.	R/A	Assembler	Q.F.	RTC	Time(sec.)	Memory
HA	artificial	SSAKE/PCAP	yes	yes	2133.29	341.41 Mb
	artificial	VELVET/PCAP	yes	yes	773.62	341.40 Mb
MM	real	EULER-SR/PHRAP	no	yes	282.1	585.43 Mb
	real	MIRA/PHRAP	no	yes	218.85	439.13 Mb
	real	PHRAP/PHRAP	no	yes	205.08	379.83 Mb
	real	SSAKE/PHRAP	no	yes	221.55	721.86 Mb
	real	VELVET/PHRAP	no	yes	211.21	576.5 Mb
	real	EULER-SR/CAP3	no	yes	8705.56	607.23 Mb
	real	MIRA/CAP3	no	yes	12448.11	2.71 Gb
	real	PHRAP/CAP3	no	no	-	-
	real	SSAKE/CAP3	no	yes	1777.23	622.92 Mb
	real	VELVET/CAP3	no	yes	1543.8	610.91 Mb
	real	EULER-SR/PCAP	yes	yes	9460.92	339.36 Mb
	real	PHRAP/PCAP	yes	yes	7218.73	339.32 Mb
	real	SSAKE/PCAP	yes	yes	1623.11	340.24 Mb
	real	VELVET/PCAP	yes	yes	1181.56	340.15 Mb
SS	artificial	EULER-SR/PHRAP	no	yes	1851.48	51.41 Gb
	artificial	MIRA/PHRAP	no	yes	179.25	373.86 Mb
	artificial	PHRAP/PHRAP	no	yes	481.17	3.69 Gb
	artificial	SSAKE/PHRAP	no	yes	552.64	836.64 Mb
	artificial	VELVET/PHRAP	no	yes	1601.59	30.96 Gb
	artificial	EULER-SR/CAP3	no	yes	21522.92	3.05 Gb
	artificial	MIRA/CAP3	no	no	-	-
	artificial	PHRAP/CAP3	no	yes	21143.62	2.80 Gb
	artificial	SSAKE/CAP3	no	yes	5657.91	1.17 Gb
	artificial	VELVET/CAP3	no	yes	4319.22	1.09 Gb
	artificial	EULER-SR/PCAP	yes	no	5289.04	433.38 Mb
	artificial	PHRAP/PCAP	yes	no	4851.75	424.82 Mb
	artificial	SSAKE/PCAP	yes	yes	2463.6	342.92 Mb
	artificial	VELVET/PCAP	yes	yes	993.76	342.82 Mb

Table B.19: Genome coverage results from assembling Illumina contigs merged with Sanger reads.

Org.	R/A	Assembler	Q.F.	RTC	GC (%)	# Contigs	MC (%)	N50	L.C
AT	artificial/real	ABYSS/PHRAP	no	yes	98.77	255	8.24	84120	284351
	artificial/real	EDENA/PHRAP	no	yes	98	277	5.42	78678	244601
	artificial/real	EULER-SR/PHRAP	no	yes	97.82	242	6.2	90785	325104
	artificial/real	VELVET/PHRAP	no	yes	95.47	423	5.2	25231	153061
	artificial/real	ABYSS/CAP3	no	no	-	-	-	-	-
	artificial/real	EDENA/CAP3	no	yes	89.7	309	6.47	39143	135663
	artificial/real	EULER-SR/CAP3	no	no	-	-	-	-	-
	artificial/real	VELVET/CAP3	no	yes	80.11	437	5.72	20683	119029
CM	artificial/real	ABYSS/PHRAP	no	yes	99.54	46	26.09	297940	590034
	artificial/real	EDENA/PHRAP	no	yes	99.84	42	33.33	353732	724635
	artificial/real	EULER-SR/PHRAP	no	yes	99.22	48	22.92	317498	443288
	artificial/real	VELVET/PHRAP	no	yes	99.62	94	13.83	87516	199628
	artificial/real	ABYSS/CAP3	no	yes	98.5	77	11.69	113721	419573
	artificial/real	EDENA/CAP3	no	yes	99.72	75	12	117434	534669
	artificial/real	EULER-SR/CAP3	no	yes	99.56	92	8.7	117117	256079
	artificial/real	VELVET/CAP3	no	yes	99.48	145	6.21	52458	160116
EC	real/artificial	ABYSS/PHRAP	no	yes	91.29	17	47.06	488389	976797
	real/artificial	EDENA/PHRAP	no	yes	99.53	32	31.25	401922	1158352
	real/artificial	EULER-SR/PHRAP	no	yes	95.17	22	54.55	439695	1005509
	real/artificial	VELVET/PHRAP	no	yes	99.61	16	56.25	610884	1263471
	real/artificial	ABYSS/CAP3	no	no	-	-	-	-	-
	real/artificial	EDENA/CAP3	no	no	-	-	-	-	-
	real/artificial	EULER-SR/CAP3	no	no	-	-	-	-	-
	real/artificial	VELVET/CAP3	no	no	-	-	-	-	-
HA	real/artificial	ABYSS/PHRAP	no	yes	99.85	13	23.08	355997	437248
	real/artificial	EDENA/PHRAP	no	yes	99.83	11	45.46	279921	453377
	real/artificial	EULER-SR/PHRAP	no	yes	99.9	10	40	216267	437306
	real/artificial	PHRAP/PHRAP	no	yes	99.35	37	13.51	197165	336577
	real/artificial	VELVET/PHRAP	no	yes	97.77	17	35.29	251525	388533
	real/artificial	ABYSS/CAP3	no	yes	99.73	169	5.33	22193	71175
	real/artificial	EDENA/CAP3	no	yes	99.14	148	4.73	29662	86883
	real/artificial	EULER-SR/CAP3	no	yes	99.87	187	3.21	29560	82343
	real/artificial	PHRAP/CAP3	no	yes	99.85	651	1.23	5345	31771
	real/artificial	VELVET/CAP3	no	yes	99.61	599	1	5396	31771
MM	artificial/real	ABYSS/PHRAP	no	yes	99.59	7	14.29	1483759	1483759
	artificial/real	EDENA/PHRAP	no	yes	99.59	7	0	1762361	1762361
	artificial/real	EULER-SR/PHRAP	no	yes	99.59	10	10	1403186	1403186
	artificial/real	VELVET/PHRAP	no	yes	99.54	12	16.67	596332	806890
	artificial/real	ABYSS/CAP3	no	no	-	-	-	-	-
	artificial/real	EDENA/CAP3	no	yes	99.77	6	16.67	1043516	1043516
	artificial/real	EULER-SR/CAP3	no	no	-	-	-	-	-
	artificial/real	VELVET/CAP3	no	yes	99.76	9	11.11	425063	686986
SS	real	ABYSS/PHRAP	no	yes	98.8	162	26.54	37147	98674
	real	EDENA/PHRAP	no	yes	98.52	198	20.2	30550	96639
	real	EULER-SR/PHRAP	no	yes	97.79	240	19.58	20387	94328
	real	PHRAP/PHRAP	no	yes	91.15	933	4.61	9107	32641
	real	SSAKE/PHRAP	no	yes	98.88	294	17.35	20989	92621
	real	VELVET/PHRAP	no	yes	98.62	158	28.48	41709	104991
	real	ABYSS/CAP3	no	yes	98.27	185	25.95	33599	97724
	real	EDENA/CAP3	no	no	-	-	-	-	-
	real	EULER-SR/CAP3	no	no	-	-	-	-	-
	real	PHRAP/CAP3	no	no	-	-	-	-	-
	real	SSAKE/CAP3	no	no	-	-	-	-	-
	real	VELVET/CAP3	no	yes	98.02	180	27.22	34444	92869

Table B.20: Execution time and Amount of memory used on results from assembling Illumina contigs merged with Sanger reads.

Org.	R/A	Assembler	Q.F.	RTC	Time(sec.)	Memory
AT	artificial/real	ABYSS/PHRAP	no	yes	1139.26	536.56 Mb
	artificial/real	EDENA/PHRAP	no	yes	1225.89	488.75 Mb
	artificial/real	EULER-SR/PHRAP	no	yes	1331.19	502.09 Mb
	artificial/real	VELVET/PHRAP	no	yes	1116.61	484.94 Mb
	artificial/real	ABYSS/CAP3	no	no	-	-
	artificial/real	EDENA/CAP3	no	yes	19087.42	3.57 Gb
	artificial/real	EULER-SR/CAP3	no	no	-	-
	artificial/real	VELVET/CAP3	no	yes	3546.33	907.49 Mb
CM	artificial/real	ABYSS/PHRAP	no	yes	287.65	417.14 Mb
	artificial/real	EDENA/PHRAP	no	yes	303.64	414.17 Mb
	artificial/real	EULER-SR/PHRAP	no	yes	322.87	387.27 Mb
	artificial/real	VELVET/PHRAP	no	yes	250.62	412.42
	artificial/real	ABYSS/CAP3	no	yes	11308.25	2.08 Gb
	artificial/real	EDENA/CAP3	no	yes	12372.13	1.65 Gb
	artificial/real	EULER-SR/CAP3	no	yes	19380.9	4.72 Gb
	artificial/real	VELVET/CAP3	no	yes	2054.96	682.89 Mb
EC	real/artificial	ABYSS/PHRAP	no	yes	1901.43	1.34 Gb
	real/artificial	EDENA/PHRAP	no	yes	1752.81	1.26 Gb
	real/artificial	EULER-SR/PHRAP	no	yes	1694.26	1.22 Gb
	real/artificial	VELVET/PHRAP	no	yes	1978.1	1.17 Gb
	real/artificial	ABYSS/CAP3	no	no	-	-
	real/artificial	EDENA/CAP3	no	no	-	-
	real/artificial	EULER-SR/CAP3	no	no	-	-
	real/artificial	VELVET/CAP3	no	no	-	-
HA	real/artificial	ABYSS/PHRAP	no	yes	571.99	7.53 Gb
	real/artificial	EDENA/PHRAP	no	yes	503.73	8.55 Gb
	real/artificial	EULER-SR/PHRAP	no	yes	367.39	3.23 Gb
	real/artificial	PHRAP/PHRAP	no	yes	412.61	2.96 Gb
	real/artificial	VELVET/PHRAP	no	yes	356.38	611.73 Mb
	real/artificial	ABYSS/CAP3	no	yes	12983.49	2.85 Gb
	real/artificial	EDENA/CAP3	no	yes	13891.64	2.69 Gb
	real/artificial	EULER-SR/CAP3	no	yes	17790.53	2.79 Gb
	real/artificial	PHRAP/CAP3	no	yes	2949.37	866.68 Mb
	real/artificial	VELVET/CAP3	no	yes	2721.53	821.18 Mb
MM	artificial/real	ABYSS/PHRAP	no	yes	206.79	631.15 Mb
	artificial/real	EDENA/PHRAP	no	yes	217.11	776.08 Mb
	artificial/real	EULER-SR/PHRAP	no	yes	217.87	798.15 Mb
	artificial/real	VELVET/PHRAP	no	yes	201.07	573.47 Mb
	artificial/real	ABYSS/CAP3	no	no	-	-
	artificial/real	EDENA/CAP3	no	yes	12741.78	2.36 Gb
	artificial/real	EULER-SR/CAP3	no	no	-	-
	artificial/real	VELVET/CAP3	no	yes	1536.46	1.85 Gb
SS	real	ABYSS/PHRAP	no	yes	197.47	363.09 Mb
	real	EDENA/PHRAP	no	yes	206.67	347.03 Mb
	real	EULER-SR/PHRAP	no	yes	206.59	355.51 Mb
	real	PHRAP/PHRAP	no	yes	215.95	446.42 Mb
	real	SSAKE/PHRAP	no	yes	216.45	381.55 Mb
	real	VELVET/PHRAP	no	yes	216.49	340.03 Mb
	real	ABYSS/CAP3	no	yes	5200.81	1.00 Gb
	real	EDENA/CAP3	no	no	-	-
	real	EULER-SR/CAP3	no	no	-	-
	real	PHRAP/CAP3	no	no	-	-
	real	SSAKE/CAP3	no	no	-	-

Continued on next page

Table B.20: Execution time and Amount of memory used on results from assembling Illumina contigs merged with Sanger reads.

Org.	R/A	Assembler	Q.F.	RTC	Time(sec.)	Memory
SS	real	VELVET/CAP3	no	yes	6291.88	638.44 Mb

Table B.21: Genome coverage results from assembling 454 contigs merged with Sanger contigs.

Org.	R/A	Assembler	Q.F.	RTC	GC (%)	# Contigs	MC (%)	N50	L.C
AT	real	PHRAP(EULER-SR/PHRAP)	no	yes	97.13	284	5.63	51679	254878
	real	PHRAP(MIRA/PHRAP)	no	yes	99.88	108	41.67	89501	303285
	real	PHRAP(PHRAP/PHRAP)	no	yes	99.59	156	17.31	72447	153566
	real	PHRAP(SSAKE/PHRAP)	no	yes	97.46	261	6.13	50070	149831
	real	PHRAP(VELVET/PHRAP)	no	yes	94.63	2167	1.02	26771	102810
CM	real	PHRAP(EULER-SR/PHRAP)	no	yes	99.85	32	9.38	364365	564014
	real	PHRAP(MIRA/PHRAP)	no	yes	99.9	29	17.24	399172	536753
	real	PHRAP(PHRAP/PHRAP)	no	yes	99.7	63	12.7	175554	636167
	real	PHRAP(SSAKE/PHRAP)	no	yes	99.82	54	5.56	195840	383418
	real	PHRAP(VELVET/PHRAP)	no	yes	99.57	360	1.94	108942	343306
EC	real/artificial	PHRAP(EULER-SR/PHRAP)	no	yes	98.92	145	30.35	87126	373237
	real/artificial	PHRAP(MIRA/PHRAP)	no	yes	99.57	120	47.5	114660	411175
	real/artificial	PHRAP(PHRAP/PHRAP)	no	yes	99.1	122	38.52	117988	530313
	real/artificial	PHRAP(SSAKE/PHRAP)	no	yes	98.59	170	40	63999	156405
	real/artificial	PHRAP(VELVET/PHRAP)	no	yes	98.24	167	30.54	61147	230307
HA	artificial	PHRAP(EULER-SR/PHRAP)	no	yes	99.73	25	32	122914	283072
	artificial	PHRAP(MIRA/PHRAP)	no	yes	97.61	32	34.38	106125	354738
	artificial	PHRAP(PHRAP/PHRAP)	no	yes	99.02	41	36.59	80696	247369
	artificial	PHRAP(SSAKE/PHRAP)	no	yes	99.74	36	27.78	97752	246739
	artificial	PHRAP(VELVET/PHRAP)	no	yes	98.45	47	23.4	106095	354738
MM	real	PHRAP(EULER-SR/PHRAP)	no	yes	99.8	5	0	1763844	1763844
	real	PHRAP(MIRA/PHRAP)	no	yes	99.81	19	31.58	1159852	1159852
	real	PHRAP(PHRAP/PHRAP)	no	yes	99.79	61	59.02	813535	955231
	real	PHRAP(SSAKE/PHRAP)	no	yes	99.8	24	16.67	941110	941110
	real	PHRAP(VELVET/PHRAP)	no	yes	99.7	15	0	951901	951901
SS	artificial/real	PHRAP(EULER-SR/PHRAP)	no	yes	98.49	78	35.9	56198	147905
	artificial/real	PHRAP(MIRA/PHRAP)	no	yes	96.76	116	32.76	40361	125959
	artificial/real	PHRAP(PHRAP/PHRAP)	no	yes	96.68	116	37.93	40361	125959
	artificial/real	PHRAP(SSAKE/PHRAP)	no	yes	99.4	87	31.04	50364	147212
	artificial/real	PHRAP(VELVET/PHRAP)	no	yes	98.28	256	14.06	37646	126272

Table B.22: Execution time and Amount of memory used on results from assembling 454 contigs merged with Sanger contigs.

Org.	R/A	Assembler	Q.F.	RTC	Time(sec.)	Memory
AT	real	PHRAP(EULER-SR/PHRAP)	no	yes	130.27	179.96 Mb
	real	PHRAP(MIRA/PHRAP)	no	yes	251.62	183.61 Mb
	real	PHRAP(PHRAP/PHRAP)	no	yes	220.51	286.92 Mb
	real	PHRAP(SSAKE/PHRAP)	no	yes	269.83	404.29 Mb
	real	PHRAP(VELVET/PHRAP)	no	yes	131.17	195.99 Mb
CM	real	PHRAP(EULER-SR/PHRAP)	no	yes	353.18	163.61 Mb
	real	PHRAP(MIRA/PHRAP)	no	yes	809.99	164.04 Mb
	real	PHRAP(PHRAP/PHRAP)	no	yes	958.76	173.98 Mb
	real	PHRAP(SSAKE/PHRAP)	no	yes	343.94	417.24 Mb
	real	PHRAP(VELVET/PHRAP)	no	yes	179.81	175.79 Mb
EC	real/artificial	PHRAP(EULER-SR/PHRAP)	no	yes	1084.32	198.27 Mb
	real/artificial	PHRAP(MIRA/PHRAP)	no	yes	1915.49	207.48 Mb
	real/artificial	PHRAP(PHRAP/PHRAP)	no	yes	2073.87	220.79 Mb
	real/artificial	PHRAP(SSAKE/PHRAP)	no	yes	456.1	207.06 Mb
	real/artificial	PHRAP(VELVET/PHRAP)	no	yes	379.27	205.17 Mb
HA	artificial	PHRAP(EULER-SR/PHRAP)	no	yes	244.43	140.68 Mb
	artificial	PHRAP(MIRA/PHRAP)	no	yes	102.99	145.93 Mb
	artificial	PHRAP(PHRAP/PHRAP)	no	yes	199.03	141.43 Mb
	artificial	PHRAP(SSAKE/PHRAP)	no	yes	115.07	161.44 Mb
	artificial	PHRAP(VELVET/PHRAP)	no	yes	107.73	162.07 Mb
MM	real	PHRAP(EULER-SR/PHRAP)	no	yes	449.9	157.45 Mb
	real	PHRAP(MIRA/PHRAP)	no	yes	521.81	161.37 Mb
	real	PHRAP(PHRAP/PHRAP)	no	yes	464.58	191.57 Mb
	real	PHRAP(SSAKE/PHRAP)	no	yes	440.11	186.13 Mb
	real	PHRAP(VELVET/PHRAP)	no	yes	447.68	228.38 Mb
SS	artificial/real	PHRAP(EULER-SR/PHRAP)	no	yes	131.6	148.74 Mb
	artificial/real	PHRAP(MIRA/PHRAP)	no	yes	47.78	143.47 Mb
	artificial/real	PHRAP(PHRAP/PHRAP)	no	yes	49.41	143.47 Mb
	artificial/real	PHRAP(SSAKE/PHRAP)	no	yes	66.34	167.16 Mb
	artificial/real	PHRAP(VELVET/PHRAP)	no	yes	57.82	162.3 Mb

APPENDIX C

GRAPHS

This appendix and the next present plots created to evaluate the performance for the different assemblies executed during this research. The plots are for results for genome coverage, misassembled contigs, N50 size, number of contigs, execution time, and amount of memory used. Some of the graphs have a two-scale x-axis. The two-scale x-axis is used when there is a large difference among the results obtained for the various organisms. For example, if the number of contigs obtained for AT are between 20-80, and for SS between 2300-14000, a two-scale axis would be more appropriate to represent the results on the graph.

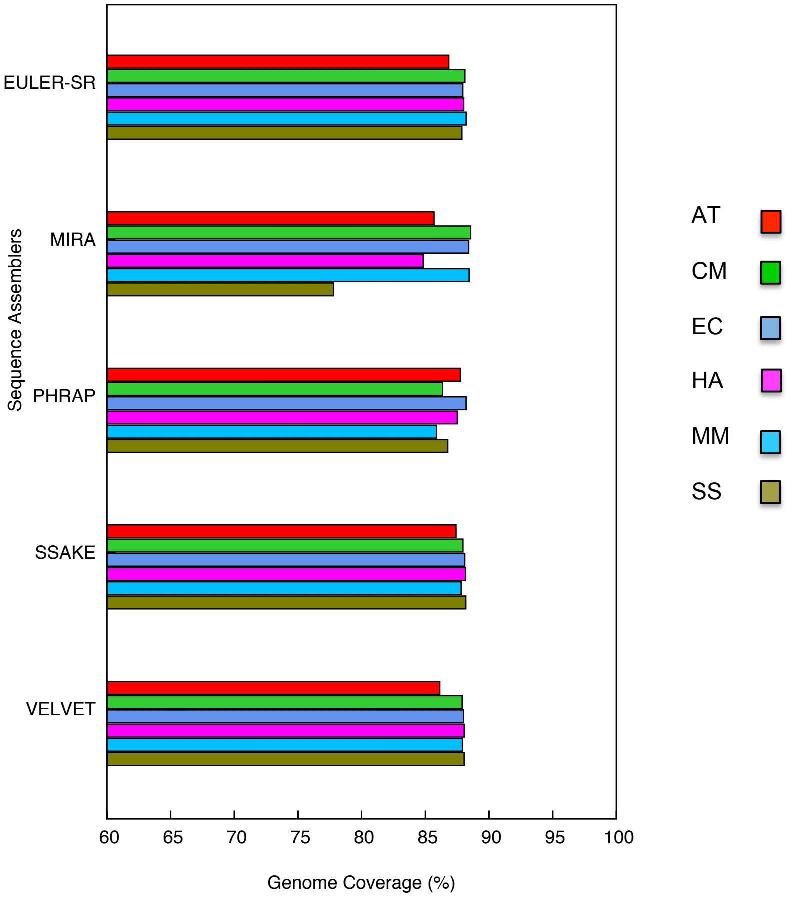


Figure C.1: Genome coverage results from running short- and long-read assemblers on 454 data.

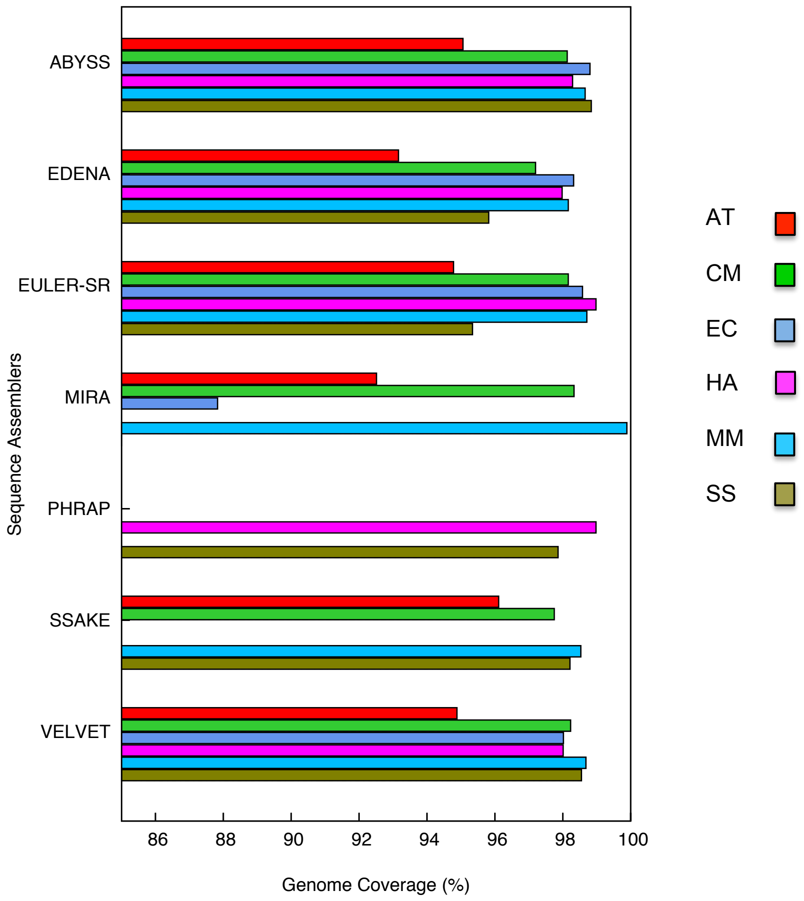


Figure C.2: Genome coverage results from running short- and long-read assemblers on Illumina data.

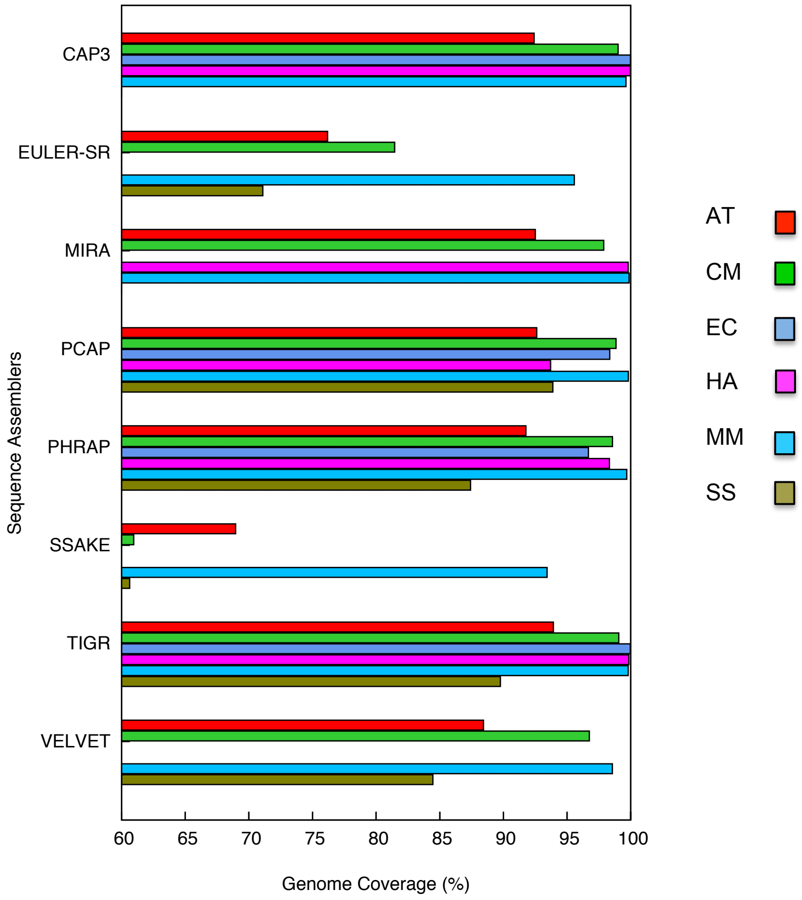


Figure C.3: Genome coverage results from running short- and long-read assemblers on Sanger data.

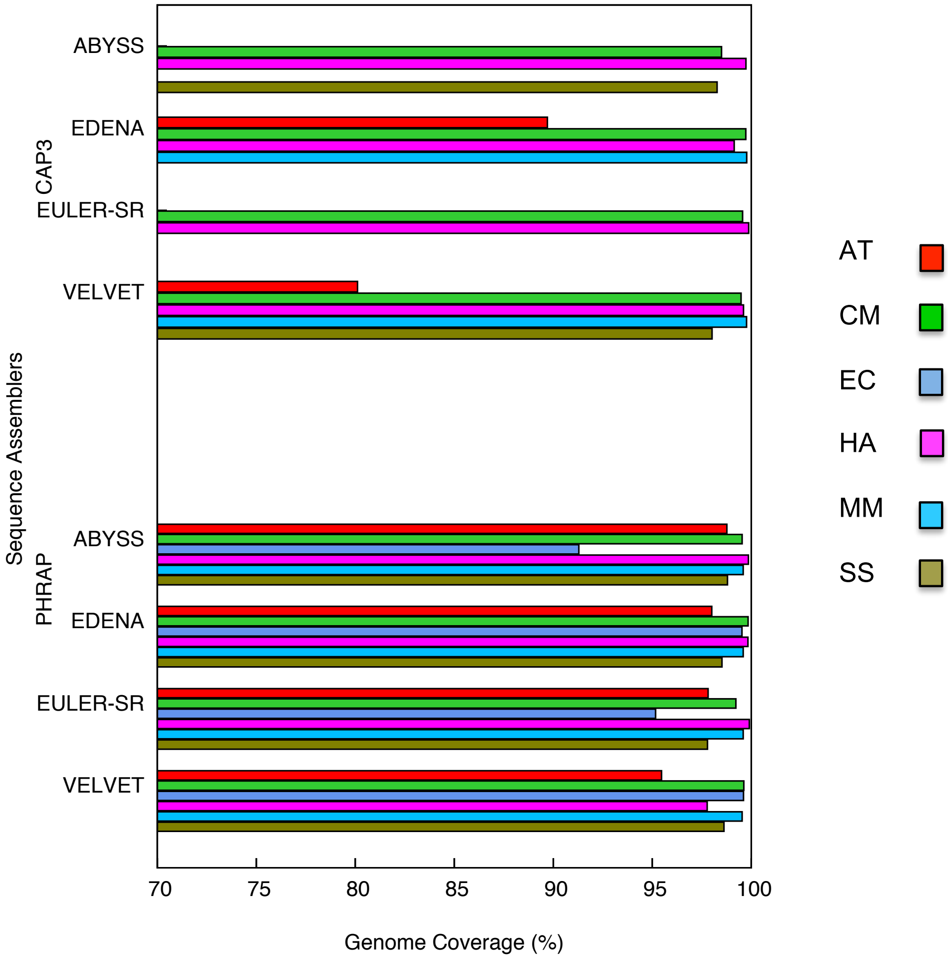


Figure C.4: Genome coverage results from running long-read assemblers on Illumina contigs + Sanger reads.

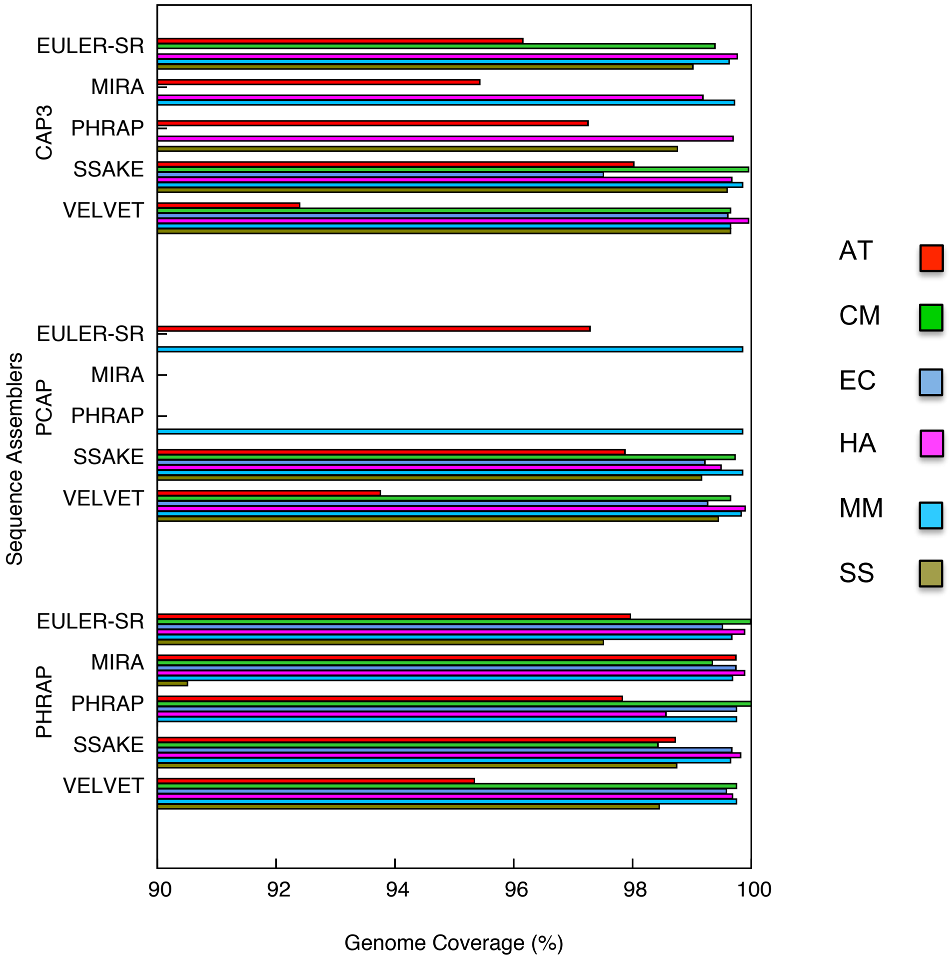


Figure C.5: Genome coverage results from running long-read assemblers on 454 contigs + Sanger reads.

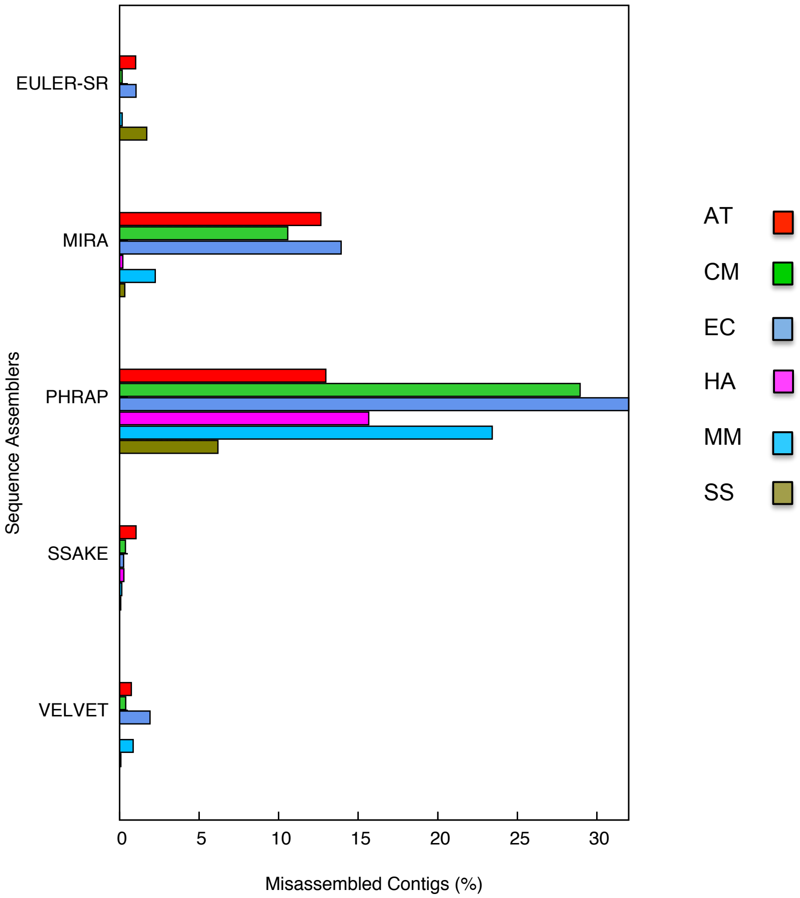


Figure C.6: Misassembled contig results from running short- and long-read assemblers on 454 data.

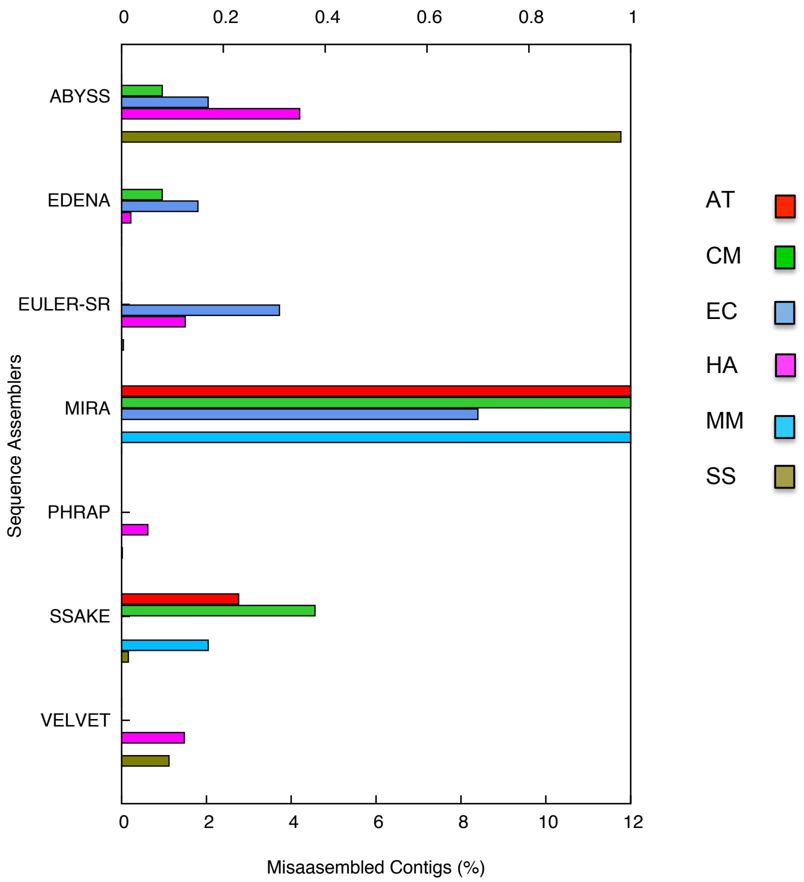


Figure C.7: Misassembled contig results from running short- and long-read assemblers on Illumina data. See appendix introduction for explanation on the use of a two-scale axis. Top x-axis: AT, CM, EC, and MM; Bottom x-axis: HA and SS.

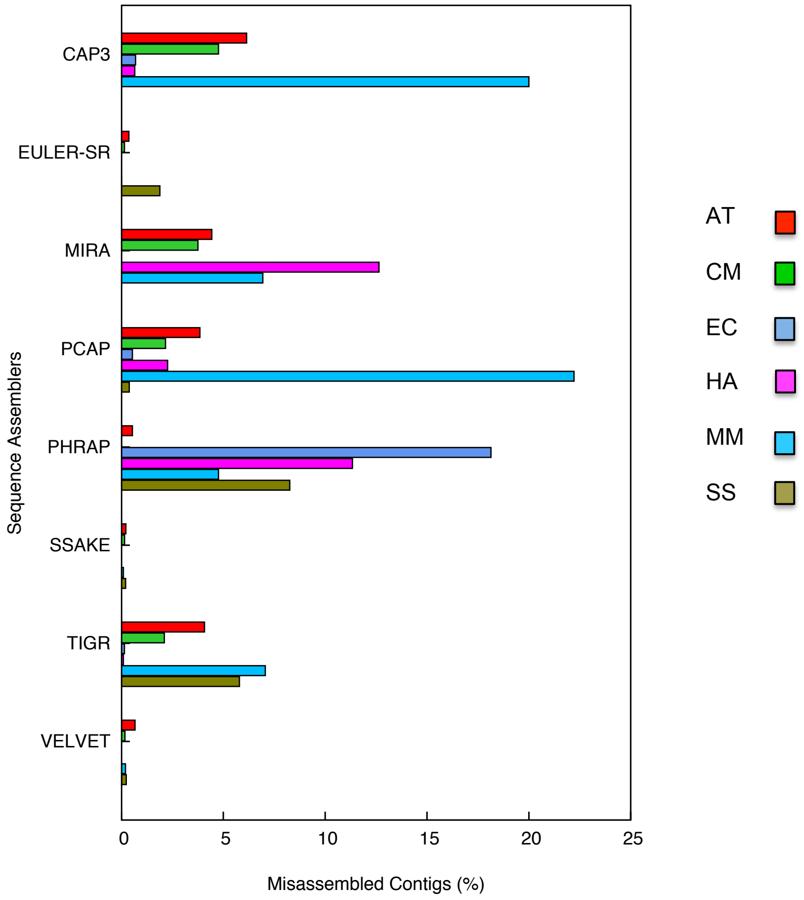


Figure C.8: Misassembled contig results from running short- and long-read assemblers on Sanger data.

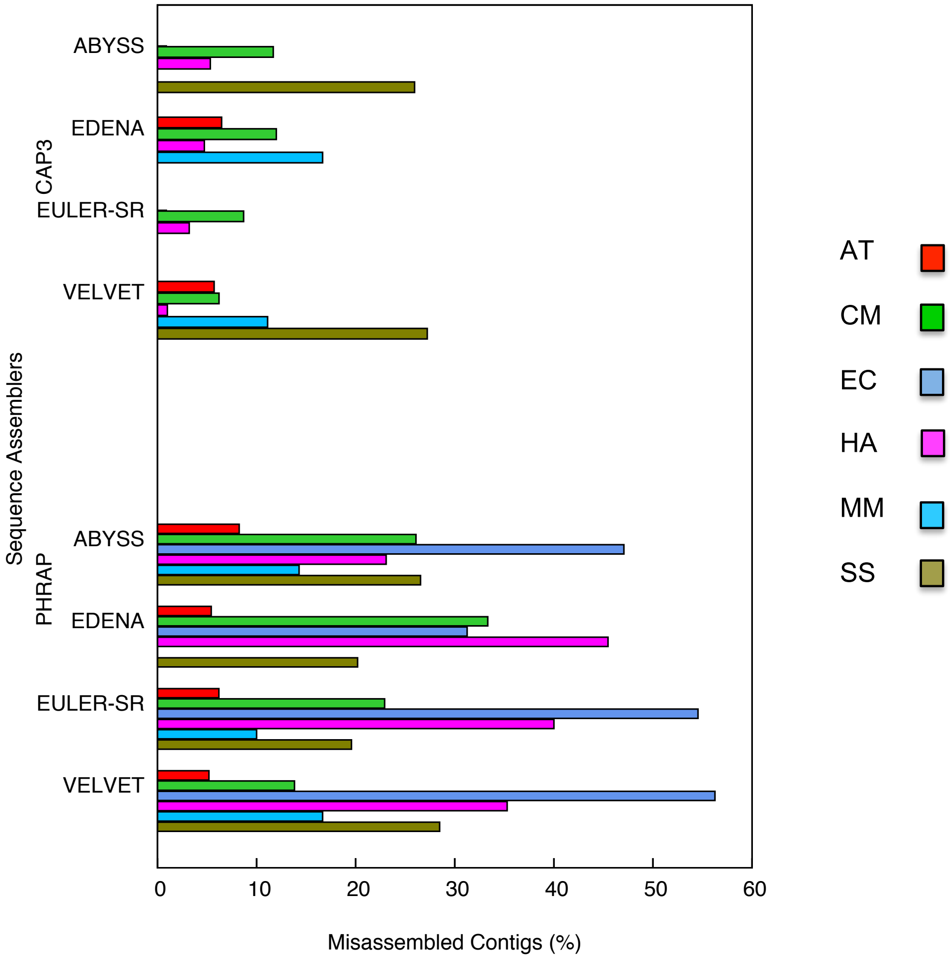


Figure C.9: Misassembled contig results from running long-read assemblers on Illumina contigs + Sanger reads.

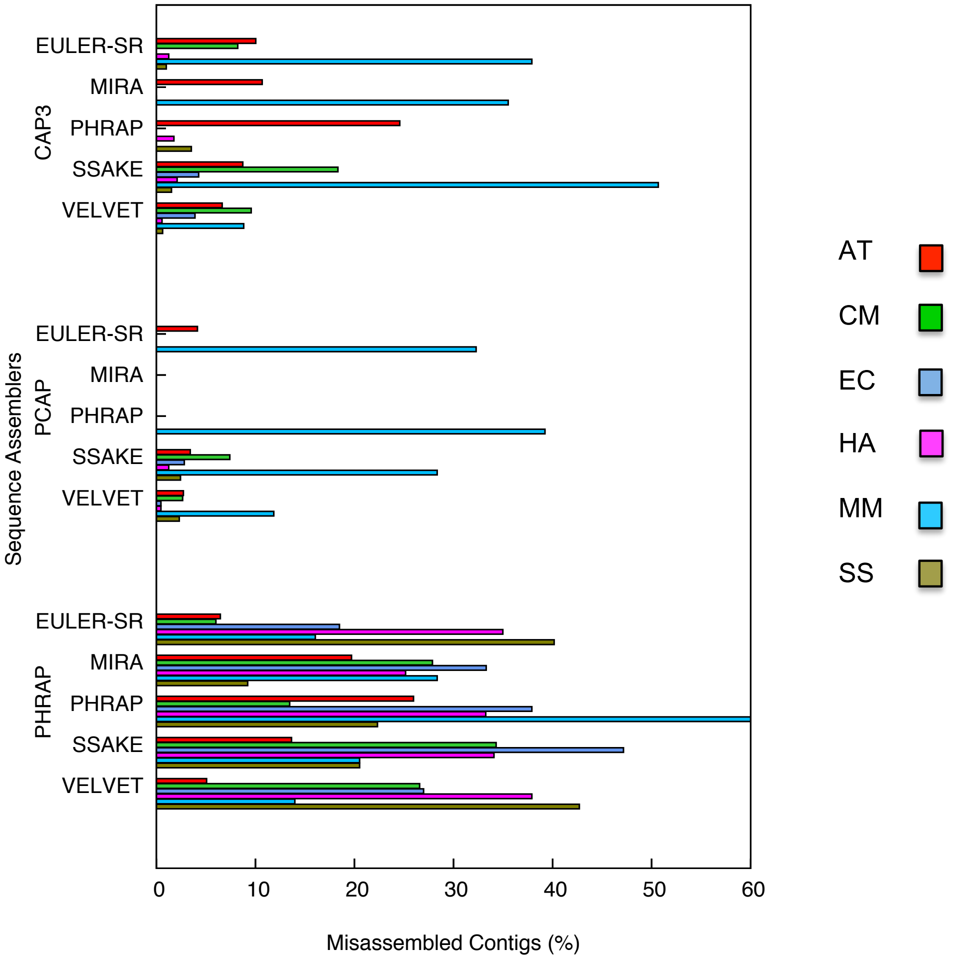


Figure C.10: Misassembled contig results from running long-read assemblers on 454 contigs + Sanger reads.

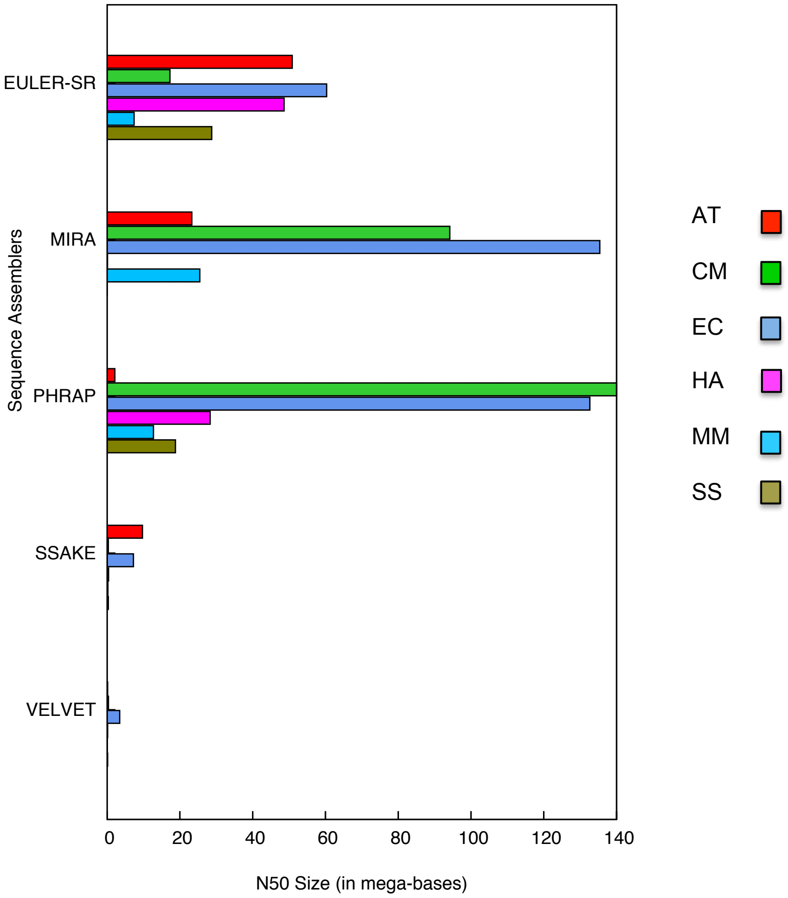


Figure C.11: N50 size results from running short- and long-read assemblers on 454 data.

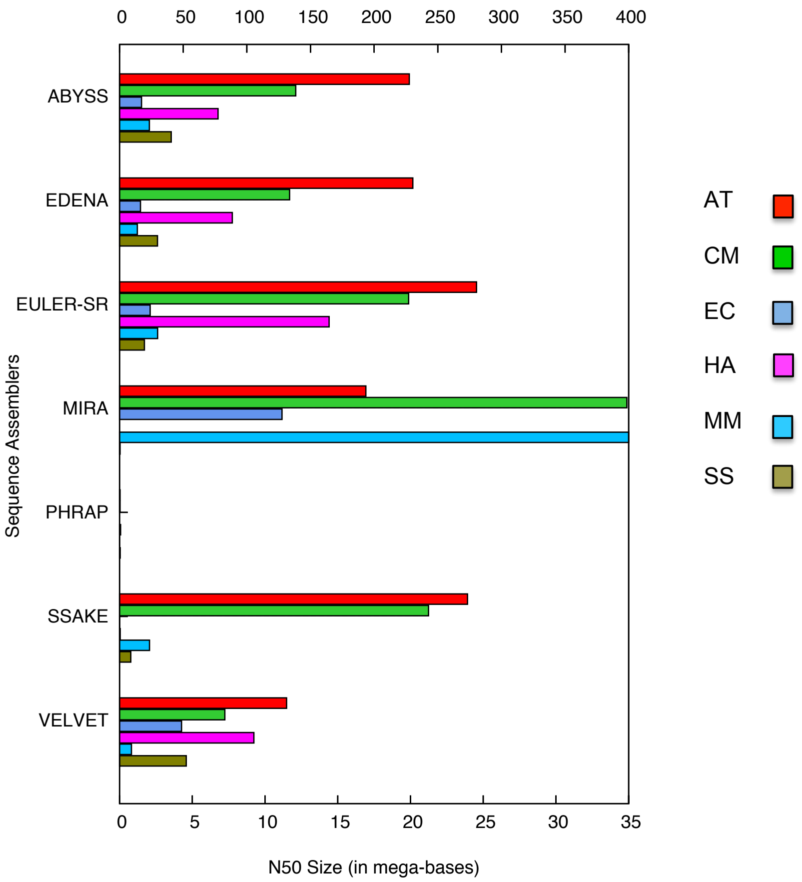


Figure C.12: N50 size results from running short- and long-read assemblers on Illumina data. See the appendix introduction for an explanation of the two-scale axis. Top x-axis: EC and MM, Bottom x-axis: AT, CM, HA, and SS.

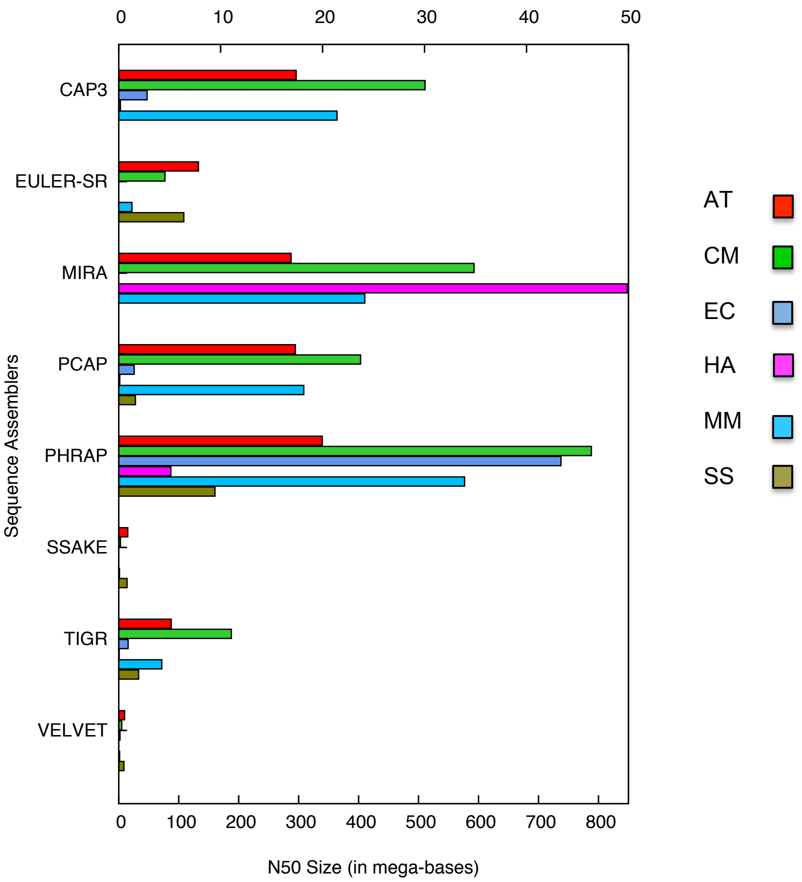


Figure C.13: N50 size results from running short- and long-read assemblers on Sanger data. See the appendix introduction for an explanation of the two-scale axis. Top x-axis: AT, CM, EC, and SS; Bottom x-axis: HA and MM.

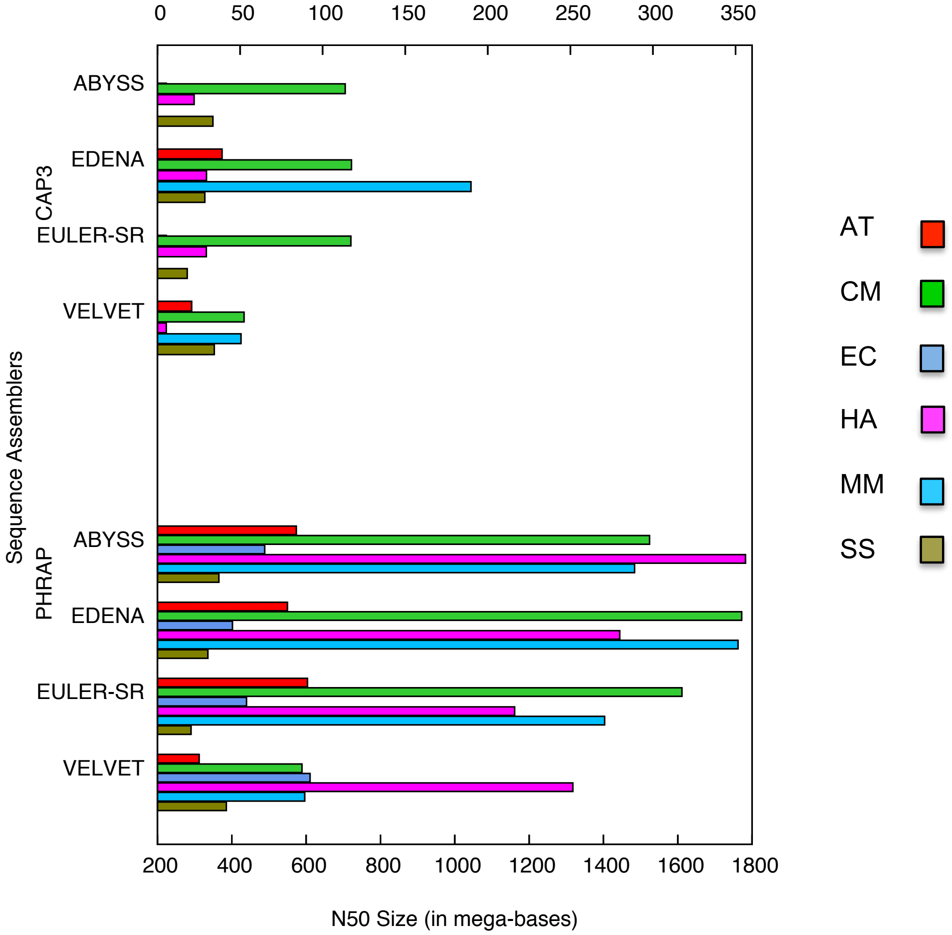


Figure C.14: N50 size results from running long-read assemblers on Illumina contigs + Sanger reads. See the appendix introduction for an explanation of the two-scale axis. Bottom x-axis: MM (CAP3), EC and MM (from PHRAP).

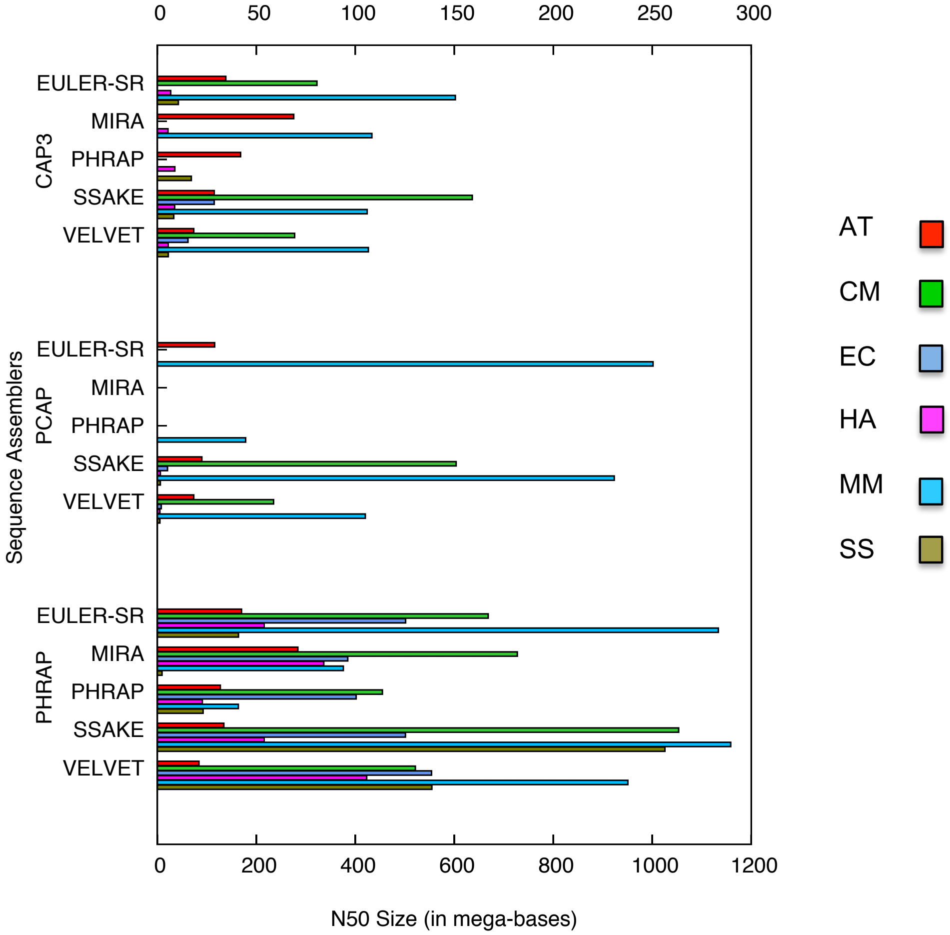


Figure C.15: N50 size results from running long-read assemblers on 454 contigs + Sanger reads. See the appendix introduction for an explanation of the two-scale axis. Bottom x-axis: MM (CAP3), EC, HA, MM, and SS (from PHRAP).

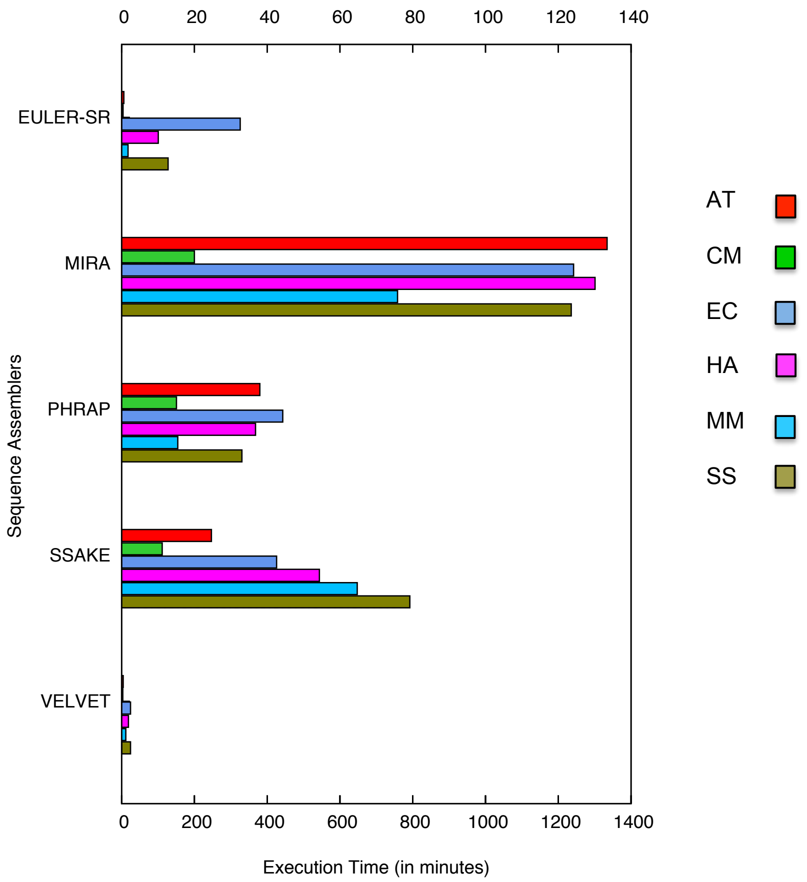


Figure C.16: Execution time results from running short- and long-read assemblers on 454 data. See the appendix introduction for an explanation of the two-scale axis. Top x-axis: AT and CM, Bottom x-axis: EC, HA, MM, and SS.

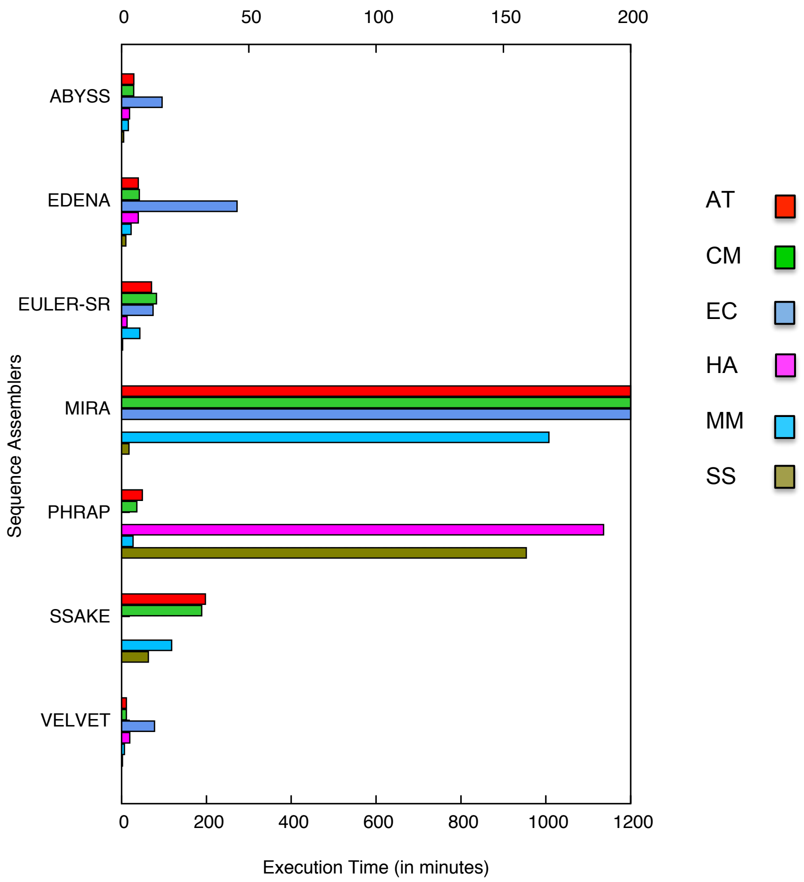


Figure C.17: Execution time results from running short- and long-read assemblers on Illumina data. See the appendix introduction for an explanation of the two-scale axis. Top x-axis: HA and SS, Bottom x-axis: AT, CM, EC, and MM.

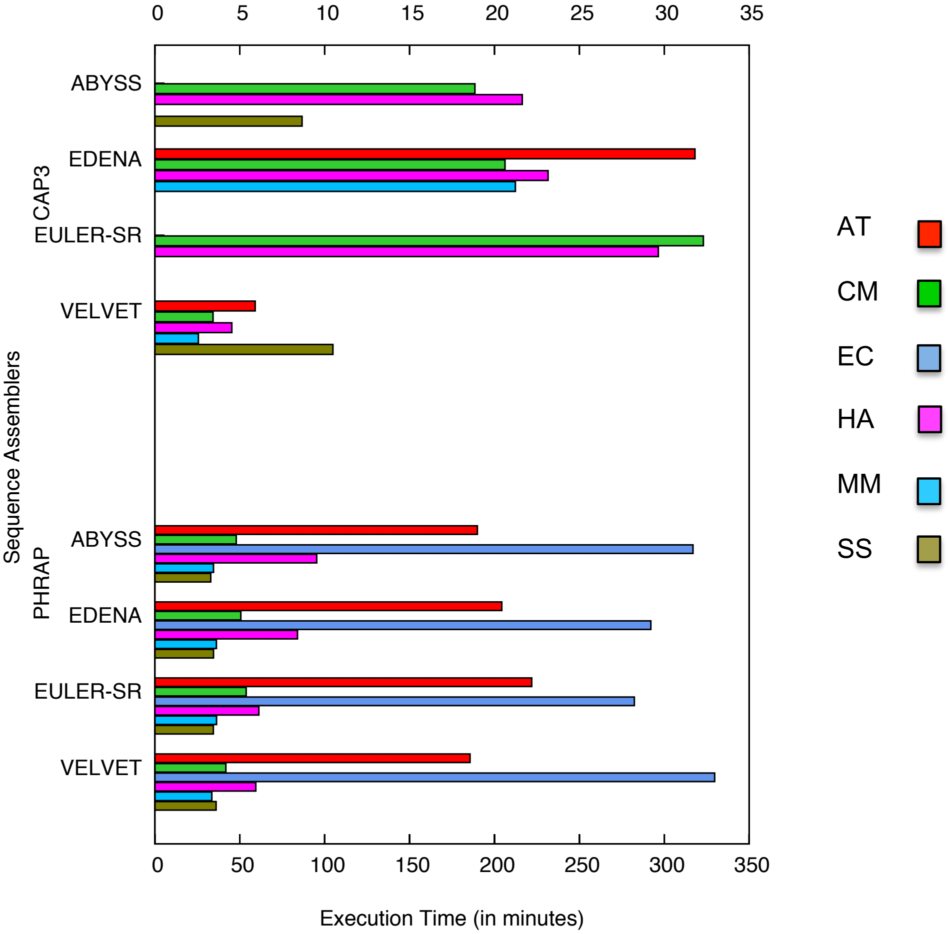


Figure C.18: Execution time results from running long-read assemblers on Illumina contigs + Sanger reads. See the appendix introduction for an explanation of the two-scale axis. Bottom x-axis: AT, CM, HA, MM, and SS (from CAP3).

APPENDIX D

GRAPHS

Some of the graphs have a two-scale x-axis. The two-scale x-axis is used when there is a large difference among the results obtained for the various organisms. For example, if the number of contigs obtained for AT are between 20-80, and for SS between 2300-14000, a two-scale axis would be more appropriate to represent the results on the graph.

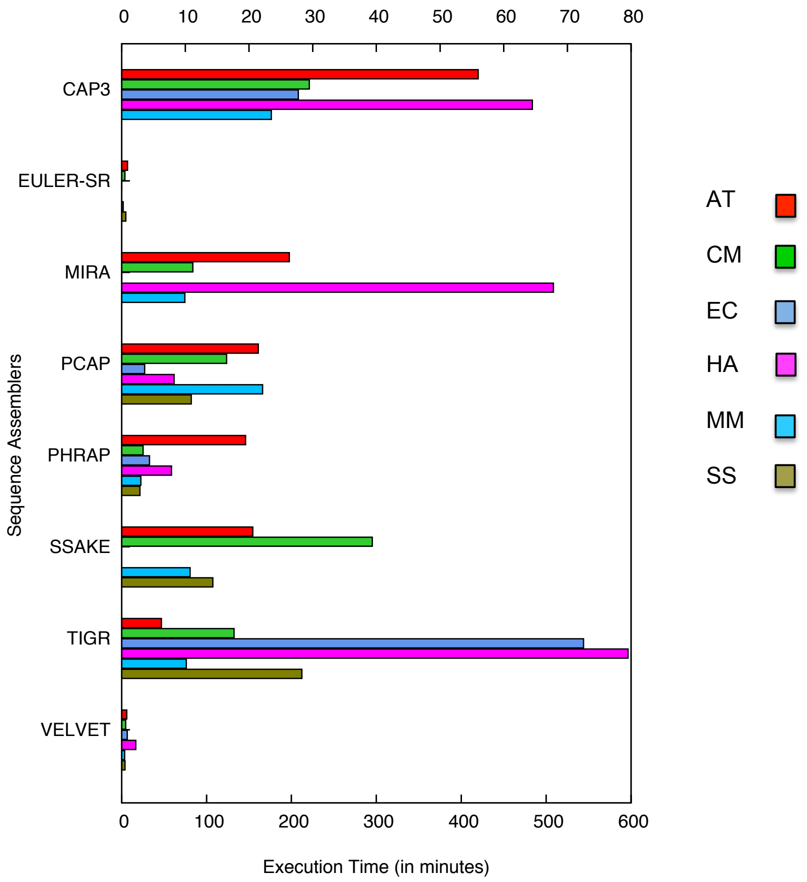


Figure D.1: Execution time results from running short- and long-read assemblers on Sanger data. See the appendix introduction for an explanation of the two-scale axis. Top x-axis: AT, CM, HA, MM, and SS; Bottom x-axis: EC.

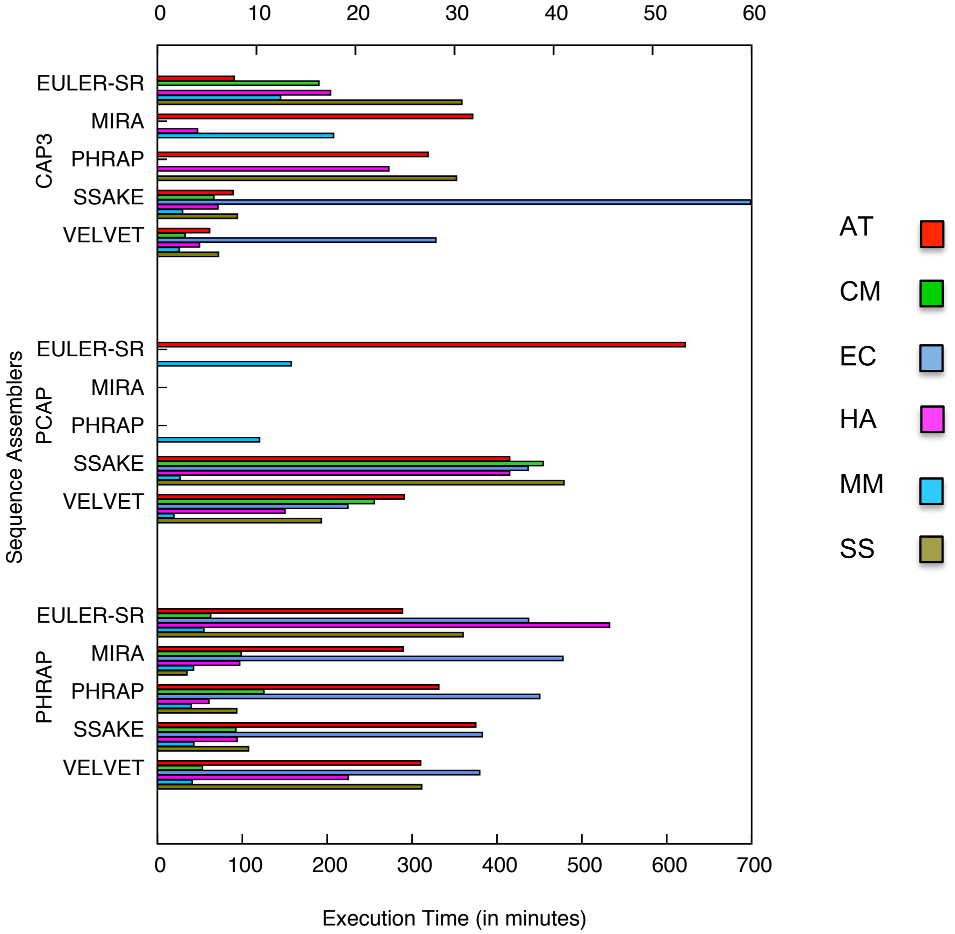


Figure D.2: Execution time results from running long-read assemblers on 454 contigs + Sanger reads. See the appendix introduction for an explanation of the two-scale axis. Bottom x-axis: All organisms (CAP3), EC and MM (from PCAP).

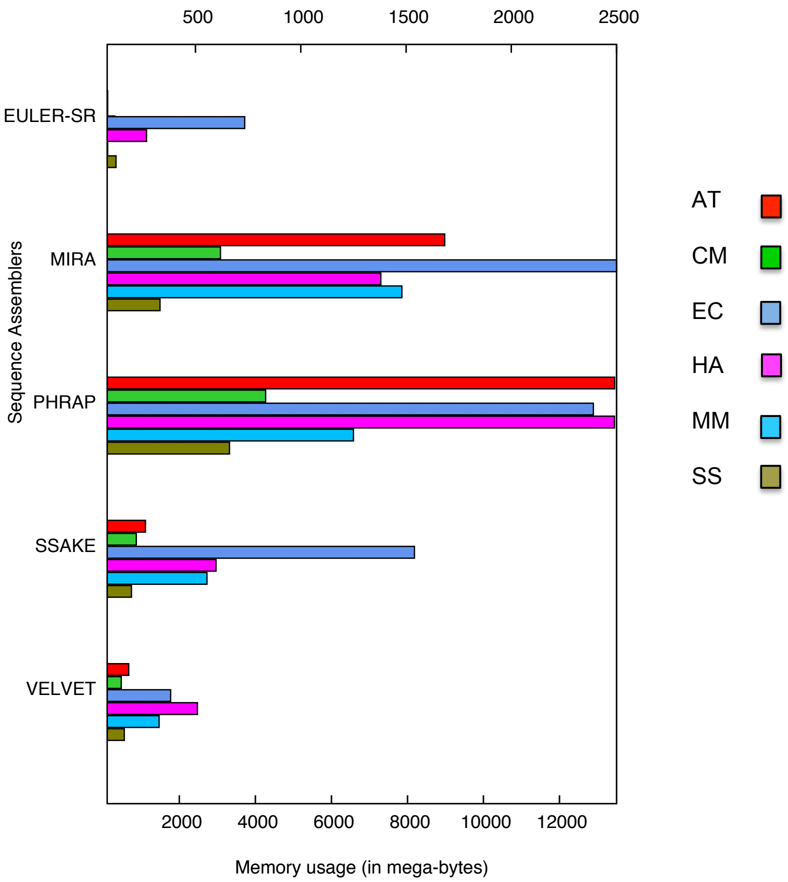


Figure D.3: Memory usage results from running short- and long-read assemblers on 454 data. See the appendix introduction for an explanation of the two-scale axis. Top x-axis: AT, CM, and SS; Bottom x-axis: HA, MM, and EC.

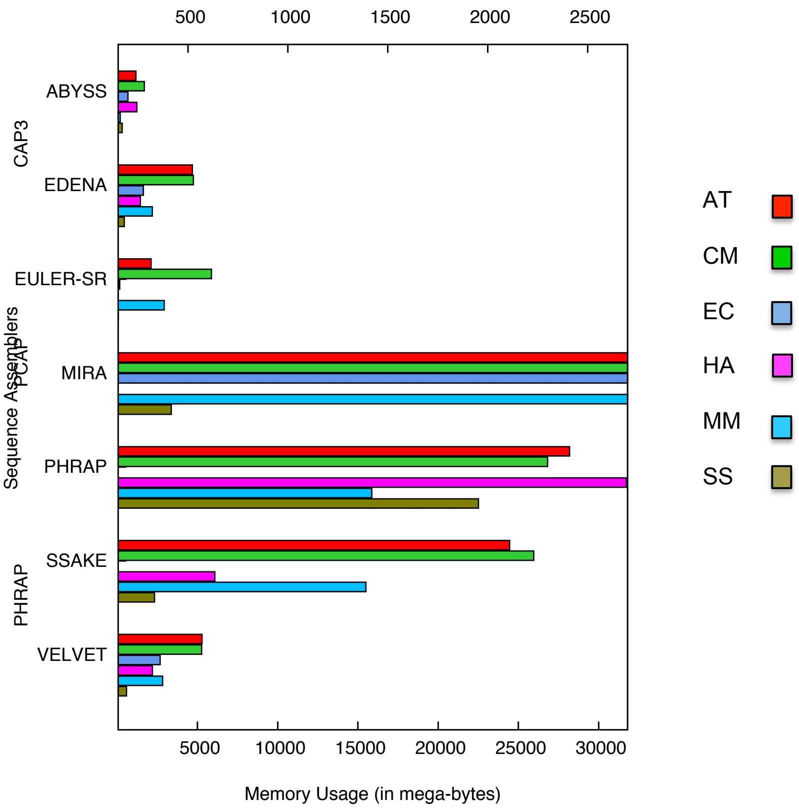


Figure D.4: Memory usage results from running short- and long-read assemblers on Illumina data. See the appendix introduction for an explanation of the two-scale axis. Top x-axis: AT, CM, and MM; Bottom x-axis: EC, HA, and SS.

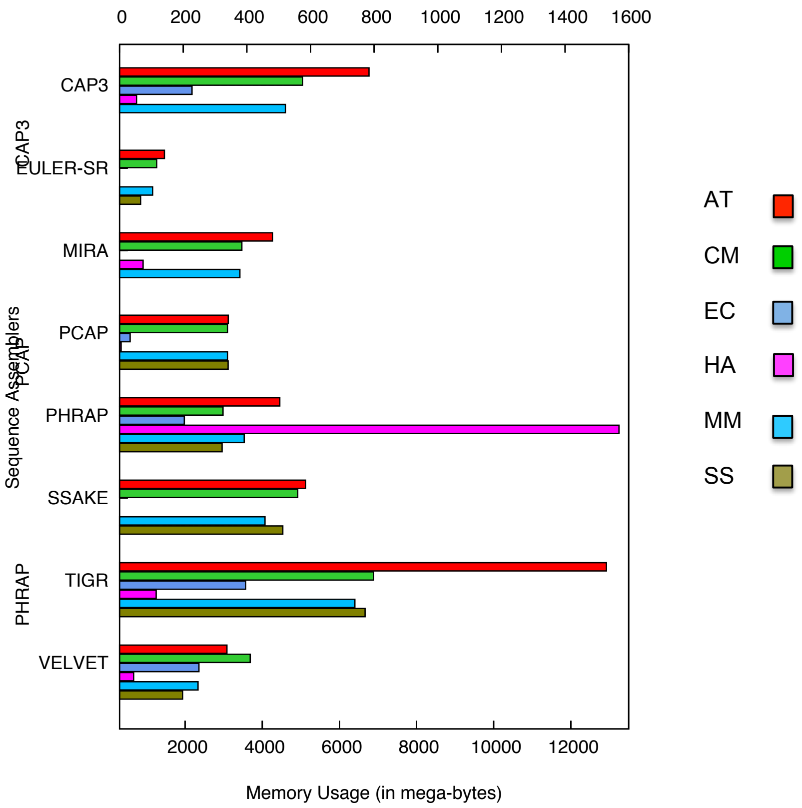


Figure D.5: Memory usage results from running short- and long-read assemblers on Sanger data. See the appendix introduction for an explanation of the two-scale axis. Top x-axis: AT, CM, MM, and SS; Bottom x-axis: EC and HA.

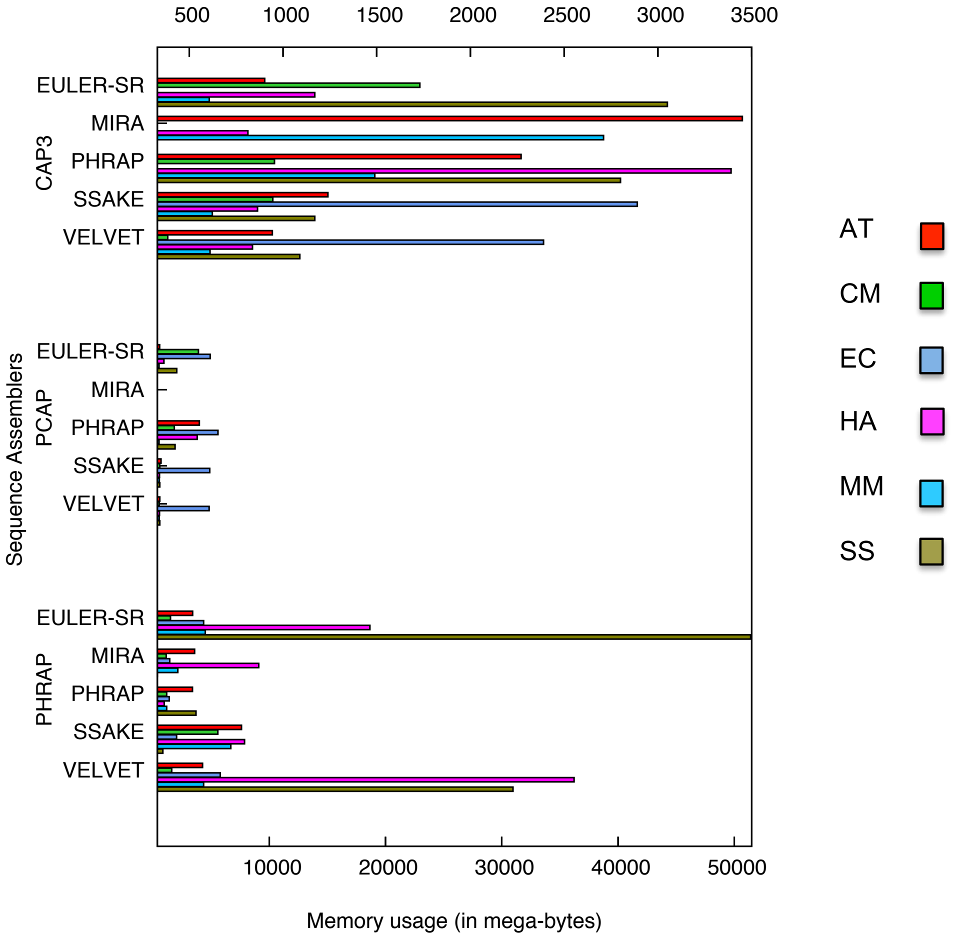


Figure D.6: Memory usage results from running long-read assemblers on 454 contigs + Sanger reads. See the appendix introduction for an explanation of the two-scale axis. Bottom x-axis: EC, HA, and SS (all from PHRAP).

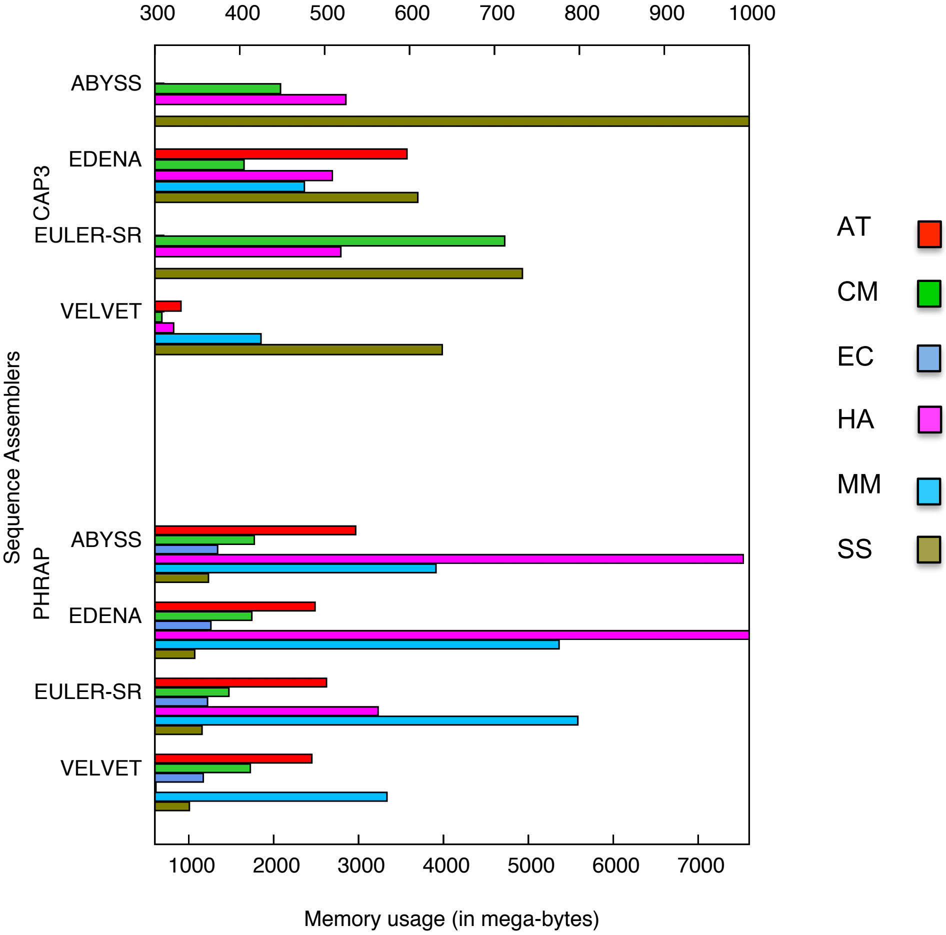


Figure D.7: Memory usage results from running long-read assemblers on Illumina contigs + Sanger reads. See the appendix introduction for an explanation of the two-scale axis. Bottom x-axis: AT, CM, HA, and MM (from CAP3); EC and HA (from PHRAP).

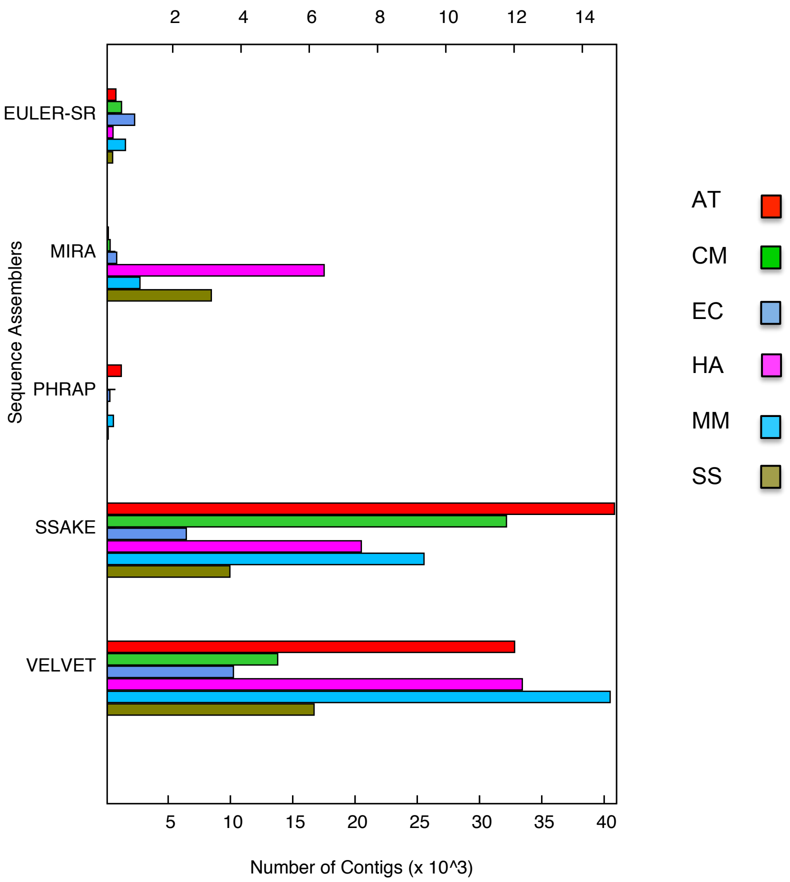


Figure D.8: Number of contigs results from running short- and long-read assemblers on 454 data. See the appendix introduction for an explanation of the two-scale axis. Top x-axis: EC, HA, and MM; Bottom x-axis: AT, CM, and SS.

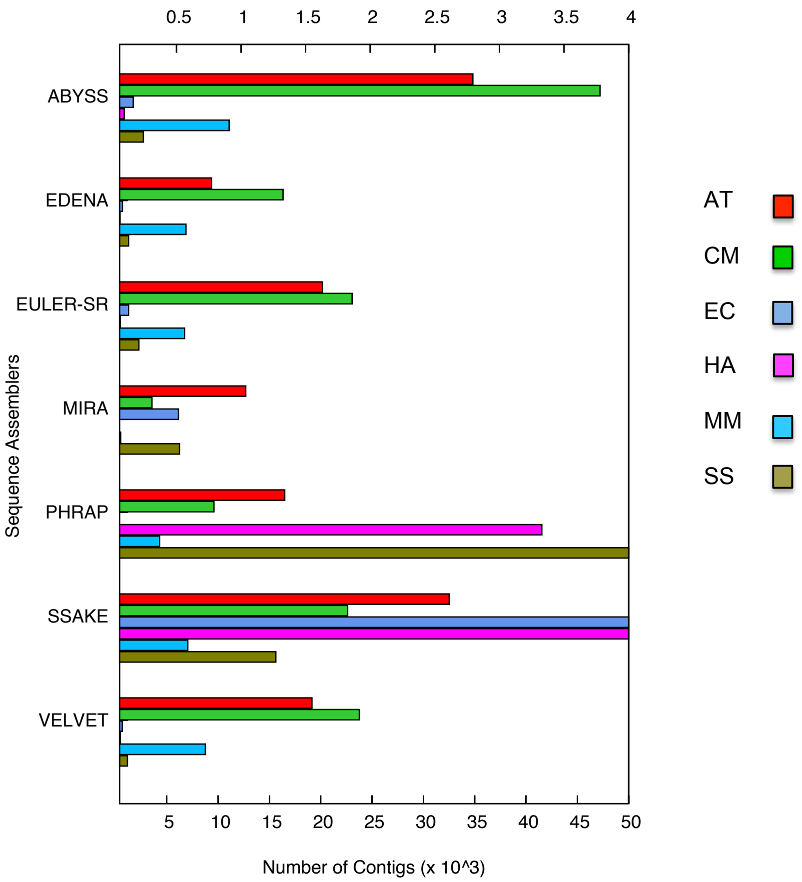


Figure D.9: Number of contigs results from running short- and long-read assemblers on Illumina data. See the appendix introduction for an explanation of the two-scale axis. Top x-axis: AT, CM, and MM; Bottom x-axis: EC, HA, and SS.

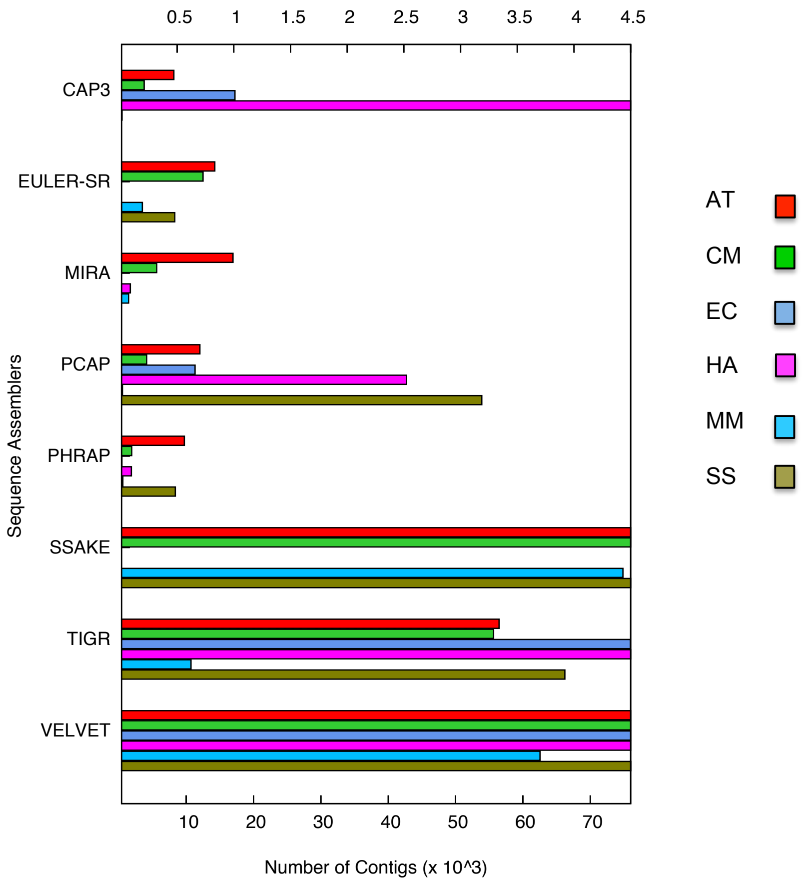


Figure D.10: Number of contigs results from running short- and long-read assemblers on Sanger data. See the appendix introduction for an explanation of the two-scale axis. Top x-axis: AT, CM, MM, HA, and SS; Bottom x-axis: EC.

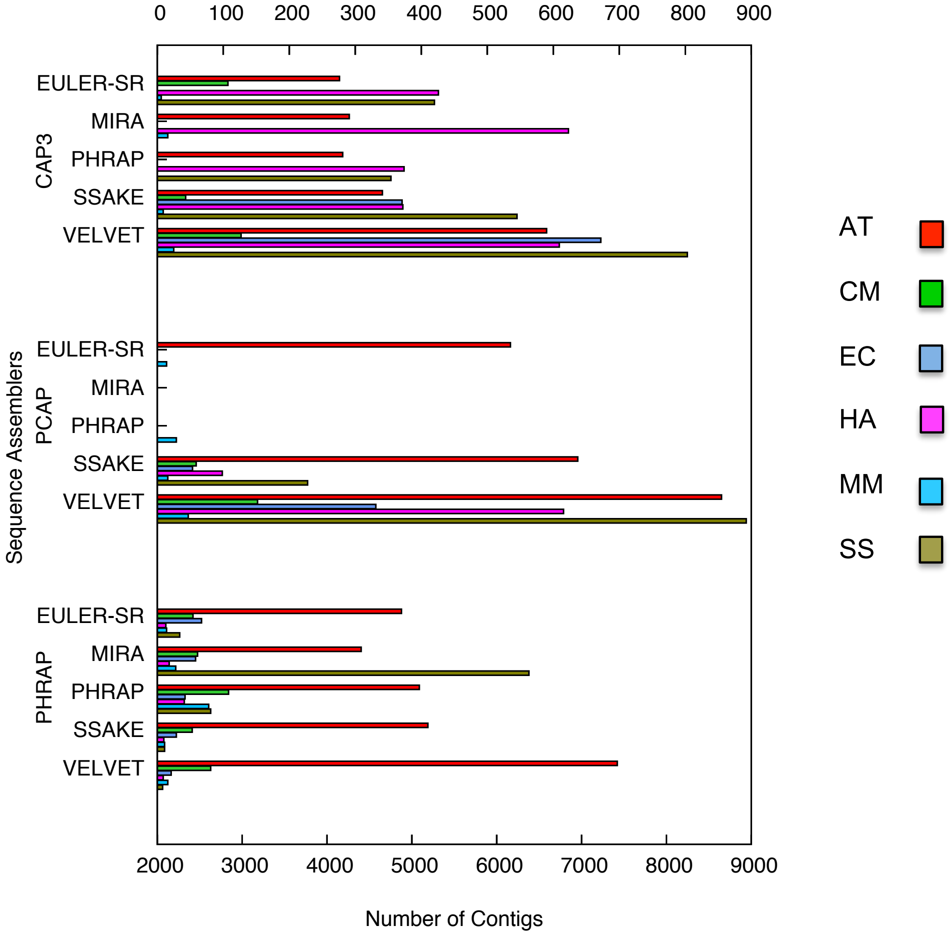


Figure D.11: Number of contigs results from running long-read assemblers on 454 contigs + Sanger reads. See the appendix introduction for an explanation of the two-scale axis. Bottom x-axis: EC and SS (from PCAP).

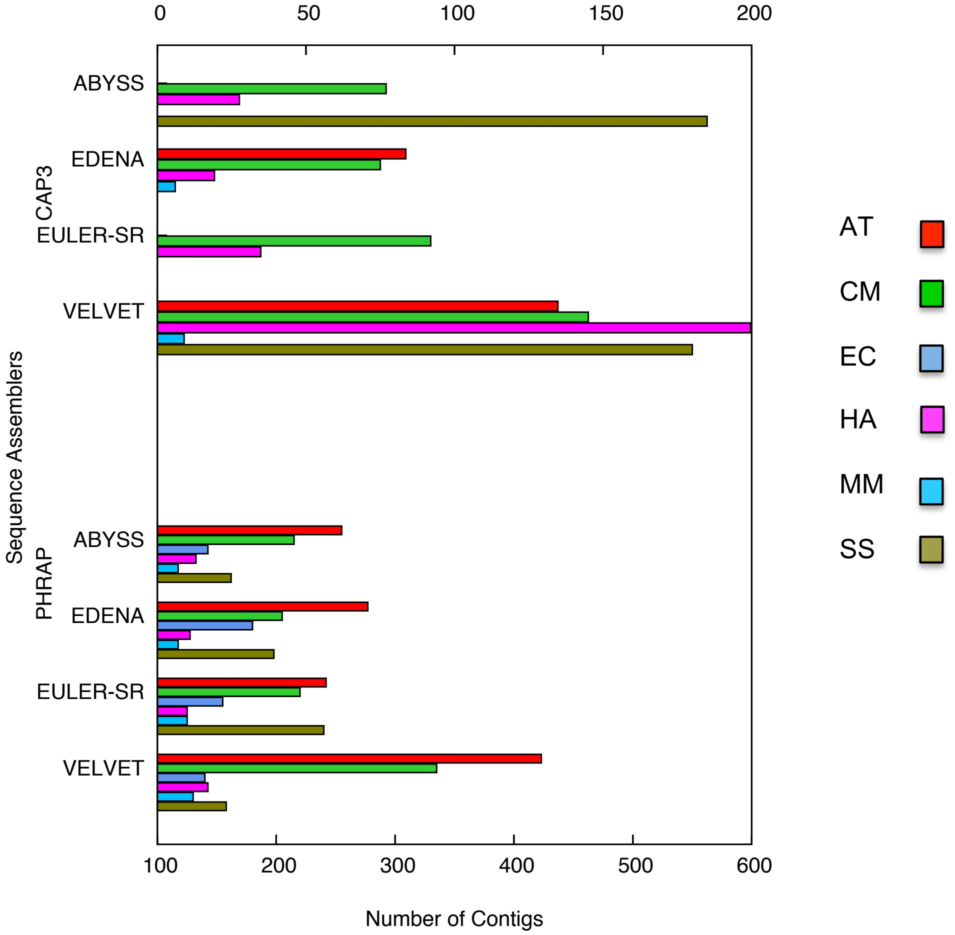


Figure D.12: Number of contigs results from running long-read assemblers on Illumina contigs + Sanger reads. See the appendix introduction for an explanation of the two-scale axis. Bottom x-axis: AT and HA (from CAP3), AT and SS (from PHRAP).

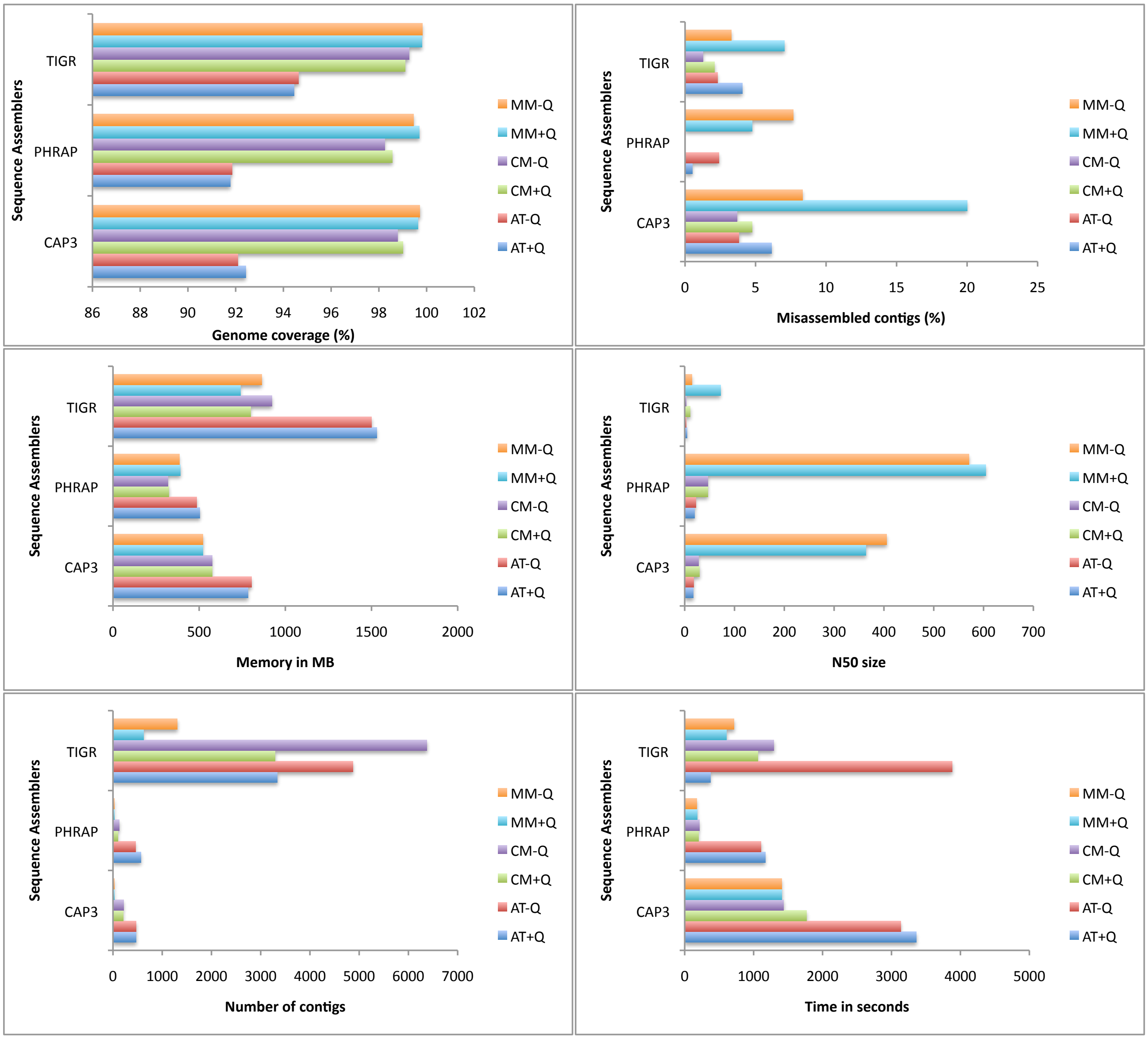


Figure D.13: Effect of the absence ("- Q") or presence ("+ Q") of quality values when assembling Sanger data.

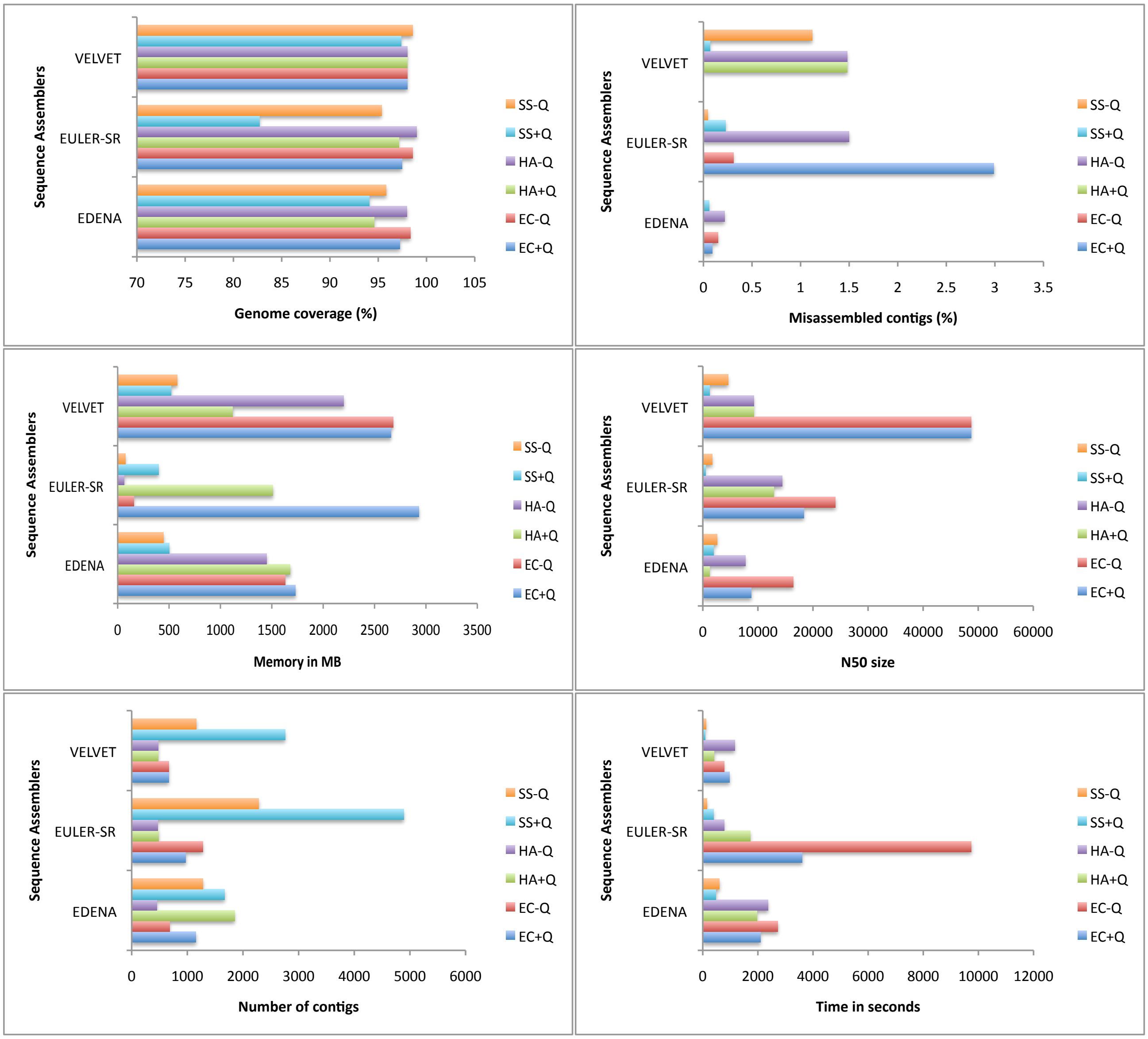


Figure D.14: Effect of the absence ("- Q") or presence ("+ Q") of quality values when assembling Illumina data.

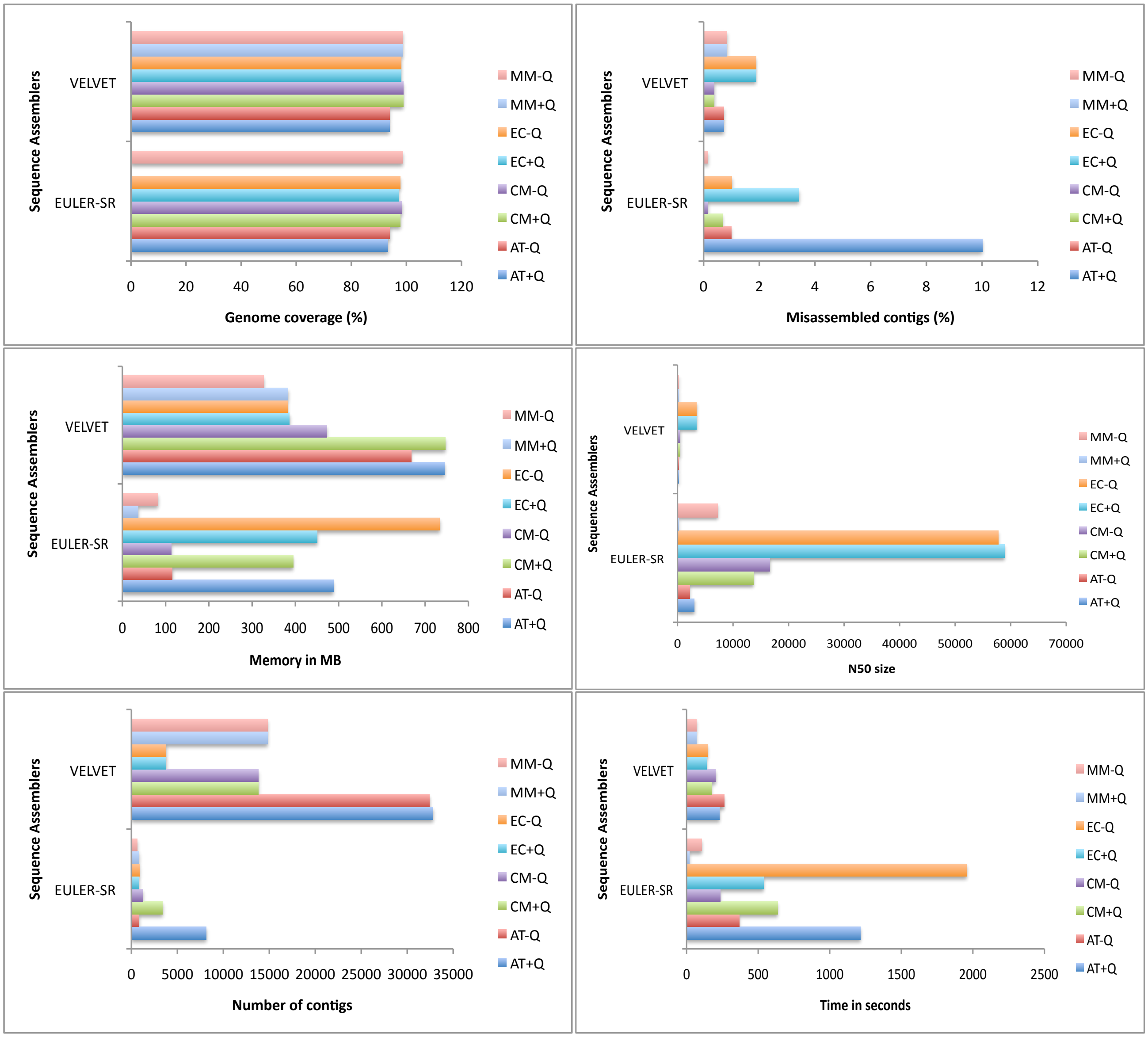


Figure D.15: Effect of the absence ("- Q") or presence ("+ Q") of quality values when assembling 454 data.

APPENDIX E

STATISTICAL RESULTS

The acronyms and abbreviations used in the column and row headings of the tables below include: GC - genome coverage, MC - misassembled contigs, NC - number of contigs, Mem. - Memory, df - degrees of freedom, Sig. - significance values, F - F-statistic, t - t-statistic, and Std. - Standard.

Table E.1: One-way ANOVA table for analyzing the metrics of the assemblers that ran on 454 reads.

		Sum of Squares	df	Mean Square	F	Sig.
NC	Between Groups	1.663E9	4	4.158E8	6.008	0.002
	Within Groups	1.730E9	25	6.920E7		
	Total	3.393E9	29			
GC	Between Groups	49.216	4	12.304	1.192	0.340
	Within Groups	247.639	24	10.318		
	Total	296.855	28			
MC	Between Groups	1858.803	4	464.701	26.510	<0.001
	Within Groups	420.705	24	17.529		
	Total	2279.508	28			
N50	Between Groups	1.436E10	4	3.589E9	2.638	0.058
	Within Groups	3.401E10	25	1.360E9		
	Total	4.836E10	29			
Time	Between Groups	2.929E8	4	7.322E7	3.583	0.020
	Within Groups	4.904E8	24	2.044E7		
	Total	7.833E8	28			
Mem.	Between Groups	7.349E7	4	1.837E7	4.212	0.010
	Within Groups	1.047E8	24	4361745.859		
	Total	1.782E8	28			

Table E.2: Group statistics for the data in Table E.1.

Number of contigs				
Assemblers	N	Subset for alpha = 0.05		
		1	2	3
EULER-SR	6	726		
MIRA	6	2804.83	2804.83	
PHRAP	6	340.67		
SSAKE	6			17000
VELVET	6		15700	15700
Sig.		0.985	0.085	0.999

Genome coverage		
Assemblers	N	Subset for alpha = 0.05
		1
EULER-SR	6	97.4033
MIRA	5	95.08
PHRAP	6	94.8317
SSAKE	6	97.7783
VELVET	6	97.6567
Sig.		0.537

Misassembled contigs			
Assemblers	N	Subset for alpha = 0.05	
		1	2
EULER-SR	6	0.6683	
MIRA	6	7.5017	
PHRAP	5		22.21
SSAKE	6	0.3517	
VELVET	6	0.6517	
Sig.		0.055	1

N50 Size		
Assemblers	N	Subset for alpha = 0.05
		1
EULER-SR	6	26343
MIRA	6	44471.17
PHRAP	6	54622
SSAKE	6	1536.83
VELVET	6	807.5
Sig.		0.116

Time in seconds		
Assemblers	N	Subset for alpha = 0.05
		1
EULER-SR	6	672.5317
MIRA	5	7845.998
PHRAP	6	6600.2467
SSAKE	6	5990.5083
VELVET	6	156.575
Sig.		0.056

Memory in MB			
Assemblers	N	Subset for alpha = 0.05	
		1	2
EULER-SR	6	274.1783	
MIRA	5	2078	2078
PHRAP	6		4528.3
SSAKE	6	901.845	
VELVET	6	484.5533	
Sig.		0.593	0.299

Table E.3: One-way ANOVA table for analyzing the metrics of the assemblers that ran on Illumina reads.

		Sum of Squares	df	Mean Square	F	Sig.
NC	Between Groups	1.861E10	6	3.101E9	1.965	0.099
	Within Groups	5.209E10	33	1.579E9		
	Total	7.070E10	39			
GC	Between Groups	1948.809	6	324.802	1.316	0.283
	Within Groups	6912.360	28	246.870		
	Total	8861.170	34			
MC	Between Groups	40.144	6	6.691	1.321	0.281
	Within Groups	141.793	28	5.064		
	Total	181.938	34			
N50	Between Groups	4.994E10	6	8.323E9	2.293	0.058
	Within Groups	1.198E11	33	3.629E9		
	Total	1.697E11	39			
Time	Between Groups	4.310E10	6	7.184E9	4.551	0.002
	Within Groups	5.051E10	32	1.578E9		
	Total	9.361E10	38			
Mem.	Between Groups	8.217E8	6	1.370E8	2.782	0.027
	Within Groups	1.575E9	32	4.922E7		
	Total	2.397E9	38			

Table E.4: Group statistics for Table E.3.

Number of contigs		
Assemblers	N	Subset for alpha = 0.05
		1
ABYSS	6	2131.67
EDENA	6	847
EULER-SR	6	1346.17
MIRA	5	2760.2
PHRAP	5	26433
SSAKE	6	61191.5
VELVET	6	1081.5
Sig.		0.172

Genome coverage		
Assemblers	N	Subset for alpha = 0.05
		1
ABYSS	6	97.9633
EDENA	6	96.7733
EULER-SR	6	97.425
MIRA	5	76.258
PHRAP	2	98.27
SSAKE	4	97.85
VELVET	6	97.7283
Sig.		0.401

Misassembled contigs		
Assemblers	N	Subset for alpha = 0.05
		1
ABYSS	6	2.7033
EDENA	6	0.075
EULER-SR	6	0.309
MIRA	5	2.382
PHRAP	2	0.32
SSAKE	4	0.235
VELVET	6	0.4333
Sig.		0.61

N50 size			
Assemblers	N	Subset for alpha = 0.05	
		1	2
ABYSS	6	13843.17	13843.17
EDENA	6	12093.17	12093.17
EULER-SR	6	19081.17	19081.17
MIRA	5		118000
PHRAP	5	46.2	
SSAKE	6	11601.83	11601.83
VELVET	6	15099.67	15099.67
Sig.		0.998	0.072

Continued on next page

Table E.4 - continued from previous page

Time in seconds			
Assemblers	N	Subset for alpha = 0.05	
		1	2
ABYSS	6	520.9633	
EDENA	6	1122.5	
EULER-SR	6	2106.1	
MIRA	5		102090
PHRAP	5	25310	
SSAKE	5	7410.6	
VELVET	6	391.6983	
Sig.		0.94	1

Memory in MB		
Assemblers	N	Subset for alpha = 0.05
		1
ABYSS	6	482.3183
EULER-SR	6	267.885
EDENA	6	815.9617
MIRA	5	10716
PHRAP	5	12072
SSAKE	5	2830
VELVET	6	1161.6617
Sig.		0.107

Table E.5: One-way ANOVA table for analyzing the metrics of the assemblers that ran on Sanger reads.

		Sum of Squares	df	Mean Square	F	Sig.
NC	Between Groups	3.783E10	7	5.404E9	1.568	0.181
	Within Groups	1.103E11	32	3.448E9		
	Total	1.481E11	39			
GC	Between Groups	2818.199	7	402.600	7.828	<0.001
	Within Groups	1542.852	30	51.428		
	Total	4361.051	37			
MC	Between Groups	311.183	7	44.455	1.457	0.220
	Within Groups	915.489	30	30.516		
	Total	1226.672	37			
N50	Between Groups	3.675E11	7	5.250E10	1.916	0.100
	Within Groups	8.770E11	32	2.741E10		
	Total	1.244E12	39			
Time	Between Groups	2.290E8	7	3.272E7	1.156	0.355
	Within Groups	9.059E8	32	2.831E7		
	Total	1.135E9	39			
Mem.	Between Groups	2.814E7	7	4020713.065	0.889	0.526
	Within Groups	1.447E8	32	4521394.399		
	Total	1.728E8	39			

Table E.6: Group statistics for the data in Table E.5.

Number of contigs		
Assemblers	N	Subset for alpha = 0.05
		1
CAP3	5	4725.6
EULER-SR	4	558.75
MIRA	4	368
PCAP	5	1332.2
PHRAP	6	288.17
SSAKE	4	7552
TIGR Assembler	6	18767.5
VELVET	6	89620.33

Genome coverage				
Assemblers	N	Subset for alpha = 0.05		
		1	2	3
CAP3	5			98.208
EULER-SR	4	81.58	81.58	
MIRA	4			97.54
PCAP	5			95.774
PHRAP	6			95.4117
SSAKE	4	70.9975		
TIGR	6			97.1617
VELVET	4		92.8875	92.8875

Misassembled contigs		
Assemblers	N	Subset for alpha = 0.05
		1
CAP3	5	6.448
EULER-SR	4	0.595
MIRA	4	6.94
PCAP	5	6.174
PHRAP	6	7.1717
SSAKE	4	0.1625
TIGR Assembler	6	3.2083
VELVET	4	0.3125

N50 size		
Assemblers	N	Subset for alpha = 0.05
		1
CAP3	5	83435
EULER-SR	4	10191.25
MIRA	4	327550.5
PCAP	5	70598.6
PHRAP	6	130482.83
SSAKE	4	818.25
TIGR Assembler	6	15323.33
VELVET	6	527.67

Continued on next page

Table E.6 - continued from previous page

Time in seconds		
Assemblers	N	Subset for alpha = 0.05
		1
CAP3	5	4578.988
EULER-SR	4	34.95
MIRA	4	1728.6375
PCAP	5	951.178
PHRAP	6	693.59
SSAKE	4	1275.485
TIGR Assembler	6	6859.1517
VELVET	6	113.0817

Memory in MB		
Assemblers	N	Subset for alpha = 0.05
		1
CAP3	5	960.614
EULER-SR	4	106.9225
MIRA	4	538.7675
PCAP	5	340.17
PHRAP	6	2795.4233
SSAKE	4	528.59
TIGR Assembler	6	1443.1983
VELVET	6	703.215

Table E.7: Table of t-test results for analyzing the metrics of the EULER-SR and VELVET assemblers that ran on 454 reads merged with Sanger reads.

Independent Samples Test						
		Levene's Test for Equality of Variances		t-test for Equality of Means		
Metrics	Equal variances	F	Sig.	t	df	Sig. (2-tailed)
NC	assumed	8.062	0.018	-2.373	10	0.039
	not assumed			-2.373	5.001	0.064
GC	assumed	16.029	0.005	3.578	7	0.009
	not assumed			2.406	2.056	0.135
MC	assumed	0.408	0.543	-1.606	7	0.152
	not assumed			-1.433	3.145	0.243
N50	assumed	11.502	0.007	4.847	10	0.001
	not assumed			4.847	5.000	0.005
Time	assumed	2.072	0.181	1.084	10	0.304
	not assumed			1.084	9.057	0.306
Mem.	assumed	4.875	0.052	-8.754	10	<0.001
	not assumed			-8.754	5.436	<0.001

Table E.8: Group statistics for the data in Table E.7.

	Assembler	N	Mean	Std. Deviation	Std. Error Mean
NC	EULER-SR	6	946.83	846.463	345.567
	VELVET	6	1.14E5	117110.898	47810.324
GC	EULER-SR	6	97.4650	1.92531	0.78601
	VELVET	3	81.2733	11.57879	6.68502
MC	EULER-SR	6	1.4900	1.26063	0.51465
	VELVET	3	3.0833	1.70735	0.98574
N50	EULER-SR	6	3.53E4	17754.780	7248.359
	VELVET	6	160.00	69.415	28.338
Time	EULER-SR	6	3.9479E2	228.98647	93.48334
	VELVET	6	2.7013E2	163.85983	66.89550
Mem.	EULER-SR	6	1.2878E2	35.52532	14.50315
	VELVET	6	7.4911E2	169.89794	69.36054

Table E.9: Table of t-test results for analyzing the metrics of the EULER-SR and VELVET assemblers that ran on Illumina reads merged with Sanger reads.

Independent Samples Test						
		Levene's Test for Equality of Variances		t-test for Equality of Means		
Metrics	Equal variances	F	Sig.	t	df	Sig. (2-tailed)
NC	assumed	7.606	0.022	-1.694	9	0.125
	not assumed			-1.869	5.113	0.119
GC	assumed	9.395	0.018	1.187	7	0.274
	not assumed			1.045	3.035	0.372
MC	assumed	1.682	0.236	-0.580	7	0.580
	not assumed			-0.539	4.279	0.617
N50	assumed	7.450	0.023	2.071	9	0.068
	not assumed			1.870	4.002	0.135
Time	assumed	15.195	0.004	1.852	9	0.097
	not assumed			1.675	4.073	0.168
Mem.	assumed	0.494	0.500	0.347	9	0.737
	not assumed			0.340	7.732	0.743

Table E.10: Group statistics for the data in Table E.9.

	Assembler	N	Mean	Std. Deviation	Std. Error Mean
NC	EULER-SR	5	1.03E4	13389.161	5987.815
	VELVET	6	1.16E5	137826.210	56267.314
GC	EULER-SR	5	97.7220	1.53436	.68619
	VELVET	4	88.3550	17.87497	8.93748
MC	EULER-SR	5	0.8100	0.79853	0.35711
	VELVET	4	1.2675	1.54131	0.77065
N50	EULER-SR	5	4.04E4	47497.461	21241.511
	VELVET	6	684.50	760.821	310.604
Time	EULER-SR	5	3.6621E3	4233.32668	1893.20125
	VELVET	6	4.7602E2	441.98674	180.44033
Mem.	EULER-SR	5	1.4583E3	1295.06386	579.17016
	VELVET	6	1.2135E3	1051.02718	429.08005

Table E.11: Table of t-test results for analyzing the metrics of the EULER-SR and VELVET assemblers that ran on Illumina reads merged with 454 reads.

Independent Samples Test						
		Levene's Test for Equality of Variances		t-test for Equality of Means		
Metrics	Equal variances	F	Sig.	t	df	Sig. (2-tailed)
NC	assumed	2.280	0.162	-3.569	10	0.005
	not assumed			-3.569	5.406	0.014
GC	assumed	0.118	0.739	-0.487	10	0.637
	not assumed			-0.487	9.999	0.637
MC	assumed	5.401	0.042	1.286	10	0.227
	not assumed			1.286	5.057	0.254
N50	assumed	10.411	0.009	2.578	10	0.027
	not assumed			2.578	5.025	0.049
Time	assumed	16.430	0.002	2.635	10	0.025
	not assumed			2.635	5.250	0.049
Mem.	assumed	0.052	0.824	-0.467	10	0.651
	not assumed			-0.467	9.709	0.651

Table E.12: Group statistics for the data in Table E.11.

	Assembler	N	Mean	Std. Deviation	Std. Error Mean
NC	EULER-SR	6	2531.50	2186.022	892.440
	VELVET	6	1.86E4	10841.674	4426.095
GC	EULER-SR	6	95.8533	2.83384	1.15691
	VELVET	6	96.6533	2.85483	1.16548
MC	EULER-SR	6	4.5700	7.63282	3.11609
	VELVET	6	0.5500	0.57630	0.23527
N50	EULER-SR	6	2.34E4	21600.705	8818.451
	VELVET	6	667.50	1074.574	438.693
Time	EULER-SR	6	2.6715E3	1974.13170	805.93589
	VELVET	6	5.2141E2	312.43572	127.55135
Mem.	EULER-SR	6	1.4854E3	1201.57615	490.54141
	VELVET	6	1.7843E3	1008.87424	411.87119

Table E.13: Table of t-test for results for analyzing the metrics of the merged EULER-SR 454 and Illumina contigs, and merged EULER-SR Illumina contigs and PHRAP 454 contigs.

Independent Samples Test						
		Levene's Test for Equality of Variances		t-test for Equality of Means		
Metrics	Equal variances	F	Sig.	t	df	Sig. (2-tailed)
NC	assumed	3.300	0.099	1.919	10	0.084
	not assumed			1.919	6.934	0.097
GC	assumed	0.411	0.536	0.214	10	0.835
	not assumed			0.214	9.138	0.836
MC	assumed	1.187	0.301	-5.718	10	0.000
	not assumed			-5.718	7.596	0.001
N50	assumed	11.906	0.006	-1.153	10	0.276
	not assumed			-1.153	5.601	0.296
Time	assumed	2.730	0.129	-0.738	10	0.478
	not assumed			-0.738	7.257	0.484
Mem.	assumed	4.197	0.068	0.772	10	0.458
	not assumed			0.772	5.179	0.474

Table E.14: Group statistics for the data in Table E.13.

	Assembler	N	Mean	Std. Deviation	Std. Error Mean
NC	EULER-SR	6	208.17	122.447	49.989
	VELVET	6	103.00	54.933	22.426
GC	EULER-SR	6	97.7867	1.09171	0.44569
	VELVET	6	97.6250	1.49937	0.61212
MC	EULER-SR	6	1.9533	2.45039	1.00037
	VELVET	6	14.1833	4.63101	1.89060
N50	EULER-SR	6	5.75E4	22014.443	8987.359
	VELVET	6	1.01E5	89617.581	36586.224
Time	EULER-SR	6	2.3139E2	293.74966	119.92280
	VELVET	6	4.3296E2	601.38815	245.51568
Mem.	EULER-SR	6	2.3502E2	286.61537	117.01024
	VELVET	6	1.4390E2	38.37013	15.66454

Table E.15: Table of t-test for results for analyzing the metrics of the EULER-SR and VELVET assemblers that ran on merged Illumina, 454, and Sanger reads.

Independent Samples Test						
		Levene's Test for Equality of Variances		t-test for Equality of Means		
Metrics	Equal variances	F	Sig.	t	df	Sig. (2-tailed)
NC	assumed	5.309	0.044	-1.536	10	0.156
	not assumed			-1.536	5.249	0.182
GC	assumed	1.638	0.236	-0.821	8	0.435
	not assumed			-0.991	6.179	0.359
MC	assumed	2.473	0.154	1.021	8	0.337
	not assumed			1.245	5.846	0.261
N50	assumed	13.929	0.004	1.952	10	0.079
	not assumed			1.952	5.000	0.108
Time	assumed	10.903	0.008	2.074	10	0.065
	not assumed			2.074	5.071	0.092
Mem.	assumed	0.465	0.511	0.333	10	0.746
	not assumed			0.333	9.145	0.746

Table E.16: Group statistics for the data in Table E.15.

	Assembler	N	Mean	Std. Deviation	Std. Error Mean
NC	EULER-SR	6	1.45E4	19366.319	7906.267
	VELVET	6	9.23E4	122615.270	50057.474
GC	EULER-SR	6	94.6283	7.58497	3.09655
	VELVET	4	97.8850	2.20140	1.10070
MC	EULER-SR	6	1.5600	2.05580	.83928
	VELVET	4	0.4700	0.49887	0.24943
N50	EULER-SR	6	1.76E4	21949.283	8960.757
	VELVET	6	146.00	30.239	12.345
Time	EULER-SR	6	4.2535E3	4444.93786	1814.63828
	VELVET	6	4.7635E2	373.46033	152.46454
Mem.	EULER-SR	6	1.5833E3	1244.38166	508.01669
	VELVET	6	1.3737E3	907.42193	370.45345

Table E.17: Games-Howell post-hoc output for analyzing the performance of each assembler that ran on 454 data based on the test metrics.

	Assemblers (I)	Assemblers (J)	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
NC	EULERSR	MIRA	-2078.833	1503.304	0.661	-8067.5	3909.83
		PHRAP	385.333	227.924	0.482	-373.66	1144.32
		SSAKE	-16313.5	6349.292	0.212	-41773.68	9146.68
		VELVET	-14967.833	3885.955	0.058	-30539.99	604.32
	MIRA	EULERSR	2078.833	1503.304	0.661	-3909.83	8067.5
		PHRAP	2464.167	1507.561	0.536	-3514.93	8443.27
		SSAKE	-14234.667	6521.832	0.306	-39363.18	10893.85
		VELVET	-12889	4161.903	0.098	-28141.08	2363.08
	PHRAP	EULERSR	-385.333	227.924	0.482	-1144.32	373.66
		MIRA	-2464.167	1507.561	0.536	-8443.27	3514.93
		SSAKE	-16698.833	6350.301	0.199	-42156.5	8758.84
		VELVET	-15353.167	3887.603	0.053	-30921.26	214.93
	SSAKE	EULERSR	16313.5	6349.292	0.212	-9146.68	41773.68
		MIRA	14234.667	6521.832	0.306	-10893.85	39363.18
		PHRAP	16698.833	6350.301	0.199	-8758.84	42156.5
		VELVET	1345.667	7441.439	1	-24149.3	26840.64
	VELVET	EULERSR	14967.833	3885.955	0.058	-604.32	30539.99
		MIRA	12889	4161.903	0.098	-2363.08	28141.08
		PHRAP	15353.167	3887.603	0.053	-214.93	30921.26
		SSAKE	-1345.667	7441.439	1	-26840.64	24149.3
GC	EULERSR	MIRA	2.32333	2.91281	0.920	-9.8992	14.5459
		PHRAP	2.57167	1.46969	0.460	-2.5431	7.6864
		SSAKE	-0.37500	0.82426	0.989	-3.2089	2.4589
		VELVET	-0.25333	1.04634	0.999	-3.7017	3.1950
	MIRA	EULERSR	-2.32333	2.91281	0.920	-14.5459	9.8992
		PHRAP	0.24833	3.10524	1.000	-11.6503	12.1470
		SSAKE	-2.69833	2.85691	0.867	-15.1198	9.7232
		VELVET	-2.57667	2.92872	0.892	-14.7530	9.5996
	PHRAP	EULERSR	-2.57167	1.46969	0.460	-7.6864	2.5431
		MIRA	-0.24833	3.10524	1.000	-12.1470	11.6503
		SSAKE	-2.94667	1.35552	0.301	-8.0138	2.1205
		VELVET	-2.82500	1.50096	0.395	-7.9837	2.3337
	SSAKE	EULERSR	.37500	0.82426	0.989	-2.4589	3.2089
		MIRA	2.69833	2.85691	0.867	-9.7232	15.1198
		PHRAP	2.94667	1.35552	0.301	-2.1205	8.0138
		VELVET	0.12167	0.87880	1.000	-2.9378	3.1812
	VELVET	EULERSR	0.25333	1.04634	0.999	-3.1950	3.7017
		MIRA	2.57667	2.92872	0.892	-9.5996	14.7530
		PHRAP	2.82500	1.50096	0.395	-2.3337	7.9837
		SSAKE	-0.12167	0.87880	1.000	-3.1812	2.9378
MC	EULERSR	MIRA	-6.83333	2.28456	0.134	-15.8948	2.2282
		PHRAP	-21.54167*	3.62381	0.018	-37.5584	-5.525
		SSAKE	0.31667	0.30561	0.832	-0.7573	1.3906
		VELVET	0.01667	0.39114	1	-1.2709	1.3043
	MIRA	EULERSR	6.83333	2.28456	0.134	-2.2282	15.8948
		PHRAP	-14.70833	4.26657	0.06	-30.0275	0.6108
		SSAKE	7.15	2.27266	0.116	-1.9389	16.2389

Continued on next page

Table E.17 – continued from previous page

						95% Confidence Interval	
	Assemblers (I)	Assemblers (J)	Mean Difference (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
MC	MIRA	VELVET	6.85	2.28574	0.133	-2.2089	15.9089
	PHRAP	EULERSR	21.54167*	3.62381	0.018	5.525	37.5584
		MIRA	14.70833	4.26657	0.06	-0.6108	30.0275
		SSAKE	21.85833*	3.61633	0.017	5.8066	37.9101
		VELVET	21.55833*	3.62456	0.018	5.5451	37.5716
	SSAKE	EULERSR	-0.31667	0.30561	0.832	-1.3906	0.7573
		MIRA	-7.15	2.27266	0.116	-16.2389	1.9389
		PHRAP	-21.85833*	3.61633	0.017	-37.9101	-5.8066
		VELVET	-0.3	0.3143	0.867	-1.4104	0.8104
	VELVET	EULERSR	-0.01667	0.39114	1	-1.3043	1.2709
		MIRA	-6.85	2.28574	0.133	-15.9089	2.2089
		PHRAP	-21.55833*	3.62456	0.018	-37.5716	-5.5451
		VELVET	0.3	0.3143	0.867	-0.8104	1.4104
N50	EULERSR	MIRA	-18128.167	23490	0.931	-103319.79	67063.46
		PHRAP	-28279	25720	0.802	-122902.61	66344.61
		SSAKE	24806.167	9090.292	0.176	-11229.24	60841.58
		VELVET	25535.5	9037.994	0.161	-10620.07	61691.07
	MIRA	EULERSR	18128.167	23490	0.931	-67063.46	103319.79
		PHRAP	-10150.833	32410	0.998	-117054.59	96752.92
		SSAKE	42934.333	21720	0.388	-43998.89	129867.56
		VELVET	43663.667	21690	0.375	-43323.44	130650.77
	PHRAP	EULERSR	28279	25720	0.802	-66344.61	122902.61
		MIRA	10150.833	32410	0.998	-96752.92	117054.59
		SSAKE	53085.167	24110	0.31	-43486.7	149657.03
		VELVET	53814.5	24100	0.3	-42806.02	150435.02
	SSAKE	EULERSR	-24806.167	9090.292	0.176	-60841.58	11229.24
		MIRA	-42934.333	21720	0.388	-129867.56	43998.89
		PHRAP	-53085.167	24110	0.31	-149657.03	43486.7
		VELVET	729.333	1228.01	0.972	-3635.75	5094.41
	VELVET	EULERSR	-25535.5	9037.994	0.161	-61691.07	10620.07
		MIRA	-43663.667	21690	0.375	-130650.77	43323.44
		PHRAP	-53814.5	24100	0.3	-150435.02	42806.02
		SSAKE	-729.333	1228.01	0.972	-5094.41	3635.75
Time	EULERSR	MIRA	-7173.46633*	1225.45	0.015	-12357.3489	-1989.5838
		PHRAP	-5927.715	3456.94	0.499	-19724.7208	7869.2908
		SSAKE	-5317.97667	1876.66	0.156	-12719.2111	2083.2578
		VELVET	515.95667	276.267	0.431	-583.415	1615.3283
	MIRA	EULERSR	7173.46633*	1225.45	0.015	1989.5838	12357.3489
		PHRAP	1245.75133	3647.06	0.996	-12315.7872	14807.2898
		SSAKE	1855.48967	2207.36	0.911	-5718.0688	9429.0482
		VELVET	7689.42300*	1194.54	0.014	2381.9323	12996.9137
	PHRAP	EULERSR	5927.715	3456.94	0.499	-7869.2908	19724.7208
		MIRA	-1245.75133	3647.06	0.996	-14807.2898	12315.7872
		SSAKE	609.73833	3914.23	1	-13052.9443	14272.421
		VELVET	6443.67167	3446.11	0.432	-7379.6728	20267.0161
	SSAKE	EULERSR	5317.97667	1876.66	0.156	-2083.2578	12719.2111
		MIRA	-1855.48967	2207.36	0.911	-9429.0482	5718.0688
		PHRAP	-609.73833	3914.23	1	-14272.421	13052.9443
		VELVET	5833.93333	1856.62	0.117	-1612.5836	13280.4502

Continued on next page

Table E.17 – continued from previous page

	Assemblers (I)	Assemblers (J)	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Time	VELVET	EULERSR	-515.95667	276.267	0.431	-1615.3283	583.415
		MIRA	-7689.42300*	1194.54	0.014	-12996.9137	-2381.9323
		PHRAP	-6443.67167	3446.11	0.432	-20267.0161	7379.6728
		SSAKE	-5833.93333	1856.62	0.117	-13280.4502	1612.5836
Mem.	EULERSR	MIRA	-1803.82167*	401.142	0.039	-3481.4474	-126.1959
		PHRAP	-4254.155	1833.94	0.274	-11593.1834	3084.8734
		SSAKE	-627.66667*	181.989	0.046	-1244.5729	-10.7605
		VELVET	-210.375	112.079	0.403	-606.6344	185.8844
	MIRA	EULERSR	1803.82167*	401.142	0.039	126.1959	3481.4474
		PHRAP	-2450.33333	1871.92	0.698	-9714.8782	4814.2115
		SSAKE	1176.155	416.957	0.158	-468.4604	2820.7704
		VELVET	1593.44667	391.529	0.063	-118.6721	3305.5654
	PHRAP	EULERSR	4254.155	1833.94	0.274	-3084.8734	11593.1834
		MIRA	2450.33333	1871.92	0.698	-4814.2115	9714.8782
		SSAKE	3626.48833	1837.47	0.389	-3704.1097	10957.0864
		VELVET	4043.78	1831.86	0.308	-3300.3634	11387.9234
	SSAKE	EULERSR	627.66667*	181.989	0.046	10.7605	1244.5729
		MIRA	-1176.155	416.957	0.158	-2820.7704	468.4604
		PHRAP	-3626.48833	1837.47	0.389	-10957.0864	3704.1097
		VELVET	417.29167	159.689	0.182	-179.7262	1014.3096
	VELVET	EULERSR	210.375	112.079	0.403	-185.8844	606.6344
		MIRA	-1593.44667	391.529	0.063	-3305.5654	118.6721
		PHRAP	-4043.78	1831.86	0.308	-11387.9234	3300.3634
		SSAKE	-417.29167	159.689	0.182	-1014.3096	179.7262

Table E.18: Tukey-HSD post-hoc output for analyzing the performance of each assembler that ran on 454 data based on the test metrics.

	Assemblers (I)	Assemblers (J)	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
NC	EULERSR	MIRA	-2078.833	4802.804	0.992	-16184.05	12026.38
		PHRAP	385.333	4802.804	1	-13719.88	14490.55
		SSAKE	-16313.500*	4802.804	0.018	-30418.72	-2208.28
		VELVET	-14967.833*	4802.804	0.034	-29073.05*	-862.62
	MIRA	EULERSR	2078.833	4802.804	0.992	-12026.38	16184.05
		PHRAP	2464.167	4802.804	0.985	-11641.05	16569.38
		SSAKE	-14234.667*	4802.804	0.047	-28339.88	-129.45
		VELVET	-12889	4802.804	0.085	-26994.22	1216.22
	PHRAP	EULERSR	-385.333	4802.804	1	-14490.55	13719.88
		MIRA	-2464.167	4802.804	0.985	-16569.38	11641.05
		SSAKE	-16698.833*	4802.804	0.015	-30804.05	-2593.62
		VELVET	-15353.167*	4802.804	0.028	-29458.38	-1247.95
	SSAKE	EULERSR	16313.500*	4802.804	0.018	2208.28	30418.72
		MIRA	14234.667*	4802.804	0.047	129.45	28339.88
		PHRAP	16698.833*	4802.804	0.015	2593.62	30804.05
		VELVET	1345.667	4802.804	0.999	-12759.55	15450.88

Continued on next page

Table E.18 – continued from previous page

						95% Confidence Interval	
	Assemblers (I)	Assemblers (J)	Mean Difference (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
NC	VELVET	EULERSR	14967.833*	4802.804	0.034	862.62	29073.05
		MIRA	12889	4802.804	0.085	-1216.22	26994.22
		PHRAP	15353.167*	4802.804	0.028	1247.95	29458.38
		SSAKE	-1345.667	4802.804	0.999	-15450.88	12759.55
GC	EULERSR	MIRA	2.32333	1.94509	0.754	-3.407	8.0536
		PHRAP	2.57167	1.85457	0.642	-2.8919	8.0353
		SSAKE	-0.375	1.85457	1	-5.8386	5.0886
		VELVET	-0.25333	1.85457	1	-5.7169	5.2103
	MIRA	EULERSR	-2.32333	1.94509	0.754	-8.0536	3.407
		PHRAP	0.24833	1.94509	1	-5.482	5.9786
		SSAKE	-2.69833	1.94509	0.641	-8.4286	3.032
		VELVET	-2.57667	1.94509	0.679	-8.307	3.1536
	PHRAP	EULERSR	-2.57167	1.85457	0.642	-8.0353	2.8919
		MIRA	-0.24833	1.94509	1	-5.9786	5.482
		SSAKE	-2.94667	1.85457	0.518	-8.4103	2.5169
		VELVET	-2.825	1.85457	0.558	-8.2886	2.6386
	SSAKE	EULERSR	0.375	1.85457	1	-5.0886	5.8386
		MIRA	2.69833	1.94509	0.641	-3.032	8.4286
		PHRAP	2.94667	1.85457	0.518	-2.5169	8.4103
		VELVET	0.12167	1.85457	1	-5.3419	5.5853
	VELVET	EULERSR	0.25333	1.85457	1	-5.2103	5.7169
		MIRA	2.57667	1.94509	0.679	-3.1536	8.307
		PHRAP	2.825	1.85457	0.558	-2.6386	8.2886
		SSAKE	-0.12167	1.85457	1	-5.5853	5.3419
MC	EULERSR	MIRA	-6.83333	2.41725	0.064	-13.9546	0.288
		PHRAP	-21.54167*	2.53524	0	-29.0106	-14.0728
		SSAKE	0.31667	2.41725	1	-6.8046	7.438
		VELVET	0.01667	2.41725	1	-7.1046	7.138
	MIRA	EULERSR	6.83333	2.41725	0.064	-0.288	13.9546
		PHRAP	-14.70833*	2.53524	0	-22.1772	-7.2394
		SSAKE	7.15000*	2.41725	0.049	0.0287	14.2713
		VELVET	6.85	2.41725	0.063	-0.2713	13.9713
	PHRAP	EULERSR	21.54167*	2.53524	0	14.0728	29.0106
		MIRA	14.70833*	2.53524	0	7.2394	22.1772
		SSAKE	21.85833*	2.53524	0	14.3894	29.3272
		VELVET	21.55833*	2.53524	0	14.0894	29.0272
	SSAKE	EULERSR	-0.31667	2.41725	1	-7.438	6.8046
		MIRA	-7.15000*	2.41725	0.049	-14.2713	-0.0287
		PHRAP	-21.85833*	2.53524	0	-29.3272	-14.3894
		VELVET	-0.3	2.41725	1	-7.4213	6.8213
	VELVET	EULERSR	-0.01667	2.41725	1	-7.138	7.1046
		MIRA	-6.85	2.41725	0.063	-13.9713	0.2713
		PHRAP	-21.55833*	2.53524	0	-29.0272	-14.0894
		SSAKE	0.3	2.41725	1	-6.8213	7.4213
N50	EULERSR	MIRA	-18128.167	21290	0.912	-80666.39	44410.06
		PHRAP	-28279	21290	0.677	-90817.23	34259.23
		SSAKE	24806.167	21290	0.771	-37732.06	87344.39
		VELVET	25535.5	21290	0.752	-37002.73	88073.73
	MIRA	EULERSR	18128.167	21290	0.912	-44410.06	80666.39

Continued on next page

Table E.18 – continued from previous page

						95% Confidence Interval	
	Assemblers (I)	Assemblers (J)	Mean Difference (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
N50	MIRA	PHRAP	-10150.833	21290	0.989	-72689.06	52387.39
		SSAKE	42934.333	21290	0.288	-19603.89	105472.56
		VELVET	43663.667	21290	0.272	-18874.56	106201.89
	PHRAP	EULERSR	28279	21290	0.677	-34259.23	90817.23
		MIRA	10150.833	21290	0.989	-52387.39	72689.06
		SSAKE	53085.167	21290	0.124	-9453.06	115623.39
		VELVET	53814.5	21290	0.116	-8723.73	116352.73
	SSAKE	EULERSR	-24806.167	21290	0.771	-87344.39	37732.06
		MIRA	-42934.333	21290	0.288	-105472.56	19603.89
		PHRAP	-53085.167	21290	0.124	-115623.39	9453.06
		VELVET	729.333	21290	1	-61808.89	63267.56
	VELVET	EULERSR	-25535.5	21290	0.752	-88073.73	37002.73
		MIRA	-43663.667	21290	0.272	-106201.89	18874.56
		PHRAP	-53814.5	21290	0.116	-116352.73	8723.73
		SSAKE	-729.333	21290	1	-63267.56	61808.89
Time	EULERSR	MIRA	-7173.46633	2737.33	0.098	-15237.7079	890.7752
		PHRAP	-5927.715	2609.94	0.189	-13616.6676	1761.2376
		SSAKE	-5317.97667	2609.94	0.279	-13006.9293	2370.976
		VELVET	515.95667	2609.94	1	-7172.996	8204.9093
	MIRA	EULERSR	7173.46633	2737.33	0.098	-890.7752	15237.7079
		PHRAP	1245.75133	2737.33	0.991	-6818.4902	9309.9929
		SSAKE	1855.48967	2737.33	0.959	-6208.7519	9919.7312
		VELVET	7689.423	2737.33	0.067	-374.8185	15753.6645
	PHRAP	EULERSR	5927.715	2609.94	0.189	-1761.2376	13616.6676
		MIRA	-1245.75133	2737.33	0.991	-9309.9929	6818.4902
		SSAKE	609.73833	2609.94	0.999	-7079.2143	8298.691
		VELVET	6443.67167	2609.94	0.132	-1245.281	14132.6243
	SSAKE	EULERSR	5317.97667	2609.94	0.279	-2370.976	13006.9293
		MIRA	-1855.48967	2737.33	0.959	-9919.7312	6208.7519
		PHRAP	-609.73833	2609.94	0.999	-8298.691	7079.2143
		VELVET	5833.93333	2609.94	0.201	-1855.0193	13522.886
	VELVET	EULERSR	-515.95667	2609.94	1	-8204.9093	7172.996
		MIRA	-7689.423	2737.33	0.067	-15753.6645	374.8185
		PHRAP	-6443.67167	2609.94	0.132	-14132.6243	1245.281
		SSAKE	-5833.93333	2609.94	0.201	-13522.886	1855.0193
Mem.	EULERSR	MIRA	-1803.82167	1264.64	0.617	-5529.4789	1921.8356
		PHRAP	-4254.15500*	1205.78	0.013	-7806.4298	-701.8802
		SSAKE	-627.66667	1205.78	0.984	-4179.9415	2924.6081
		VELVET	-210.375	1205.78	1	-3762.6498	3341.8998
	MIRA	EULERSR	1803.82167	1264.64	0.617	-1921.8356	5529.4789
		PHRAP	-2450.33333	1264.64	0.326	-6175.9906	1275.3239
		SSAKE	1176.155	1264.64	0.882	-2549.502	4901.8122
		VELVET	1593.44667	1264.64	0.717	-2132.2106	5319.1039
	PHRAP	EULERSR	4254.15500*	1205.78	0.013	701.8802	7806.4298
		MIRA	2450.33333	1264.64	0.326	-1275.3239	6175.9906
		SSAKE	3626.48833*	1205.78	0.044	74.2135	7178.7631
		PHRAP	4043.78000*	1205.78	0.02	491.5052	7596.0548
	SSAKE	EULERSR	627.66667	1205.78	0.984	-2924.6081	4179.9415
		MIRA	-1176.155	1264.64	0.882	-4901.8122	2549.5022

Continued on next page

Table E.18 – continued from previous page

						95% Confidence Interval	
	Assemblers (I)	Assemblers (J)	Mean Difference (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
Mem.	SSAKE	PHRAP	-3626.48833*	1205.78	0.044	-7178.7631	-74.2135
		VELVET	417.29167	1205.78	0.997	-3134.9831	3969.5665
	VELVET	EULERSR	210.375	1205.78	1	-3341.8998	3762.6498
		MIRA	-1593.44667	1264.64	0.717	-5319.1039	2132.2106
		PHRAP	-4043.78000*	1205.78	0.02	-7596.0548	-491.5052
		SSAKE	-417.29167	1205.78	0.997	-3969.5665	3134.9831

Table E.19: Tukey-HSD post-hoc output for analyzing the performance of each assembler that ran on Illumina data based on the test metrics.

						95% Confidence Interval	
	Assemblers (I)	Assemblers (J)	Mean Difference (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
NC	ABYSS	EDENA	1284.667	22940	1	-70675.71	73245.04
		EULERSR	785.5	22940	1	-71174.87	72745.87
		MIRA	-628.533	24060	1	-76101.21	74844.14
		PHRAP	-24301.333	24060	0.948	-99774.01	51171.34
		SSAKE	-59059.833	22940	0.167	-131020.21	12900.54
		VELVET	1050.167	22940	1	-70910.21	73010.54
	EDENA	ABYSS	-1284.667	22940	1	-73245.04	70675.71
		EULERSR	-499.167	22940	1	-72459.54	71461.21
		MIRA	-1913.2	24060	1	-77385.88	73559.48
		PHRAP	-25586	24060	0.934	-101058.68	49886.68
		SSAKE	-60344.5	22940	0.149	-132304.87	11615.87
		VELVET	-234.5	22940	1	-72194.87	71725.87
	EULERSR	ABYSS	-785.5	22940	1	-72745.87	71174.87
		EDENA	499.167	22940	1	-71461.21	72459.54
		MIRA	-1414.033	24060	1	-76886.71	74058.64
		PHRAP	-25086.833	24060	0.94	-100559.51	50385.84
		SSAKE	-59845.333	22940	0.156	-131805.71	12115.04
		VELVET	264.667	22940	1	-71695.71	72225.04
	MIRA	ABYSS	628.533	24060	1	-74844.14	76101.21
		EDENA	1913.2	24060	1	-73559.48	77385.88
		EULERSR	1414.033	24060	1	-74058.64	76886.71
		PHRAP	-23672.8	25130	0.963	-102501.44	55155.84
		SSAKE	-58431.3	24060	0.219	-133903.98	17041.38
		VELVET	1678.7	24060	1	-73793.98	77151.38
	PHRAP	ABYSS	24301.333	24060	0.948	-51171.34	99774.01
		EDENA	25586	24060	0.934	-49886.68	101058.68
		EULERSR	25086.833	24060	0.94	-50385.84	100559.51
		MIRA	23672.8	25130	0.963	-55155.84	102501.44
		SSAKE	-34758.5	24060	0.774	-110231.18	40714.18
		VELVET	25351.5	24060	0.937	-50121.18	100824.18
	SSAKE	ABYSS	59059.833	22940	0.167	-12900.54	131020.21
		EDENA	60344.5	22940	0.149	-11615.87	132304.87
		EULERSR	59845.333	22940	0.156	-12115.04	131805.71
		MIRA	58431.3	24060	0.219	-17041.38	133903.98

Continued on next page

Table E.19 – continued from previous page

	Assemblers (I)	Assemblers (J)	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
NC	SSAKE	PHRAP	34758.5	24060	0.774	-40714.18	110231.18
		VELVET	60110	22940	0.152	-11850.37	132070.37
	VELVET	ABYSS	-1050.167	22940	1	-73010.54	70910.21
		EDENA	234.5	22940	1	-71725.87	72194.87
		EULERSR	-264.667	22940	1	-72225.04	71695.71
		MIRA	-1678.7	24060	1	-77151.38	73793.98
		PHRAP	-25351.5	24060	0.937	-100824.18	50121.18
		SSAKE	-60110	22940	0.152	-132070.37	11850.37
GC	ABYSS	EDENA	1.19	9.07138	1	-27.5856	29.9656
		EULERSR	0.53833	9.07138	1	-28.2373	29.3139
		MIRA	21.70533	9.51415	0.288	-8.4748	51.8854
		PHRAP	-0.30667	12.82887	1	-41.0015	40.3882
		SSAKE	0.31333	10.14212	1	-31.8588	32.4854
		VELVET	0.235	9.07138	1	-28.5406	29.0106
	EDENA	ABYSS	-1.19	9.07138	1	-29.9656	27.5856
		EULERSR	-0.65167	9.07138	1	-29.4273	28.1239
		MIRA	20.51533	9.51415	0.35	-9.6648	50.6954
		PHRAP	-1.49667	12.82887	1	-42.1915	39.1982
		SSAKE	-0.87667	10.14212	1	-33.0488	31.2954
		VELVET	-0.955	9.07138	1	-29.7306	27.8206
	EULERSR	ABYSS	-0.53833	9.07138	1	-29.3139	28.2373
		EDENA	0.65167	9.07138	1	-28.1239	29.4273
		MIRA	21.167	9.51415	0.315	-9.0131	51.3471
		PHRAP	-0.845	12.82887	1	-41.5399	39.8499
		SSAKE	-0.225	10.14212	1	-32.3971	31.9471
		VELVET	-0.30333	9.07138	1	-29.0789	28.4723
	MIRA	ABYSS	-21.70533	9.51415	0.288	-51.8854	8.4748
		EDENA	-20.51533	9.51415	0.35	-50.6954	9.6648
		EULERSR	-21.167	9.51415	0.315	-51.3471	9.0131
		PHRAP	-22.012	13.14568	0.638	-63.7118	19.6878
		SSAKE	-21.392	10.54	0.42	-54.8262	12.0422
		VELVET	-21.47033	9.51415	0.299	-51.6504	8.7098
	PHRAP	ABYSS	0.30667	12.82887	1	-40.3882	41.0015
		EDENA	1.49667	12.82887	1	-39.1982	42.1915
		EULERSR	0.845	12.82887	1	-39.8499	41.5399
		MIRA	22.012	13.14568	0.638	-19.6878	63.7118
		SSAKE	0.62	13.60708	1	-42.5434	43.7834
		VELVET	0.54167	12.82887	1	-40.1532	41.2365
	SSAKE	ABYSS	-0.31333	10.14212	1	-32.4854	31.8588
		EDENA	0.87667	10.14212	1	-31.2954	33.0488
		EULERSR	0.225	10.14212	1	-31.9471	32.3971
		MIRA	21.392	10.54	0.42	-12.0422	54.8262
		PHRAP	-0.62	13.60708	1	-43.7834	42.5434
		VELVET	-0.07833	10.14212	1	-32.2504	32.0938
	VELVET	ABYSS	-0.235	9.07138	1	-29.0106	28.5406
		EDENA	0.955	9.07138	1	-27.8206	29.7306
		EULERSR	0.30333	9.07138	1	-28.4723	29.0789
		MIRA	21.47033	9.51415	0.299	-8.7098	51.6504
		PHRAP	-0.54167	12.82887	1	-41.2365	40.1532

Continued on next page

Table E.19 – continued from previous page

						95% Confidence Interval	
	Assemblers (I)	Assemblers (J)	Mean Difference (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
GC	VELVET	SSAKE	0.07833	10.14212	1	-32.0938	32.2504
MC	ABYSS	EDENA	2.62833	1.29924	0.423	-1.493	6.7497
		EULERSR	2.39433	1.29924	0.532	-1.727	6.5157
		MIRA	0.32133	1.36265	1	-4.0012	4.6438
		PHRAP	2.38333	1.8374	0.848	-3.4451	8.2118
		SSAKE	2.46833	1.45259	0.622	-2.1395	7.0761
		VELVET	2.27	1.29924	0.592	-1.8513	6.3913
	EDENA	ABYSS	-2.62833	1.29924	0.423	-6.7497	1.493
		EULERSR	-0.234	1.29924	1	-4.3553	3.8873
		MIRA	-2.307	1.36265	0.626	-6.6295	2.0155
		PHRAP	-0.245	1.8374	1	-6.0735	5.5835
		SSAKE	-0.16	1.45259	1	-4.7678	4.4478
		VELVET	-0.35833	1.29924	1	-4.4797	3.763
	EULERSR	ABYSS	-2.39433	1.29924	0.532	-6.5157	1.727
		EDENA	0.234	1.29924	1	-3.8873	4.3553
		MIRA	-2.073	1.36265	0.73	-6.3955	2.2495
		PHRAP	-0.011	1.8374	1	-5.8395	5.8175
		SSAKE	0.074	1.45259	1	-4.5338	4.6818
		VELVET	-0.12433	1.29924	1	-4.2457	3.997
	MIRA	ABYSS	-0.32133	1.36265	1	-4.6438	4.0012
		EDENA	2.307	1.36265	0.626	-2.0155	6.6295
		EULERSR	2.073	1.36265	0.73	-2.2495	6.3955
		PHRAP	2.062	1.88277	0.924	-3.9104	8.0344
		SSAKE	2.147	1.50958	0.785	-2.6416	6.9356
		VELVET	1.94867	1.36265	0.781	-2.3738	6.2712
	PHRAP	ABYSS	-2.38333	1.8374	0.848	-8.2118	3.4451
		EDENA	0.245	1.8374	1	-5.5835	6.0735
		EULERSR	0.011	1.8374	1	-5.8175	5.8395
		MIRA	-2.062	1.88277	0.924	-8.0344	3.9104
		SSAKE	0.085	1.94886	1	-6.097	6.267
		VELVET	-0.11333	1.8374	1	-5.9418	5.7151
	SSAKE	ABYSS	-2.46833	1.45259	0.622	-7.0761	2.1395
		EDENA	0.16	1.45259	1	-4.4478	4.7678
		EULERSR	-0.074	1.45259	1	-4.6818	4.5338
		MIRA	-2.147	1.50958	0.785	-6.9356	2.6416
		PHRAP	-0.085	1.94886	1	-6.267	6.097
		VELVET	-0.19833	1.45259	1	-4.8061	4.4095
	VELVET	ABYSS	-2.27	1.29924	0.592	-6.3913	1.8513
		EDENA	0.35833	1.29924	1	-3.763	4.4797
		EULERSR	0.12433	1.29924	1	-3.997	4.2457
		MIRA	-1.94867	1.36265	0.781	-6.2712	2.3738
		PHRAP	0.11333	1.8374	1	-5.7151	5.9418
		SSAKE	0.19833	1.45259	1	-4.4095	4.8061
N50	ABYSS	EDENA	1750	34780	1	-107361.11	110861.11
		EULERSR	-5238	34780	1	-114349.11	103873.11
		MIRA	-104121.433	36480	0.094	-218558.13	10315.27
		PHRAP	13796.967	36480	1	-100639.73	128233.67
		SAKE	2241.333	34780	1	-106869.78	111352.44
		VELVET	-1256.5	34780	1	-110367.61	107854.61

Continued on next page

Table E.19 – continued from previous page

	Assemblers (I)	Assemblers (J)	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
N50	EDENA	ABYSS	-1750	34780	1	-110861.11	107361.11
		EULERSR	-6988	34780	1	-116099.11	102123.11
		MIRA	-105871.433	36480	0.085	-220308.13	8565.27
		PHRAP	12046.967	36480	1	-102389.73	126483.67
		SSAKE	491.333	34780	1	-108619.78	109602.44
		VELVET	-3006.5	34780	1	-112117.61	106104.61
	EULERSR	ABYSS	5238	34780	1	-103873.11	114349.11
		EDENA	6988	34780	1	-102123.11	116099.11
		MIRA	-98883.433	36480	0.127	-213320.13	15553.27
		PHRAP	19034.967	36480	0.998	-95401.73	133471.67
		SSAKE	7479.333	34780	1	-101631.78	116590.44
		VELVET	3981.5	34780	1	-105129.61	113092.61
	MIRA	ABYSS	104121.433	36480	0.094	-10315.27	218558.13
		EDENA	105871.433	36480	0.085	-8565.27	220308.13
		EULERSR	98883.433	36480	0.127	-15553.27	213320.13
		PHRAP	117918.4	38100	0.055	-1606.83	237443.63
		SSAKE	106362.767	36480	0.082	-8073.93	220799.47
		VELVET	102864.933	36480	0.101	-11571.77	217301.63
	PHRAP	ABYSS	-13796.967	36480	1	-128233.67	100639.73
		EDENA	-12046.967	36480	1	-126483.67	102389.73
		EULERSR	-19034.967	36480	0.998	-133471.67	95401.73
		MIRA	-117918.4	38100	0.055	-237443.63	1606.83
		SSAKE	-11555.633	36480	1	-125992.33	102881.07
		VELVET	-15053.467	36480	1	-129490.17	99383.23
	SSAKE	ABYSS	-2241.333	34780	1	-111352.44	106869.78
		EDENA	-491.333	34780	1	-109602.44	108619.78
		EULERSR	-7479.333	34780	1	-116590.44	101631.78
		MIRA	-106362.767	36480	0.082	-220799.47	8073.93
		PHRAP	11555.633	36480	1	-102881.07	125992.33
		VELVET	-3497.833	34780	1	-112608.94	105613.28
	VELVET	ABYSS	1256.5	34780	1	-107854.61	110367.61
		EDENA	3006.5	34780	1	-106104.61	112117.61
		EULERSR	-3981.5	34780	1	-113092.61	105129.61
		MIRA	-102864.933	36480	0.101	-217301.63	11571.77
		PHRAP	15053.467	36480	1	-99383.23	129490.17
		SSAKE	3497.833	34780	1	-105613.28	112608.94
Time	ABYSS	EDENA	-601.57667	22937.8	1	-72699.0322	71495.8789
		EULERSR	-1585.14	22937.8	1	-73682.5955	70512.3155
		MIRA	-1.01571E5*	24057.4	0.003	-177190	-25954.0594
		PHRAP	-24788.67267	24057.4	0.943	-100410	50827.7766
		SSAKE	-6889.68667	24057.4	1	-82506.136	68726.7626
		VELVET	129.265	22937.8	1	-71968.1905	72226.7205
	EDENA	ABYSS	601.57667	22937.8	1	-71495.8789	72699.0322
		EULERSR	-983.56333	22937.8	1	-73081.0189	71113.8922
		MIRA	-1.00969E5*	24057.4	0.003	-176590	-25352.4827
		PHRAP	-24187.096	24057.4	0.949	-99803.5453	51429.3533
		SSAKE	-6288.11	24057.4	1	-81904.5593	69328.3393
		VELVET	730.84167	22937.8	1	-71366.6139	72828.2972
	EULERSR	ABYSS	1585.14	22937.8	1	-70512.3155	73682.5955

Continued on next page

Table E.19 – continued from previous page

						95% Confidence Interval	
	Assemblers (I)	Assemblers (J)	Mean Difference (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
Time	EULERSR	EDENA	983.56333	22937.8	1	-71113.8922	73081.0189
		MIRA	-9.99854E4*	24057.4	0.004	-175600	-24368.9194
		PHRAP	-23203.53267	24057.4	0.958	-98819.982	52412.9166
		SSAKE	-5304.54667	24057.4	1	-80920.996	70311.9026
		VELVET	1714.405	22937.8	1	-70383.0505	73811.8605
	MIRA	ABYSS	1.01571E5*	24057.4	0.003	25954.0594	177186.958
		EDENA	1.00969E5*	24057.4	0.003	25352.4827	176585.3813
		EULERSR	99985.36867*	24057.4	0.004	24368.9194	175601.818
		PHRAP	76781.836	25127.1	0.061	-2196.9695	155760.6415
		SSAKE	94680.82200*	25127.1	0.011	15702.0165	173659.6275
		VELVET	1.01700E5*	24057.4	0.003	26083.3244	177316.223
	PHRAP	ABYSS	24788.67267	24057.4	0.943	-50827.7766	100405.122
		EDENA	24187.096	24057.4	0.949	-51429.3533	99803.5453
		EULERSR	23203.53267	24057.4	0.958	-52412.9166	98819.982
		MIRA	-76781.836	25127.1	0.061	-155760	2196.9695
		SSAKE	17898.986	25127.1	0.991	-61079.8195	96877.7915
		VELVET	24917.93767	24057.4	0.942	-50698.5116	100534.387
	SSAKE	ABYSS	6889.68667	24057.4	1	-68726.7626	82506.136
		EDENA	6288.11	24057.4	1	-69328.3393	81904.5593
		EULERSR	5304.54667	24057.4	1	-70311.9026	80920.996
		MIRA	-9.46808E4*	25127.1	0.011	-173660	-15702.0165
		PHRAP	-17898.986	25127.1	0.991	-96877.7915	61079.8195
		VELVET	7018.95167	24057.4	1	-68597.4976	82635.401
	VELVET	ABYSS	-129.265	22937.8	1	-72226.7205	71968.1905
		EDENA	-730.84167	22937.8	1	-72828.2972	71366.6139
		EULERSR	-1714.405	22937.8	1	-73811.8605	70383.0505
		MIRA	-1.01700E5*	24057.4	0.003	-177320	-26083.3244
		PHRAP	-24917.93767	24057.4	0.942	-100530	50698.5116
		SSAKE	-7018.95167	24057.4	1	-82635.401	68597.4976
Mem.	ABYSS	EDENA	-333.64333	4050.54	1	-13065.1912	12397.9046
		EULERSR	214.43333	4050.54	1	-12517.1146	12945.9812
		MIRA	-10233.68167	4248.24	0.228	-23586.6418	3119.2784
		PHRAP	-11589.68167	4248.24	0.124	-24942.6418	1763.2784
		SSAKE	-2347.68167	4248.24	0.998	-15700.6418	11005.2784
		VELVET	-679.34333	4050.54	1	-13410.8912	12052.2046
	EDENA	ABYSS	333.64333	4050.54	1	-12397.9046	13065.1912
		EULERSR	548.07667	4050.54	1	-12183.4712	13279.6246
		MIRA	-9900.03833	4248.24	0.261	-23252.9984	3452.9218
		PHRAP	-11256.	4248.24	0.145	-24608.9984	2096.9218
		SSAKE	-2014.03833	4248.24	0.999	-15366.9984	11338.9218
		VELVET	-345.7	4050.54	1	-13077.2479	12385.8479
	EULERSR	ABYSS	-214.43333	4050.54	1	-12945.9812	12517.1146
		EDENA	-548.07667	4050.54	1	-13279.6246	12183.4712
		MIRA	-10448.115	4248.24	0.208	-23801.0751	2904.8451
		PHRAP	-11804.115	4248.24	0.112	-25157.0751	1548.8451
		SSAKE	-2562.115	4248.24	0.996	-15915.0751	10790.8451
		VELVET	-893.77667	4050.54	1	-13625.3246	11837.7712
	MIRA	ABYSS	10233.68167	4248.24	0.228	-3119.2784	23586.6418
		EDENA	9900.03833	4248.24	0.261	-3452.9218	23252.9984

Continued on next page

Table E.19 – continued from previous page

						95% Confidence Interval	
	Assemblers (I)	Assemblers (J)	Mean Difference (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
Mem.	MIRA	EULERSR	10448.115	4248.24	0.208	-2904.8451	23801.0751
		PHRAP	-1356	4437.15	1	-15302.712	12590.712
		SSAKE	7886	4437.15	0.572	-6060.712	21832.712
		VELVET	9554.33833	4248.24	0.299	-3798.6218	22907.2984
	PHRAP	ABYSS	11589.68167	4248.24	0.124	-1763.2784	24942.6418
		EDENA	11256.03833	4248.24	0.145	-2096.9218	24608.9984
		EULERSR	11804.115	4248.24	0.112	-1548.8451	25157.0751
		MIRA	1356	4437.15	1	-12590.712	15302.712
		SSAKE	9242	4437.15	0.386	-4704.712	23188.712
		VELVET	10910.33833	4248.24	0.17	-2442.6218	24263.2984
	SSAKE	ABYSS	2347.68167	4248.24	0.998	-11005.2784	15700.6418
		EDENA	2014.03833	4248.24	0.999	-11338.9218	15366.9984
		EULERSR	2562.115	4248.24	0.996	-10790.8451	15915.0751
		MIRA	-7886	4437.15	0.572	-21832.712	6060.712
		PHRAP	-9242	4437.15	0.386	-23188.712	4704.712
		VELVET	1668.33833	4248.24	1	-11684.6218	15021.2984
	VELVET	ABYSS	679.34333	4050.54	1	-12052.2046	13410.8912
		EDENA	345.7	4050.54	1	-12385.8479	13077.2479
		EULERSR	893.77667	4050.54	1	-11837.7712	13625.3246
		MIRA	-9554.33833	4248.24	0.299	-22907.2984	3798.6218
		PHRAP	-10910.33833	4248.24	0.17	-24263.2984	2442.6218
		SSAKE	-1668.33833	4248.24	1	-15021.2984	11684.6218

Table E.20: Games-Howell post-hoc output for analyzing the performance of each assembler that ran on Illumina data based on the test metrics.

						95% Confidence Interval	
	Assemblers (I)	Assemblers (J)	Mean Difference (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
NC	ABYSS	EDENA	1284.667	498.268	0.275	-793.55	3362.89
		EULERSR	785.5	559.176	0.787	-1325.6	2896.6
		MIRA	-628.533	1488.019	0.999	-7342.04	6084.98
		PHRAP	-24301.333	17330	0.783	-110647.43	62044.76
		SSAKE	-59059.833	39180	0.736	-234393.47	116273.8
		VELVET	1050.167	527.927	0.489	-1022.81	3123.15
	EDENA	ABYSS	-1284.667	498.268	0.275	-3362.89	793.55
		EULERSR	-499.167	331.02	0.735	-1788.52	790.19
		MIRA	-1913.2	1418.132	0.806	-8900.21	5073.81
		PHRAP	-25586	17320	0.75	-111966.48	60794.48
		SSAKE	-60344.5	39180	0.72	-235686.44	114997.44
		VELVET	-234.5	274.963	0.971	-1265.08	796.08
	EULERSR	ABYSS	-785.5	559.176	0.787	-2896.6	1325.6
		EDENA	499.167	331.02	0.735	-790.19	1788.52
		MIRA	-1414.033	1440.661	0.936	-8290.52	5462.45
		PHRAP	-25086.833	17320	0.763	-111456.39	61282.72
		SSAKE	-59845.333	39180	0.726	-235184.64	115493.97
		VELVET	264.667	374.18	0.989	-1107.86	1637.19

Continued on next page

Table E.20 – continued from previous page

						95% Confidence Interval	
	Assemblers (I)	Assemblers (J)	Mean Difference (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
NC	MIRA	ABYSS	628.533	1488.019	0.999	-6084.98	7342.04
		EDENA	1913.2	1418.132	0.806	-5073.81	8900.21
		EULERSR	1414.033	1440.661	0.936	-5462.45	8290.52
		PHRAP	-23672.8	17380	0.8	-109727.67	62382.07
		SSAKE	-58431.3	39200	0.744	-233693.33	116830.73
		VELVET	1678.7	1428.822	0.876	-5252.84	8610.24
	PHRAP	ABYSS	24301.333	17330	0.783	-62044.76	110647.43
		EDENA	25586	17320	0.75	-60794.48	111966.48
		EULERSR	25086.833	17320	0.763	-61282.72	111456.39
		MIRA	23672.8	17380	0.8	-62382.07	109727.67
		SSAKE	-34758.5	42830	0.976	-205896.6	136379.6
		VELVET	25351.5	17320	0.756	-61023.82	111726.82
	SSAKE	ABYSS	59059.833	39180	0.736	-116273.8	234393.47
		EDENA	60344.5	39180	0.72	-114997.44	235686.44
		EULERSR	59845.333	39180	0.726	-115493.97	235184.64
		MIRA	58431.3	39200	0.744	-116830.73	233693.33
		PHRAP	34758.5	42830	0.976	-136379.6	205896.6
		VELVET	60110	39180	0.723	-115230.69	235450.69
	VELVET	ABYSS	-1050.167	527.927	0.489	-3123.15	1022.81
		EDENA	234.5	274.963	0.971	-796.08	1265.08
		EULERSR	-264.667	374.18	0.989	-1637.19	1107.86
		MIRA	-1678.7	1428.822	0.876	-8610.24	5252.84
		PHRAP	-25351.5	17320	0.756	-111726.82	61023.82
		VELVET	-60110	39180	0.723	-235450.69	115230.69
GC	ABYSS	EDENA	1.19	1.00724	0.885	-2.5329	4.9129
		EULERSR	0.53833	0.96271	0.997	-2.9933	4.0699
		MIRA	21.70533	18.51629	0.876	-70.531	113.9416
		PHRAP	-0.30667	0.9245	1	-6.4858	5.8725
		SSAKE	0.31333	0.79983	1	-2.7605	3.3872
		VELVET	0.235	0.82921	1	-2.7697	3.2397
	EDENA	ABYSS	-1.19	1.00724	0.885	-4.9129	2.5329
		EULERSR	-0.65167	1.11363	0.996	-4.6909	3.3876
		MIRA	20.51533	18.52475	0.898	-71.6717	112.7024
		PHRAP	-1.49667	1.08076	0.791	-6.8967	3.9033
		SSAKE	-0.87667	0.97625	0.963	-4.6252	2.8719
		VELVET	-0.955	1.00047	0.952	-4.6616	2.7516
	EULERSR	ABYSS	-0.53833	0.96271	0.997	-4.0699	2.9933
		EDENA	0.65167	1.11363	0.996	-3.3876	4.6909
		MIRA	21.167	18.52238	0.886	-71.0338	113.3678
		PHRAP	-0.845	1.03939	0.969	-6.3154	4.6254
		SSAKE	-0.225	0.93025	1	-3.7831	3.3331
		VELVET	-0.30333	0.95563	1	-3.8158	3.2091
	MIRA	ABYSS	-21.70533	18.51629	0.876	-113.9416	70.531
		EDENA	-20.51533	18.52475	0.898	-112.7024	71.6717
		EULERSR	-21.167	18.52238	0.886	-113.3678	71.0338
		PHRAP	-22.012	18.52044	0.87	-114.2245	70.2005
		SSAKE	-21.392	18.51463	0.882	-113.638	70.854
		VELVET	-21.47033	18.51593	0.88	-113.7088	70.7681
	PHRAP	ABYSS	0.30667	0.9245	1	-5.8725	6.4858

Continued on next page

Table E.20 – continued from previous page

						95% Confidence Interval	
	Assemblers (I)	Assemblers (J)	Mean Difference (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
GC	PHRAP	EDENA	1.49667	1.08076	0.791	-3.9033	6.8967
		EULERSR	0.845	1.03939	0.969	-4.6254	6.3154
		MIRA	22.012	18.52044	0.87	-70.2005	114.2245
		SSAKE	0.62	0.89064	0.98	-6.3173	7.5573
		VELVET	0.54167	0.91711	0.991	-5.7263	6.8096
	SSAKE	ABYSS	-0.31333	0.79983	1	-3.3872	2.7605
		EDENA	0.87667	.97625	0.963	-2.8719	4.6252
		EULERSR	0.225	0.93025	1	-3.3331	3.7831
		MIRA	21.392	18.51463	0.882	-70.854	113.638
		PHRAP	-0.62	0.89064	0.98	-7.5573	6.3173
		VELVET	-0.07833	0.79128	1	-3.1248	2.9682
	VELVET	ABYSS	-0.235	0.82921	1	-3.2397	2.7697
		EDENA	0.955	1.00047	0.952	-2.7516	4.6616
		EULERSR	0.30333	0.95563	1	3.2091	3.8158
		MIRA	21.47033	18.51593	0.88	-70.7681	113.7088
		PHRAP	-0.54167	0.91711	0.991	-6.8096	5.7263
		SSAKE	0.07833	0.79128	1	-2.9682	3.1248
MC	ABYSS	EDENA	2.62833	1.93566	0.804	-6.0327	11.2893
		EULERSR	2.39433	1.95052	0.859	-6.2222	11.0108
		MIRA	0.32133	2.23598	1	-8.2685	8.9112
		PHRAP	2.38333	1.9584	0.863	-6.22	10.9867
		SSAKE	2.46833	1.93595	0.84	-6.1917	11.1284
		PHRAP	2.27	1.95515	0.884	-6.3338	10.8738
	EDENA	ABYSS	-2.62833	1.93566	0.804	-11.2893	6.0327
		EULERSR	-0.234	0.24622	0.947	-1.3142	0.8462
		MIRA	-2.307	1.12059	0.495	-7.8883	3.2743
		PHRAP	-0.245	0.30241	0.949	-8.745	8.255
		SSAKE	-0.16	0.06346	0.29	-0.4221	0.1021
		VELVET	-0.35833	0.28058	0.839	-1.5948	0.8782
	EULERSR	ABYSS	-2.39433	1.95052	0.859	-11.0108	6.2222
		EDENA	0.234	0.24622	0.947	-0.8462	1.3142
		MIRA	-2.073	1.14606	0.596	-7.5262	3.3802
		PHRAP	-0.011	0.38623	1	-2.6759	2.6539
		SSAKE	0.074	0.24849	1	-1.0014	1.1494
		VELVET	-0.12433	0.36938	1	-1.4676	1.219
	MIRA	ABYSS	-0.32133	2.23598	1	-8.9112	8.2685
		EDENA	2.307	1.12059	0.495	-3.2743	7.8883
		EULERSR	2.073	1.14606	0.596	-3.3802	7.5262
		PHRAP	2.062	1.15943	0.609	-3.3807	7.5047
		SSAKE	2.147	1.12109	0.554	-3.4314	7.7254
		VELVET	1.94867	1.15393	0.651	-3.4725	7.3698
	PHRAP	ABYSS	-2.38333	1.9584	0.863	-10.9867	6.22
		EDENA	0.245	0.30241	0.949	-8.255	8.745
		EULERSR	0.011	0.38623	1	-2.6539	2.6759
		MIRA	-2.062	1.15943	0.609	-7.5047	3.3807
		SSAKE	0.085	0.30426	1	-7.9715	8.1415
		VELVET	-0.11333	0.40899	1	-2.5589	2.3322
	SSAKE	ABYSS	-2.46833	1.93595	0.84	-11.1284	6.1917
		EDENA	0.16	0.06346	0.29	-0.1021	0.4221

Continued on next page

Table E.20 – continued from previous page

	Assemblers (I)	Assemblers (J)	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
MC	SSAKE	EULERSR	-0.074	0.24849	1	-1.1494	1.0014
		MIRA	-2.147	1.12109	0.554	-7.7254	3.4314
		PHRAP	-0.085	0.30426	1	-8.1415	7.9715
		VELVET	-0.19833	0.28257	0.987	-1.4301	1.0334
	VELVET	ABYSS	-2.27	1.95515	0.884	-10.8738	6.3338
		EDENA	0.35833	0.28058	0.839	-0.8782	1.5948
		EULERSR	0.12433	0.36938	1	-1.219	1.4676
		MIRA	-1.94867	1.15393	0.651	-7.3698	3.4725
		PHRAP	0.11333	0.40899	1	-2.3322	2.5589
		SSAKE	0.19833	0.28257	0.987	-1.0334	1.4301
N50	ABYSS	EDENA	1750	4063.441	0.999	-13105.86	16605.86
		EULERSR	-5238	5147.176	0.938	-24127.73	13651.73
		MIRA	-104121.433	76430	0.8	-484541.28	276298.42
		PHRAP	13796.967	3157.244	0.055	-334.52	27928.45
		SSAKE	2241.333	5973.709	1	-20302.33	24785
		VELVET	-1256.5	7488.992	1	-30859.55	28346.55
	EDENA	ABYSS	-1750	4063.441	0.999	-16605.86	13105.86
		EULERSR	-6988	4802.99	0.762	-25084.98	11108.98
		MIRA	-105871.433	76400	0.79	-486421.88	274679.02
		PHRAP	12046.967*	2558.023	0.041	597.56	23496.37
		SSAKE	491.333	5679.832	1	-21668.28	22650.94
		VELVET	-3006.5	7256.741	0.999	-32610.41	26597.41
	EULERSR	ABYSS	5238	5147.176	0.938	-13651.73	24127.73
		EDENA	6988	4802.99	0.762	-11108.98	25084.98
		MIRA	-98883.433	76470	0.829	-479054.73	281287.87
		PHRAP	19034.967*	4065.138	0.042	839.8	37230.13
		SSAKE	7479.333	6499.413	0.897	-16303.86	31262.53
		VELVET	3981.5	7914.68	0.998	-26065.18	34028.18
	MIRA	ABYSS	104121.433	76430	0.8	-276298.42	484541.28
		EDENA	105871.433	76400	0.79	-274679.02	486421.88
		EULERSR	98883.433	76470	0.829	-281287.87	479054.73
		PHRAP	117918.4	76360	0.72	-262883.09	498719.89
		SSAKE	106362.767	76530	0.788	-273463.44	486188.97
		VELVET	102864.933	76660	0.809	-276209.02	481938.89
	PHRAP	ABYSS	-13796.967	3157.244	0.055	-27928.45	334.52
		EDENA	-12046.967*	2558.023	0.041	-23496.37	-597.56
		EULERSR	-19034.967*	4065.138	0.042	-37230.13	-839.8
		MIRA	-117918.4	76360	0.72	-498719.89	262883.09
		SSAKE	-11555.633	5071.206	0.393	-34253.88	11142.62
		VELVET	-15053.467	6790.944	0.416	-45449.13	15342.2
	SSAKE	ABYSS	-2241.333	5973.709	1	-24785	20302.33
		EDENA	-491.333	5679.832	1	-22650.94	21668.28
		EULERSR	-7479.333	6499.413	0.897	-31262.53	16303.86
		MIRA	-106362.767	76530	0.788	-486188.97	273463.44
		PHRAP	11555.633	5071.206	0.393	-11142.62	34253.88
		VELVET	-3497.833	8475.491	0.999	-34728.38	27732.72
	VELVET	ABYSS	1256.5	7488.992	1	-28346.55	30859.55
		EDENA	3006.5	7256.741	0.999	-26597.41	32610.41
		EULERSR	-3981.5	7914.68	0.998	-34028.18	26065.18

Continued on next page

Table E.20 – continued from previous page

	Assemblers (I)	Assemblers (J)	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
N50	VELVET	MIRA	-102864.933	76660	0.809	-481938.89	276209.02
		PHRAP	15053.467	6790.944	0.416	-15342.2	45449.13
		SSAKE	3497.833	8475.491	0.999	-27732.72	34728.38
Time	ABYSS	EDENA	-601.57667	485.145	0.857	-2586.2041	1383.0508
		EULERSR	-1585.14	1539.9	0.926	-8408.4408	5238.1608
		MIRA	-101571	47520.3	0.465	-338550	135405.8646
		PHRAP	-24788.67267	15368.4	0.688	-101420	51842.0929
		SSAKE	-6889.68667	5222.83	0.818	-32906.1307	19126.7574
		VELVET	129.265	252.962	0.998	-789.9489	1048.4789
	EDENA	ABYSS	601.57667	485.145	0.857	-1383.0508	2586.2041
		EULERSR	-983.56333	1596.93	0.993	-7689.9912	5722.8646
		MIRA	-100969	47522.2	0.47	-337930	135996.3719
		PHRAP	-24187.096	15374.2	0.706	-100780	52409.5481
		SSAKE	-6288.11	5239.93	0.866	-32206.7759	19630.5559
		VELVET	730.84167	492.844	0.748	-1251.4696	2713.1529
	EULERSR	ABYSS	1585.14	1539.9	0.926	-5238.1608	8408.4408
		EDENA	983.56333	1596.93	0.993	-5722.8646	7689.9912
		MIRA	-99985.36867	47544.7	0.478	-336820	136848.2576
		PHRAP	-23203.53267	15443.5	0.738	-99406.64	52999.5747
		SSAKE	-5304.54667	5439.92	0.939	-30355.6596	19746.5663
		VELVET	1714.405	1542.34	0.901	-5101.9291	8530.7391
	MIRA	ABYSS	101571	47520.3	0.465	-135410	338546.8819
		EDENA	100969	47522.2	0.47	-136000	337934.2359
		EULERSR	99985.36867	47544.7	0.478	-136850	336818.995
		PHRAP	76781.836	49943.1	0.722	-150210	303777.9257
		SSAKE	94680.822	47805.9	0.525	-140680	330046.0534
		VELVET	101700	47520.4	0.464	-135280	338675.6809
	PHRAP	ABYSS	24788.67267	15368.4	0.688	-51842.0929	101419.4383
		EDENA	24187.096	15374.2	0.706	-52409.5481	100783.7401
		EULERSR	23203.53267	15443.5	0.738	-52999.5747	99406.64
		MIRA	-76781.836	49943.1	0.722	-303780	150214.2537
		SSAKE	17898.986	16229.9	0.903	-55312.0512	91110.0232
		VELVET	24917.93767	15368.6	0.685	-51711.3882	101547.2635
	SSAKE	ABYSS	6889.68667	5222.83	0.818	-19126.7574	32906.1307
		EDENA	6288.11	5239.93	0.866	-19630.5559	32206.7759
		EULERSR	5304.54667	5439.92	0.939	-19746.5663	30355.6596
		MIRA	-94680.822	47805.9	0.525	-330050	140684.4094
		PHRAP	-17898.986	16229.9	0.903	-91110.0232	55312.0512
		VELVET	7018.95167	5223.55	0.808	-18993.285	33031.1883
	VELVET	ABYSS	-129.265	252.962	0.998	-1048.4789	789.9489
		EDENA	-730.84167	492.844	0.748	-2713.1529	1251.4696
		EULERSR	-1714.405	1542.34	0.901	-8530.7391	5101.9291
		MIRA	-101700	47520.4	0.464	-338680	135276.1336
		PHRAP	-24917.93767	15368.6	0.685	-101550	51711.3882
		SSAKE	-7018.95167	5223.55	0.808	-33031.1883	18993.285
Mem.	ABYSS	EDENA	-333.64333	285.432	0.89	-1390.5803	723.2936
		EULERSR	214.43333	187.607	0.896	-512.7742	941.6408
		MIRA	-10233.68167	6151.24	0.665	-40885.0208	20417.6574
		PHRAP	-11589.68167	6316.09	0.589	-43063.7102	19884.3469

Continued on next page

Table E.20 – continued from previous page

	Assemblers (I)	Assemblers (J)	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Mem.	ABYSS	SSAKE	-2347.68167	847.98	0.264	-6409.6525	1714.2891
		VELVET	-679.34333	442.543	0.722	-2466.1382	1107.4515
	EDENA	ABYSS	333.64333	285.432	0.89	-723.2936	1390.5803
		EULERSR	548.07667	247.967	0.398	-463.8626	1560.0159
		MIRA	-9900.03833	6153.38	0.69	-40538.8969	20738.8202
		PHRAP	-11256.03833	6318.17	0.612	-42717.9076	20205.8309
		SSAKE	-2014.03833	863.343	0.381	-6015.34	1987.2633
		VELVET	-345.7	471.312	0.986	-2150.7633	1459.3633
	EULERSR	ABYSS	-214.43333	187.607	0.896	-941.6408	512.7742
		EDENA	-548.07667	247.967	0.398	-1560.0159	463.8626
		MIRA	-10448.115	6149.62	0.649	-41108.9818	20212.7518
		PHRAP	-11804.115	6314.5	0.574	-43287.424	19679.194
		SSAKE	-2562.115	836.114	0.211	-6683.1104	1558.8804
		VELVET	-893.77667	419.357	0.443	-2705.4775	917.9242
	MIRA	ABYSS	10233.68167	6151.24	0.665	-20417.6574	40885.0208
		EDENA	9900.03833	6153.38	0.69	-20738.8202	40538.8969
		EULERSR	10448.115	6149.62	0.649	-20212.7518	41108.9818
		PHRAP	-1356	8813.37	1	-35009.4789	32297.4789
		SSAKE	7886	6204.97	0.839	-22469.3794	38241.3794
		VELVET	9554.33833	6162.67	0.717	-21030.9361	40139.6127
	PHRAP	ABYSS	11589.68167	6316.09	0.589	-19884.3469	43063.7102
		EDENA	11256.03833	6318.17	0.612	-20205.8309	42717.9076
		EULERSR	11804.115	6314.5	0.574	-19679.194	43287.424
		MIRA	1356 88	13.37	1	-32297.4789	35009.4789
		SSAKE	9242	6368.42	0.761	-21942.753	40426.753
		VELVET	10910.33833	6327.21	0.638	-20499.2903	42319.967
	SSAKE	ABYSS	2347.68167	847.98	0.264	-1714.2891	6409.6525
		EDENA	2014.03833	863.343	0.381	-1987.2633	6015.34
		EULERSR	2562.115	836.114	0.211	-1558.8804	6683.1104
		MIRA	-7886	6204.97	0.839	-38241.3794	22469.3794
		PHRAP	-9242	6368.42	0.761	-40426.753	21942.753
		VELVET	1668.33833	927.218	0.59	-2219.2381	5555.9147
	VELVET	ABYSS	679.34333	442.543	0.722	-1107.4515	2466.1382
		EDENA	345.7	471.312	0.986	-1459.3633	2150.7633
		EULERSR	893.77667	419.357	0.443	-917.9242	2705.4775
		MIRA	-9554.33833	6162.67	0.717	-40139.6127	21030.9361
		PHRAP	-10910.33833	6327.21	0.638	-42319.967	20499.2903
		SSAKE	-1668.33833	927.218	0.59	-5555.9147	2219.2381

Table E.21: Games-Howell post-hoc output for analyzing the performance of each assembler that ran on Sanger data based on the test metrics.

	(I) Assemblers	(J) Assemblers	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
NC	CAP3	EULERSR	4166.85	3304.997	0.877	-12966.7	21300.4
		MIRA	4357.6	3308.977	0.857	-12751.46	21466.66

Continued on next page

Table E.21 – continued from previous page

	Assemblers (I)	Assemblers (J)	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
NC	CAP3	PCAP	3393.4	3363.394	0.951	-13418.66	20205.46
		PHRAP	4437.433	3303.38	0.847	-12706.19	21581.05
		SSAKE	-2826.4	3791.56	0.991	-19119.29	13466.49
		TIGR	-14041.9	12480	0.927	-68944.3	40860.5
		VELVET	-84894.733	59450	0.816	-360864.48	191075.02
	EULERSR	CAP3	-4166.85	3304.997	0.877	-21300.4	12966.7
		MIRA	190.75	258.7	0.991	-993.69	1375.19
		PCAP	-773.45	655.754	0.906	-4016.47	2469.57
		PHRAP	270.583	172.948	0.757	-490.42	1031.58
		SSAKE	-6993.25	1869.097	0.182	-18591.63	4605.13
		TIGR	-18208.75	12030	0.779	-74208.69	37791.19
		VELVET	-89061.583	59360	0.785	-365345.14	187221.97
	MIRA	CAP3	-4357.6	3308.977	0.857	-21466.66	12751.46
		EULERSR	-190.75	258.7	0.991	-1375.19	993.69
		PCAP	-964.2	675.528	0.816	-4142.92	2214.52
		PHRAP	79.833	237.142	1	-1111.89	1271.55
		SSAKE	-7184	1876.126	0.17	-18698.74	4330.74
		TIGR	-18399.5	12030	0.772	-74395.65	37596.65
		VELVET	-89252.333	59360	0.783	-365535.12	187030.45
	PCAP	CAP3	-3393.4	3363.394	0.951	-20205.46	13418.66
		EULERSR	773.45	655.754	0.906	-2469.57	4016.47
		MIRA	964.2	675.528	0.816	-2214.52	4142.92
		PHRAP	1044.033	647.552	0.734	-2237.51	4325.58
		SSAKE	-6219.8	1970.517	0.234	-16908.16	4468.56
		TIGR	-17435.3	12050	0.808	-73379.81	38509.21
		VELVET	-88288.133	59360	0.791	-364560.3	187984.04
	PHRAP	CAP3	-4437.433	3303.38	0.847	-21581.05	12706.19
		EULERSR	-270.583	172.948	0.757	-1031.58	490.42
		MIRA	-79.833	237.142	1	-1271.55	1111.89
		PCAP	-1044.033	647.552	0.734	-4325.58	2237.51
		SSAKE	-7263.833	1866.236	0.167	-18897.39	4369.73
		TIGR	-18479.333	12030	0.769	-74480.82	37522.15
		VELVET	-89332.167	59360	0.783	-365616.03	186951.7
	SSAKE	CAP3	2826.4	3791.56	0.991	-13466.49	19119.29
		EULERSR	6993.25	1869.097	0.182	-4605.13	18591.63
		MIRA	7184	1876.126	0.17	-4330.74	18698.74
		PCAP	6219.8	1970.517	0.234	-4468.56	16908.16
		PHRAP	7263.833	1866.236	0.167	-4369.73	18897.39
		TIGR	-11215.5	12180	0.97	-66775.5	44344.5
		VELVET	-82068.333	59390	0.836	-358251.39	194114.73
	TIGR	CAP3	14041.9	12480	0.927	-40860.5	68944.3
		EULERSR	18208.75	12030	0.779	-37791.19	74208.69
		MIRA	18399.5	12030	0.772	-37596.65	74395.65
		PCAP	17435.3	12050	0.808	-38509.21	73379.81
		PHRAP	18479.333	12030	0.769	-37522.15	74480.82
		SSAKE	11215.5	12180	0.97	-44344.5	66775.5
		VELVET	-70852.833	60570	0.913	-343595.42	201889.75
	VELVET	CAP3	84894.733	59450	0.816	-191075.02	360864.48
		EULERSR	89061.583	59360	0.785	-187221.97	365345.14

Continued on next page

Table E.21 – continued from previous page

						95% Confidence Interval	
	Assemblers (I)	Assemblers (J)	Mean Difference (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
NC	VELVET	MIRA	89252.333	59360	0.783	-187030.45	365535.12
		PCAP	88288.133	59360	0.791	-187984.04	364560.3
		PHRAP	89332.167	59360	0.783	-186951.7	365616.03
		SSAKE	82068.333	59390	0.836	-194114.73	358251.39
		TIGR	70852.833	60570	0.913	-201889.75	343595.42
GC	CAP3	EULERSR	16.628	5.9167	0.319	-17.3454	50.6014
		MIRA	0.668	2.27165	1	-8.977	10.313
		PCAP	2.434	2.07441	0.919	-5.775	10.643
		PHRAP	2.79633	2.44518	0.929	-6.6673	12.26
		SSAKE	27.2105	7.85763	0.208	-19.5651	73.9861
		TIGR	1.04633	2.24685	1	-7.5882	9.6809
		VELVET	5.3205	3.32471	0.739	-11.0857	21.7267
	EULERSR	CAP3	-16.628	5.9167	0.319	-50.6014	17.3454
		MIRA	-15.96	5.99299	0.35	-49.3684	17.4484
		PCAP	-14.194	5.92104	0.432	-48.13	19.742
		PHRAP	-13.83167	6.0609	0.462	-46.6915	19.0282
		SSAKE	10.5825	9.61758	0.934	-32.329	53.494
		TIGR	-15.58167	5.98363	0.366	-48.9904	17.8271
		VELVET	-11.3075	6.46601	0.671	-42.8441	20.2291
	MIRA	CAP3	-0.668	2.27165	1	-10.313	8.977
		EULERSR	15.96	5.99299	0.35	-17.4484	49.3684
		PCAP	1.766	2.28295	0.989	-7.9048	11.4368
		PHRAP	2.12833	2.62442	0.987	-8.3081	12.5647
		SSAKE	26.5425	7.91524	0.219	-19.6883	72.7733
		TIGR	0.37833	2.4407	1	-9.4818	10.2385
		VELVET	4.6525	3.45867	0.85	-11.6978	21.0028
	PCAP	CAP3	-2.434	2.07441	0.919	-10.643	5.775
		MIRA	14.194	5.92104	0.432	-19.742	48.13
		PCAP	-1.766	2.28295	0.989	-11.4368	7.9048
		PHRAP	0.36233	2.45568	1	-9.1333	9.8579
		SSAKE	24.7765	7.860	0.257	-21.9656	71.5186
		TIGR	-1.38767	2.25826	0.998	-10.0641	7.2888
		VELVET	2.8865	3.33244	0.976	-13.4984	19.2714
	PHRAP	CAP3	-2.79633	2.44518	0.929	-12.26	6.6673
		EULERSR	13.83167	6.0609	0.462	-19.0282	46.6915
		MIRA	-2.12833	2.62442	0.987	-12.5647	8.3081
		PCAP	-0.36233	2.45568	1	-9.8579	9.1333
		SSAKE	24.41417	7.96678	0.265	-21.3131	70.1414
		TIGR	-1.75	2.60298	0.996	-11.5524	8.0524
		VELVET	2.52417	3.57504	0.993	-13.435	18.4833
	SSAKE	CAP3	-27.2105	7.85763	0.208	-73.9861	19.5651
		EULERSR	-10.5825	9.61758	0.934	-53.494	32.329
		MIRA	-26.5425	7.91524	0.219	-72.7733	19.6883
		PCAP	-24.7765	7.8609	0.257	-71.5186	21.9656
		PHRAP	-24.41417	7.96678	0.265	-70.1414	21.3131
		TIGR	-26.16417	7.90816	0.227	-72.4301	20.1018
		VELVET	-21.89	8.27915	0.343	-65.6485	21.8685
	TIGR	CAP3	-1.04633	2.24685	1	-9.6809	7.5882
		EULERSR	15.58167	5.98363	0.366	-17.8271	48.9904

Continued on next page

Table E.21 – continued from previous page

						95% Confidence Interval	
	Assemblers (I)	Assemblers (J)	Mean Difference (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
GC	TIGR	MIRA	-0.37833	2.4407	1	-10.2385	9.4818
		PCAP	1.38767	2.25826	0.998	-7.2888	10.0641
		PHRAP	1.75	2.60298	0.996	-8.0524	11.5524
		SSAKE	26.16417	7.90816	0.227	-20.1018	72.4301
		VELVET	4.27417	3.44243	0.888	-11.7959	20.3442
	VELVET	CAP3	-5.3205	3.32471	0.739	-21.7267	11.0857
		EULERSR	11.3075	6.46601	0.671	-20.2291	42.8441
		MIRA	-4.6525	3.45867	0.85	-21.0028	11.6978
		PCAP	-2.8865	3.33244	0.976	-19.2714	13.4984
		PHRAP	-2.52417	3.57504	0.993	-18.4833	13.435
		SSAKE	21.89	8.27915	0.343	-21.8685	65.6485
		TIGR	-4.27417	3.44243	0.888	-20.3442	11.7959
MC	CAP3	EULERSR	5.853	3.58575	0.725	-12.4793	24.1853
		MIRA	-0.492	4.09257	1	-18.0617	17.0777
		PCAP	0.274	5.39096	1	-21.1529	21.7009
		PHRAP	-0.72367	4.5471	1	-18.6727	17.2254
		SSAKE	6.2855	3.55941	0.666	-12.2036	24.7746
		TIGR	3.23967	3.75352	0.978	-14.3904	20.8698
		VELVET	6.1355	3.56121	0.685	-12.3423	24.6133
	EULERSR	CAP3	-5.853	3.58575	0.725	-24.1853	12.4793
		MIRA	-6.345	2.06627	0.269	-18.4822	5.7922
		PCAP	-5.579	4.0722	0.837	-26.472	15.314
		PHRAP	-6.57667	2.86294	0.425	-19.6448	6.4914
		SSAKE	0.4325	0.43556	0.948	-2.2777	3.1427
		TIGR	-2.61333	1.26858	0.514	-8.0309	2.8043
		VELVET	0.2825	0.45003	0.995	-2.2823	2.8473
	MIRA	CAP3	0.492	4.09257	1	-17.0777	18.0617
		EULERSR	6.345	2.06627	0.269	-5.7922	18.4822
		PCAP	0.766	4.52486	1	-19.1155	20.6475
		PHRAP	-0.23167	3.47678	1	-14.0117	13.5484
		SSAKE	6.7775	2.02022	0.235	-5.8648	19.4198
		TIGR	3.73167	2.34539	0.744	-7.1077	14.571
		VELVET	6.6275	2.02339	0.246	-5.9752	19.2302
	PCAP	CAP3	-0.274	5.39096	1	-21.7009	21.1529
		EULERSR	5.579	4.0722	0.837	-15.314	26.472
		MIRA	-0.766	4.52486	1	-20.6475	19.1155
		PHRAP	-0.99767	4.93977	1	-20.9429	18.9476
		SSAKE	6.0115	4.04903	0.789	-15.0212	27.0442
		TIGR	2.96567	4.22069	0.992	-17.2452	23.1765
		VELVET	5.8615	4.05061	0.805	-15.1613	26.8843
	PHRAP	CAP3	0.72367	4.5471	1	-17.2254	18.6727
		EULERSR	6.57667	2.86294	0.425	-6.4914	19.6448
		MIRA	0.23167	3.47678	1	-13.5484	14.0117
		PCAP	0.99767	4.93977	1	-18.9476	20.9429
		SSAKE	7.00917	2.82988	0.365	-6.1611	20.1795
		TIGR	3.96333	3.07047	0.877	-8.8259	16.7526
		VELVET	6.85917	2.83215	0.383	-6.3033	20.0217
	SSAKE	CAP3	-6.2855	3.55941	0.666	-24.7746	12.2036
		EULERSR	-0.4325	0.43556	0.948	-3.1427	2.2777

Continued on next page

Table E.21 – continued from previous page

	Assemblers (I)	Assemblers (J)	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
MC	SSAKE	MIRA	-6.7775	2.02022	0.235	-19.4198	5.8648
		PCAP	-6.0115	4.04903	0.789	-27.0442	15.0212
		PHRAP	-7.00917	2.82988	0.365	-20.1795	6.1611
		TIGR	-3.04583	1.19209	0.339	-8.5917	2.5001
		VELVET	-0.15	0.1197	0.876	-0.8451	0.5451
	TIGR	CAP3	-3.23967	3.75352	0.978	-20.8698	14.3904
		EULERSR	2.61333	1.26858	0.514	-2.8043	8.0309
		MIRA	-3.73167	2.34539	0.744	-14.571	7.1077
		PCAP	-2.96567	4.22069	0.992	-23.1765	17.2452
		PHRAP	-3.96333	3.07047	0.877	-16.7526	8.8259
		SSAKE	3.04583	1.19209	0.339	-2.5001	8.5917
		VELVET	2.89583	1.19746	0.383	-2.6323	8.424
	VELVET	CAP3	-6.1355	3.56121	0.685	-24.6133	12.3423
		EULERSR	-0.2825	0.45003	0.995	-2.8473	2.2823
		MIRA	-6.6275	2.02339	0.246	-19.2302	5.9752
		PCAP	-5.8615	4.05061	0.805	-26.8843	15.1613
		PHRAP	-6.85917	2.83215	0.383	-20.0217	6.3033
		SSAKE	0.15	0.1197	0.876	-0.5451	0.8451
		TIGR	-2.89583	1.19746	0.383	-8.424	2.6323
N50	CAP3	EULERSR	73243.75	70440	0.943	-291394.22	437881.72
		MIRA	-244115.5	208000	0.90	-1360770.37	872539.37
		PCAP	12836.4	92230	1	-354745.75	380418.55
		PHRAP	-47047.833	114200	1	-487389.47	393293.81
		SSAKE	82616.75	70330	0.905	-282725.33	447958.83
		TIGR	68111.667	71250	0.961	-291892.85	428116.19
		VELVET	82907.333	70330	0.904	-282435.51	448250.18
	EULERSR	CAP3	-73243.75	70440	0.943	-437881.72	291394.22
		MIRA	-317359.25	195800	0.732	-1542424.79	907706.29
		PCAP	-60407.35	59800	0.949	-369555.27	248740.57
		PHRAP	-120291.583	90030	0.854	-538629.13	298045.96
		SSAKE	9373	4008.106	0.463	-15570.23	34316.23
		TIGR	-5132.083	12110	1	-57130.16	46865.99
		VELVET	9663.583	4005.986	0.441	-15305.82	34632.99
	MIRA	CAP3	244115.5	208000	0.905	-872539.37	1360770.37
		EULERSR	317359.25	195800	0.732	-907706.29	1542424.79
		PCAP	256951.9	204700	0.876	-881824.82	1395728.62
		PHRAP	197067.667	215500	0.969	-881839.68	1275975.01
		SSAKE	326732.25	195800	0.713	-898847.05	1552311.55
		TIGR	312227.167	196100	0.744	-909177.91	1533632.25
		VELVET	327022.833	195800	0.712	-898557.02	1552602.69
	PCAP	CAP3	-12836.4	92230	1	-380418.55	354745.75
		EULERSR	60407.35	59800	0.949	-248740.57	369555.27
		MIRA	-256951.9	204700	0.876	-1395728.62	881824.82
		PHRAP	-59884.233	107900	0.999	-482255.76	362487.3
		SSAKE	69780.35	59670	0.907	-240193.31	379754.01
		TIGR	55275.267	60760	0.97	-248624.62	359175.15
		VELVET	70070.933	59670	0.906	-239903.62	380045.49
	PHRAP	CAP3	47047.833	114200	1	-393293.81	487389.47
		EULERSR	120291.583	90030	0.854	-298045.96	538629.13

Continued on next page

Table E.21 – continued from previous page

						95% Confidence Interval	
	Assemblers (I)	Assemblers (J)	Mean Difference (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
N50	PHRAP	MIRA	-197067.667	215500	0.969	-1275975.01	881839.68
		PCAP	59884.233	107900	0.999	-362487.3	482255.76
		SSAKE	129664.583	89950	0.811	-288977.57	548306.74
		TIGR	115159.5	90670	0.879	-301130.19	531449.19
		VELVET	129955.167	89950	0.809	-288687.32	548597.65
	SSAKE	CAP3	-82616.75	70330	0.905	-447958.83	282725.33
		EULERSR	-9373	4008.106	0.463	-34316.23	15570.23
		MIRA	-326732.25	195800	0.713	-1552311.55	898847.05
		PCAP	-69780.35	59670	0.907	-379754.01	240193.31
		PHRAP	-129664.583	89950	0.811	-548306.74	288977.57
		TIGR	-14505.083	11440	0.879	-67717.67	38707.51
		VELVET	290.583	328.471	0.978	-1072.62	1653.79
	TIGR	CAP3	-68111.667	71250	0.961	-428116.19	291892.85
		EULERSR	5132.083	12110	1	-46865.99	57130.16
		MIRA	-312227.167	196100	0.744	-1533632.25	909177.91
		PCAP	-55275.267	60760	0.97	-359175.15	248624.62
		PHRAP	-115159.5	90670	0.879	-531449.19	301130.19
		SSAKE	14505.083	11440	0.879	-38707.51	67717.67
		VELVET	14795.667	11440	0.87	-38419.49	68010.83
	VELVET	CAP3	-82907.333	70330	0.904	-448250.18	282435.51
		EULERSR	-9663.583	4005.986	0.441	-34632.99	15305.82
		MIRA	-327022.833	195800	0.712	-1552602.69	898557.02
		PCAP	-70070.933	59670	0.906	-380045.49	239903.62
		PHRAP	-129955.167	89950	0.809	-548597.65	288687.32
		SSAKE	-290.583	328.471	0.978	-1653.79	1072.62
		TIGR	-14795.667	11440	0.87	-68010.83	38419.49
Time	CAP3	EULERSR	4544.038	2030.09	0.471	-6001.6152	15089.6912
		MIRA	2850.3505	2186.24	0.867	-7153.4301	12854.1311
		PCAP	3627.81	2036.87	0.659	-6876.1591	14131.7791
		PHRAP	3885.398	2051.87	0.608	-6532.3597	14303.1557
		SSAKE	3303.503	2065.75	0.74	-7045.3272	13652.3332
		TIGR	-2280.16367	5580.03	1	-25831.5057	21271.1784
		VELVET	4465.90633	2030.94	0.485	-6074.3783	15006.1909
	EULERSR	CAP3	-4544.038	2030.09	0.471	-15089.6912	6001.6152
		MIRA	-1693.6875	811.5	0.549	-6772.5827	3385.2077
		PCAP	-916.22800*	166.434	0.041	-1778.4019	-54.0541
		PHRAP	-658.64	298.406	0.464	-2046.5829	729.3029
		SSAKE	-1240.535	382.364	0.252	-3632.2682	1151.1982
		TIGR	-6824.20167	5197.66	0.863	-31016.0801	17367.6767
		VELVET	-78.13167	60.12018	0.868	-353.3695	197.1062
	MIRA	CAP3	-2850.3505	2186.24	0.867	-12854.1311	7153.4301
		EULERSR	1693.6875	811.5	0.549	-3385.2077	6772.5827
		PCAP	777.4595	828.307	0.96	-4113.1669	5668.0859
		PHRAP	1035.0475	864.545	0.897	-3572.249	5642.344
		SSAKE	453.1525	896.991	0.999	-4047.2686	4953.5736
		TIGR	-5130.51417	5260.61	0.961	-29128.4415	18867.4131
		VELVET	1615.55583	813.637	0.586	-3436.7386	6667.8503
	PCAP	CAP3	-3627.81	2036.87	0.659	-14131.7791	6876.1591
		EULERSR	916.22800*	66.434	0.041	54.0541	1778.4019

Continued on next page

Table E.21 – continued from previous page

						95% Confidence Interval	
	Assemblers (I)	Assemblers (J)	Mean Difference (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
Time	PCAP	MIRA	-777.4595	828.307	0.96	-5668.0859	4113.1669
		PHRAP	257.588	341.476	0.991	-1109.6849	1624.8609
		TIGR	-5907.97367	5200.31	0.922	-30090.6888	18274.7414
		SSAKE	-324.307	416.848	0.986	-2452.1554	1803.5414
		VELVET	838.09633*	176.562	0.046	18.148	1658.0447
	PHRAP	CAP3	-3885.398	2051.87	0.608	-14303.1557	6532.3597
		EULERSR	658.64	298.406	0.464	-729.3029	2046.5829
		MIRA	-1035.0475	864.545	0.897	-5642.344	3572.249
		PCAP	-257.588	341.476	0.991	-1624.8609	1109.6849
		SSAKE	-581.895	484.879	0.907	-2639.3928	1475.6028
		TIGR	-6165.56167	5206.21	0.907	-30328.1596	17997.0363
		VELVET	580.50833	304.171	0.592	-790.4934	1951.5101
	SSAKE	CAP3	-3303.503	2065.75	0.74	-13652.3332	7045.3272
		EULERSR	1240.535	382.364	0.252	-1151.1982	3632.2682
		MIRA	-453.1525	896.991	0.999	-4953.5736	4047.2686
		PCAP	324.307	416.848	0.986	-1803.5414	2452.1554
		PHRAP	581.895	484.879	0.907	-1475.6028	2639.3928
		TIGR	-5583.66667	5211.69	0.939	-29728.0816	18560.7483
		VELVET	1162.40333	386.88	0.288	-1176.0623	3500.869
	TIGR	CAP3	2280.16367	5580.03	1	-21271.1784	25831.5057
		EULERSR	6824.20167	5197.66	0.863	-17367.6767	31016.0801
		MIRA	5130.51417	5260.61	0.961	-18867.4131	29128.4415
		PCAP	5907.97367	5200.31	0.922	-18274.7414	30090.6888
		PHRAP	6165.56167	5206.21	0.907	-17997.0363	30328.1596
		SSAKE	5583.66667	5211.69	0.939	-18560.7483	29728.0816
		VELVET	6746.07	5197.99	0.868	-17444.6486	30936.7886
	VELVET	CAP3	-4465.90633	2030.94	0.485	-15006.1909	6074.3783
		EULERSR	78.13167	60.12018	0.868	-197.1062	353.3695
		MIRA	-1615.55583	813.637	0.586	-6667.8503	3436.7386
		PCAP	-838.09633*	176.562	0.046	-1658.0447	-18.148
		PHRAP	-580.50833	304.171	0.592	-1951.5101	790.4934
		SSAKE	-1162.40333	386.88	0.288	-3500.869	1176.0623
		TIGR	-6746.07	5197.99	0.868	-30936.7886	17444.6486
Mem.	CAP3	EULERSR	853.6915	309.194	0.308	-747.9902	2455.3732
		MIRA	421.8465	333.736	0.882	-1098.4706	1942.1636
		PCAP	620.444	308.8	0.561	-983.6954	2224.5834
		PHRAP	-1834.80933	2129.74	0.979	-11571.5703	7901.9516
		SSAKE	432.024	310.078	0.827	-1164.3108	2028.3588
		TIGR	-482.58433	541.333	0.979	-2590.4315	1625.2629
		VELVET	257.399	457.891	0.999	-1501.5192	2016.3172
	EULERSR	CAP3	-853.6915	309.194	0.308	-2455.3732	747.9902
		MIRA	-431.845	127.54	0.225	-1212.6241	348.9341
		PCAP	-233.24750*	15.62389	0.004	-330.9021	-135.5929
		PHRAP	-2688.50083	2107.29	0.876	-12496.1875	7119.1858
		SSAKE	-421.66750*	32.18048	0.001	-575.8713	-267.4637
		TIGR	-1336.27583	444.891	0.222	-3404.7485	732.1968
		VELVET	-596.2925	338.454	0.663	-2168.6732	976.0882
	MIRA	CAP3	-421.8465	333.736	0.882	-1942.1636	1098.4706
		EULERSR	431.845	127.54	0.225	-348.9341	1212.6241

Continued on next page

Table E.21 – continued from previous page

						95% Confidence Interval	
	Assemblers (I)	Assemblers (J)	Mean Difference (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
Mem.	MIRA	PCAP	198.5975	126.581	0.753	-593.738	990.933
		PHRAP	-2256.65583	2111.03	0.94	-12051.5819	7538.2702
		TIGR	-904.43083	462.284	0.567	-2933.3729	1124.5112
		SSAKE	10.1775	129.67	1	-748.5722	768.9272
		VELVET	-164.4475	361.012	1	-1701.1967	1372.3017
	PCAP	CAP3	-620.444	308.8	0.561	-2224.5834	983.6954
		EULERSR	233.24750*	15.62389	0.004	135.5929	330.9021
		MIRA	-198.5975	126.581	0.753	-990.933	593.738
		PHRAP	-2455.25333	2107.23	0.913	-12263.1407	7352.634
		SSAKE	-188.42000*	28.14178	0.042	-364.4971	-12.3429
		TIGR	-1103.02833	444.617	0.363	-3172.4464	966.3898
		VELVET	-363.045	338.094	0.939	-1936.6634	1210.5734
	PHRAP	CAP3	1834.80933	2129.74	0.979	-7901.9516	11571.5703
		EULERSR	2688.50083	2107.29	0.876	-7119.1858	12496.1875
		MIRA	2256.65583	2111.03	0.94	-7538.2702	12051.5819
		PCAP	2455.25333	2107.23	0.913	-7352.634	12263.1407
		SSAKE	2266.83333	2107.42	0.938	-7540.4024	12074.0691
		TIGR	1352.225	2153.63	0.996	-8321.4823	11025.9323
		VELVET	2092.20833	2134.18	0.96	-7631.4361	11815.8527
	SSAKE	CAP3	-432.024	310.078	0.827	-2028.3588	1164.3108
		EULERSR	421.66750*	32.18048	0.001	267.4637	575.8713
		MIRA	-10.1775	129.67	1	-768.9272	748.5722
		PCAP	188.42000*	28.14178	0.042	12.3429	364.4971
		PHRAP	-2266.83333	2107.42	0.938	-12074.0691	7540.4024
		VELVET	-174.625	339.262	0.999	-1744.32	1395.07
		TIGR	-914.60833	445.506	0.529	-2980.9981	1151.7814
	TIGR	CAP3	482.58433	541.333	0.979	-1625.2629	2590.4315
		EULERSR	1336.27583	444.891	0.222	-732.1968	3404.7485
		MIRA	904.43083	462.284	0.567	-1124.5112	2933.3729
		PCAP	1103.02833	444.617	0.363	-966.3898	3172.4464
		PHRAP	-1352.225	2153.63	0.996	-11025.9323	8321.4823
		SSAKE	914.60833	445.506	0.529	-1151.7814	2980.9981
		VELVET	739.98333	558.562	0.869	-1387.2604	2867.2271
	VELVET	CAP3	-257.399	457.891	0.999	-2016.3172	1501.5192
		EULERSR	596.2925	338.454	0.663	-976.0882	2168.6732
		MIRA	164.4475	361.012	1	-1372.3017	1701.1967
		PCAP	363.045	338.094	0.939	-1210.5734	1936.6634
		PHRAP	-2092.20833	2134.18	0.96	-11815.8527	7631.4361
		SSAKE	174.625	339.262	0.999	-1395.07	1744.32
		TIGR	-739.98333	558.562	0.869	-2867.2271	1387.2604

Table E.22: Table of t-test results for analyzing the metrics of the assemblers when run with Sanger reads with (+ QUAL) and without (- QUAL) quality values.

Independent Samples Test							
			Levene's Test for Equality of Variances		t-test for Equality of Means		
Assembler (+QUAL vs -QUAL)	Metric	Equal variances	F	Sig.	t	df	Sig. (2-tailed)
CAP3	NC	assumed	0	0.988	-0.002	4	0.999
		not assumed			-0.002	4	0.999
	GC	assumed	0.01	0.925	0.047	4	0.965
		not assumed			0.047	3.993	0.965
	MC	assumed	6.497	0.063	0.983	4	0.381
		not assumed			0.983	2.388	0.414
	N50	assumed	0.109	0.758	-0.078	4	0.941
		not assumed			-0.078	3.946	0.941
	Time	assumed	0.007	0.936	0.226	4	0.832
		not assumed			0.226	3.991	0.832
	Mem.	assumed	0.038	0.855	-0.054	4	0.96
		not assumed			-0.054	3.977	0.96
PHRAP	NC	assumed	0.402	0.56	0.133	4	0.901
		not assumed			0.133	3.791	0.901
	GC	assumed	0.017	0.904	0.046	4	0.966
		not assumed			0.046	3.991	0.966
	MC	assumed	0.618	0.476	-0.587	4	0.589
		not assumed			-0.587	3.474	0.593
	N50	assumed	0.002	0.971	0.005	4	0.996
		not assumed			0.005	3.999	0.996
	Time	assumed	0.035	0.861	0.044	4	0.967
		not assumed			0.044	3.983	0.967
	Mem.	assumed	0.023	0.887	0.153	4	0.886
		not assumed			0.153	3.978	0.886
TIGR	NC	assumed	1.027	0.368	-1.01	4	0.37
		not assumed			-1.01	3.266	0.381
	GC	assumed	0.004	0.955	-0.054	4	0.96
		not assumed			-0.054	3.998	0.96
	MC	assumed	1.895	0.241	1.36	4	0.246
		not assumed			1.36	2.631	0.276
	N50	assumed	9.581	0.036	1.064	4	0.347
		not assumed			1.064	2.144	0.392
	Time	assumed	7.614	0.051	-1.292	4	0.266
		not assumed			-1.292	2.17	0.317
	Mem.	assumed	0.382	0.57	-0.223	4	0.834
		not assumed			-0.223	3.814	0.835

Table E.23: Group statistics for the data in Table E.22.

	Assemblers	N	Mean	Std. Deviation	Std. Error Mean
NC	CAP3 + QUAL	3	232.33	229.317	132.396
	CAP3 - QUAL	3	232.67	229.454	132.476
	PHRAP + QUAL	3	228.33	293.367	169.376
	PHRAP - QUAL	3	199.67	230.994	133.365
	TIGR + QUAL	3	2417.67	1554.412	897.44
	TIGR - QUAL	3	4185.67	2602.686	1502.661
GC	CAP3 + QUAL	3	97.02	3.99576	2.30695
	CAP3 - QUAL	3	96.8633	4.16046	2.40204
	PHRAP + QUAL	3	96.6767	4.28604	2.47455
	PHRAP - QUAL	3	96.52	4.08934	2.36098
	TIGR + QUAL	3	97.7833	2.91589	1.68349
	TIGR - QUAL	3	97.91	2.84522	1.64269
MC	CAP3 + QUAL	3	10.3	8.42874	4.86633
	CAP3 - QUAL	3	5.2867	2.63641	1.52213
	PHRAP + QUAL	3	1.7633	2.608687	1.50612
	PHRAP - QUAL	3	3.3633	3.93447	2.27157
	TIGR + QUAL	3	4.41	2.49742	1.44189
	TIGR - QUAL	3	2.2967	1.0052	0.58035
N50	CAP3 + QUAL	3	137000	196568.12	113488.66
	CAP3 - QUAL	3	151000	220978.71	127582.12
	PHRAP + QUAL	3	214000	314196.09	181401.198
	PHRAP - QUAL	3	213000	309801.76	178864.127
	TIGR + QUAL	3	29400	36984.619	21353.08
	TIGR - QUAL	3	6258	7026.817	4056.935
Time	CAP3 + QUAL	3	2180.1	1037.4719	589.98468
	CAP3 - QUAL	3	1992.9	989.33202	571.19111
	PHRAP + QUAL	3	517.64	563.51161	325.34358
	PHRAP - QUAL	3	497.98	527.71079	304.67397
	TIGR + QUAL	3	681.07	347.63445	200.70685
	TIGR - QUAL	3	1964.7	1685.95404	973.38602
Mem.	CAP3 + QUAL	3	627.1	138.47712	79.9498
	CAP3 - QUAL	3	633.42	149.4654	86.29389
	PHRAP + QUAL	3	406.8	90.16323	52.05577
	PHRAP - QUAL	3	395.92	83.75823	48.35784
	TIGR + QUAL	3	1022.5	440.45028	254.29409
	TIGR - QUAL	3	1095.2	351.82148	203.12423

Table E.24: Table of t-test results for analyzing the metrics of the assemblers when run with Illumina reads with (+ QUAL) and without (- QUAL) quality values.

Independent Samples Test							
			Levene's Test for Equality of Variances		t-test for Equality of Means		
Assembler (+QUAL vs -QUAL)	Metric	Equal variances	F	Sig.	t	df	Sig. (2-tailed)
EDENA	NC	assumed	0.125	0.742	2.335	4	0.08
		not assumed			2.335	3.895	0.082
	GC	assumed	0.33	0.597	-1.652	4	0.174
		not assumed			-1.652	3.808	0.177
	MC	assumed	2.451	0.193	-1.046	4	0.354
		not assumed			-1.046	2.647	0.381
	N50	assumed	0.728	0.442	-1.044	4	0.356
		not assumed			-1.044	3.267	0.368
	Time	assumed	0.346	0.588	-0.459	4	0.67
		not assumed			-0.459	3.797	0.671
	Mem.	assumed	0.067	0.808	0.236	4	0.825
		not assumed			0.236	3.971	0.825
EULER-SR	NC	assumed	4.78	0.094	0.519	4	0.631
		not assumed			0.519	2.556	0.645
	GC	assumed	8.77	0.041	-1.042	4	0.356
		not assumed			-1.042	2.224	0.397
	MC	assumed	3.55	0.133	0.43	4	0.69
		not assumed			0.43	2.829	0.698
	N50	assumed	0.07	0.807	-0.339	4	0.752
		not assumed			-0.339	3.844	0.753
	Time	assumed	6.74	0.06	-0.51	4	0.637
		not assumed			-0.51	2.361	0.654
	Mem.	assumed	4.65	0.097	2.065	4	0.108
		not assumed			2.065	2.006	0.175
VELVET	NC	assumed	7.253	0.054	0.701	4	0.522
		not assumed			0.701	2.31	0.547
	GC	assumed	0.322	0.601	-1.407	4	0.232
		not assumed			-1.407	3.847	0.235
	MC	assumed	0.072	0.802	-0.533	4	0.622
		not assumed			-0.533	3.976	0.622
	N50	assumed	0.011	0.923	-0.055	4	0.959
		not assumed			-0.055	3.991	0.959
	Time	assumed	0.089	0.781	-0.484	4	0.654
		not assumed			-0.484	3.898	0.654
	Mem.	assumed	0.001	0.98	-0.429	4	0.69
		not assumed			-0.429	4	0.69

Table E.25: Group statistics for the data in Table E.24.

	Assemblers	N	Mean	Std. Deviation	Std. Error Mean
NC	EDENA + QUAL	3	1556.7	361.11	208.487
	EDENA - QUAL	3	803.67	426.176	246.053
	EULER-SR + QUAL	3	2114.67	2413.127	1393.219
	EULER-SR - QUAL	3	1341.33	908.763	524.675
	VELVET + QUAL	3	1299.33	1266.885	731.436
	VELVET - QUAL	3	766.67	353.906	204.328
GC	EDENA + QUAL	3	95.297	1.70371	0.98364
	EDENA - QUAL	3	97.373	1.35593	0.78284
	EULER-SR + QUAL	3	92.4333	8.41351	4.85755
	EULER-SR - QUAL	3	97.6333	1.99613	1.15247
	VELVET + QUAL	3	97.7967	0.3782	0.21835
	VELVET - QUAL	3	98.1933	0.30892	0.17836
MC	EDENA + QUAL	3	0.05	0.04583	0.02646
	EDENA - QUAL	3	0.1233	0.1124	0.06489
	EULER-SR + QUAL	3	1.0733	1.66386	0.96063
	EULER-SR - QUAL	3	0.618	0.77533	0.44764
	VELVET + QUAL	3	0.5167	0.835	0.48209
	VELVET - QUAL	3	0.8667	0.77184	0.44562
N50	EDENA + QUAL	3	4022.3	4175.241	2410.577
	EDENA - QUAL	3	8927.3	6987.303	4034.121
	EULER-SR + QUAL	3	10600	9148.569	5281.929
	EULER-SR - QUAL	3	13400	11222.778	6479.474
	TIGR + QUAL	3	19700	25412.76	14672.064
	TIGR - QUAL	3	20900	24247.277	13999.172
Time	EDENA + QUAL	3	1515.7	897.01295	517.89067
	EDENA - QUAL	3	1898.8	1134.87535	655.22059
	EULER-SR + QUAL	3	1908.7	1618.62443	934.51325
	EULER-SR - QUAL	3	3558.8	5364.22061	3097.03421
	VELVET + QUAL	3	493.09	446.31848	257.68209
	VELVET - QUAL	3	685.8	525.57648	303.44172
Mem.	EDENA + QUAL	3	1304	695.04832	401.28633
	EDENA - QUAL	3	1175.6	637.53807	368.08278
	EULER-SR + QUAL	3	1612.6	1269.16618	732.75344
	EULER-SR - QUAL	3	98.2633	50.88475	29.37833
	VELVET + QUAL	3	1433.8	1103.23506	636.95306
	VELVET - QUAL	3	1819.7	1100.7918	635.54244

Table E.26: Table of t-test results for analyzing the metrics of the assemblers when run with 454 reads with (+ QUAL) and without (- QUAL) quality values.

Independent Samples Test							
			Levene's Test for Equality of Variances		t-test for Equality of Means		
Assembler (+QUAL vs -QUAL)	Metric	Equal variances	F	Sig.	t	df	Sig. (2-tailed)
EULER-SR	NC	assumed	5.547	0.057	1.318	6	0.216
		not assumed			1.381	3.036	0.26
	GC	assumed	8.135	0.029	-1.044	6	0.337
		not assumed			-1.044	3.012	0.373
	MC	assumed	4.637	0.075	1.282	6	0.247
		not assumed			1.282	3.068	0.288
	N50	assumed	0.024	0.881	-0.11	6	0.916
		not assumed			-0.11	5.966	0.916
	Time	assumed	1.394	0.282	-0.13	6	0.901
		not assumed			-0.13	4.739	0.902
	Mem.	assumed	0.752	0.419	0.43	6	0.682
		not assumed			0.43	5.188	0.684
VELVET	NC	assumed	0	1	0	6	1
		not assumed			0	6	1
	GC	assumed	0	1	0	6	1
		not assumed			0	6	1
	MC	assumed	0	1	0	6	1
		not assumed			0	6	1
	N50	assumed	0	1	0	6	1
		not assumed			0	6	1
	Time	assumed	0.218	0.657	-0.334	6	0.75
		not assumed			-0.334	5.798	0.75
	Mem.	assumed	3.065	0.131	0.805	6	0.452
		not assumed			0.805	5.445	0.455

Table E.27: Group statistics for the data in Table E.26.

	Assemblers	N	Mean	Std. Deviation	Std. Error Mean
NC	EULER-SR + QUAL	4	3285.75	3455.8	1727.899
	EULER-SR - QUAL	4	892	267.959	133.979
	VELVET + QUAL	4	16300	12082.851	6041.426
	VELVET - QUAL	4	16300	12082.851	6041.426
GC	EULER-SR + QUAL	4	72.0775	48.0923	24.04613
	EULER-SR - QUAL	4	97.2025	2.19353	1.09676
	VELVET + QUAL	4	97.365	2.36517	1.18259
	VELVET - QUAL	4	97.365	2.36517	1.18259
MC	EULER-SR + QUAL	4	3.5275	4.56749	2.28374
	EULER-SR - QUAL	4	0.5825	0.48801	0.24401
	VELVET + QUAL	4	0.9575	0.64552	0.32276
	VELVET - QUAL	4	0.9575	0.64552	0.32276
N50	EULER-SR + QUAL	4	18900	27251.5	13625.768
	EULER-SR - QUAL	4	21000	25264	12632.004
	TIGR + QUAL	4	1065.8	1591.488	795.744
	TIGR - QUAL	4	1065.8	1591.488	795.744
Time	EULER-SR + QUAL	4	601.6	489.736	244.86778
	EULER-SR - QUAL	4	666.36	866.643	433.32163
	VELVET + QUAL	4	152.41	67.8605	33.93025
	VELVET - QUAL	4	170.18	81.95236	40.97618
Mem.	EULER-SR + QUAL	4	342.35	207.564	103.78208
	EULER-SR - QUAL	4	261.17	315.418	157.70922
	VELVET + QUAL	4	565.44	208.51504	104.25752
	VELVET - QUAL	4	462.14	149.75848	74.87924