Methods for Transcriptome Assembly in the Allopolyploid *Brassica napus*

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

By
Miles R. Buchwaldt

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

Canada is the world's largest producer of canola and the trend of production is ever increasing with an annual growth rate of 9.38% according to FAOSTAT [1]. In 2017, canola acreage surpassed wheat in Saskatchewan, the highest producer of both crops in Canada. Country-wide, the total farming area of canola increased by 9.9% to 22.4 million acres while wheat area saw a slight decline to 23.3 million acres [2]. While Canada is the highest producer of the crop, yields are lower than other countries [1]. To maximize the benefit of this market, canola cultivation could be made more efficient with further characterization of the organism's genes and their involvement in plant robustness. Such studies using transcriptome analysis have been successful in organisms with relatively small and simple genomes. However, such analyses in *B. napus* are complicated by the allopolyploid genome structure resulting from ancestral whole genome duplications in the species' evolutionary history. Homeologous gene pairs originating from the orthology between the two *B. napus* progenitor species complicate the process of transcriptome assembly. Modern assemblers: Trinity [3], Oases [4] and SOAPdenovo-Trans [5] were used to generate several *de novo* transcriptome assemblies for *B. napus*. A variety of metrics were used to determine the impact that the complex genome structure has on transcriptome studies. In particular, the most important questions for transcriptome assembly in *B. napus* were how does varying the k-mer parameter effect assembly quality, and to what extent do similar genes resulting from homeology within *B. napus* complicate the process of assembly. These metrics used for evaluating the assemblies include basic assembly statistics such as the number of contigs and contig lengths (via N25, N50 and N75 statistics); as well as more involved investigation via comparison to annotated coding DNA sequences; evaluation softwares scores for *de novo* transcriptome assemblies and finally; quantification of homeolog differentiation by alignment to previously identified pairs of homeologous genes. These metrics provided a picture of the trade-offs between assembly softwares and the k-parameter determining the length of subsequences used to build de Bruijn graphs for *de novo* transcriptome assembly. It was shown that shorter k-mer lengths produce fewer, and more complete contigs due to the shorter required overlap between read sequences; while longer k-mer lengths increase the sensitivity of an assembler to sequence variation between similar gene sequences. The Trinity assembler outperformed Oases and SOAPdenovo-Trans when considering the total breadth of evaluation metrics, generating longer transcripts with fewer chimers between homeologous gene pairs.

Acknowledgements

For the support I have received throughout my studies, I would like to express the utmost gratitude to both of my co-supervisors: Dr. Isobel Parkin and Dr. Matthew Links. In addition to the staggering knowledge and experience they both have shared, their dedication to my personal and academic development has been a crucial positive factor for this thesis. Dr. Isobel Parkin's vast breadth of knowledge has been a great inspiration to me, as well as a safety net at times. Dr. Matthew Links' reassuring patience and calm approach the countless problems I put forth, whether they were trivial or monumentally complex, has always kept me focused on the important things. I would also like to thank the members of my committee: Dr. Ian McQuillan and Dr. Tony Kusalik who have provided invaluable feedback on my thesis. Special thanks to the Saskatoon Research and Development Centre (SRDC) colleagues whose expertise I have relied on. In particular, Erin Higgins for generating the sequencing data this project utilized and Wayne Clarke for the countless free lessons in Perl he provided.

I also must thank my family, Lone and Andreas, who have supported and advised my personal and professional growth. Finally, I thank my partner Erin Hopkins, who inspires me with unrivaled generosity, intelligence, and work ethic. You have been the highlight of my life, enhancing every experience we have shared together.

Funding for my Master's degree was provided by the Agriculture and Agri-Food Canada (AAFC) Canadian Crops Genomics Initiative and SaskCanola.

To my father and role-model, Dr. Roger Rimmer, whose legacy motivates and inspires me.

Contents

List of Tables

List of Figures

List of Abbreviations

bp	base pairs
BLAST	Basic Local Alignment Search Tool
BLAT	BLAST Like Alignment Tool
BVS	biojs-vis-sequence
CDS	Coding DNA Sequence
D3	Data Driven Documents
DBG	De Bruijn Graph
DNA	Deoxyribonucleic Acid
dNTP	Deoxynucleotide Triphosphate
Gb	Giga basepairs
GVART	Graph Visualization of Assembled RNA Transcripts
mRNA	Messenger RNA
OLC	Overlap Layout Consensus
ORF	Open Reading Frame
RNA	Ribonucleic Acid
SBS	Sequencing By Synthesis
SDT	SOAPdenovo-Trans
SNP	Single Nucleotide Polymorphism
SRDC	Saskatoon Research and Development Centre
UTR	Untranslated Region
TSS	Transcription Start Site
WGD	Whole Genome Duplication
ZMW	Zero-mode Waveguide

Chapter 1

Introduction

Since the beginning of DNA sequencing 40 years ago, technology has advanced significantly, allowing unprecedented volumes of data to be generated from labs of all sizes. Early methods for DNA sequencing by the Sanger method involved long, labour intensive wet lab protocols. These approaches did not scale well to whole-genome shotgun sequencing, leading to the large costs observed for genome assembly experiments such as the Human Genome Project [6]. The research cost of the first human genome has fallen from roughly $100 million in 2001 to nearly $1000 as of 2015 [7]. This has facilitated genome-wide studies of a great variety of organisms, resulting in the assembly and publication of full genomes, transcriptomes and proteomes. As the cost for this data decreases, analysis and man-power becomes the bottleneck for conducting genome-wide studies. While there are more genomes available than ever, there are still many organisms that are of interest to researchers, whose genome is too complicated to commit the resources to a genome assembly. For example, many plant genomes are much larger, and more repetitive than animals, which complicates the process of determining the origin of reads corresponding to short segments of the genome.

Transcriptome assembly is a less resource intensive, genome-wide study that can be done in non-model organisms. Using *de novo* (without a reference) assemblers, RNA-seq data sampled from an organism can be assembled to generate a snapshot of the transcripts present in the sample. The full transcriptome can provide a wealth of information about the state of an organism: for example, the set of genes that are expressed, or unexpressed and the relative expression levels of these genes. When coupled with tissue-specific sampling or growth condition studies, this data provides insight by identifying those genes that respond to environmental changes perceived by the plant.

There are some difficulties however, in producing a transcriptome with no reference genome to compare to. Determining the quality of a set of transcripts can be difficult when there is little information about the true set of genes in an organism. In that case, comparison to a closely related organism is the only means of confirmation. For organisms with large, complex, or repetitive genomes, assembling transcripts is further complicated. Polyploid organisms, which contain multiple sets of homologous chromosomes, can contain many copies of genes. Sequence data derived from genes with a close evolutionary origin, but distinct genomic loci are difficult for computational techniques to differentiate. This phenomenon can result in the mis-assembly of distinct transcripts into chimeric transcripts. Additionally, the great variation in expression level of transcripts compounded with errors in sequencing data can be confused for single nucleotide polymorphisms

of orthologous gene copies. Reference-based methods can rely too heavily on the known genome sequence, limiting the study to only the sequences that appear in the reference. Even the most thoroughly annotated reference genomes do not represent the complete set of genes expressed for a species, as has been shown in recent studies [8]. These potentially unknown genes can be particularly lucrative to breeders, if they are responsible for a trait of interest, such as yield or stress tolerance.

Brassica napus (*B. napus*, rapeseed) is a cruciferous plant of the Brassicaceae family. The term canola refers to the crops developed as part of a breeding project in the 1970s using rapeseed, producing a group of oilseed crops defined by a fatty acid profile with less than 2% erucic acid and low levels of seed glucosinolates (< 30 μM). The *Brassica napus* species, in its evolutionary history has undergone several whole genome duplication (WGD) events that have resulted in an allopolyploid AACC genome structure. As a result, the genome contains many homeologous gene pairs between its A and C sub-genomes which originate from *Brassica rapa* and *Brassica oleracea* orthologous pairs. This is a source of ambiguity during *de novo* assembly where RNA sequences originating from different homeolog pairs (transcribed from distinct genes) share similar sequences of nucleotides. Assembled transcripts can then be constructed from a mixture of RNA sequences deriving from distinct genomic locations. The problem is exacerbated when the RNA sequences are divided further into subsequences called k-mers, the basis for the de Bruijn graph method of assembly. When a sequence read is broken into smaller piece, the read continuity of the sequence is lost.

The two major questions that this thesis focuses on are: what are the effects of k-mer length on de Bruijn graph assembly, and to what extent do orthologous gene pairs present in *B. napus* complicate the assembly process. These are addressed by assembling multiple transcriptomes using current *de novo* assemblers: Trinity, Oases and SOAPdenovo-Trans for varying length of k-mers. The performance of these assemblers was evaluated and compared using a variety of metrics covering a breadth of assembly qualities. The most basic criteria used for assembly evaluation are the quantity of transcripts and their lengths, measured by mean contig lengths and N25, N50 and N75 scores. To determine the ability of an assembler to recreate actual gene sequences, assembled transcripts were mapped against a dataset of predicted coding DNA sequences (CDSs) from the *B. napus* genome. The lengths of these alignments as well as the proportion of unique CDSs mapped by each assembly were considered. Finally, the extent to which related gene pairs, or orthologs, are erroneously collapsed together into single transcripts was determined by sequence comparison to known gene pairs in *B. napus*.

A method for visualization and exploration of the de Bruijn graphs computed by Trinity was developed in the form of a web-based application called GVART (Graph Visualization of Assembled RNA Transcripts). This tool allows a researcher to see the branches and paths followed within the simplified de Bruijn graph created by Trinity to produce the resulting assembled transcript, as well as some highlighted features of the reported transcript sequence.

The groundwork concepts and technologies that this thesis builds on are covered in Chapter 2. Chapter 3 explains the specific problems and goals of the research project. The data files and procedures used to

compare *de novo* transcriptome assemblers, as well as the implementation of the graph visualization tool are discussed in Chapter 4. Chapter 5 presents the results of evaluating several *de novo* transcriptome assemblies with a variety of quality metrics. The de Bruijn graph visualization tool, GVART, developed as part of this project is presented in Chapter 6. Finally, Chapter 7 contains a discussion of the implications of some of our findings, as well as some possible directions for future study.

Chapter 2

Background

2.1 *Brassica napus*

The cruciferous plant *Brassica napus* is an oilseed crop that is seeing increased production around the world, being used for human edible oils, a protein source for animal feed, and as a feedstock in the production of biodiesel [9]. A major breeding project by Keith Downey and Baldur Stefansson undertaken in the 1970s used rapeseed to produce the first variety of canola [10], a group of oilseed crops that are defined by a fatty acid profile with less than 2% erucic acid and low levels of seed glucosinolates (< 30 μM). Since then, the production of canola has steadily increased throughout North America, China and northern Europe with the top producers being Canada, China, India, Germany and France respectively [1]. In Canada, canola almost surpassed wheat in acres farmed, reaching 22.4 million versus wheat's 23.3 million acres in 2017 [2]. More recent studies of *B. napus* include the publication of the genome of the French variety 'Darmor-bzh' as well as homology analysis of *napus's* 19 chromosomes to the chromosomes of it's progenitor species and their ancestor species; *Arabidopsis thaliana* and the more distant *Vitis vinifera* and *Amborella trichopoda* [11]. These studies have shown that the genome of *B. napus* has undergone a 72-fold duplication of it's genome since the first angiosperms [11, 12].

B. napus has been the focus of many breeding projects, first to reduce erucic acid and glucosinolate content to make the oils better for human consumption. Current areas of crop development are focused on increasing yield, either through increasing abiotic stress tolerance or by plant pathogen resistance. In either case, information about important genes and nearby markers are crucial for breeders to genotype their crops and relate genomic information to phenotypic traits. These research interests have sparked the need for transcriptome analysis studies in *B. napus* to provide growers with improved crop qualities that meet modern demands.

Polyploid genomes are defined as those that possess three or more complete sets of chromosomes. Polyploidy is common in eukaryotes, and plays a important role in plant speciation [13]. The *B. napus* genome is a polyploid genome of 19 chromosomes, 10 originating from *Brassica rapa* and 9 from *Brassica oleracea* [14]. The formation of the hybrid *B. napus* genome occurred roughly 7500 years ago [11] by the hybridization of the two related progenitors species. The *B. napus* genome, formed from the joining of two different species, is an example of allopolyploidy. Allopolyploids differ from autopolyploids in that they represent polyploids formed

from the hybridization of two species, rather than a polyploid arising from a cross between populations of a single species resulting in a genome doubling [13]. This genome is an example of tetraploidy (containing four genomes), resulting from the hybridization of two diploid genomes each containing two genomes. The fact that the two progenitors of *B. napus* are so closely related has lead to relatively large proportion of gene duplication within the compound genome of *B. napus*. Gene duplication introduces problems in many types of studies, such as genome assembly, annotation and expression studies, where gene or transcript sequences that originate from different loci (such as transcripts derived from homeologs in each sub-genome) may share common sequences as a result of their shared evolutionary origin. These gene relationships are discussed in more detail in Section 2.3.

2.2 Genomics and Transcriptomics

DNA plays an information containing role and allows an organism to pass on its genetic information to the next generation by reproduction. By the central dogma of molecular biology, DNA provides the sequence for genes (as well as mechanisms for their expression) that are transcribed to RNA transcripts. These RNA transcripts encode the amino-acid sequence of a protein, the functional molecule of cells. Proteins have a wide range of functional roles: as catalysts for chemical reactions, structural roles (for example as microtubules in muscle cells), regulatory roles controlling the expression of other genes under certain circumstances and many more. The intermediary molecule between DNA and protein is RNA. The central dogma of molecular biology outlines the transfer of information from DNA to RNA by transcription and RNA to protein by translation. This process is present in all living organisms and the complex interactions between these three classes of molecules are the focus of a vast majority of biological research today.

Initially with genomics, the various -omics fields have become widely popular in the world of biological research. The most well-known example is the human genome project, an international research project to sequence the entire 3.3 billion base pairs of human DNA. This project took roughly 13 years to complete and cost $2.7 billion [6]. As technology has advanced, the cost and time requirement of such projects has dramatically decreased. Complete genomic sequences have been developed for many other organisms such as the fruit fly *Drosophila melanogaster* [15], yeast [16] and the model plant *Arabidopsis thaliana* [17]. Using the complete sequence of an organisms genome, it is possible for a researcher to study the genes and pathways that give an organism its properties.

Following the success of the major genome assemblies mentioned previously and the ensuing genomic analysis, other -omics fields have gained traction. An -omics approach is the study of the complete complement of a type of molecule present in an organism. If genomics is the study of the genome of an organism, transcriptomics is therefore the study of the RNA transcripts, and proteomics the study of the proteins expressed. These types of experiments have been facilitated by recent developments lowering the cost of entry to high-throughput sequencing methods for nucleic acids. These advancements are further discussed in

Section 2.4. Such technology allows for the generation of large datasets for a specific type of molecule isolated from biological samples. For transcriptomics; RNA molecules can be sampled from cell tissue, isolated using centrifugation and DNase treatment, transformed into cDNA via a reverse transcriptase and sequenced as DNA molecules. These tasks can all be performed with relative ease in a modern molecular-biology lab.

By the process of transcription, small regions of DNA within chromosomes called genes are used as a template to produce RNA transcripts. A transcriptome is the complete set of all transcripts, and their relative quantities, expressed from the genes of an organism. There are several forms of RNA including messenger RNA (mRNA), ribosomal RNA (rRNA), amino acid binding tRNA and other forms of non-coding RNA involved with gene regulation (miRNA, siRNA, etc.). Typically, transcriptome studies are most interested in messenger RNA, which are eventually used to produce proteins which carry out cellular activities. The goal of transcriptomics is to discover the function of these molecules and how they interact with other macromolecules to create a specific trait, or phenotype, in an individual organism.

In eukaryotes, genes have many of the same features between different organisms. A typical protein-coding gene contains promoter regions involved with expression regulation, a transcription start site (TSS), untranslated regions (UTRs) at the 5‘ and 3‘ ends, and several exons and introns between the UTRs. Codons are groups of 3 nucleotides in the DNA or RNA sequence that code for one of 20 amino acids or signal the start or end of translation. This is the method by which proteins are encoded by nucleic acids. During transcription, the nucleotides of DNA from the start of the 5‘UTR up until the end of the 3‘UTR are used as a template to produce a pre-mRNA molecule. A 5‘ cap (modified guanine nucleotide) is added to the “front”, or 5’ end of the transcript shortly after transcription begins. The pre-mRNA contains both intronic and exonic sequence. This molecule is processed by splicing machinery to remove the introns to create the final messenger RNA (mRNA) molecule that is ready for translation into protein. Alternative splicing may also occur, allowing for multiple combinations of the exons to be included in the final transcript. By this process, a single gene can code for multiple transcripts (isoforms) and by extension, different proteins [18].

Model organisms are used extensively in research for several reasons: they often have short life-cycles; undemanding living-requirements; and a well annotated, relatively simple genomes that are potentially closely related to other organisms of interest. In model organisms such as *Arabidopsis thaliana*, the complete genome along with extensive annotation is available for researchers to use as a groundwork for their analyses. This makes identifying the source and function of a transcript possible with a single sequence alignment search using a method such as the popular Basic Local Alignment Search Tool (BLAST) [19]. This paradigm of research is starting to decline however, as recent studies have shown that single genomes are not totally representative of a species. A genome analysis of multiple strains of *Streptococcus agalactiae* done by Herv Tettelin et al. [8] revealed that not only did the genomes differ significantly in sequence, but new genes specific to each strain were identified. The study also indicated that even after sequencing hundreds of genomes from strains of the same species, novel genes would still be discovered [8]. These findings suggest that reference-based techniques relying on a genome of a single individual are not as representative as once

believed.

2.3 Gene Sequence Homology

A pair of two genes, or sometimes more, that are similar in sequence are often concluded to be related by evolution from a common gene ancestor. Homologous genes are sets of genes that evolved independently from the same gene ancestor. Gene paralogs are a more specific type of homolog that are created from a gene duplication event within an organism [20]. These genes often evolve new functions within the organism, related or not to the original function. Orthologs are homologs that are the result of a speciation event, whereby related genes in two organisms share a common ancestral gene. For example, when a species diverges into descendant species who are reproductively independent, the two gene sequences accrue random variation independently. By this process, genes that once originated from the same sequence diverge into similar, but different genes that often retain the same function. The term homology can be applied to sequences of genes (DNA), proteins, genomic regions or even entire chromosomes. The term has also been used in the context of phenotypic traits, for example in comparative anatomy of the bone structure of winged mammals compared to a human hand containing the same number of joints and bone segments is an example of paralogous phenotypes. The two appendages evolved from a common ancestor, but have diverged to perform different functions.

Due to the evolutionary history of *B. napus*, many homeologous genes exist within the plant's own genome. *B. napus* has been shown to be a product of a hybridization event between the two closely related species *Brassica rapa* and *Brassica oleracea*. These two progenitor genomes, because of their close relatedness, share many orthologous gene pairs that have become homeologous pairs when the genomes were combined to create the *B. napus* genome. The term homeolog, refers to a special case of gene orthologs where the two homologous genes are present within a single genome. The genome of *B. napus* can be thought of as containing two subgenomes, A and C, with the majority of genes within one subgenome being similar in sequence to another gene within the other subgenome. An example of homeology in *B. napus* wuold be gene within the A subgenome related to a gene within the C subgenome whereby the two genes originate from *B. oleracea* and *B. rapa* respectively. The two ancestral genes from the progenitors are orthologs, and developed into homeologs within *B. napus* when the two genomes hybridized. These orthologous genes resulting from gene duplication is an important force for evolution. After a duplication event, the two genes undergo functionalization in one of a number of ways. By conservation, the two genes maintain the function of the original ancestor. Neofunctionalization results in one gene mutating to perform a new function distinct from the ancestor while the other maintains the original function of the ancestor. Subfunctionalization causes both gene copies to develop new functions in the organism while maintain the original function of the ancestral gene in tandem. Finally, specialization results in both genes evolving new functions distinct from one another and the ancestral gene [21]. The *Brassica* family of plants is a well-studied model for these processes due

to the genome triplication observed since the family's divergence from the model plant *Arabidopsis thaliana* [22].

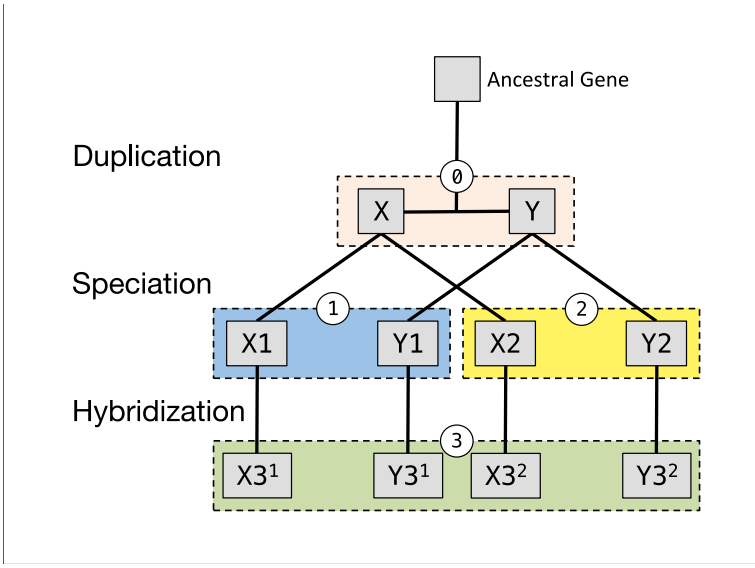


Figure 2.1: Diagram of the homologous relationships between a gene and its descendants after duplication, speciation and hybridization events throughout a hypothetical organism's evolutionary history. Dotted line boxes and numbered circles represent distinct organisms or species resulting from the three evolutionary events (duplication, speciation and hybridization). Square gray boxes represent genes sharing a common ancestral gene and are all homologs. Adapted from Walter M. Fitch, 2000 [23].

Figure 2.1 provides an example hypothetical organism's evolutionary history very similar to the one observed for *B. napus*. Due to the common ambiguous usage of terminology for different relationships of homology, this diagram will be used to explain the usage of terms in this thesis. Firstly, homology is the broadest term for any two genes or traits that share a common ancestor. All pairs of genes within the diagram are homologs, as they all share the very top-most gene as an ancestor. The term ortholog is used to identify homologous genes originating from a single gene resulting from a speciation event. In the diagram, genes $X1$ and $X2$ as well as $Y1$ and $Y2$ are examples of pairs of orthologous genes. Paralogs are homologous genes resulting from a duplication event, examples of which would be X and Y, $X1$ and $Y1$, and $X2$ and $Y2$. Another important subset of homologous genes are homeologs (alternative spelling: homoeolog) which are unique to allopolyploid organisms. Homeologs are the result of a polyploidization, or hybridization event that joins the genetic material of two related organisms. Genes $X3^a$, $Y3^a$, $X3^b$ and $Y3^b$ would all be considered homeologous as they result from the combination of homologous genes from two related progenitors into a single genome.

2.4 Nucleic Acid Sequencing

2.4.1 Sanger Sequencing

DNA sequencing has been an ever evolving technique since the 1970s. The first breakthrough in the field was with chain-terminating inhibitor sequencing. This method was developed by Frederick Sanger [24], from whom it gets its more common name of Sanger sequencing. By this method, four iterations of the DNA synthesis reaction are carried out. Each reaction includes the DNA template to be sequenced, DNA primers to initiate synthesis, a DNA polymerase, the four normal deoxynucleoside triphosphates (dNTPs) and one altered nucleotide; di-deoxynucleosidetriphosphate (ddNTPs) for each reaction which terminates the synthesis reaction upon incorporation to the growing DNA strand. In each reaction, many DNA templates are copied and the incorporation of the terminating ddNTPs occur at random points in the reaction, causing multiple DNA copies of various length, all of which end with the incorporation of the base pair corresponding the the ddNTP present. Determining the lengths of the resulting fragments yields the base-pair position of the corresponding nucleotide in the sequence. The fragments were originally separated by size using gel-electrophoresis in four lanes (one per reaction) whereby the sequence can be read by observing the bands from the bottom to the top. Advancements were made to introduce capillaries with the capability to sequence up to 384 molecules at a time. This technique allows for sequencing DNA fragments up to 1000 bps but is quite labour intensive and time consuming compared to more modern techniques. The first completed human genome in 2001 used the whole-genome shotgun approach along with Sanger sequencing [25, 26]. This involved breaking the genome into random fragments of DNA and sequencing the smaller fragments. These sequences, called reads, are then overlapped and pieced together into the full genome using assembly software. More recent genome assembly projects make use of the longer reads generated by Sanger sequencing in tandem with next-generation short-read methods. This allows for longer repeat regions to be correctly assembled and contigs and/or scaffolds to be joined into full chromosomes.

Building upon the foundations of DNA sequencing, RNA sequencing techniques provide novel methods for studying organisms, particularly gene expression studies. RNA sequencing however, does introduce a few unique complications to overcome. One difficulty is the occurrence of RNA-ase enzymes produced by humans as a defense against foreign RNA. These enzymes degrade RNA quickly and make sample preparation more difficult. The presence of ribosomal RNA and tRNA, which are not of particular interest to most transcriptome analysis, makes up the vast majority of RNA. These molecules must be removed during sample preparation, so as not to drown out the observation of mRNAs. As there are no current techniques for direct sequencing of RNA, it must first be converted back into DNA using a reverse-transcriptase enzyme reaction before being sequenced. This adds an extra step not present for normal DNA sequencing experiments.

2.4.2 Next-Generation Sequencing

The next major improvements were made in the late 1990s. New techniques, termed "next-generation" sequencing methods, automated the process of sequencing and provided improvements in speed and cost. 454 pyrosequencing, Illumina sequencing and Ion-torrent semiconductor methods were invented. These techniques all use a massively parallelized approach where a large number of DNA fragments are sequenced simultaneously. They vary in how the DNA template is prepared and stored and how the addition of nucleotides to the the growing template are detected and identified [27]. While there were a few competitors at the onset of the next-generation sequencing boom, as the technologies matured, several promising methods died out while others flourished. For example, the pyrosequencing method, while it was the first technology to provide relatively long insert lengths for paired reads, the technology ultimately could not compete with the cost-effectiveness and throughput of the Illumina sequencing method. When Illumina also began introducing methods for sequencing longer inserts, it became the obvious choice. More recently, the very long sequences produced using the single-molecule real-time sequencing developed by PacBio has been somewhat overshadowed by the very new, Oxford Nanopore sequencing technology, which provides relatively the same length of sequences in a more cost-effective, and simple way.

2.4.3 Solexa, Illumina

Developed by two Cambridge scientists, Shankar Balasubramanian and David Klenerman, the sequencing by synthesis (SBS) technology [28] has become one of the most widely used sequencing techniques. The two researchers went on to form the Solexa company in 1998 which would later be acquired by Illumina to further improve the sequencing technique. The high-throughput generation of large quantities of DNA sequence data at relatively low cost has made SBS a popular choice for genomics research.

Current iterations of Illumina sequencers operate on fragments of DNA between 200 and 8000 bps generated by a library preparation step. Adapter sequences are added to the ends of each DNA fragment, one of which contains a cleavage site. These adapters are complimentary to adapters affixed to the flowcell of an Illumina sequencer to immobilize the DNA templates. Once the templates are bound to the flowcell, the first copy is synthesized using the adapters as primers. The original template is washed away while the first copy is covalently bound to the surface of the flowcell. Amplification of these fragments into clusters occurs by the iteration of three buffer washes that catalyze the denaturation, annealing and extension of the DNA. The process is repeated to generate many copies of each DNA fragment localized to clusters where the original library molecule was present. Sequencing is done on the clusters one base at a time by repeating cycles. Fluorescently labeled dNTPs (each base with a different colour) are added and incorporated into the growing DNA strand. The labeled bases also act as reversible-terminating inhibitors, similar in principle to the Sanger method to cease DNA replication upon their addition. Once the labelled dNTPs are added and detected, the reversible-terminator is deactivated and the next cycle of extension is carried out [29]. Parallelism is achieved

by the presence of many clusters undergoing this process simultaneously on one flowcell. Images capturing the dNTP fluorescance of the entire flowcell keep track of the coordinates for each observation, corresponding to a cluster, or DNA fragment being sequenced.

There is a chance that an incorporation event does not occur correctly for one or more of the DNA fragments. Either extra extensions (pre-phasing) or missing an extension (phasing) will occur in a portion of the extending DNA molecules. As the replication continues producing longer and longer sequences, these errors accumulate causing noise and leading to a drop in base-calling quality in the 3' end of the sequences. These are typically trimmed using read quality control software, such as Trimmomatic [30] and FastX-toolkit [31].

Reads produced by an Illumina sequencer are now typically 125 or 250 bp paired-end sequences. Paired-end reads correspond to two sequence reads derived from sequencing both ends of a fragment of DNA. This fragment can vary in size from 200 up to 700 bps in paired-end experiments. Mate-pair experiments generate paired reads from longer fragments, approximately 2-8 kb, by first Biotynilating the ends of the long fragment, joining the ends to form circular DNA, fragmenting into shorter fragments and selecting the Biotynilated DNA for sequencing [32]. HiSeq machines are capable of generating 3 billion reads per run in high output mode, yielding up to 600 gigabases (1,000,000,000 base pairs) of sequence data in roughly 12 days [33]. The error rate of all Illumina platforms are roughly 0.1% [34], measured as the percentage of erroneous base calls within single reads of the maximum length.

2.4.4 Ion Torrent

The Ion Torrent sequencing technique is another SBS method, however detection of nucleotides makes use of the liberation of a H+ ion after the incorporation of dNTPs during DNA synthesis. Since there is no difference in pH change between the incorporation of the different nucleotides, each nucleotide is supplied to the reaction in alternating washes, with a boolean signal for whether or not extension occurred in the growing DNA sequence. The change in pH is relative to the number of bases incorporated, so if there are multiple repeating bases in the template, the incorporation of multiple bases in a single wash can be detected by a change in pH proportional to the number of repeated nucleotides.

A typical run takes 4 hours and produces 4 million single-end reads of 200-400 bps with reported error rates of 0.46% up to 2.4% [34].

2.4.5 Single Molecule, Real-Time Sequencing (SMRT)

Pacific Biosciences developed a method for observing the addition of nucleotides in real time as a DNA template is being replicated by a DNA polymerase [35]. This technique does not belong to the NGS wave of sequencing technologies, but has gained interest much more recently. Their method uses a zero-mode waveguide (ZMW) to observe the incorporation of each fluorescently labeled nucleotide. SMRT sequencers utilize a single DNA polymerase immobilized to the bottom of a ZMW bound to the DNA template. As

each phospholinked nucleotide is incorporated, the ZMW is excitated by laser light from below, emitting wavelengths of light corresponding to the four possible nucleotides. The inclusion of numerous ZMWs and therefore multiple simultaneous reactions allows for high-throughput sequencing of many molecules in parallel. The ability to sequence molecules as they are extended in real-time is a vast improvement over other techniques, which rely on either termination of extension, followed by reactivating extension, or alternating washes of single nucleotide solutions [29].

The SMRT Chip technology allows for very long read lengths, reporting up to 50% of the reads being longer than 10,000 bps. These long reads are especially exciting for RNA sequencing, as a single read can account for an entire RNA molecule, forgoing the requirement of assembly. A single run can generate 0.8 million reads or 5 Gb of read data. Reported error rates are less than 1% [34].

2.4.6 Maturity of Sequencing Technologies

While there are a variety of options for sequencing methods available, several stand out for use in genome and transcriptome assembly experiments. The early days of NGS included many competing technologies with equal promise. Now however, a few technologies have stood out and become dominant methods in the field. The most important factors for an assembly experiment are error-rates, and cost per base pair. Assembling a large quantity of data requires many reads, in order to ensure that each piece of the puzzle is sufficiently represented to appear in the final assembly. The more cost-effective sequencers provide the means to generate more bases for less money, and when research projects are funded by grants, this becomes an important consideration for the scientist. The error-rate of the technology is also of concern, as single base pair errors can look a lot like single nucleotide polymorphisms when the coverage is low. When assembling reads together, errors can mean the difference between linking two reads or not, or even joining two scaffolds. For these two reasons, the Illumina platform has come out ahead of the other competing methods for high-throughput short-read sequencing (Ion torrent and 454 pyrosequencing). The next wave of sequencing technologies, with a shift towards parallel sequencing of relatively much longer reads includes PacBio's SMRT technology, and the Oxford Nanopore sequencers. The Oxford Nanopore aims to provide the same long-read, real-time sequencing as PacBio's SMRT chips, but for cheaper and with fewer errors [36], although it is still in the early stages of public use. It is yet to be seen which of these two will become the standard.

2.5 Sequence Assembly

Early sequence assemblers were built upon the foundations of sequence alignment programs in order to assemble large quantities of reads to form a larger contiguous sequence (contig). Initially, these assemblers only dealt with small genomes of viruses and bacteria due to limitations of the sequencing technology of the time. As sequencing technology advanced, assemblers were required to process larger quantities of data. To be able to sequence eukaryotes and eventually the human genome, software tools became more sophisticated

to handle larger quantities of data containing repetitive sequences and sequencing errors. The genome of *Drosophila melanogaster* was an early example of a relatively complicated genome project. The genome was completed using the Celera assembler [15]. A year later, the Celera assembler was used to create the first draft of the human genome.

While there are many assembly tools, they all fall into one of two main categories: reference, or *de novo*. Reference algorithms require a template sequence to map new reads to. A reference genome allows for faster and less memory intensive aligning of the transcript reads to their position in the genome. In *de novo* methods, each transcript read has to be compared to all other reads in order to determine which reads come from the same transcript.

While full genomic sequences for new organisms are now being published constantly, the demand for research in non-reference species is still common. Whether an organism is newly discovered and not well understood, or the funding is not sufficient to justify a full genome assembly project, researchers are now undertaking many projects without the existence of a reference genome dataset. Transcriptome assembly is one such area of research, whereby the sequences of expressed RNA transcripts can be generated using RNA-seq data without comparison to the genome. These methods rely on determining the overlaps between short sequences, either reads or subsequences derived from reads, to determine the underlying biological sequences. There are three approaches that have been used over the years to accomplish this task. One benefit of *de novo* assembly is the potential to discover new genes. Even in species with a reference-genome available, the published genome likely does not exhaustively represent the full set of genes for the species. *De novo* experiments allows the researcher to discover novel sequences outside the reference-genome.

2.5.1 Greedy Assembly

Early assemblers such as phrap which was used as part of the human genome project, and TIGR [37] used in the assembly of *Haemophilus influenzae* [38] and *Mycoplasma genitalium* [39] used a greedy algorithm for assembling sequence reads. These approaches do not exhaustively compare reads, but iteratively choose the best overlapping read for the current extending contig from a subset of the total readset. This process fails to utilize global information carried by paired-reads in order to increase performance to deal with large quantities of data in a reasonable time frame. This is particularly problematic for repetitive regions. As of writing this thesis, there are no popular contemporary transcriptome assemblers that use this approach for assembly.

2.5.2 Overlap-Layout-Consensus (OLC)

These algorithms exhaustively search the read space and find overlaps between reads. During the overlap step, each read is compared pairwise to every other read. An overlap graph is then built, representing reads with nodes, and connecting the nodes with edges where sufficient overlap has been found to be considered biologically relevant. In the layout step, stretches of the overlap graph are condensed into contigs, whereby

some edges that are extraneous and can be represented by combinations of other edges are removed to create linear stretches of graph. Finally, the consensus step considers the reads that represent a contig, and report the nucleotides in the sequence with the best representation in the read data.

Due to the nature of pairwise comparison of each read, the performance of the overlap-layout-consensus (OLC) approach is heavily dependent on the quantity of read data used. This makes the time required for building the overlap graph scale exponentially with the number of reads. The overlap graphs also scale in size with the number of reads. The current trend of high-throughput short-read sequencing, such as with Illumina HiSeq machines, has made assembly by this method very time-consuming and memory intensive. The Celera Assembler [15] created by Eugene Myers et al. used a variant of this approach to achieve the whole genome assembly of the fruit fly *Drosophila melanogaster*. This assembly made use of 3.2 million reads obtained by sequencing the ends of Bacterial artificial chromosome (BAC) vectors to obtain mate-pair reads [15]. The OLC method of assembly has seen a decline in popularity as modern sequencing experiments using next-generation sequencing technologies can result in read sets containing anywhere from tens to hundreds of millions of sequences.

2.5.3 de Bruijn Graph (DBG)

The de Bruijn Graph (DBG) is a way to represent overlaps between strings of letters or symbols via a directed graph. Nodes of the graph represent each unique string of a specific length, k, using an alphabet of m symbols. This allows for a possible graph size of m^k nodes. Sequence overlaps are represented by the edges connecting two nodes which are present only when the rightmost $k-1$ symbols of a node, N_1, are identical to the leftmost $k-1$ symbols of another, N_2, resulting in the edge: $N_1 \rightarrow N_2$. A visual comparison of OLC and DBG assembly algorithms is shown in Figure 2.2.

The de Bruijn graph can be applied to transcriptome assembly by breaking up the RNA-seq reads into their constituent sequences of length k. These subsequences are referred to as k-mers. The alphabet for such a sequence graph would have four symbols corresponding to the four nucleotides that make up the sequence data, A, T, C and G. The existence of each edge can be computed efficiently by storing each individual k-mer in a hash using the sequence as a key. Since there are only four symbols, there are only four possible outgoing edges from each node. These can be searched for by simply determining if the corresponding four sequences are present in the hash. Once the existence of each edge is determined, the connected components of the graph are then traversed to build transcripts starting with the sequence of the first root node, and adding one base for each additional node visited. Assuming an ideal dataset where distinct genes do not share sequences of length k or longer, each connected component of the de Bruijn graph can be assumed as a separate transcript. In practice however, sequences of length k or more in common between genes cause separate genic regions to be represented in single connected components. Sequencing errors may also introduce k-mers that erroneously join reads derived from different loci. Additionally, there is no implicit method in the DBG algorithm to follow the k-mers deriving from a single read until it's conclusion, which

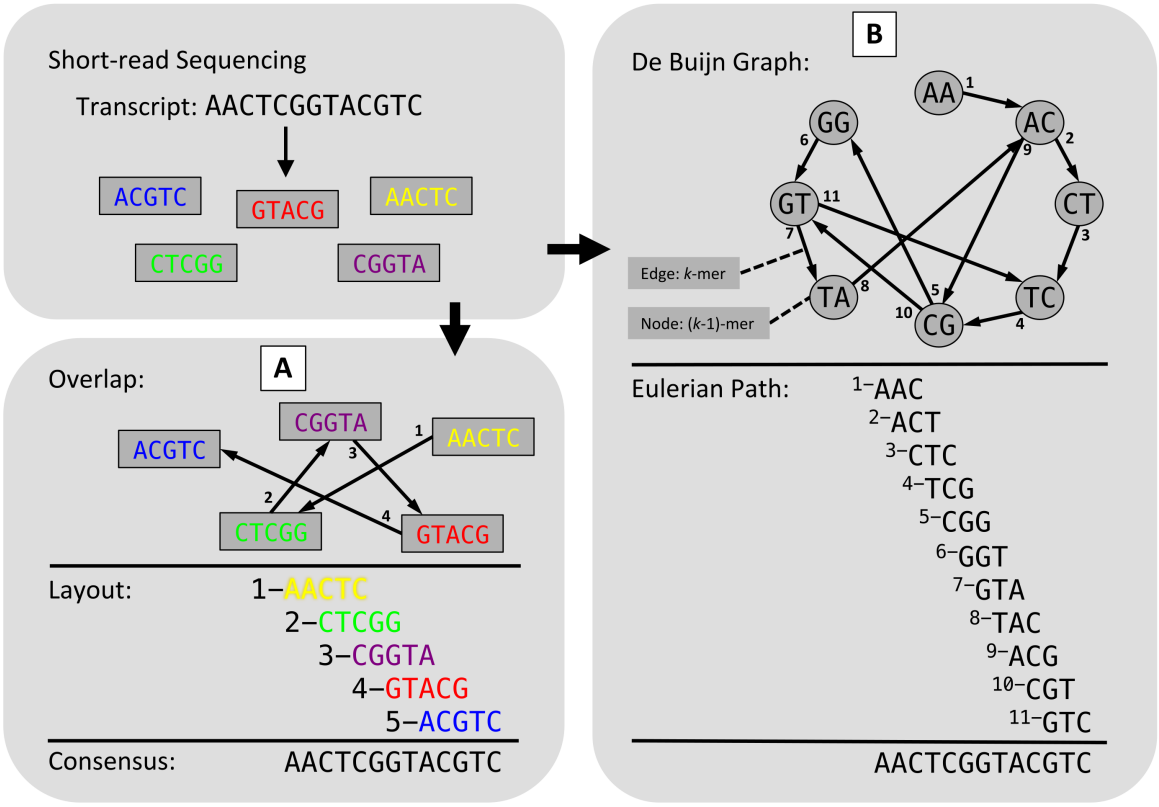


Figure 2.2: Comparison of the two techniques for *de novo* assembly of transcriptomes from short-read sequences. (A) Overlap-layout-consensus techniques directly aligning reads, while (B) de Bruijn graph assemblers divide reads into k-mer subsequences, build a graph using k-mers as edges between the $k-1$mer prefix and suffix. The transcript is determined by computing the Eulerian path (path traversing all edges) of the graph, adding one base to the growing transcript with each edge. Note that multiple Eulerian paths exist in the shown example, a problem that is less prevalent with longer k-mers. Adapted from Phillip E.C. Compeau et al., 2011 [40]

results in a loss of read coherence. Additional methods for identifying these errors are necessary to make DBG algorithms biologically meaningful, and are often what differentiates DBG-utilizing softwares from each other.

The properties of de Bruijn graphs have several implications for their application to genome and transcriptome assembly. Most importantly, DBGs eliminate redundant data resulting from repetitive sequences of DNA. The size of de Bruijn Graphs are bounded by the number of unique k-mers (m^k), which for biological sequences is 4^k.Though usually, the complete set of possible k-mers is not present in a read-set. The length of the k-mer subsequences used is always an odd integer, to prevent any k-mer sequence being the reverse compliment of itself. These subsequences can theoretically be as short as 3 bps, and as long as the length of reads used. Though these extremes are not useful, as very short k-mer lengths (< 21 bp) lose a great deal of read coherence and very long k-mers are prohibitively strict on the length of read overlaps. OLC algorithms on the other hand, require the comparison of each read to all others for a thorough search. Therefore, as the read-depth of a sequencing experiment increases, the performance benefits of DBG algorithms over OLC algorithms increase. This is especially important in the current age of sequencing where short-read, high-throughput sequencing techniques dominate the market (such as the Illumina platform).

2.6 Evaluation of Transcriptome Assemblies

2.6.1 Contig Metrics

The primary goal of sequence assembly is to combine the reads into contiguous sequences that are representative of the actual genomic content of the organism of interest. Since the true assembly is not known, the quality of each assembly is approximated by several different metrics. Most frequently the lengths of assembled sequences is used as a method for comparing assemblies. In the case of genome assembly, the algorithm attempts to generate entire chromosomes and the number of chromosomes may be known. Genome assemblies then attempt to assemble sequences that are as long as possible and try to assign them to a chromosome. Thus, genome assemblies are often aiming to assemble the longest contiguous sequences while keeping the number of sequences as low as possible. Most commonly, the contig length statistic of $N50$ is used to assess a genome assembly, with larger values being better. The $N50$ statistic is a weighted mean of the lengths of assembled sequences. Simply put, the $N50$ is the length of contig such that 50% of the assembled bases are in contigs longer than the $N50$ value, calculated by ordering the contigs by descending length and summing their lengths until 50% of the total assembly base pairs is reached. The length of the final contig summed when this threshold is reached is the $N50$ value. This process can be used for other cutoffs, for which the corresponding statistic is termed Nx where x is the cutoff percentage of assembled bp. Common lengths, such as $N25$ and $N75$ are used in conjunction with $N50$ to provide a distribution of the contig lengths.

There are several extensions of the Nx statistics that aim to better explain the quality of an assembly.

NAx statistics incorporate potential for mis-assembly of sequences into the standard Nx statistics by aligning the assembled contigs to a reference genome and counting the lengths of blocks of correctly assembled sequence. This process requires a high-quality reference genome to determine the portions of each contig that represent aligned blocks [41]. Another statistic is the NGx statistic, which aims to correct for bias towards assemblies that gain artificial benefits to Nx scores by filtering short contigs. This is done by summing contig lengths in descending length order until reaching a threshold of the organism's genome size, rather than the total assembly size [42]. It is possible to combine these two approaches, by summing the alignment blocks up to a threshold of the genome size, which is termed the $NGAx$ statistic, solving both the issue of filtering short contigs and mis-assembly of sequences simultaneously. Another variant of the Nx statistic that is specific to transcriptome assembly is the $ExN50$ statistic created by the creators of the Trinity [3] assembler. This measure is computed by summing the lengths of the most highly expressed transcripts up to a percentage of the total normalized expression data [43]. There are several tools available to compute these scores for an assembly, including the *de novo* transcriptome assembly evaluation software TransRate [44] that is further discussed in section 2.6.3.

Transcriptome assembly is complicated by the unknown distribution of sequence lengths in the true assembly. It is difficult to know how many expressed genes there are at the time of sampling, because not all genes will be expressed at one time for any tissue type. The lengths of transcript sequences also vary greatly due to mRNA processing which splices out introns, and the great range of transcript lengths for any given organism. Therefore, it is necessary to operate under several assumptions about the transcriptome: 1, depending on the tissue type used for sampling, only a portion of the genes will be expressed in the sample, 2, the lengths of RNA transcripts will be a range of values, 3, the presence or absence of non-coding RNA sequences depends on the type of library preparation and sequencing protocols used. For example, total RNA-seq experiments contain all type of RNA, including rRNA and tRNA, PolyA selection includes only coding RNA, and rRNA depletion techniques leave both coding and non-coding RNA in the sample.

2.6.2 Read Alignment and Sequence Comparisons

Another popular method for evaluating assembled sequences is by comparison to either known sequences, or to sequences from a closely related organism. This can be done by searching a database of sequences, such as with the popular tool BLAST (Basic Local Alignment Search Tool) [19]. BLAST allows a researcher to search against a public database containing sequences from many organisms or the creation of a local database from a collection of sequences. An alternative to BLAST is the BLAST-Like Alignment Tool (BLAT) [45], which is a very fast method for many-to-many sequence alignments.

2.6.3 Evaluation Software

Due to the popularity of whole transcriptome assembly in non-model organisms, new techniques for assessing assembly quality are being developed. To date, two software tools for reference-free transcriptome assembly

evaluation have been published, RSEM-EVAL [46] and TransRate [44]. Both tools use the read data as input to provide an assembly score. They differ in how the scores are calculated, which affects how the resulting scores can be used.

DETONATE RSEM-EVAL:

The RSEM-EVAL approach developed by Li et al. [46], aims to provide a score that corresponds to the probability that the input set of reads is explained by an assembly. The RSEM-EVAL assembly score is the sum of three components: an assembly prior score, a likelihood estimate, and a Bayesian information criterion (BIC) penalty. In practice though, the largest factor in the RSEM-EVAL score is the likelihood estimate which has the largest effect on the final assembly score. The contribution of each contig to the full assembly score is also reported, which allows contigs to be filtered based on the criteria of RSEM-EVAL to remove problematic or unsupported contigs.

The first part of the RSEM-EVAL assembly score, the assembly prior, penalizes contigs with lengths that are not likely given the read coverage over the transcript it originated from. The prior score also favors shorter assemblies in general, that account for the reads in as few bases as possible.

The likelihood component uses the RSEM model [47] to generate a probability distribution over the dataset, D, of RNA-seq reads using an assembly, A, in place of the full-length transcripts T. A repeating process of generating reads from the assembly, using the contigs as full length transcripts, and then determining if the reads completely cover the contigs is done, resulting in a likelihood of the form:

$$P(D|A, \Lambda_{MLE}) = \frac{P_{RSEM}(D|T = A, \Theta^c_{MLE})}{P_{RSEM}(C = 1|T = A, \Theta^c_{MLE})} \tag{2.1}$$

where P_{RSEM} is the probability under the RSEM model; $C = 1$ is the event whereby each position of the assembly is covered by reads that overlap by at least w bases and Θ^c_{MLE} is the equivalent parameter values of the RSEM model given the maximum likelihood expected read coverage values, Λ_{MLE} [46].

Finally, the BIC penalty penalizes assemblies based on the total number of contigs, favouring assemblies with fewer total sequences. This is proportional to the logarithm of the size of the read data set N and the product of the number of free parameters; which are the expected coverages of each contig, M, plus one parameter for the expected number of reads from RSEM's noise model [46].

The assembly scores calculated by RSEM-EVAL as part of the DETONATE software package are heavily tied to the RNA-seq dataset and are therefore not comparable between assemblies using different RNA-seq data sets. There is also no possible way to determine what constitutes a "good" or "good enough" threshold for DETONATE scores. This makes assessing an assembly based on the RSEM-EVAL score alone very difficult. The scores provided by RSEM-EVAL are more suited to comparing multiple assembly softwares, or parameter sets when refining an assembly over multiple iterations to rank assemblies.

TransRate:

The TransRate software package is another means of assessing the quality of a *de novo* transcriptome assembly using only the RNA-seq reads as input. This is done by calculating and reporting two scores:

the contig score for each assembled transcript and the total assembly score. Each contig score is a product of four components: (1) $s(C_{nuc})$ the bases of the assembled transcript correspond to those in the true transcript, evaluated by examining alignments between the read data and the assembled transcripts; (2) $s(C_{cov})$ the number of bases in the assembled contig is the same as the true transcript, which is evaluated by measuring the proportion of bases with no coverage in the read-data; (3) $s(C_{ord})$ the order of bases in the assembled transcript corresponds to the order present in the true transcript, evaluated using read-pairing data to determine if the reads align to contigs while preserving their paired orientation; and (4) $s(C_{seg})$ the assembled transcript represents a single true transcript, evaluated by the probability that the coverage of reads over the transcript is univariate.

The assembly score (T) is calculated by the geometric mean of the mean contig scores and the proportion of successfully mapped read pairs of the form:

$$T = \sqrt{(\prod_{c=1}^{n} s(C))^{\frac{1}{2}} R_{valid}} \tag{2.2}$$

where $s(C)$ is the product of the four contig score components:

$$s(C) = s(C_{nuc})s(C_{cov})s(C_{ord})s(C_{seg}) \tag{2.3}$$

Calculation of the four components are done as follows. The $s(C_{nuc})$ score is determined by calculating the "edit distance", e, between each contig and the reads that align to it. The maximum edit distance, e', is determined by the threshold for read alignment. Finally, support values for each contig are calculated as $1 - \frac{e}{e'}$. Support values for each read that align to the contig are averaged to produce the $s(C_{nuc})$ score. The $s(C_{cov})$ score is calculated by the fraction of bases in the assembled contig with at least one overlapping mapped read. The $s(C_{ord})$ makes use of paired information in the read data. First, 1% of the reads are used to determine the fragment size mean and standard deviation. The full set of read pairs is then evaluated, classifying correct pairs if both reads align to the same assembled contig, the orientation of the reads is consistent with the library preparation and the relative position of the reads is consistent with the previously determined mean and standard deviation fragment lengths. $s(C_{ord})$ is calculated as the proportion of correct read pairs is the data-set. Finally, the $s(C_{seg})$ score is determined by examining the coverage of reads across each contig. Using a Bayesian segmentation algorithm, the probability that the coverage of each nucleotide derives from a single Dirichlet distribution is evaluated. This probability is used as the $s(C_{nuc})$ score.

The developers of TransRate have identified several major assembly errors and classified them into types. They purport that the four contig component scores can identify and penalize these common errors. For example, chimeric transcripts consisting of the assembly of reads originating from two separate transcripts into one assembled contig would potentially exhibit a noticeable difference in read coverage of the contig. These chimeric assembled transcripts would be penalized by the $s(C_{seg})$ contig score component.

Using the contig scores, TransRate also provides an optimized assembly score by determining the contig cut-off score in order to trim contigs such that the assembly score is maximized. The developers do warn

that this process has the potential to filter out accurately assembled contigs with low abundance that may be incompletely covered.

In order to contextualize the assembly scores provided by the TransRate software, the developers have evaluated 155 published transcriptomes, where the read-data was available. They have indicated that an assembly score of ≥ 0.22 (or ≥ 0.35 optimized assembly score) would represent a better assembly than 50% of the published assemblies. They also advise that quality of the read-data has the largest impact on assembly score, more so than the quantity of reads or assembly methods.

Chapter 3

Research Goal

The overall objective of this thesis is to evaluate the current state of transcriptome assembly software in the context of a complex allopolyploid genome. The highly duplicated genome of *B. napus* provides a number of complications for *de novo* assembly of RNA seq data into transcript sequences. The occurrence of multiple homeologous regions within the genome can result in reads from different genomic locations being overlapped during assembly, collapsing duplicated genes into chimers. A variety of evaluation metrics were used, based on assembly statistics like quantities and lengths of contigs, comparison to annotated *B. napus* genes, transcriptome evaluation software scores, and finally, the quantification of assembled transcripts including fragments of homeologous gene pairs originating from distinct gene loci.

The secondary objective was to provide a means to better observe the underlying mechanisms and procedures behind de Bruijn graph assembly. To accomplish this, scripts were created to parse Trinity assemblies, reconstruct and simplify the de Bruijn graphs and format them for display in an explorative, interactive web-tool.

Three prominent *de novo* assemblers were identified from the literature: Trinity [3], Oases [4] and SOAPdenovo-Trans [5]. These assemblers were used to generate multiple transcriptome assemblies using varying parameters. Specifically, the k-mer length parameter was identified as an important factor in assembling short read data for de Bruijn graph based algorithms. For shorter values of k, the probability of erroneous overlaps increases as there are fewer distinct k-mer sequences. Increasing the value of k should, in theory, decrease the number of erroneous overlapping reads as the required overlap length is longer. In order to test this theory, we generated assemblies over the range of possible k-mer lengths for each assembler.

3.1 Evaluation of Transcriptome Assembly Software

Several metrics were employed to get the whole picture of transcriptome assembly quality. Starting with the basic metrics of quantity and lengths of transcripts produced by each assembler, a basic comparison of the distribution of transcript lengths was produced. While these metrics are much more important for genome assemblies, they are still an important indicator for transcriptomes, to determine if an assembler is producing a reasonable quantity of transcripts at expected lengths.

To evaluate the accuracy of assembled transcripts, we can compare the sequences to a set of annotated *B.*

napus coding DNA sequences (CDS) created from the reference genome. By aligning assembled transcripts to known sequences, it is possible to evaluate the number of genes expressed by quantifying the unique CDSs represented by each assembly. The length at which these sequences were represented was used to determine which assemblers or parameter sets result in the most fully formed transcript sequences. Furthermore, the unique sets of CDSs represented in each assembly allowed the comparison of the exhaustiveness of an assembly.

Existing methods for transcriptome evaluation exists, whereby the reads are aligned to the assemblies and used to generate numeric ratings. Two such softwares, DETONATE RSEM-EVAL and TransRate, were identified and used to rate assemblies. These techniques use mathematical models to rate an assembly, often on several criteria.

Finally, the presence of homeologous gene pairs within the hybrid *B. napus* genome is a cause of substantial ambiguity during assembly. Sequences that originate from such genes may be erroneously overlapped or mis-assembled during assembly due to the common similarity between them. To quantify this problem, homeologous gene pairs were identified by sequence comparison of *B. napus* morphotype 'DH12075' genes to the two progenitors, and by using an ortholog table generated for another morphotype, 'Darmor-bzh'. Determining the pairs within 'DH12075' that correspond to orthologs in *B. oleracea* and *B. rapa* then allow the comparison of assembled transcripts to CDS sequences, determining if the assemblers correctly assign gene pairs to separate gene loci or not.

3.2 Visualization of de Bruijn Graphs

In order to better examine assembled transcripts, especially ones that resembled genes of interest in *B. napus*, a graph visualization tool was created to display Trinity results. In particular, we wanted to be able to compare the transcript isoforms that originate from a single connected component, which Trinity calls a "gene". To make the de Bruijn graphs created by the assemblers meaningful and understandable to humans, long stretches of unbranching graph were simplified into nodes of sequence longer than k. This was necessary to compress the sometimes very large k-mer graphs into a more comprehensible visual representation. The tool also aimed to provide useful interactivity functions, as well as sequence display and capture for assembled transcripts or custom selections of nodes.

Chapter 4

Data and Methodology

This section describes the techniques used to carry out the experiments and analysis, as well as the implementation of the tools built as part of the thesis work. Firstly, the read data used for assembly, as well as the reference genome data used, are described. Software versions and parameters for the three *de novo* transcriptome assemblers are outlined. Evaluation techniques involving the calculation of metrics and statistical analysis are described, as well as the protocol for analyzing assemblies using existing software tools. The implementation of the de Bruijn graph visualization tool, GVART, developed as part of this research project, is described in Chapter 6.

4.1 Data

4.1.1 DH12075 RNA-Seq Data

The sequences used for assembly were 100 bp, paired-end, RNA-seq reads produced by an Illumina HiSeq machine. RNA was extracted and purified from *B. napus* morphotype 'DH12075' leaf tissue. The reads were post-processed for quality trimming and adapter removal using the Trimmomatic software (version 0.32) [30]. Trimming steps were done in the following order: `ILLUMINACLIP:Trimmomatic-0.32/adapters/TruSeq2-PE.fa:2:40:15 LEADING:15 TRAILING:15 SLIDINGWINDOW:4:15 MINLEN:55`. Of the 10,355,910 total reads pairs, 9,423,704 (91.00%) were both surviving, 567,088 (5.48%) were forward only surviving, 196,420 (1.9%) were backward only surviving, and 168,698 (1.63%) were eliminated entirely.

Histograms of k-mer repetition were created for lengths used to identify any peaks which may represent repeating genomic sequence of length k. Jellyfish [48] `count` and `histo` commands were used to generate histogram data files which were plotted in R [49] using the ggplot2 [50] package.

4.1.2 DH12075 v3.1 Reference Data

The 'DH12075' v3.1 reference genome was created at the Saskatoon Research and Development Centre (SRDC) using a method similar to the recently published *B. oleracea* genome [22]. Illumina read data was checked and filtered using the Illumina-Pipeline to remove low-quality reads and leftover adapter sequences. The remaining reads were assembled using SOAPdenovo version 1.04 [51]. Additional reads from the 454

platform of 20 kbp-span paired-reads as well as BAC-end reads of 105 kbp-span were used to extend scaffolds by aligning with BLAST (identity $\geq 95\%$, alignment length ≥ 100 bp and $\geq 60\%$ coverage). The reference genome was used to generate the ref-based Cufflinks assembly as well as being the groundwork for annotation to produce the coding DNA sequence (CDS) dataset by *ab initio* gene prediction software MAKER [52].

4.2 Transcriptome Assembly

This section describes the protocols for generating *de novo* transcriptome assemblies using the three software tools Trinity v2.0.6 [3], Oases v0.2.8 [4] and SOAPdenovo-Trans v1.03 [5]. All assemblies were submitted to a parallel compute server using SunGrid Engine. A reference based assembly was also created as a control using the Cufflinks protocol.

4.2.1 Trinity

Trinity assemblies were generated for odd values of k-mer length beginning at 21 bp up to the maximum allowable k value of 31 bp. Assembly was done on a compute server using 8 threads supplying 32 GB to the jellyfish software used for k-mer counting. Trinity version 2.0.6 was used and all parameters were otherwise default.

4.2.2 Oases

Merged Oases assemblies were generated using the bundled oases˙pipeline.py script with version 0.2.8. Assemblies were computed for odd k-mer lengths from 21 bp up to 51 bp and 3 merged assemblies were generated using assemblies 21 to 31, 31 to 41 and 41 to 51 using the supplied Oases python script. Parameters used were `-shortpaired`, `-separate`, with an insert length of 350 bp.

4.2.3 SOAPdenovo-Trans

SOAPdenovo-Trans assemblies were created for similar k-mer lengths to Trinity and Oases due to the availability of a broader range of possible k-mer lengths. Assemblies were generated for k-mer lengths of 25, 27, 29, 31, 41, 51, 61 and 71 bps. Parameters specified in the configuration file were as follows: `max_rd_len=100`, `rd_len_cutoff=100`, `avg_ins=350`, `reverse_seq=0`, `asm_flags=3`, and `map_len=3`. Software version 1.03 was used.

4.2.4 Cufflinks Ref-based Control

A reference-based assembly was created using the Cufflinks pipeline [53]. Both paired and unpaired reads were aligned to the *B. napus* 'DH12075' reference genome using TopHat (version 2.1.0) and Bowtie2 [54]

version 2.2.4. Read alignments were sorted and indexed with SAMtools (version 0.1.19) and Cufflinks version 2.2.1 was used to generate a reference based assembly using default parameters.

4.3 Assembly Evaluation

This section outlines the computation of assembly metrics and the intermediate data-files used. All data was plotted in R [49] using the ggplot2 [50] package.

4.3.1 Contig Length Statistics

Transcript length statistics—mean, median, %GC, $N25$, $N50$ and $N75$—were calculated using a Perl script written by Joseph Fass from the Bioinformatics Core at UC Davis Genome Center [55]. All metrics were calculated using the output .FASTA files from the assemblers. In some cases, header lines needed to be converted to conform to the expectations of these scripts.

4.3.2 Coding DNA Sequence Representation

Assemblies were compared to a dataset of predicted Coding DNA Sequences (CDS) to determine the percentage of expressed sequences that were represented. Transcripts were aligned to the CDSs using BLAT [45] to find high quality alignments using the default BLAT parameters. Filtering alignments for unique CDS names yielded the number of sequences that were represented. For each assembly, histograms for the percentage of the CDS covered by an alignment were counted using a custom Perl script for 1% wide bins from 85% up to 100% length.

Venn diagrams were computed by a Perl script, reading lists of unique CDSs represented by several assemblies. Booleans representing the status of representation of each CDS were stored in a hash in 3-bits. The hashes were then parsed for the total counts of each quadrant of the Venn diagram.

4.3.3 Assembly Evaluation Software

Two transcriptome assembly evaluation softwares were identified, DETONATE [46] and TransRate [44].

The DETONATE RSEM-EVAL (version 1.8.1) software was used to evaluate the assembled transcripts of each assembly by aligning the trimmed, paired-end reads. RSEM-EVAL made use of the Bowtie2 sequence aligner [54] with a fragment length of 350 bp.

TransRate scores were calculated using version 1.0.3 for each assembly supplying the trimmed paired-end read files (`--left` and `--right` reads). The software was run using 3 threads with all other parameters default.

4.3.4 Generation of the DH12075 Ortholog Table

To determine the extent to which orthologs were collapsed into chimeric sequences, an ortholog table was generated for the *B. napus* morphotype 'DH12075' CDS dataset. The ortholog table for morphotype 'Darmor' was available, pairing known gene orthologs between the two subgenomes labelled A and C in the 'Darmor' genome assembly for the two progenitors, *Brassica rapa* and *Brassica oleracea* respectively. First, genes were reciprocally aligned between 'Darmor' and 'DH12075' CDS datasets using BLAT and a minIdentity (98%). This dataset was then intersected with an ortholog table for 'Darmor' containing pairs of known orthologs to get the equivalent 'DH12075' gene pairs. This resulted in 27,469 gene pairs analogous to the 'Darmor' ortholog table. Reciprocal matches not corresponding to a pair in the 'Darmor' ortholog table were also included if the two genes were not located on the same chromosome, increasing the number of orthologous gene pairs to 29,492 pairs.

4.3.5 Sequence-based Ortholog Differentiation

While each assembler defines the relatedness of transcripts slightly differently, each assembler has some characterization of which transcripts belong to a single locus. Trinity transcripts are given a TR designation, as well as 3 integer designations for read cluster (c), gene (g) and isoform (i). The combination of c and g designations are to be used to identify a "gene id." For our study, we considered isoforms to be representative of sequences from a single locus. Therefore, we required two transcripts that have a unique TR and gene id (c and g) designation to be considered separate genes. For Oases, transcripts are given a locus number and transcript number. We required transcript numbers be unique for separated transcripts in our analysis. SOAPdenovo-Trans assemblies use a unique name for each transcript produced, only specifying loci in extra fields which were not recorded by BLAT. For this reason, only ortholog pairs that both matched the same transcript sequence were marked as a collapsed pair, leading to an artificially improved result for this study. Despite the obvious advantage of being based on the reference-sequence, as the sequence of the homeologs is available for comparison, Cufflinks was included. The Cufflinks assembly was used as a reasonable ceiling for achievable differentiation of orthologous gene pairs during assembly.

To calculate the number of gene pairs that fall into either category for an assembly, Perl scripts were written to parse the table of orthologous pairs and the results of BLAT alignment of the transcripts to the coding DNA sequences. For each BLAT hit, if the hit is a better scoring alignment than the current best, it is saved as the representative transcript for that *B. napus* gene. Alignment score is determined by the following formula:

$$SCORE = (M + M_{rep}) - N - Q_{ins} - T_{ins} \tag{4.1}$$

where M is the number of matches between sequences not in repeats, M_{rep} is the number of matches within repeats, N is the number of mismatches between sequences, Q_{ins} is the number of insertions to the query sequence and T_{ins} is the number of insertions to the target sequence. Each of these scores is provided by the

default BLAT output. Once the BLAT hits have been recorded, the script iterated over gene pairs, comparing the representative transcripts ID strings to determine if they were assigned to different loci (separated) or not (collapsed) by the assembler.

Chapter 5

Results

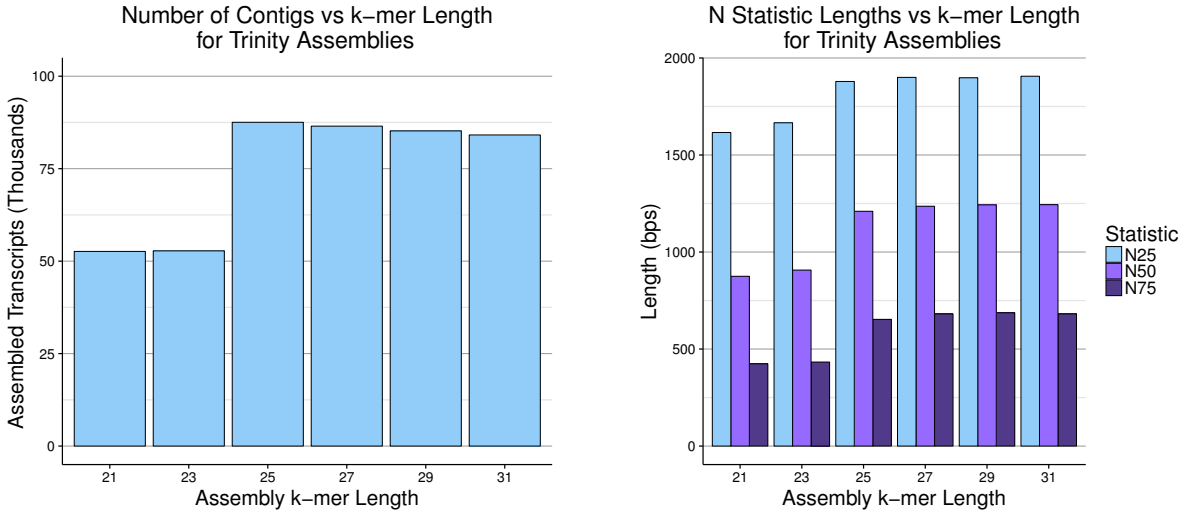
5.1 Evaluation of transcriptome assemblers

This section details the comparison and evaluation of de Bruijn graph (DBG), *de novo* transcriptome assemblies created by multiple software tools and varying k-mer lengths. Three *de novo* assemblers using the DBG approach were identified as the most prominently used in current research: Trinity [3], Oases [4] and SOAPdenovo-Trans [51]. A reference-based assembly was also generated using the Cufflinks protocol [53] using a draft *Brassica napus* (*B. napus*) genome. As well as using basic assembly statistics such as contig lengths and counts, transcripts were compared to a known dataset of coding DNA sequences (CDSs). Two relatively new softwares for transcriptome assembly evaluation were also used to rate assemblies.

5.1.1 Number of Transcripts Assembled and Length Statistics

In genome assembly studies, the number and length of contigs generated is an important indicator for assembly quality. Since the number of true biological sequences being assembled in genome (chromosome number) assembly is very low compared to transcriptome (unique RNA molecules), minimizing the number of contigs and maximizing length is an easy way to determine the quality of an assembly. Additionally, chromosome sizes are often known for genome studies, whereas the number of expressed transcripts and the distribution of their lengths is not known for *de novo* transcriptome studies. This limits the ability for contig length metrics to determine which assembly is better on their own. Extensions of the Nx statistics such as NGx and $NGAx$ statistics are potentially able to account for the size of the organisms genome as well as penalize misassemblies by breaking contigs where they misalign to the genome. These are not used in this study however, as they rely on the presence of a high-quality reference genome that would be unavailable to true *de novo* transcriptome assembly studies. The contigs of a transcriptome assembly can be thought of as the assembled transcripts, therefore in this study, the terms contig and assembled transcript are used interchangeably.

Assembled transcript length statistics were calculated for each assembly generated using a script written by Joseph Fass from the UC Davis Genome Centre[55]. These statistics include the total number of reported transcripts, mean transcript length, N25, N50 and N75 statistics. Table 5.1 displays these statistics for each assembly method. Contig lengths metrics are used instead to compare the ability of the multiple assemblers



(a) Number of Transcripts produced

(b) N statistics

Figure 5.1: Assembly statistics for the Trinity assembler. Number of assembled transcripts produced (a) and N25, N50, and N75 statistics (b) for each varied k-mer length assembly.

to generate an adequate number of transcripts considering the proportion of expressed gene in the sample tissue-type, at reasonable expected lengths. The latest research in *B. napus* suggests approximately 100,000 genes [11], roughly two thirds of which would be expected to be expressed in leaf tissue RNA-seq data. Accounting for alternative splicing would put the estimate for a reasonable number of expected transcripts between 50,000 and 100,000. The length of these transcripts are expected to be between a few hundred to the low-thousands base pairs with an even distribution of the three N statistics: N25, N50 and N75.

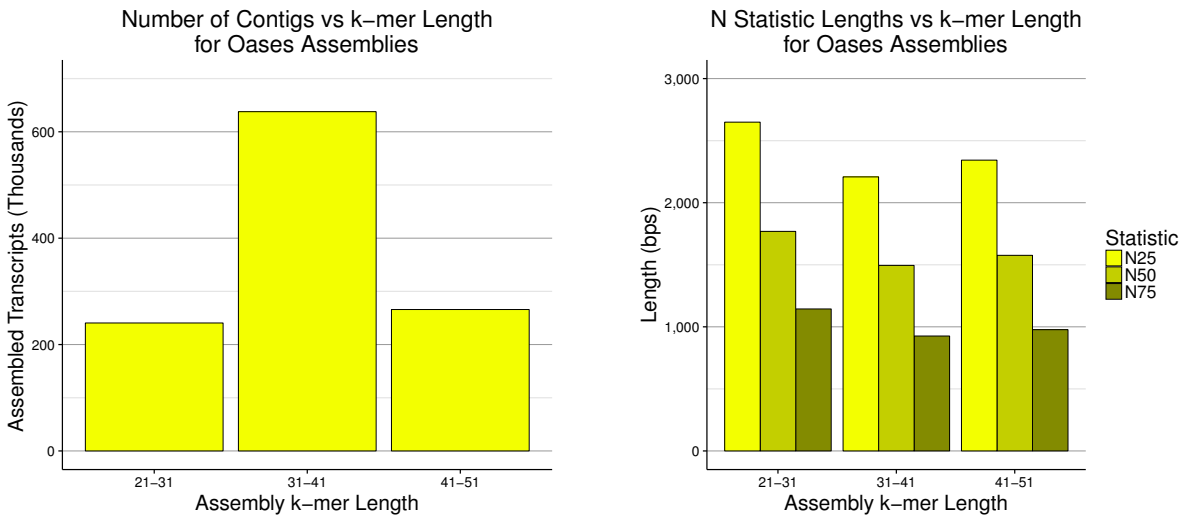
Trinity:

The developers of the Trinity assembler advise that the default k-mer length of 25 bp is optimal for assembly [3]. In it's current implementation, Trinity allows for k values up to 31 due to limitations with the data structures and memory usage of the software. In total, six assemblies were generated using values for k-mer length of 21, 23, 25, 27, 29 and 31. All other parameters for the assembler were default for these six assemblies. Some other parameters were tested which change the final "butterfly" step of Trinity to mimic other assemblers (`--CuffFly` for Cufflinks and `--PasaFly` for PASA) or use known sequences as a guide (`--genome_guided_bam`). The `--CuffFly` option aims to report minimum transcripts imitating the Cufflinks algorithm, while `--PasaFly` reports maximally supported isoforms. These three parameters did not improve assembly in any of the metrics used in this study, and displayed very negative results in the evaluation software. They are therefore only reported in appendices figures.

Figure 5.1 shows the number of assembled transcripts produced for each assembly, as well as the N25, N50 and N75 contig length statistics. Assemblies using a k value below the default 25 bp show a large drop-off in both the number and length of produced transcripts, while values between 25 and 31 are comparable with a slight decreasing trend in number of contigs produced as k-mer lengths increase. The expected

Table 5.1: Transcript length statistics for *de novo* assemblies varying the *k*-mer length parameter and one reference-based assembly.

Assembly Method	Total Transcripts	Mean Length (bp)	N25 (bp)	N50 (bp)	N75 (bp)
Trinity					
21 bp *k*-mer	52,629	646.6	1,616	875	424
23 bp *k*-mer	52,779	660.4	1,666	907	433
25 bp *k*-mer	87,545	830.5	1,879	1,210	653
27 bp *k*-mer	86,506	851.3	1,900	1,236	681
29 bp *k*-mer	85,219	857.5	1,898	1,243	687
31 bp *k*-mer	84,120	856.9	1,906	1,244	681
Oases (merged)					
21–31 bp *k*-mers	240,559	1,249.0	2,649	1,769	1,144
31–41 bp *k*-mers	637,797	1,058.5	2,208	1,495	926
41–51 bp *k*-mers	265,849	1,129.1	2,343	1,576	977
SOAPdenovo-Trans					
25 bp *k*-mer	119,121	403.8	1,545	848	333
27 bp *k*-mer	129,544	399.4	1,538	852	330
29 bp *k*-mer	138,937	396.9	1,535	857	327
31 bp *k*-mer	146,036	395.2	1,551	866	321
41 bp *k*-mer	195,182	228.5	479	258	155
51 bp *k*-mer	161,477	338.1	1,318	639	217
61 bp *k*-mer	169,051	301.9	1,046	492	187
71 bp *k*-mer	195,182	228.5	479	258	155
Cufflinks					
ref-based	65,055	1,269.4	2,290	1,569	1,031



(a) Number of Transcripts produced

(b) N statistics

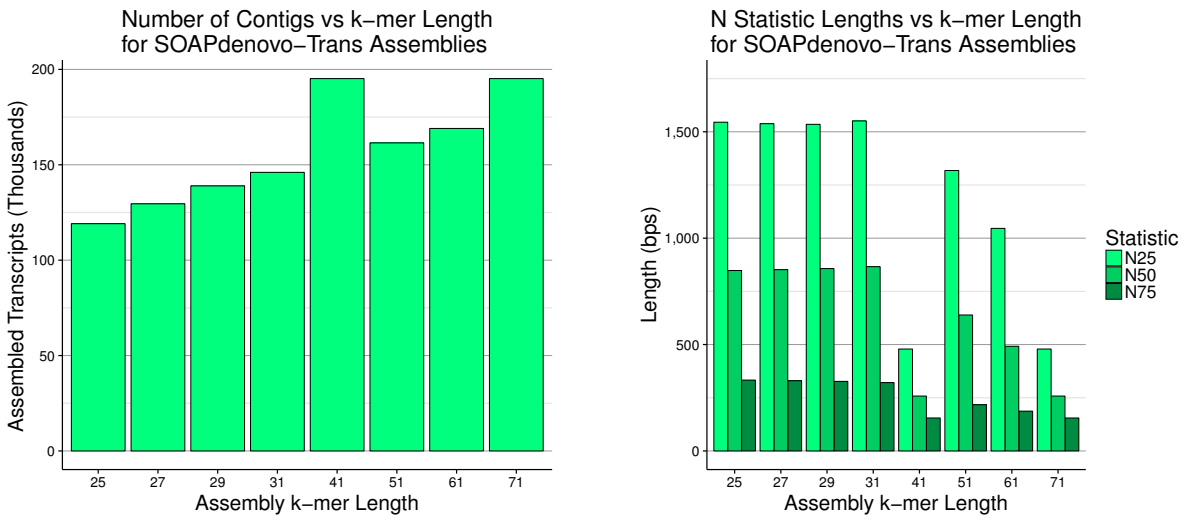
Figure 5.2: Assembly statistics for the Oases assembler. Number of assembled transcripts produced (a) and N25, N50, and N75 statistics (b) for the three multiple-k merged assemblies.

number of transcripts observable in the leaf-tissue sample material is greater than 60% of the 100,000 known genes considering alternative splicing genes creating multiple distinct transcripts from a single gene. Trinity assemblies using 25 to 31 bp k-mers produce a quantity of transcripts closer to the expected number while 21 bp and 23 bp assemblies produce fewer than expected transcripts. The contig length N statistics display a similar trend with the two shortest k-mer length assemblies under-performing compared to the other four, producing shorter transcripts in general. The 25, 27, 29 and 31 bp k-mer length assemblies performed similarly with regards to weighted median transcript lengths. While these metrics alone do not indicate an optimal k value, the recommended k-mer length of 25 bps is supported, and in fact, 25 – 31 seems reasonable for future studies in *B. napus*.

Oases:

Oases uses a different approach for transcript assembly that is based on the same core DBG algorithm, whereby multiple assemblies generated with a single k-mer are merged together for the final assembly. As such, assemblies generated in Oases are labelled with a lower and upper bound of the single-k assemblies used in the final merged assemblies, which include the only odd integers in between. Oases was used for three experimental ranges of k generated from sixteen total single k-mer length assemblies using ranges: 21–31, 31–41 and 41–51. Number of transcripts and their N25, N50 and N75 lengths are shown for the unmerged, single k-mer length assemblies in Appendix A.

Figure 5.2 summarizes the number of transcripts produced and the contig length N statistics for each of these three assemblies. There is a large disparity in the number of transcripts produced between the middle range 31–41 bp k-mer merged assembly, which yielded 637,797 contigs, and the two neighbouring ranges of 21–31 bp and 41–51 bp which produced 240,559 and 265,849 contigs respectively. The same assembly



(a) Number of Transcripts produced

(b) N statistics

Figure 5.3: Assembly statistics for the SOAPdenovo-Trans assembler. Number of assembled transcripts produced (a) and N25, N50, and N75 statistics (b) for each assembly varying k-mer length.

produced a drop in assembled transcript lengths as shown in the N statistics (Fig. 5.2b). This observation could indicate either a disproportionate number of repetitive expressed genomic elements of length 31–41 bp in the RNA-Seq data, or a problem with the assembler for the specific k-mer lengths. Further investigation into the interaction between the Oases assembler and k-mer lengths 31 to 41 are discussed in Section 5.1.3.

SOAPdenovo-Trans:

SOAPdenovo-Trans (SDT) provides a much broader range of k-mer values than the other two assemblers. The software allows for k-mer values up to 127 bp. The maximum values were not used however, as k-mer lengths approaching the read length of 100 bps lose the benefits of DBG assembly over the overlap-layout-consensus approach and k-mers greater than the read length are not possible. Assemblies were generated for k-mer lengths of 25, 27, 29 and 31 for comparison with Trinity, and values of 31, 41, 51, 61 and 71 for comparison to Oases' larger compliment of allowable values and to observe any larger trends related to increased k-mer length. Limitations of time and storage space prohibited generating transcriptomes for all possible k-mer lengths for SDT.

The number of transcripts assembled and contig N statistics for SOAPdenovo-Trans transcripts are shown in Figure 5.3. For most of the SDT assemblies, the quantity of transcripts increases as k value does. This is likely caused by a longer required overlap between reads to join them during assembly. For transcript regions that are only represented in the sequence data with two reads overlapping, if the overlap is shorter than the k-mer length, the two reads will not be joined during assembly. This would result in a trend of increased frequency of fragmented and incomplete transcripts as the k-mer length is increased. Larger quantities of unjoined transcript sequences is confirmed by the larger quantities of produced contigs and a decrease in the length of the contigs as k-mer increases, which is confirmed in Figure 5.3.

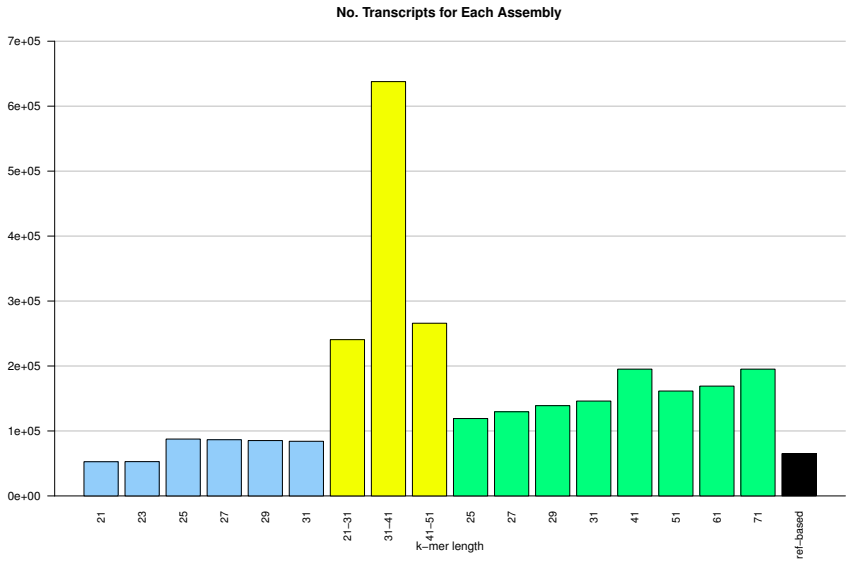


Figure 5.4: Number of Transcripts produced by each of the three *de novo* assemblers used; Trinity (blue), Oases (yellow) and SOAPdenovo-Trans (green) and one ref-based assembler Cufflinks (black).

A spike in number of transcripts was observed for the 41 bp *k*-mer assembly, which yielded 195,182 contigs compared to the neighbouring 31 bp and 51 bp assemblies which produced 146,036 and 161,477 contigs respectively. This observation mimics the Oases assemblies where the median merged 31–41 bp assembly produced a drastically increased number of contigs than the two neighbouring assemblies. Both the SDT 41 bp assembly and the Oases 31–41 bp assembly also produced shorter contigs when compared to the two neighbouring assemblies. In particular, the SDT assembly with k=41 bp produced drastically shorter contigs with an N50 of 258, compared to the two surrounding assemblies' 866 and 639 bp for k=31 and k=51 bp respectively. This lead us to believe that the spike was caused by an attribute of the genome sequence rather than a software specific problem. As with the Trinity assembler, *k*-mer length values for the SDT assembler between 25 and 31 bps produced assemblies with the longest transcripts. Longer *k*-mer assemblies produced shorter and shorter transcripts as *k* increased.

Due to the use of RNA-seq data derived from a single tissue type sample, the full complement of genes are not expected to be represented in these transcriptome assemblies, thus the expected number of transcripts present in the sampled RNA should be around 2/3 or greater than the total number of genes, accounting for alternative splicing isoforms. The most recent estimate for the number of genes in *B. napus* is 101,040 [11]. The number of transcripts produced for each assembly across the different assemblers is shown in Figure 5.4. The Trinity assembler produced around 80,000 transcripts for *k*-mer values 25, 27, 29 and 31, which was within the range of acceptable numbers of transcripts present. Both SOAPdenovo-Trans and Oases produced an excess number of transcripts, indicating a lower level of contig joining, resulting in many more incomplete transcript fragments. It could also indicate an increase in assembling chimeric transcripts, whereby sequence variations greater than the *k*-mer length are not correctly paired.

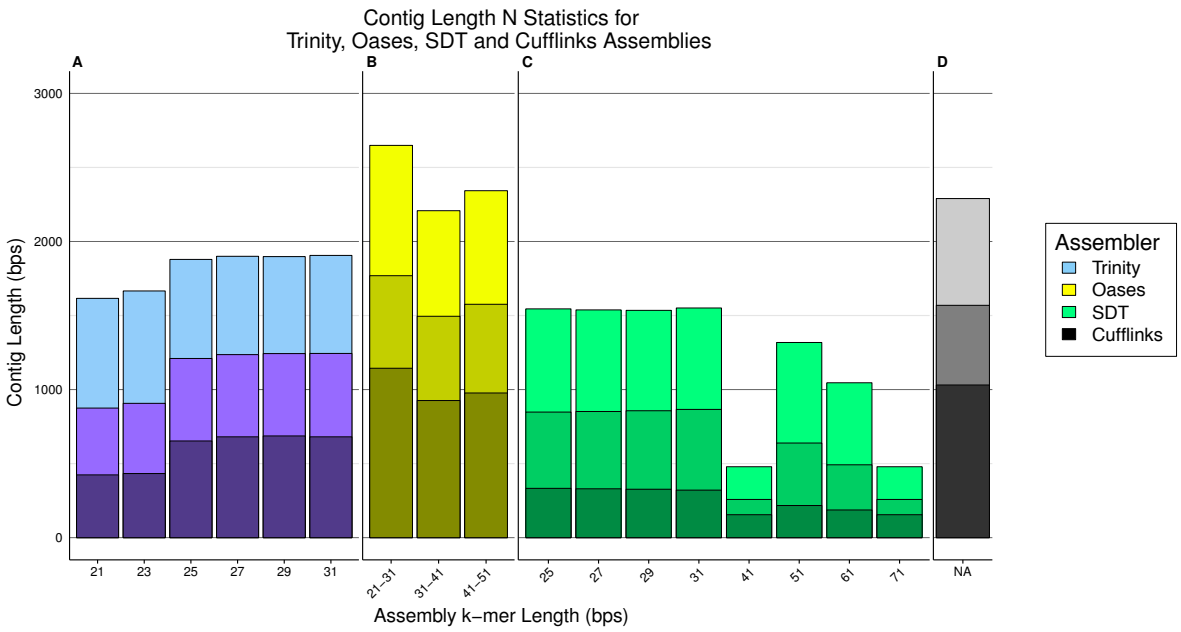


Figure 5.5: Assembly N statistics for all assemblies generated by the three *de novo* assemblers Trinity (blue), Oases (yellow), and SOAPdenovo-Trans (green) and the reference-based Cufflinks assembly pipeline (black and gray). Overlapping bars display the N25, N50 and N75 lengths for each assembly, with light, medium and dark shades respectively.

For all assemblers, N25 N50 and N75 statistics (shown in figure 5.5) were well distributed, suggesting a reasonable spread of transcript lengths over the assemblies. For Trinity and Oases assemblies, the range of transcripts was roughly 200 bps for the shortest, and 10,000 bps for the longest. These ranges, are consistent with the expected distributions for transcript lengths. SOAPdenovo-Trans assemblies included transcripts as low as 100 bps which likely contributed to their lower N statistics compared to the other assemblers.

5.1.2 Coding DNA Sequence Representation

The availability of a dataset of coding DNA sequences (CDSs) for *B. napus* morphotype 'DH12075' facilitated the verification of assembled transcripts by comparison to annotated expressed DNA sequences. CDSs are sequences of DNA derived from putative genes predicted from alignment of expressed sequences, either protein or RNA to the genome. These sequences are essentially composed of the exons of a gene with the 5' untranslated region (UTR) and introns removed. Using the sequence alignment software BLAT, assembled transcripts were aligned to the CDS dataset and the number of unique CDSs represented in each assembly were recorded. The CDS dataset contained 111,382 sequences, roughly two thirds of which would be expected to be present in the RNA sample from a single tissue type.

For each assembly, the number of CDSs that were matched by alignments to assembled transcripts using the BLAT aligner. BLAT reports hits between two sequences if the two sequences have an identity of $\geq 95\%$, where large inserts are allowable. To determine the number of CDSs represented at nearly full length, these

alignment hits were filtered to count the number of CDSs that matched an assembled transcript where at least 90% of the CDS bases matched the transcript. Surviving BLAT hits were considered as assembled transcripts representing putative genes at full length. The number of CDSs matched under these two sets of requirements is displayed for the three *de novo* assemblers and the reference-based Cufflinks assembly in Table 5.2. Both stringencies were included due to the nature of the CDS dataset consisting of predicted sequences built from exons of an annotation dataset. These sequences may not represent all the expressed RNA present in nature, and therefore high-scoring BLAT hits of shorter alignment lengths may also be biologically relevant. To better evaluate the proportion of transcripts that are full length within each assembly, the proportion of full length transcripts is shown relative to the total CDSs represented by each assembly as well. This data gives a more stratified view of how assemblies differ in producing full transcripts.

BLAT hits were split into bins based on the percentage of CDS bases matched by a transcript to more thoroughly investigate the distribution of assembled transcript lengths. These counts were converted to percentage of the total transcripts for each assembly to better compare across assemblers, as the total number of transcripts generated varied greatly between the softwares. The proportion of transcripts matching CDSs was plotted for alignment lengths in sixteen 1% wide bins from 85% up to 99%. This shows the distribution in the upper range of alignment lengths between assembled transcripts and annotated *B. napus* genes for each assembly.

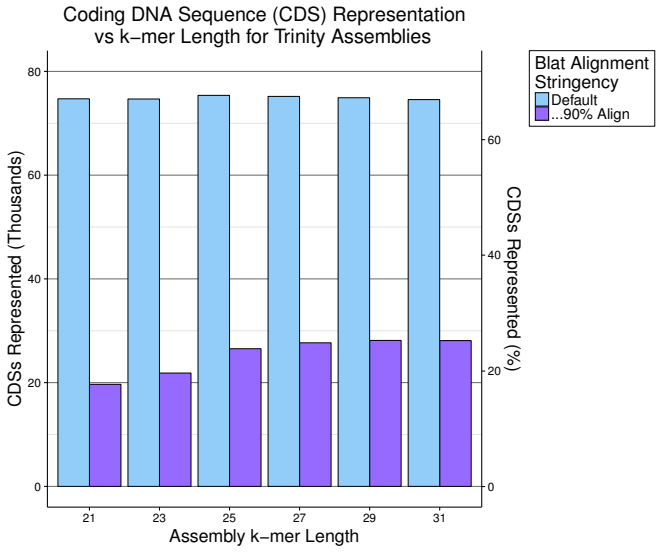
The sets of unique CDSs represented fully were compared triple-wise between several selections of assemblies by means of Venn Diagrams (Figures 5.10 – 5.12). These diagrams illustrate the number of CDSs missed by certain softwares or parameter sets, and also demonstrates the potential for combining the results of multiple assemblies.

Figure 5.6a shows the representation of CDS sequences at two BLAT stringencies for the *de novo* Trinity assemblies. Overall representation of the CDS dataset did not vary greatly between the six assemblies for the default BLAT stringency as all assemblies fell within 1% of each other. Filtering for $\geq$90% alignment length BLAT hits showed that assemblies using short k-mers (21 and 23 bp) showed a lower representation of CDSs, 17.7% and 19.6%, than the other four assemblies, using k = 25, 27, 29 and 31, which produced 23.8%, 24.9% 25.3% and 25.3% of the total CDSs at full length. This is consistent with the contig length metrics which also demonstrated lower k values produced fewer transcripts and at shorter average lengths. The default 25 bp k-mer length did not assemble the most transcripts representative of known CDSs. Longer k-mer length assemblies (29 and 31 bps) performed slightly better than the default, yielding an increase of 1.43% and 1.40% respectively of the total unique CDSs fully mapped by transcripts. The proportion of represented genes for each assembler that were fully mapped (Figure 5.6b show a more clear trend of improving assemblies as k increases up to the maximum 31 bps).

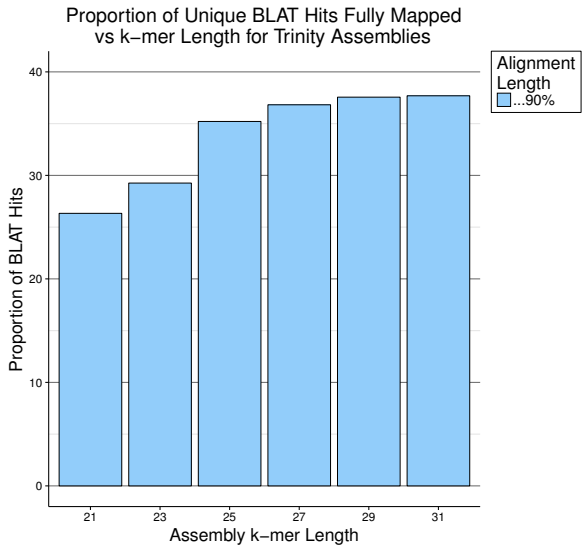
Trinity assemblies shown in Figure 5.6c exhibit a trend of higher k-mer lengths producing longer transcripts from 85% alignment length up to 87%. The assemblies diverge into two groups at the 88% alignment length and higher. For these lengths, the four higher k-mer length assemblies: 25, 27, 29 and 31 bps outper-

Table 5.2: Proportion of the coding DNA sequences represented in each of the assemblies at varying stringencies of alignment length. The columns represent the total number and % of CDSs matched by assembled transcripts, and the number of CDSs and % of all CDSs represented at >90% alignment length, respectively.

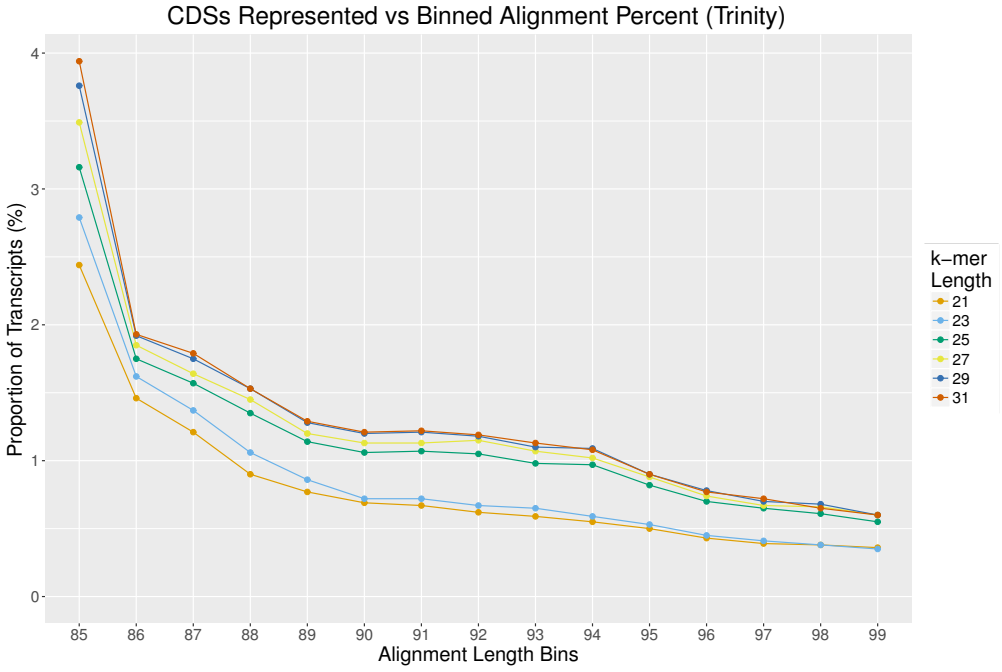
	All Alignments		>90% Alignment Lengths	
Assembly Method	CDS Matches	% of Total CDSs	CDS Matches	% of Total CDSs
Trinity				
21 bp *k*-mer	74,718	67.08	19,675	17.66
23 bp *k*-mer	74,679	67.05	21,849	19.62
25 bp *k*-mer	75,381	67.68	26,542	23.83
27 bp *k*-mer	75,179	67.50	27,681	24.85
29 bp *k*-mer	74,921	67.26	28,139	25.26
31 bp *k*-mer	74,562	66.94	28,102	25.23
Oases (merged)				
21–31 bp *k*-mers	72,578	65.16	32,853	29.50
31–41 bp *k*-mers	74,519	66.90	22,052	19.80
41–51 bp *k*-mers	70,605	63.39	28,609	25.69
SOAPdenovo-Trans				
25 bp *k*-mer	77,422	69.51	9,112	8.18
27 bp *k*-mer	77,676	69.74	9,481	8.51
29 bp *k*-mer	78,128	70.14	9,903	8.89
31 bp *k*-mer	78,369	70.36	10,315	9.26
41 bp *k*-mer	79,133	71.05	3,849	3.46
51 bp *k*-mer	78,720	70.68	7,877	7.07
61 bp *k*-mer	78,790	70.74	7,165	6.43
71 bp *k*-mer	79,133	71.05	3,849	3.46
Cufflinks				
ref-based	71,570	64.26	38,390	34.47



(a) CDS Representation



(b) Propotion of Assembly Full Length



(c) Binned Alignment Length

Figure 5.6: Representation of coding DNA sequences for the Trinity *de novo* assembler. (a) Quantity and percentage of the total CDSs representation by assembled transcripts with two BLAT stringencies: the default parameter set (light blue) and full length BLAT alignments where the contig covered at least 90% of the CDS bases with matches (dark blue). (b) CDSs represented at full length as a proportion of the CDSs represented in total by each assembly. (c) Proportion of the assembled transcripts aligning to CDSs at sequence identities in 1% bins.

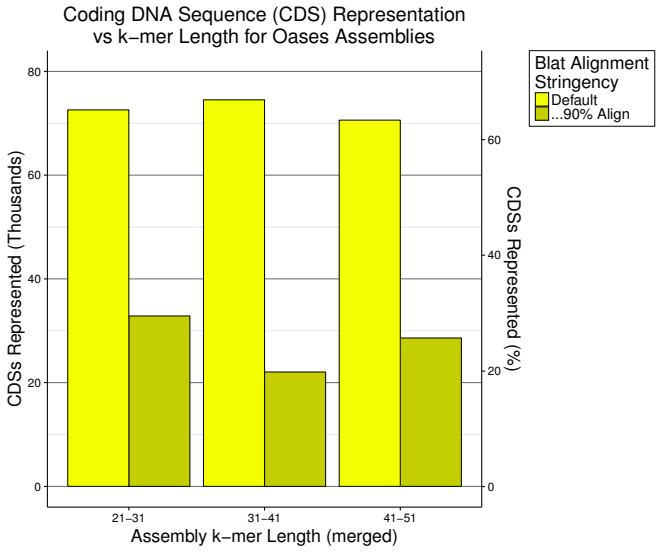
formed the lower k-mer assemblies of 21 and 23 bps as was the case with CDS representation by two BLAT stringencies. Within these two groups, longer k-mers in general outperformed shorter k-mers in terms of assembling transcripts accurately at full length.

Assemblies generated by Oases had the largest disparity between varying k-mer lengths (Figure 5.7). The representation of CDS sequences by Oases assemblies varied greatly between the three merged k-mer-length assemblies, with no linear trend observed. The middle assembly, using k-mers from 31 up to 41 bps had a noticeable drop in fully represented CDSs, and a minor increase in CDSs represented at shorter lengths. This observation, coupled with the large increase in transcripts produced at shorter average lengths for this assembly suggests that this range of k-mers is more prone to fragmented or incomplete contigs. For the 21 to 31 bp k-mer range, the Oases assembler was able to produce the highest number of transcripts represented at $\geq 90\%$ length of all the *de novo* assemblers. Finally, the longer k-mer length assembly, merging k-mers from 41 to 51 bps performed slightly worse than the merged 21–31 bp assembly. The proportion of CDSs represented at full length (Figure 5.7b also shows the 21–31 bp k-mer range producing a higher ratio of longer transcripts produced.

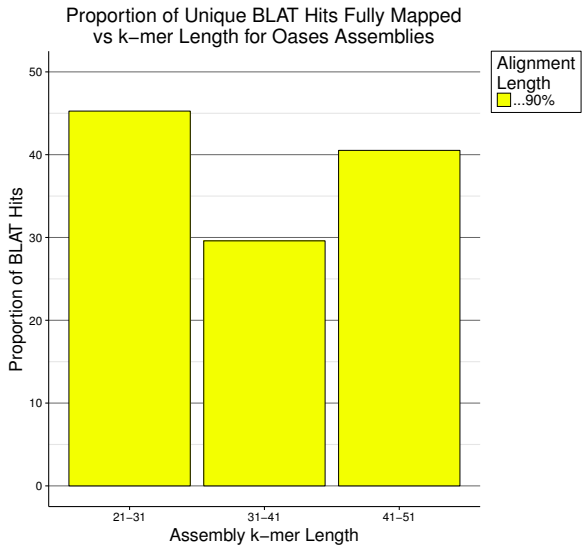
Figure 5.7c displays the distribution of alignment lengths for transcripts representing CDSs. The larger disparity observed between assemblies is likely a result of the assemblies being merged from wider ranges of k-mer lengths. The 21–31 bps k-mer merged assembly greatly outperformed the other two in this study. This assembly also produced an unexpected trend with an increasing proportion of CDSs assembled up to 88% sequence length, before transitioning to the expected downward trend. All other assemblies, including those from other assemblers, exhibited a downward trend throughout the bins as alignment length increases.

The wide range of allowable k-mer lengths during assembly using SOAPdenovo-Trans provided a unique insight into the interaction between de Bruijn graph k-mer lengths and the sensitivity and specificity of the assembler to real transcripts. The number and proportion of CDSs that were represented in SDT assemblies is shown in Figure 5.8a. Varying the k parameter in SDT had a very negligible effect on the proportion of CDSs matching assembled transcripts using the BLAT aligner at default parameters. There were however, noticeable drops when considering only alignments that span at $\geq 90\%$ of the CDSs length. All assemblies over 31 bp performed worse in this metric than those at 31 bp or below. The 41 bp k-mer length assembly in particular performing poorly.

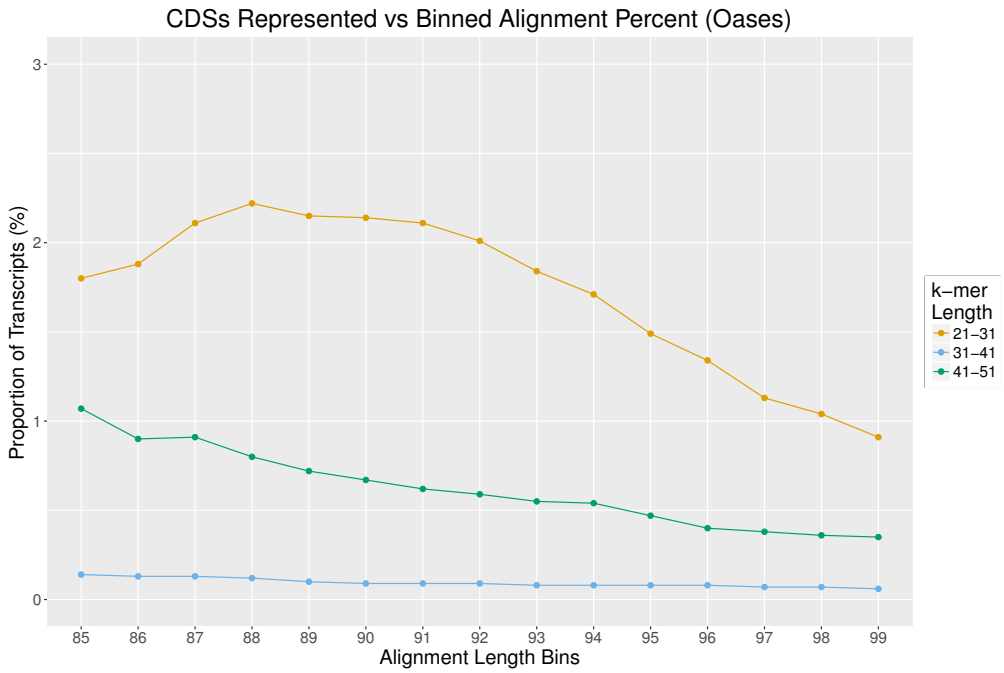
This trend was also observed in the Oases assemblies, as the merged 31–41 bp assembly, and also the single k-mer assemblies from 33 up to 39 performed poorly when compared to the surrounding trends of k-mer length. These assemblies also exhibit a large increase in the number of transcripts produced. The fact that these observation occurred in both Oases and SDT assemblers for similar values of k-mer length suggested that there may be significant increase in repetitive sequences in the read data, or genome of same length as k-mers used. This would cause an increase of chimeric transcripts resulting from these repeats. This possibility was investigated further by examining k-mer repetition in the read sequences (Appendix C), although, no abnormal repetition was found matching those k-mer lengths.



(a) CDS Representation

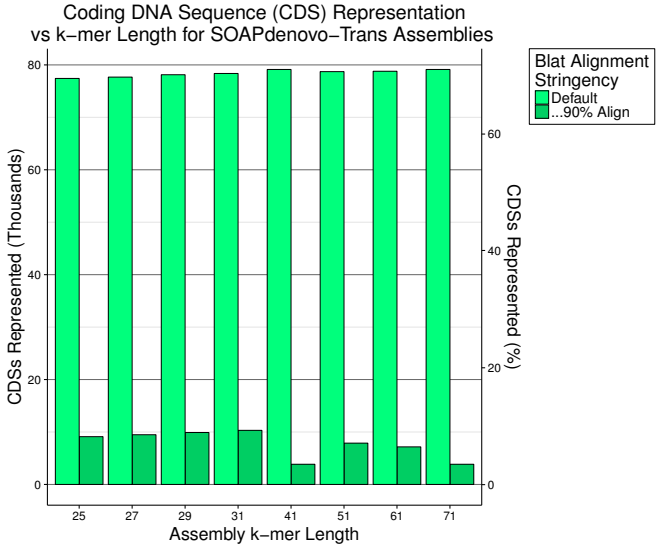


(b) Propotion of Assembly Full Length

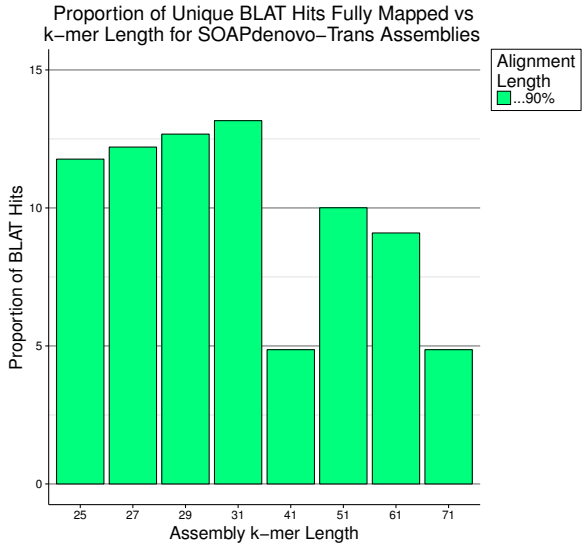


(c) Binned Alignment Length

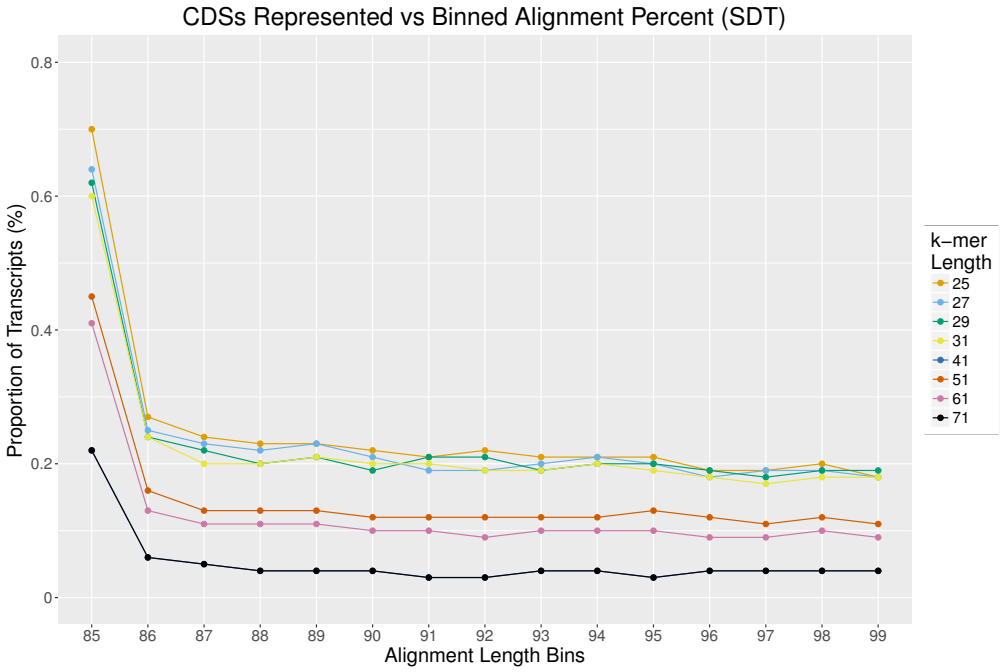
Figure 5.7: Representation of coding DNA sequences for the Oases *de novo* assembler. (a) Quantity and percentage of the total CDSs representation by assembled transcripts with two BLAT stringencies: the default parameter set (light yellow) and full length BLAT alignments where the contig covered at least 90% of the CDS bases with matches (dark yellow). (b) CDSs represented at full length as a proportion of the CDSs represented in total by each assembly. (c) Proportion of the assembled transcripts aligning to CDSs at sequence identities in 1% bins from 85% sequence identity up to 99%.



(a) CDS Representation



(b) Propotion of Assembly Full Length



(c) Binned Alignment Length

Figure 5.8: Representation of coding DNA sequences for the SOAPdenovo-Trans *de novo* assembler. (a) Quantity and percentage of the total CDSs representation by assembled transcripts with two BLAT stringencies: the default parameter set (light green) and full length BLAT alignments where the contig covered at least 90% of the CDS bases with matches (dark green). (b) CDSs represented at full length as a proportion of the CDSs represented in total by each assembly. (c) Proportion of the assembled transcripts aligning to CDSs at sequence identities in 1% bins. (the k=41 assembly is not visible because it is exactly behind the k=71 assembly line).

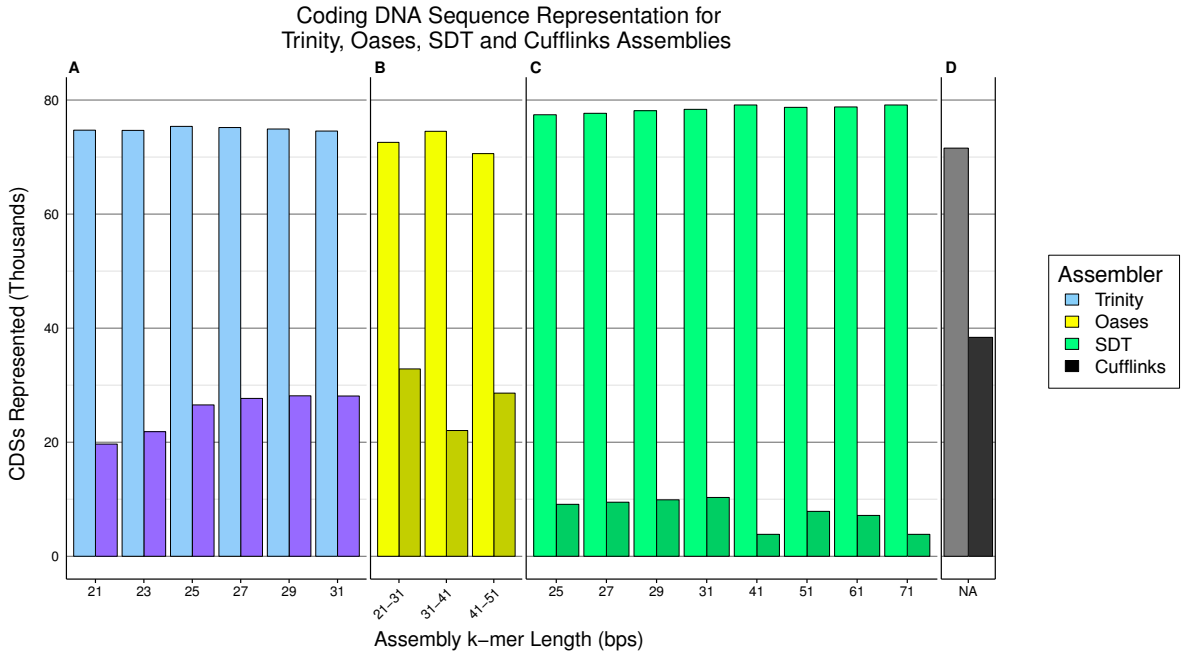
The SOAPdenovo-Trans binned CDS representation plot in Figure 5.8c shows that SDT assemblies exhibit a very even representation of CDSs over the varied alignment lengths for each assembly. Further, the assemblies were stratified into three clusters: 25-31 bp k-mers, 51 and 61 bp k-mers and the 41 and 71 bp k-mer assemblies that have performed poorly in all other studies. Within these groupings, generally the shorter k-mer assembly perform the best with slight variations between bins.

A reference based assembly generated by Cufflinks was produced to set a reasonable goal for optimal performance. While the assembly still is not considered biologically perfect, it sets a standard for what a reasonably well-founded assembly might look like. For the metric of CDS representation, it shows that roughly 75% of the *B. napus* genes are present in the RNA-Seq reads and just under 40% can be assembled at 90% length or greater (Table 5.2, Appendix A). Interestingly, the Cufflinks assembly yielded fewer total unique CDS hits at the default BLAT stringency, suggesting the other assemblies potentially generated transcripts that are similar to unexpressed genes, or chimeric misassembled transcripts that would be eliminated upon comparison to the reference genome. This supports the hypothesis that transcripts derived from different homeologous genes could be confused during transcriptome assembly where no reference is present for verification.

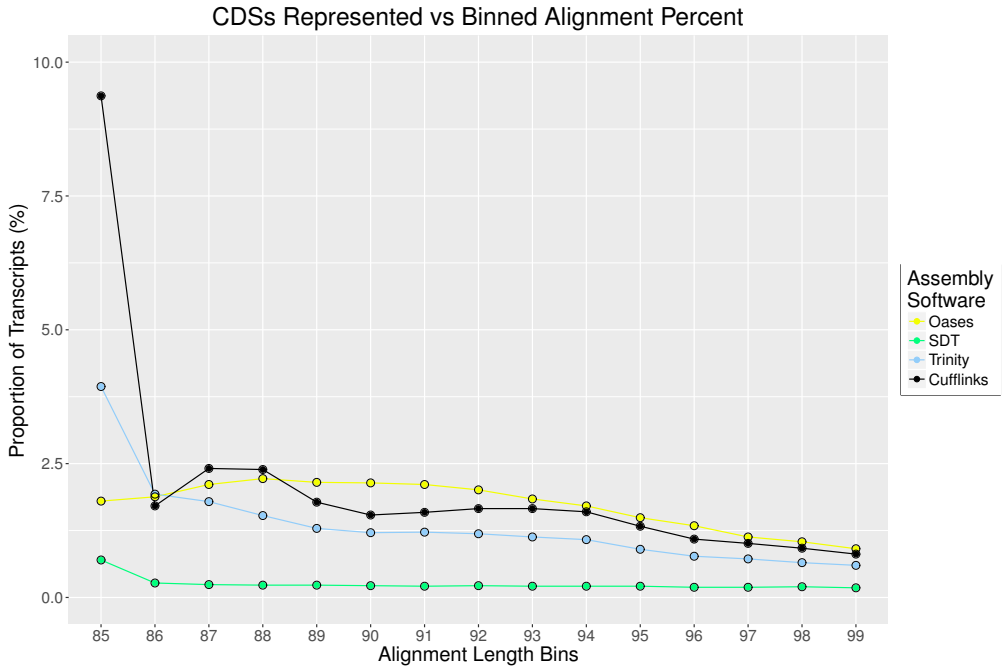
Figure 5.9a illustrates CDS representation between the three transcriptome assemblers as well as the reference-based Cufflinks assembly. For all three *de novo* assemblers, over 70% of the CDS sequences were represented with BLAT hits over all k values. This suggests that each of the assemblers were able to represent sequences similar to all of the present transcripts. This number of CDSs was possibly in excess considering the expected number of transcripts present in a single tissue type. When combined with the number of transcripts produced in each assembly however, it is apparent that there was repetition in the assemblies. This is especially true for the Oases assembler using k-mer values 31–41 bp. In this assembly, 637,797 transcripts were produced corresponding to only 74,519 distinct CDSs. This assembly also produced the lowest proportion of coding DNA sequences represented at $\geq$90% alignment length of the Oases assemblies.

The comparison between the best performing assemblies from each software for binned alignment lengths is shown in Figure 5.9b. This more detailed view illustrates the Oases assembler's superiority for constructing full length transcripts that match real coding DNA sequences. The 21–31 bp merged assembly from Oases even performed better in this regard than the reference-based Cufflinks assembly for alignment lengths above 89% sequence identity.

While SOAPdenovo-Trans was able to recreate the largest proportion of CDS sequences at a basic level, it performed poorly at assembling transcripts in full. Since it is not expected for nearly 80% of *B. napus* genes to be expressed in leaf tissue, such a large number of represented CDSs is more likely a result of incomplete assembled segments matching unexpressed gene homeologs. The merged Oases assembly using k-mer lengths 21 up to 31 bp on the other hand, produced fewer unique transcripts corresponding to CDSs but generated nearly five-fold more full transcripts than SOAPdenovo-Trans. This trade-off is in favor of the Oases assembler, as the >75% proportion of CDS sequences found in SDT assemblies is higher than the expected



(a) CDS Representation



(b) Binned Alignment Length

Figure 5.9: Representation of coding DNA sequences for the three *de novo* assembler: Trinity, Oases and SOAPdenovo-Trans and one reference-based Cufflinks assembly. (a) Representation by aligning the assembled transcripts to the CDS dataset at two BLAT stringencies: the default parameter set (light colours) and BLAT alignments where the contig covered at least 90% of the CDS bases with matches (dark colours). (b) Proportion of the assembled transcripts aligning to CDSs at binned sequence identity for the Trinity K31, Oases 31-31 bp merged, SDT K25 and Cufflinks assemblies.

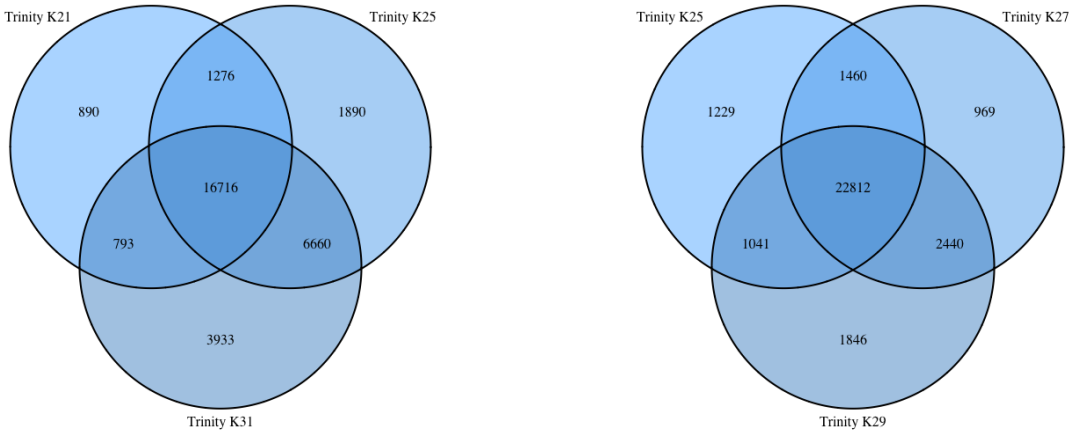
proportion of expressed genes in a single tissue type. Additionally, fully formed and long transcripts are of much greater importance in downstream analysis, for example, in determining gene function by comparison to similar genes. Overall, the Oases 21–31 bp merged assembly shows the most promise when comparing to known coding DNA sequences.

Although the merged assemblies from the Oases software may not be directly comparable to Trinity and SOAPdenovo-Trans, there is evidence to support k-mer length ranges of 25 to 31 bp as optimal for similar genomes. Assemblies using k-mer lengths within this range outperformed the others consistently in the number of transcripts matching CDSs at $\geq 90\%$ length for all three *de novo* transcriptome assemblers. These results may not translate to other organisms though, as genome structure and complexity no doubt play a large role in transcriptome assembly. The sharp decline in assemblies using k-mers longer than 31 bps was also not confirmed to be specific to *B. napus* or an intricacy of the assemblers. In our studies, the Oases assembler performed the best out of the three *de novo* assemblers. Trinity also performed well while, SOAPdenovo-Trans was not able to represent the CDS sequences at full lengths as well as the other assemblers.

In order to further demonstrate the disparity in representation of known *B. napus* CDSs, Venn diagrams were generated to determine the number of sequences represented in multiple assemblies and the number uniquely represented by others. In Figure 5.10, three assemblies and their intersection are shown in two Venn diagrams. These plots illustrate a number of things: firstly, the number of CDSs that are recreated uniquely by certain assemblers or assembly parameters, secondly, the potential CDS sequences to be gained by combining different combinations of assemblies together.

For the six Trinity assemblies, three assemblies spanning the total allowable range of k-mer lengths in the Trinity software, and another comparing three assemblies near the default k-mer length which were the better performing assemblies in terms of representation of the known CDSs. A similar approach was used for SOAPdenovo-Trans, comparing three assemblies across the wider range of allowable k-mer lengths, and three assemblies that performed the best. Finally, the three Oases merged k-mer assemblies were compared along with a selection of the most promising assemblies from each of the three assembly softwares.

The Venn diagram showing the unique CDSs represented for the allowable range of k-mer lengths using Trinity shown in Figure 5.10a further illustrated the poor performance of short k-mer lengths. Of the total 32,158 unique CDSs represented by the K21, K25 and K31 assemblies, 890 (2.77%), 1,890 (5.88%) and 3,933 (12.23%) were uniquely assembled in the K21, K25 and K31 assemblies respectively. As a single assembly, the default value of k for Trinity assemblies potentially miss as much as 13.00% CDSs present in the read data. In this dataset, the 31 bp k-mer length assembly was able to represent the most true sequences uniquely when compared to other Trinity assemblies. While multiple assemblies could be potentially used to increase the number of true sequences assembled, the vast majority (16716 or 51.98%) of unique sequences were assembled by all three Trinity assemblies. It also suggests that using multiple k-mer lengths is a valid way to ensure maximal representation of the set of true sequences, as all assemblies include some CDSs undetected by other



(a) Trinity **(b)** Trinity

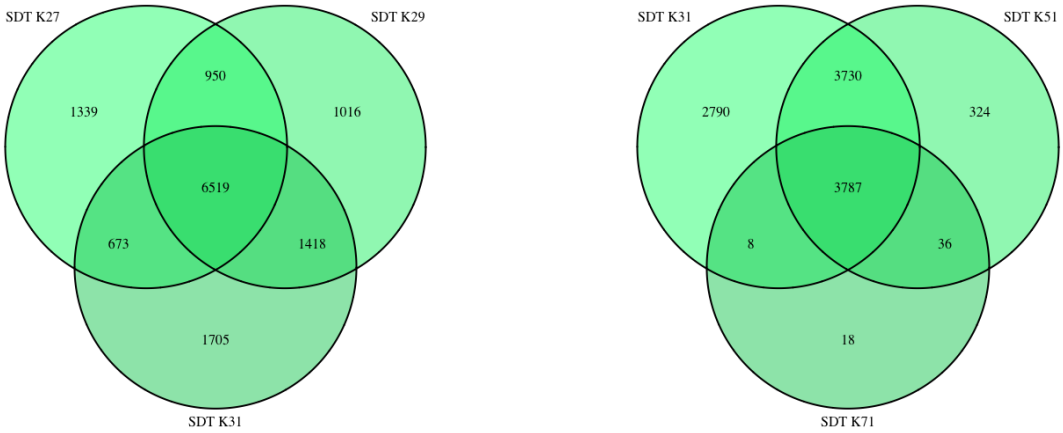
Figure 5.10: Venn diagrams outlining the number of CDS sequences represented in Trinity assemblies and the overlaps of CDS representation between assemblies across different k-mer lengths. Plots compare CDSs represented in a wide distribution of the k-parameter (a) and for the best performing assemblies (b).

assemblies.

In addition to the wide range of Trinity k-mer length assemblies, we compared a closer range near to the default Trinity k-mer length including the 25, 27 and 29 bp assemblies shown in Figure 5.10b. These three assemblies represented a combined total of 31,797 unique coding DNA sequences. For the three assemblers present in this comparison, 1229 (3.87%), 969 (3.05%) and 1846 (5.81%) CDSs were unique to each of the K25, K27 and K29 assemblies respectively. These results suggest that the default k-mer length of 25 bps does not provide the complete picture of genes present in an RNA-Seq dataset. There are 4,726 CDSs represented in the K31 assembly (Figure 5.10a) and a total 5,255 CDSs found in the K27 and K29 assemblies (Figure 5.10b not present in the default K25 assembly. Varying the k-mer length is an effective method for increasing the sensitivity of the Trinity assembler.

The SOAPdenovo-Trans assembler followed a similar trend to Trinity, with K27. k29 and K31 bp k-mer assemblies performing very comparably (Figure 5.11a). Out of the 13,619 total CDSs represented by these three assemblies, 1339 (9.83%) were uniquely assembled in the K27 assembly, 1016 (7.46%) in the K29 assembly, and 1705 (12.52%) in the K31 assembly. While the number of unique sequences to each assembly is similar to the Trinity assemblies, the relatively low total number of unique CDSs represented in SDT assemblies make these numbers a more significant proportion and increase the justification for combining assembled transcript sets across multiple k-mer lengths.

For the wider range of SDT assemblies up to 71 bp k-mers shown in Figure 5.11b, there was a sharp decline in the number of CDSs present in the final assemblies. This was evident in the Venn diagrams as well, with the three assemblies producing a total of 10,693 unique CDSs represented, 10,315 (96.46%) of which were represented in the K31 assembly alone. The 51 and 71 bp k-mer assemblies produced very few uniquely mapped CDSs, 324 (3.03%) and 18 (0.17%) respectively. Together, the two assemblies only produced 378



(a) SOAPdenovo-Trans **(b)** SOAPdenovo-Trans

Figure 5.11: Venn diagrams comparing CDSs represented in multiple SOAPdenovo-Trans assemblies over a range of k values comparable to Trinity's allowable range (a) and over a wider spread of the SOAPdenovo-Trans range of k values (b).

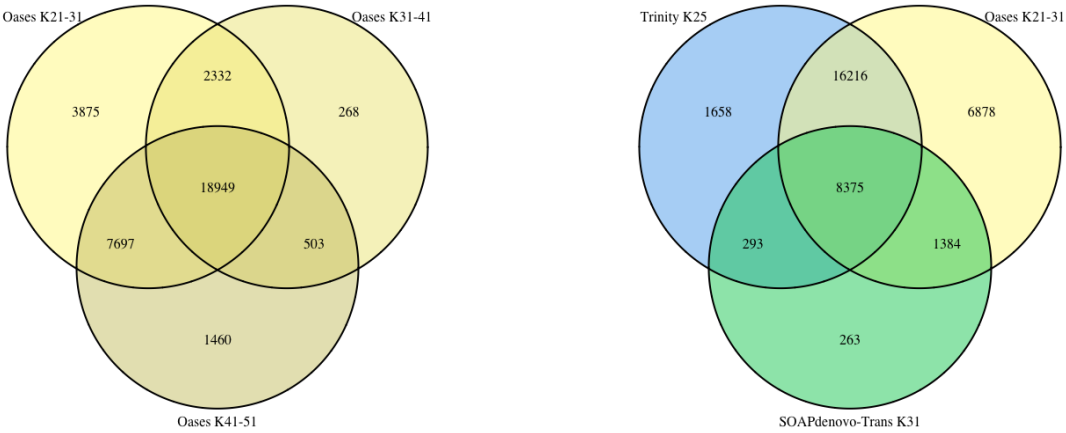
(3.54%) sequences not found in the K31 assembly.

The three Oases assemblies generated from merged single k assemblies 21–31 bp, 31–41 bp and 41–51 bp are compared in Figure 5.12a. A total of 35,084 combined unique CDSs were represented by Oases. The 21–31 bp, 31–41 bp and 41–51 bp merged assemblies each uniquely represented 3875 (11.04%), 268 (0.76%), 1460 (4.16%) CDSs respectively. Similarly to the CDS representation statistics for Oases, the merged 21–31 bp k-mer assembly greatly outperformed the other assemblies by Oases.

In order to compare across assemblers, we selected three assemblies, one from each software that was representative shown in Figure 5.12b. Assemblies compared were the default K25 assembly from Trinity, 21–31 from Oases and K31 from SOAPdenovo-Trans, because they performed the best in preliminary comparisons. In total, 35,067 of the coding DNA sequences were present in at least on of the assemblies. Oases yielded the highest number of unique CDSs represented with 6,878 (19.6%) sequences not found in the other two. Trinity and SOAPdenovo-Trans uniquely produced 1,658 (4.7%) and 263 (0.75%) CDSs. The low number of unique CDSs represented by SDT as well as the large number of sequences only present in the other two assemblies (24,752 or 70.6%) indicates that the SOAPdenovo-Trans assembler does not perform nearly as well as Trinity and Oases at assembling the putative transcripts at longer lengths.

5.1.3 Transcriptome Assembly Evaluation Software

Two software tools have purported to evaluate transcriptome assemblies and assign a numerical score using only the RNA-seq reads as input to determine assembly quality and plausibility. The DETONATE software provides a tool called "RSEM-EVAL" [46] which aligns reads to transcripts, assessing how many of the assembled transcripts are represented in the read data and provides a score for the probability that the assembly would be correct based on the reads. The software also provides scores for each transcript, for the



(a) Oases **(b)** All Assemblies

Figure 5.12: Venn diagrams of CDS representation in the three Oases assemblies (a) and comparing assemblies from each of the *de novo* transcriptome assemblers using the default or best performing assemblies from each assembler (b).

purposes of trimming low scoring transcripts to create a more concise transcriptome assembly. TransRate [44] is another software tool which purports the ability to compare assemblies not deriving from the same read set. The software also provides two scores: a general assembly score, and an optimized score, resulting from removing low scoring transcripts that TransRate deems extraneous. TransRate score range from 0 to 1, with higher scores denoting better assemblies. These two packages were used to score the 17 *de novo* assemblies from Trinity, Oases and SOAPdenovo-Trans as well as the reference-based assembly created by Cufflinks.

Both the DETONATE RSEM-EVAL and the TransRate softwares were used to produce transcriptome assembly scores for each of the three *de novo* assemblers: Trinity, Oases, and SOAPdenovo-Trans as well as the reference-based Cufflinks assembly. The scores produced are shown in table 5.3.

The DETONATE package [46] contains two methods for evaluating assemblies. The *de novo* method, called RSEM-EVAL, takes as input the reads used during assembly as well as the completed assembly in fasta format. The result is a probability score that the reads are explained by the assembly, as well as scores for each of the transcripts. The authors suggest that the transcript individual scores can be used to remove extraneous transcripts that are not as well represented by the reads to improve an assembly.

The assembly scores provided by the RSEM-EVAL software reflected much of the same results as our previous metrics. For Trinity assemblies, the four assemblies using k-mer lengths of 25, 27, 29 and 31 bps yielded the best scores out of all the *de novo* assemblies. These four assemblies were even competitive with the reference-based assembly generated by Cufflinks. The k21 and k23 assemblies were markedly worse than the other Trinity assemblies, but still outperformed the other two de Bruijn assemblers.

The merged assembly of the longest k-mer lengths performed the best for the Oases assembly which was surprising due to the larger quantity and shorter length of transcripts produced, which are generally

Table 5.3: Assembly scores calculated by reference-free evaluation software DETONATE RSEM-EVAL and TransRate taking into consideration the read data used for assembly.

Assembly Method	RSEM-EVAL	TransRate	TransRate Optimized
Trinity			
21 bp k-mer	-1.073×10^9	0.446	0.485
23 bp k-mer	-1.050×10^9	0.474	0.513
25 bp k-mer	-8.189×10^8	0.065	0.189
27 bp k-mer	-8.021×10^8	0.067	0.190
29 bp k-mer	-7.256×10^8	0.070	0.192
31 bp k-mer	-7.392×10^8	0.076	0.201
Oases (merged)			
21–31 bp k-mers	-1.727×10^9	0.001	0.020
31–41 bp k-mers	-2.634×10^9	0.001	0.012
41–51 bp k-mers	-1.569×10^9	0.001	0.012
SOAPdenovo-Trans			
25 bp k-mer	-2.104×10^9	0.011	0.077
27 bp k-mer	-2.070×10^9	0.003	0.032
29 bp k-mer	-2.045×10^9	0.010	0.077
31 bp k-mer	-2.037×10^9	0.000	0.000
41 bp k-mer	-2.215×10^9	0.001	0.013
51 bp k-mer	-2.045×10^9	0.037	0.168
61 bp k-mer	-2.058×10^9	0.000	0.001
71 bp k-mer	-2.215×10^9	0.038	0.143
Cufflinks			
ref-based	-5.854×10^8	0.127	0.280
threshold for TransRate		$\geq$0.220	$\geq$0.350

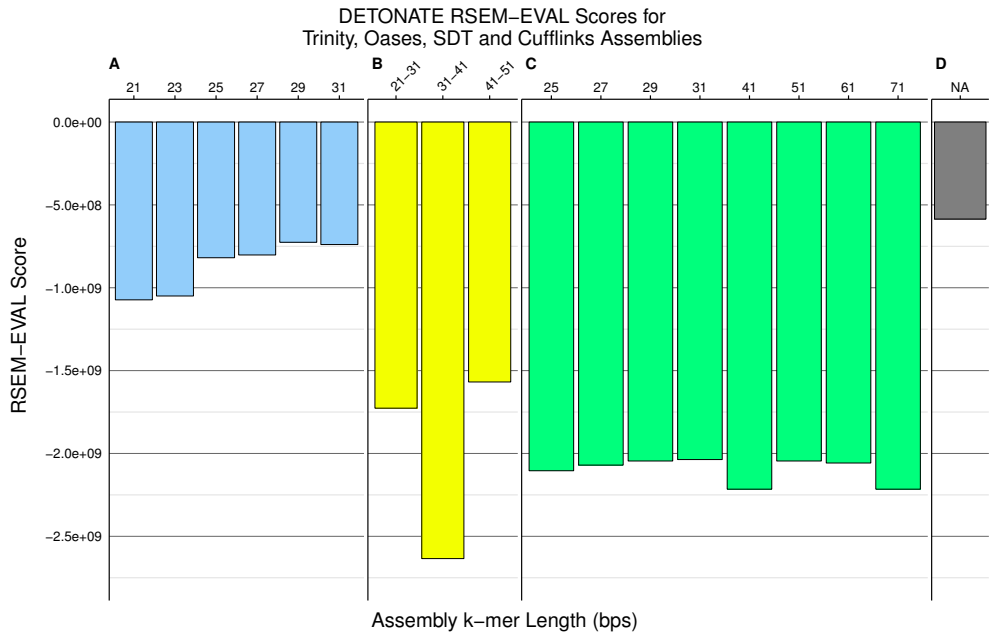


Figure 5.13: Transcriptome assembly evaluation scores provided by the DETONATE RSEM-EVAL tool for the *de novo* assemblers: Trinity (A), Oases (B), SOAPdenovo-Trans (C), and the reference-based assembler Cufflinks (D). Scores are large negative numbers, where scores closer to zero are better.

unfavoured for DETONATE scores. The merger k31–41 bp assembly, as in all other tests yielded very poor results, likely resulting from the excessively large number of transcripts produced at relatively shorter lengths. The CDS representation of this assembly was also considerable worse than the other two Oases assemblies. The cause for this extreme drop in assembly quality for specific k-mer lengths was further investigated using DETONATE to evaluate the single k-mer assemblies produced by Oases.

For the most part, SOAPdenovo-Trans had very little difference between DETONATE scores for the eight assemblies. SDT exhibits the same trend as Trinity for the relevant k-mer lengths of 25, 27, 29 and 31 bps; where longer k-mer lengths produce a slight increase in assembly quality according to the RSEM-EVAL algorithm. Although, these assemblies performed much worse than their Trinity counterparts, k-mer lengths of 51 and 61 base pairs yielded similar scores to the better assemblies of lower k-mer lengths (29 and 31), while the two 41 bp and 71 bp k-mer assemblies resulted in much lower DETONATE scores, consistent with the results from our other metrics.

The Trinity assembler was able to produce the best scoring transcriptome assemblies, with all six assemblies scoring higher than the other *de novo* assemblies. The two highest scoring Trinity assemblies (29 and 31 bp k-mers) were comparable to the reference-based Cufflinks assembly. Oases and SOAPdenovo-Trans assemblies were not rated as highly by the DETONATE software, likely due to the large number of transcripts produced by Oases, and the relatively shorter SDT transcripts. Except for the merged 31–41 bp Oases assembly, SDT yielded the lowest overall DETONATE scores of all three assemblies.

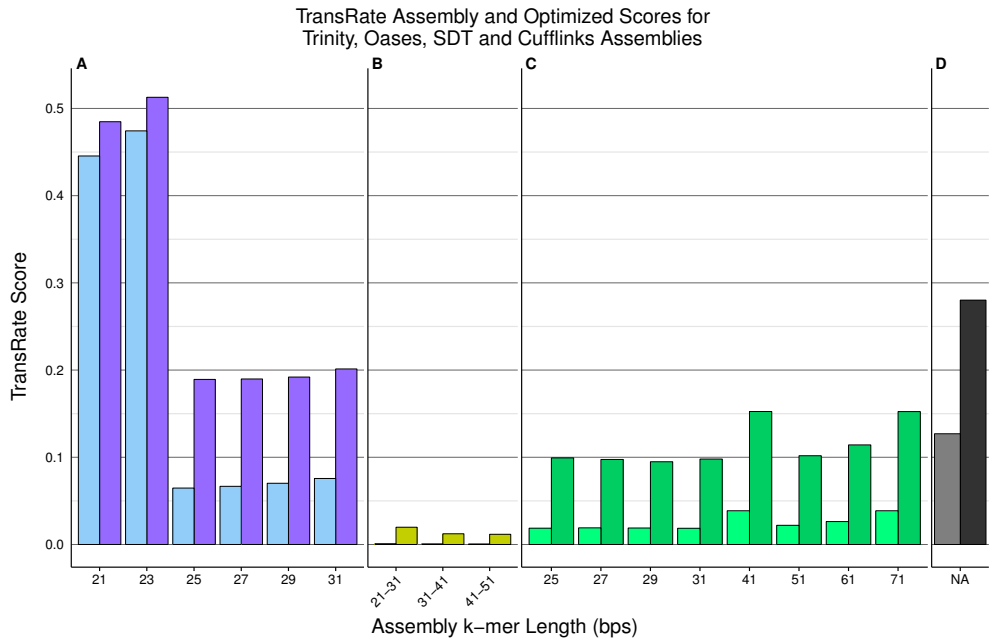


Figure 5.14: Transcriptome assembly evaluation scores provided by TransRate for *de novo* assemblers: Trinity (A), Oases (B), SOAPdenovo-Trans (C), and the reference-based assembler Cufflinks (D). Scores are decimals between 0 and 1, with 1 being a the best possible assembly score.

Another recent software developed for evaluating transcriptome assemblies without a reference-genome to compare to is the TransRate package [44]. From the TransRate developer's study of 155 published transcriptomes, the median score was 0.22 or optimized score of 0.35, which gives a rough idea of what constitutes an acceptable assembly according the the TransRate algorithm.

The TransRate assembly scores and optimized assembly scores are displayed in Figure 5.14 for the three *de novo* assemblers and the reference-based Cufflinks assembly. Overall, the results from the TransRate software was quite surprising. In contrast with all the other metrics used so far, TransRate puts the two shortest k-mer length Trinity assemblies well above all the other assemblies, including the Cufflinks assembly using a reference genome. Additionally, Transrate scored the SOAPdenovo-Trans k41 and k71 assemblies higher than the rest, contrary to the metrics used thus far. Oases assemblies also scored much worse than all other assemblies, where they have been comparable or better in other studies. Other trends such as improving scores from Trinity and SDT assemblies from k25 up to k31 did correspond with our other tests.

While the evaluation software tools do allow for comparisons between assemblies, there is reason to be skeptical of the scores presented for these assemblies. Both of the described methods use sequence alignment to assign reads to assembled transcripts and score an assembly based on the proportion of transcripts that appear to be represented in the read set. This can be misleading when the transcriptome contains several similar gene homeologs, where one copy may "soak-up" the reads corresponding to several related genes. This would result in only one of several genes being reported as present in the read data, erroneously penalizing the more correct assembly. Thus, in an allopolyploid organism, these scores may fail to account for the

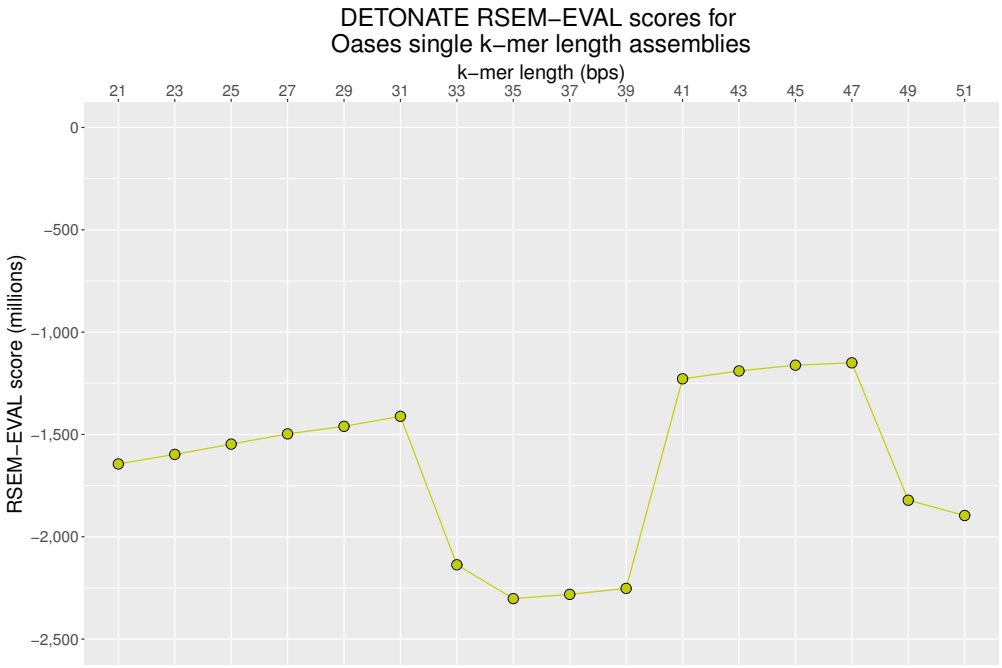


Figure 5.15: Transcriptome assembly evaluation scores provided by the DETONATE RSEM-EVAL tool for the sixteen single k-mer-length Oases assemblies used to generate the final merged assemblies. Scores are large negative numbers, where scores closer to zero are better.

possibility of several similar transcripts each being biologically relevant. For example, consider two genomes, one with a single gene for each group of homologs and another containing all the various homologs and recreating the repetition that is present in nature. Obviously the latter is the more biologically representative assembly, but when assigning reads to transcripts, the evaluation software would favor fewer assembly bases representing the reads. The DETONATE software includes an assembly prior component which specifically favors parsimonious assemblies, or those that explain the read data in a minimal set of transcripts. This may be more biologically appropriate for animals, where extraneous DNA is generally more unfavourable to the phenotype than in plants.

Investigation of k-mer Range 31–41 bp

A noticeable drop in assembly quality was observed for the Oases 31–41 bp merged assembly, as well as the SDT 41 bp k-mer lengths assembly. The observed quality drop occurring in two of the software sparked the hypothesis that an interaction between these specific k-mer lengths and a characteristic of the read-data could be the cause for poor assembly quality, rather than a quirk of the assembly software. This was investigated further with DETONATE RSEM-EVAL scores for each of the un-merged Oases assemblies from 21 up to 51 bps, as well as an analysis of k-mer repetition for a similar range of k values.

Figure 5.15 displays the DETONATE RSEM-EVAL scores for the single un-merged Oases assemblies. There is a clear drop in assembly quality for specific k-mer lengths of 33-39 bps, as well as 49 and 51 bps.

The rest of the k-mer lengths appear to follow a trend of increasing assembly quality as k increases. One hypothesis for the cause of these specific quality drops was that repetitive sequences near these lengths may be more common in either the genome or the read data. To test this, k-mer repetition was plotted for odd k values of 25 up to 51 and shown in Appendix C. These histograms show the number of unique k-mers that occur n times where $1 \leq n \leq 10,000$. These graphs should indicate an abnormal increase in repeats for the corresponding k-mer lengths if there is an abundance of repetitive sequences in the read data of length k, although no abnormal repetition was found for any of the lengths of k.

Another hypothesis for the drop in quality for assemblies using 41 bp k-mers was the drop in Illumina per-base quality scores roughly 40-50 bps into each 100 bp read. Using the FastQC tool [56], the Illumina trimmed reads used for assembly were examined for per-base quality, shown in Appendix B. There is an observable drop in base quality around base pair 50 out of 100, reaching a minimum phred score of 30 corresponding to an error rate of 1 in 1,000 (99.9% accuracy). This may result in some erroneous read joining between reads origination from similar sequences (homeologs) but is unlikely to be the sole cause for a systematic problem of assembling k-mers of 33 to 41 bps.

5.1.4 Sequence-based Ortholog Differentiation

The large quantity of homeologous gene-pairs in *Brassica napus* originating from ortholog pairs between the two progenitor genomes present a problem for sequence assembly, especially in *de novo* studies where a reference genome is not available for comparison. This section describes the methods used to determine the extent to which these homeologous gene pairs are misassembled as chimeric transcripts containing reads derived from two separate loci.

First, an orthology table of DH12075 CDS gene pairs was generated using the method of reciprocal best hit searching. This method is commonly used for determining the most likely orthologous pairs between two datasets [20, 57]. This was done based on reciprocal alignment of 'DH12075' CDSs to a set of *B. napus* morphotype 'Darmor' genes, where an orthology table was available. This allowed us to convert 'Darmor' orthologous gene pairs into the equivalent 'DH12075' gene pairs. As well as those genes verified by the 'Darmor' orthology table, gene pairs that produced both reciprocal BLAT hits that were not located on the same subgenome were included, to account for the possibility of genes unique to the 'DH12075' morphotype. In total, 27,469 gene-pairs verified by the 'Darmor' orthology table and 2,023 additional genes were identified for a total of 29,492 pairs of homeologous genes.

To determine whether or not the pairs were represented in the assemblies as separate genes or collapsed into genes from a single locus, we used the highest scoring BLAT hits to each gene and compared the gene identifier assigned by the assembler between two genes of a homeologous pair. For Trinity transcripts, identifiers with the same TR, read cluster and gene id were considered collapsed into a single loci. Oases provides a locus number used to differentiate gene loci. SOAPdenovo-Trans provides a unique ID for each transcript, resulting in only genes matching the same transcript sequence being reported as collapsed. The

Table 5.4: Ortholog gene pairs in DH12075 determined to be collapsed into single assembled transcripts, correctly identified as separate genes, or not assembled.

Assembly Method	Collapsed Pairs	Separate Pairs	Not Assembled
Trinity			
21 bp *k*-mer	10,952	6,157	12,383
23 bp *k*-mer	11,091	6,016	12,385
25 bp *k*-mer	11,714	5,884	11,894
27 bp *k*-mer	11,596	5,985	11,911
29 bp *k*-mer	11,627	5,925	11,940
31 bp *k*-mer	11,454	5,949	12,089
Oases (merged)			
21–31 bp *k*-mers	12,365	4,300	12,827
31–41 bp *k*-mers	12,024	5,226	12,242
41–51 bp *k*-mers	11,673	4,494	13,325
SOAPdenovo-Trans			
25 bp *k*-mer	10,468	7,239	11,785
27 bp *k*-mer	10,484	7,304	11,704
29 bp *k*-mer	10,344	7,767	11,381
31 bp *k*-mer	10,340	7,861	11,291
41 bp *k*-mer	10,205	8,204	11,083
51 bp *k*-mer	10,187	8,065	11,240
61 bp *k*-mer	10,253	8,039	11,200
71 bp *k*-mer	10,205	8,204	11,083
Cufflinks			
ref-based	6,726	9,155	13,611

number of gene pairs corresponding to transcripts within a single locus, separate loci and where at least one of the two genes was not represented (not assembled) are shown in Table 5.4.

Results of this categorization can be found in Figure 5.16. For the Trinity assembler, very little variation in ortholog differentiation is observable between *k*-mer lengths. Interestingly, the default and recommended *k*-mer length of 25 bp performed slightly worse than all other Trinity assemblies. The merged 31–41 bp Oases assembly correctly separated more orthologous pairs than the other two assemblies from the same software. This is the only metric where this assembly outperformed the other Oases assemblies. This may be due to the sheer volume of transcripts produced for that assembly, increasing the number of separated pairs just by reducing the probability of orthologs matching the same transcript. While SOAPdenovo-Trans appears to produce the largest number of correctly separated orthologous gene pairs, this is likely due to

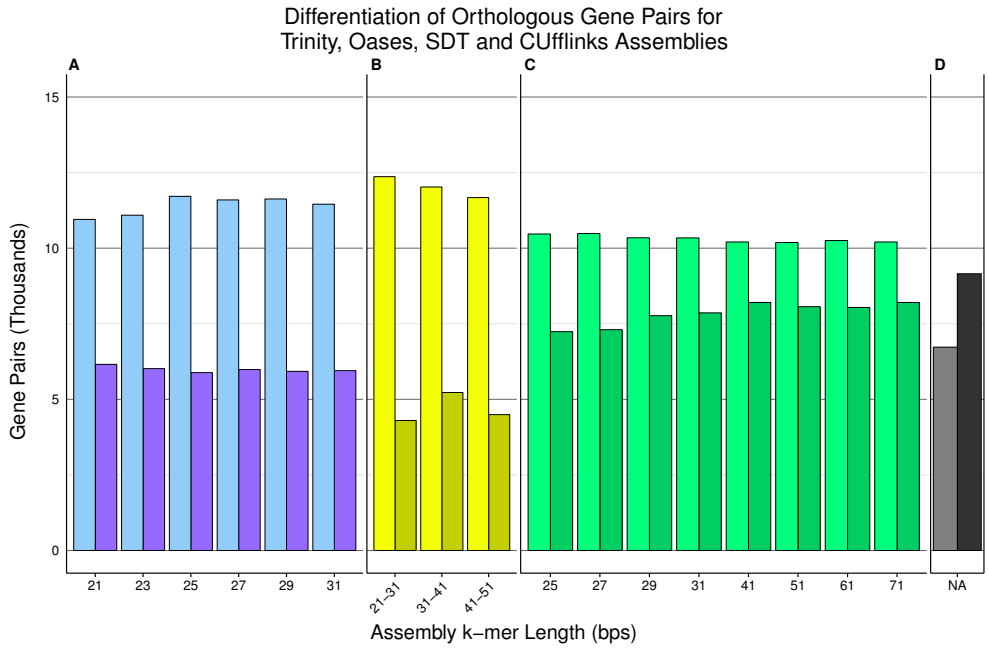


Figure 5.16: Differentiation of orthologous gene pairs for three *de novo* assemblers Trinity (A), Oases (B), SOAPdenovo-Trans (C) and one reference-based assembler Cufflinks (D). Bars represent the number of 'collapsed' gene pairs whose highest BLAT hit were to the same transcript (left/lighter bars) or to transcripts designated as different genes by the assembler, classified as 'separated' (right/darker bars).

the naming convention of the SDT transcripts, which do not include a locus designation in the transcript name. Due to this, reported SDT collapsed pairs are only reported when both CDS of an orthologous pair matched the same transcript twice. This would artificially inflate the number of separated pairs for SDT assemblies. The longer k-mer length assemblies for SDT did also outperform the lower, an observation also unique to this metric. The reference-based Cufflinks assembly was the only one to produce a majority of correctly separated homeologous gene-pairs, illustrating the advantage that a reference genome provides for differentiation between duplicated genes in transcriptome studies.

The overall trend observed for the differentiation of similar homeologous genes in *B. napus* is that longer k-mer lengths increase the number of gene pairs that are correctly separated. This supports the hypothesis that longer k-mer lengths increase the specificity to sequence variance, allowing the assembler to see a larger portion of the sequence data at a time. Since the de Bruijn graph only guarantees that sequence variants that exists within k bps of each other are assembled together, longer k-mers increase the allowable distance between variation where the correct pairing of two variants are captured within a single k-mer, eliminating the ambiguity. This comes at the cost of losing some read overlaps, since the required overlap between two single reads must be at least k bps, increasing the number of read overlaps that are missed as k increases.

Chapter 6

Graph Visualization of Assembled RNA Transcripts (GVART)

The complexity of the de Bruijn graph algorithm and the large-scale data output often complicate interpretation of the results of *de novo* transcriptome assembly experiments. Assembly statistics such as N50 or evaluation software scores, are also of little use when taking a fine-grained look at the underlying data. Reported transcripts can be ambiguous in how they are related and performing multiple sequence alignments can be time consuming. The Graph Visualization of Assembled RNA Transcripts (GVART) tool was developed as a web-based application for examining the graph structures and resulting transcripts that is compatible with the current most popular transcriptome assembler: Trinity.

This section outlines the Graph Visualization of Assembled RNA Transcripts (GVART) tool developed over the course of this research project. GVART was implemented as a web tool for providing a means for researchers to view the de Bruijn graphs built by Trinity during an assembly experiment. It is portable enough for a scientist to use as a tool to explore their own assemblies, and robust enough to be hosted publicly to allow other researchers to explore a published or work-in-progress assembly. The tool uses Perl to parse assembly output, GVART to layout the graph and javascript libraries: D3, SlickGrid and Biojs-vis-sequence to present the data in an interactive and explorative manner.

6.1 Introduction

De Bruijn Graphs are a type of directed graph that define the overlaps between sequences of letters or symbols. A more in-depth description of DBG properties and their implications is provided in section 2.5.3. For DBGs created in transcriptome assembly, typically a k-mer length of $\geq$21 bp is used and can be increased to various lengths depending on the software limitations. This results in a total of 4^{21} unique k-mers, or nearly 4.40×10^{12} (4.4 trillion). Additionally, nucleotide sequences of 21 base pairs or longer are not easily readable or comprehensible to humans. For these reasons, de Bruijn graphs are inscrutable without simplification. Thankfully, large linear paths are common for DBGs based in biological sequence strings. These linear paths can be compressed into superstrings of the associated combined k-mers.

In polyploid organisms, multiple copies of genes called homeologs are present. Because of their similarity in sequence, homeologous genes can often be assembled together in chimeric sequences whereby the resulting

transcript contains sequence from two gene from different loci; or be confused as alternative splicing isoforms during assembly. Homeologs occur in different frequencies depending on the organism, and are most prevalent in plants where polyploidy is common, resulting in large genomes relative to animals. For example, the wheat and canola genomes have very large repetitive genomes resulting from whole-genome duplication events and their high ploidy (tetraploid or hexaploid for domestic wheat and diploid for canola). The existence of many sequences that repeat throughout these genomes complicates many aspects of genome-wide studies, such as assembly and annotation of sequences.

Repetitive sequences in the genome present a larger problem for DBG algorithms than OLC. When breaking up read sequences into k-mers, only variations that are at most k bases apart are represented in the graph. For example, if two singe nucleotide polymorphisms are spaced $k+1$ bases apart, there is no way for a DBG algorithm to identify which variation of these two SNPs came from the same sequence. This opens up the possibility for chimeric transcripts to be reported, inflating the number of transcripts assembled. One way to combat this problem is by increasing the length of k for the algorithm. This would increase the threshold for spacing between variations and lower the number of chimers produced. This also increases the search space of the algorithm, as it depends on the number of unique k-mers. For example, increasing the k-mer length from 25 to 31 bps, for biological sequence data would increase the number of unique k-mers from 4^{25}=3.36x10^7 to 4^{31}=4.61x10^{18}, an 8 magnitude difference. For modern assemblers, this translates to higher memory requirement rather than a time complexity one, but may still hinder some experiments.

Identifying problems with DBG assemblies can be difficult due to the inherent complexities of the algorithm. De Bruijn graphs are very large, and incomprehensible to humans. Because the graphs are made up of very long sequences of only a few words, humans are not well trained to read or compare these sequences. There are tools that allow the visualization of graphs created by DBG assembly methods, however these existing tools have limitations that hinder their usefulness. Bandage [58] is a tool that generates interactive images of the graphs produced by several DBG assemblers, including Velvet [59], SPAdes [60] and Trinity [3]. While this software is functionally similar to GVART, it is significantly more heavy-weight, and suffers from performance issues from loading entire assemblies at a time. Additionally, it does not allow for hosting externally for collaboration between researchers. Another tool for visualizing DBGs is ABySS-Explorer [61], which produces very striking, stylized visualizations that illustrate aspects of the assembly such as lengths of linear paths, overlaps between nodes and branches. The tool is designed for visualizing genome assemblies, simplifying a great deal of sequence information into a single image. Transcriptome assemblies are inherently more fragmented and do not translate well to this style of visualization. The developers of ABySS-Explorer have identified these shortcomings and have expressed their desire to adapt their tool for this use-case.

GVART displays a scalable vector graphic (SVG) image representing a simplified connected component of the de Bruijn graph used to assemble RNA-seq data into transcript sequences. It allows for the user to drag nodes and zoom the graph, as well as select and highlight multiple nodes to create custom assemblies. Assembled transcripts are displayed in a table below the graph image and are selectable to highlight the

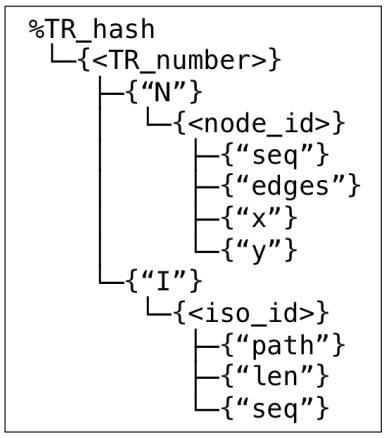


Figure 6.1: Hash data structure used in `parse_trinity_assemply.pl` to store the assembly data for conversion to GVART usable format. Each level represents a sub-hash within the above hash. Text within triangular braces correspond to a variable representing a certain part of the assembly. Text within quotes are accessed using the quoted string as the key and return a data point corresponding to the string.

corresponding nodes and path through the graph. Selecting a sequence either by transcript or a collection of nodes displays the corresponding sequence in a third pane, allowing highlighting of features such as start and stop codons and continuous open reading frames (ORF) in each of the 6 possible reading frames. The number of amino acids or codons of the longest ORF is displayed for each reading frame for quick comparison of sequences.

6.2 Implementation

Due to its robust plain text manipulation functionality, Perl is used to read in a Trinity assembly and format it for the web-app. The custom script “`parse_trinity_assembly.pl`”, reads in the Trinity.fasta output file, building a data structure of each unique TR. These TRs are each connected components of the larger de Bruijn graph generated by the whole assembly, and are of ideal size to be visualized in a single instance. A multi-level Perl hash is used to store the entire assembly, with the hierarchy shown in Figure 6.1. From the header of each transcript, node IDs, lengths, sequences and outgoing edges are read and stored for each transcript. Each assembled transcripts is stored at the isoform level, as an array of node ids, with the length and sequence of the transcript stored as well. The hash format is useful as it allows for redundant information to be eliminated, as well as iteration over the entire structure for printing to data files.

Layout of the nodes into a reasonable area with minimal edge overlaps is done with dot software from GraphViz [62]. Once the nodes of a connected component and their connections are stored, the Perl script formats the data into a format accepted by the GraphViz dot software. The IPC::Open2 Perl module is used to send output to the external GraphViz software and receive output. This is done to reduce the overall number of intermediate data files. Output from the dot program is received in plain-text format with

coordinates provided for the nodes. These are used in conjunction with the javascript scale function to size the graph according to the dimensions of the browser window.

Once the node positions are determined and the graph data is stored, the perl script formats and prints a javascript file containing the variables that the main web-app will read and display. Each node is stored in an array, with fields corresponding to the name, id, coordinates, length and sequence. Edges are stored as pairs of references to the nodes array. Finally, transcripts are stored with name, length and a list of nodes IDs corresponding to the path through the graph. One datafile for each TR is output, as well as one file listing the directory of TR javascript data files for the web-app to provide selection to the user. This way, the web-app only has to process one portion of the assembly at a time, as it is requested by the user.

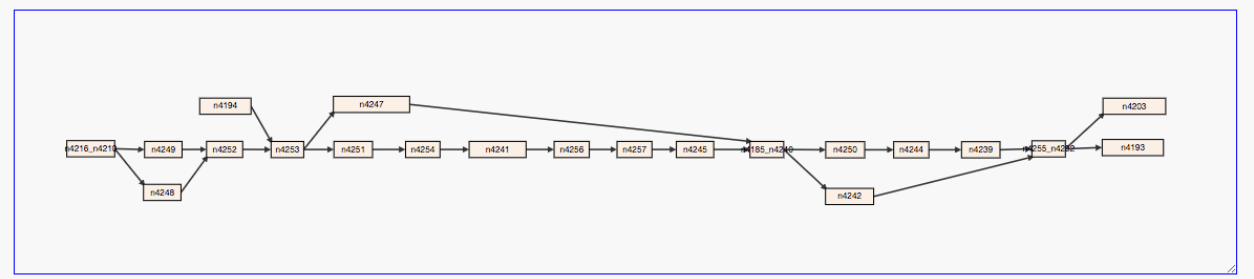
The main feature of the data driven documents (D3) library [63] is to provide interactivity to the graph and underlying data. When the page is first loaded, the javascript graph data passed to the web-page is loaded and used to generate an SVG image of the nodes, and edges of the graph with node IDs as labels and arrows to show the directionality of the paths (Fig. 6.2). SVG elements are created and formatted using the d3 "append" and "attr" functions respectively. Interactivity with the graph is accomplished via listeners for certain events, which update the corresponding constant variables, and then calling the updateGraph function. The graph is then redrawn to display the new graph state. Functionality of the graph includes zooming and dragging the graph to reposition it, moving nodes via dragging, and selecting one or more nodes.

Transcripts are listed in a table format using the SlickGrid [64] javascript library, displaying each unique ID from the assembly fasta file, their length, a confidence value for Oases assemblies and the list of nodes in the path. This table is sortable by name, length or confidence value (if it is present). Selecting a transcript by clicking will highlight the corresponding nodes in the graph pane, and animate the edges from the first node to the last using dashed-line arrows.

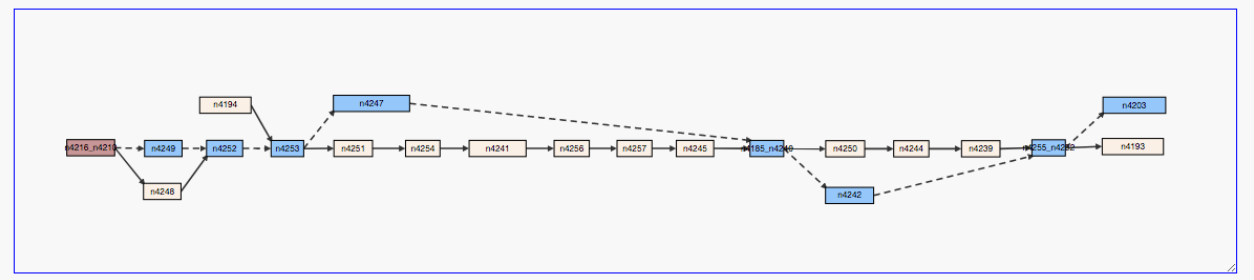
GVART also provides the sequence of transcripts or nodes of interest. Selecting one or more nodes, or a transcript will result in the corresponding sequence to be shown in the bottom most pane of the tool (Fig. 6.4). This pane is toggle-able due to some performance loss for longer sequences slowing the selection of graph elements. The sequences associated with the selected nodes or transcript is displayed in the third pane from the top using the javascript library biojs-vis-sequence [65]. Here, the sequence can be shown in a variety of formats with different options for highlighting features. The sequence can be displayed in codata, fasta, pride or raw format and also supports selecting and copying highlighted sequence.

6.3 Future Development

Currently, the GVART tool only supports assemblies generated from the Trinity assembler where the path is output as part of the header of transcript sequences. If interest in the tool is sufficient, we would look to expand this to allow parsing of intermediate Trinity files for versions where the path is not output. Additionally, the Oases assembler output is sufficient to allow for a version of GVART that supports Oases



(a) No selection



(b) Transcript selected

Figure 6.2: An SVG graph formatted by the GVART web-app from a Trinity assembly corresponding to a single connected component of the de Bruijn graph called a TR. Nodes are represented by rectangular boxes and correspond to a linear path of any number of k-mers. The length of the sequence associated with each node corresponds to the node's width. Edges between nodes are drawn with arrows and show the potential paths through the graph that correspond to overlaps between sequences in the read data. (a) default graph visible on page load, (b) selected path with root node highlighted in red, subsequent nodes in blue and the path edges in dotted lines.

Name	Length	Confidence Val	Nodes in Path
c0_g1_i1	967		4216,4249,4252,4253,4247,4185,4242,4255,4203
c0_g1_i2	955		4216,4248,4252,4253,4251,4254,4241,4256,4257,4245,4185,4250,4244,...
c0_g1_i3	920		4194,4253,4247,4185,4242,4255,4203

Figure 6.3: The GVART table of reported transcripts for the current TR provided by SlickGrid. The table is sortable alphabetically by name or numerically by length. Nodes are displayed in sequence order, corresponding to nodes shown in the graph pane. Confidence values are an Oases specific value only reported for relevant assemblies.

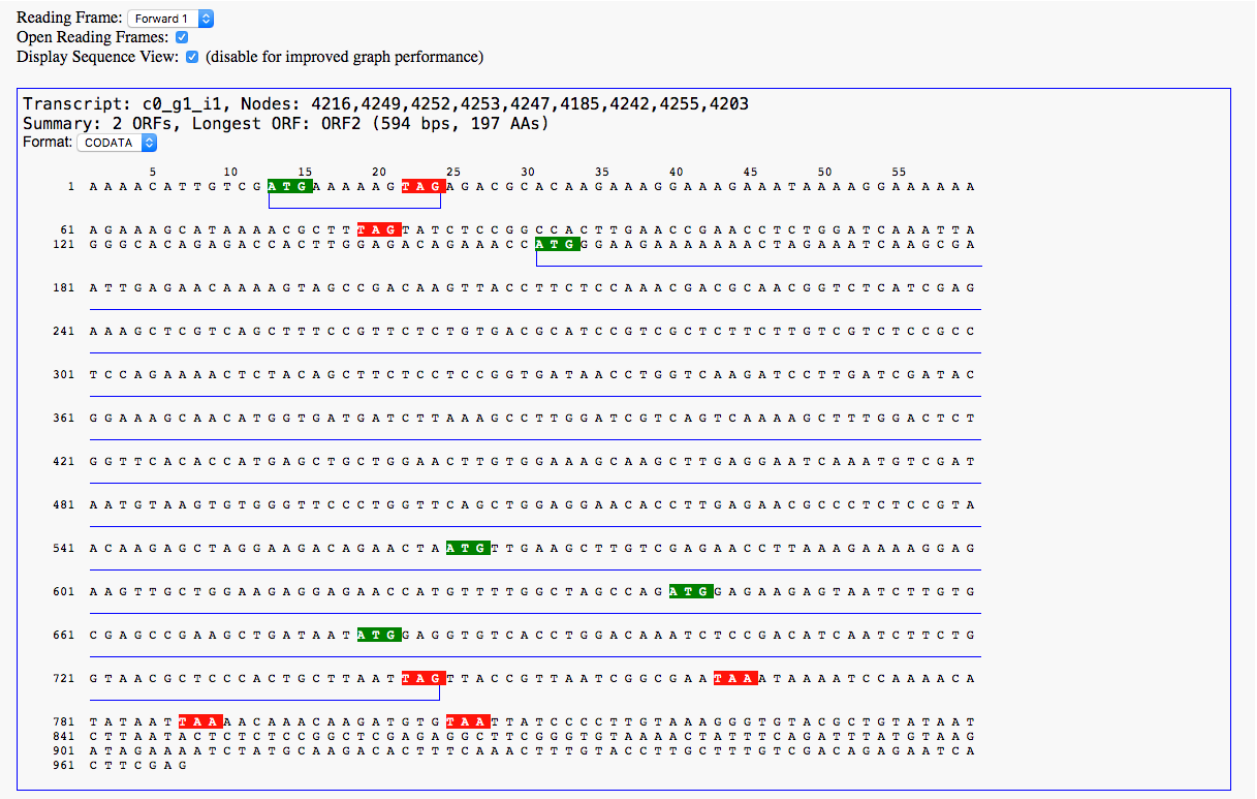


Figure 6.4: The GVART sequence pane displaying the nucleotides of the selected transcript (or series of nodes) is formatted and displayed using biojs-vis-sequence. Annotation is possible in each of the 6 reading frames, selected by the drop-down menu. Open reading frames, as well as the entire sequence view are toggle-able via check-boxes.

assemblies as well. Initial tests of displaying Oases graphs have illuminated problems with conflicting node IDs as well as a difficulty in displaying the multi-directionality of some edge sequences. There has been interest in aligning known sequences to the nodes and paths of the de Bruijn graphs. Currently, there is no clear solution on how to display a sequence alignment over top of the simplified de Bruijn graphs of GVART.

6.4 Availability

GVART is available for download via GitHub at: https://www.github.com/coadunate/GVART

Chapter 7

Discussion and Future Work

7.1 Current Transcriptome Assembly Software

Using a variety of metrics, *de novo* assemblies generated from three different softwares, Trinity, Oases and SOAPdenovo-Trans, were evaluated and compared. In addition to using several different current assembly softwares, the k-mer length parameter was also varied, dictating the length of subsequences the RNA-seq read data is divided into to create the de Bruijn graphs. This chapter outlines the observed trends within, and across the three softwares, as well as providing some advice to researchers who intend to conduct transcriptome studies within a polyploid organism.

The metrics used in this study were chosen due to their availability to researchers who aim to evaluate their transcriptome assemblies without the presence of a reference genome. While NGx and $NGAx$ statistics were a potential option due to the presence of the 'DH12075' reference genome, they were not included to better model the situation when a reference genome is not available. Additionally, the comparison techniques to known sequences used could be done using data from closely-related species via database searching.

7.1.1 Effect of k-mer Length on Assembly Statistics

In general, a favour towards k-mer lengths of 25-31 bps was observed for all three *de novo* assemblers. For Trinity, this constitutes the upper limit to the k-parameter and resulted in better assemblies for all metrics except for TransRate score. The four assemblies in this range were competitive for contig statistics, CDS representation and DETONATE scores, with a slight favor towards the two longest k-mer lengths of 29 and 31 bps. Short k-mer Trinity assemblies K21 and K23 performed worse in all metrics but the TransRate evaluation scores. The Oases assembly encompassing merged single length assemblies for this range also performed the best, yielding a reasonable number of transcripts with generally longer lengths. This assembly represented the most CDS sequences at longer lengths, and included a large number of transcripts representative of CDSs that the other Oases assemblies did not produce.

There was an observed drop in assembly quality specifically for assemblies using k-mer lengths in the $33 \leq k \leq 41$ range. Both the merged Oases assembly in this range (31-41 bp k-mers) and the K41 SDT assembly were the worst performing assemblies for their software. These assemblies consistently produced excessive quantities of transcripts with relatively shorter lengths. These two assemblies also did not represent

known coding DNA sequences well. Examining the DETONATE scores for single k-mer assemblies of the merged Oases assembly for this range revealed that assemblies 33, 35, 37 and 39 showed uncharacteristically poor assembly quality, where a clear trend existed around them as shown in Figure 5.15. Due to the Oases assembly showing expected results while SDT performed poorly, the cause of the issue was not believed to be related to the specific k-mer length of 41 base pairs, but rather specific to the software. This illustrates the benefit of examining the single k-mer assemblies of Oases prior to merging assemblies in order to achieve the best possible assembly.

The two major implications that k-mer lengths have on de Bruijn graph assembly are: (1) the required overlap between two reads is $\geq k$ bases in order to join them during de Bruijn graph assembly, and (2) only sequence variants that are $\leq k$ bases apart are guaranteed to be captured in the k-mer sequences. These two factors create a trade-off between sensitivity to read-overlaps and specificity to sequence variation. Increasing the k-mer length for de Bruijn graph assemblers should, in theory, produce less sensitive but more specific assemblies. This should decrease the number of chimeric transcripts observed when joining sequences derived from two separate gene loci. We therefore expected to observe more accurate assemblies as we increased the k-mer length, while also observing a decrease in contig lengths and an increase in transcripts produced. This represents a trade-off between sensitivity and specificity. Increasing the k-mer length produces assemblies that are more specific, i.e. fewer false transcripts are reported by the assembler resulting in increased specificity. Increasing k also results in less sensitive assemblies, as fewer read overlaps are present in the DBGs, leading to more missed true transcripts (false negatives). Lower values of k result in more sensitive and less specific assemblies. The difficulty in transcriptome assembly evaluation lies in testing the validity of assembled transcripts, as the true sequences cannot be known with 100% certainty. Using sequence comparison to putative transcripts to determine ground truth creates the problem of determining the thresholds for how similar sequences must be, especially when dealing with homeologous genes that can be up to 98% identical while still originating from distinct loci [66]. At this level of sequence similarity, even the error rate of the sequencing technology can impact the decision of the truth of an assembled transcript.

For the most part, this hypothesis was affirmed for Oases and SOAPdenovo-Trans, where a wider range of k-mer lengths was allowed. For these assemblies, lower k-mer lengths produced assemblies with fewer contigs, but at longer lengths which was shown in the contig statistics. The trend was most obvious in the SOAPdenovo-Trans assemblies as they did not use merged assemblies, resulting in more assemblies over a wide range of k-mers. Shorter k-mer lengths also resulted in transcripts that represented CDSs at full lengths. We believe this is due to the BLAT stringency of 95% not being high enough to eliminate chimeric transcripts. In this case, assembled transcripts, whether they contain variants originating from both homeologs of a pair, would be reported by BLAT as a hit. Analysis of ortholog differentiation however, showed that longer k-mer lengths resulted in transcripts that more accurately represented the variation originating from a single gene locus, instead of a mixture of variations from both gene from a homeologous pair. For Oases and SDT assemblies, where a larger range of k-mers were allowable, increasing k-mer length resulted in more gene

pairs being correctly differentiated as homeologs. We believe that the differentiation analysis of ortholog pairs is a more accurate metric than CDS alignment lengths for assessing the specificity to sequence variance (discussed further in section 7.2).

7.1.2 Comparison of Assemblers

Considering the range of results we observed over a range of assemblies from each software, the Trinity assembler was the better performing transcriptome assembly software based on overall performance for each of the metrics used. While Oases was able to produce longer transcripts according to the N statistics, and more transcripts aligned at $\geq 90\%$ alignment length, Trinity was able to more accurately represent the repetitive sequences within *B. napus*. Additionally, the evaluation softwares both placed Trinity higher than the other two assemblers. SOAPdenovo-Trans, for the majority of studies undertaken here, performed very poorly. For it's entire range of k-mer lengths, SDT produced shorter transcripts, failed to create transcripts representative of real CDSs and was rated poorly by both evaluation softwares.

The results observed here differ to a recent evaluation of the same assemblers using the killifish *Fundulus heteroclitus* where the Oases assembler performed better than Trinity[67]. The Oases merging approach in combination with the difference in ploidy between the two organisms (2n for *Fundulus heteroclitus* versus 4n for *Brassica napus*) were the most likely causes of this discrepancies. In a relatively more simple transcriptome, the merging approach in Oases may be helpful in recreating the largest number of unique transcripts. In highly duplicated transcriptomes, merging multiple assemblies results in an assembly containing many more misassembled transcripts that are representative of chimers between homeologous gene variances. For this reason, the often recommended solution for improving transcriptome assemblies of merging multiple assemblies may not be useful for transcriptomes including duplicated genes.

7.2 Ortholog Differentiation in *Brassica napus*

For all three *de novo* assemblies, the majority of assembled transcripts representative of genes within homeologous pairs were collapsed into genes within the same locus. The six Trinity assemblies produced similar proportions of gene pairs correctly separated vs collapsed. The disparity between collapsed and separated gene pairs was largest in Oases, where the 21-31 bp k-mer length performed the worst at differentiation of gene pairs. SOAPdenovo-Trans appeared to separate the most gene pairs, but this was likely due to every transcript being reported as a unique locus. The Cufflinks reference-based assembly was the only assembler able to separate the majority of gene pairs present in the read-data. The fact that there were still genes collapsed when the reference was available for verification illustrates the difficulty of assembly algorithms to assemble transcripts when multiple similar sequences exist derived from different genomic locations. For all *de novo* assemblies, the majority of transcripts corresponding to homeologs were not identified as separate genes, suggesting that for transcriptome assembly in genomes with many related gene pairs, transcripts

reported as isoforms should not be assumed to represent genes from the same locus.

7.3 Recommendations for Transcriptome Assembly of Polyploid Organisms

After comparison of the three most prominent transcriptome assemblers, using a wide breadth of evaluation metrics for a duplicated polyploid organism, the Trinity assembler produced the most accurate transcripts, while still representing a high number of real gene sequences. Trinity provides a great balance of specificity to similar genes resulting from homeology of the two subgenomes, and sensitivity to the large quantity of gene sequences present in the organism. The approach used by Trinity, whereby the full reads are used in downstream simplification of the de Bruijn graphs seems to be a better approach for solving the read-coherance problem of DBG assemblers than the Oases multiple k-mer length approach. While Oases managed to recreate more CDS sequences with transcripts aligning using BLAT, they were revealed to be more likely a result of chimeric assembly of reads originating from distinct genomic loci, corresponding to homeologous gene pairs. Oases assemblies also contained an excessive number of contigs, which complicate downstream analysis just by the sheer volume of sequences. The multiple k-mer assembly merging of Oases resulted in assemblies including a great deal of misassembled and chimeric transcripts due to the approach of including unique transcripts produced across the range of merged assemblies. Many of the transcripts that differ between varied k assemblies, and would thus be included in the merged assembly, are likely the result of chimerism. Therefore, any benefit gained by increasing k is lost by merging shorter k assemblies. SDT produced generally short, fragmented transcripts that did not represent a comparable set of CDS sequences compared to the other two. Evaluation software also placed Trinity well above the other two softwares.

The k-parameter to DBG assemblers also proved to be a useful method for assembly tuning. The hypothesized trade-off between specificity and sensitivity was supported in the varied k-mer assemblies. In general, increasing the k-mer length provided more accurate transcripts by increasing the read-coherence. Shorter k-mer length assemblies resulted in more complete transcripts though, as the stringency for read overlaps was lessened. For the Trinity assembler, values of k above the default 25 bps were some of the best performing assemblies. For this reason, it is recommended to use a variety of k-mer lengths for assembly from the default up to the maximum value of 31 bps if possible. Even basic metrics of contig quantity and length can provide a quick indication of whether or not the extra assemblies are meaningful to pursue further.

7.4 Graph Visualization Tool - GVART

Assembled transcripts that represented well-studied genes in *B. napus* were isolated and graphed using the Graph Visualization tool. Looking at several such graphs for Trinity assemblies done using different values of k-mer length allowed the understanding of how the branching patterns of the gene differed under different

assembly conditions. In combination with sequence searching tools such as BLAST, the tool will allow a researcher to determine TRs relating to important phenotypes. In particular, GVART will highlight shared nodes between transcripts of a TR, potentially bringing forward the protein-coding portion of gene that is active in important biological mechanisms. Analysis tools that highlight the most important and relevant subsets of large datasets are crucial to research in this age of high-throughput data-generation. GVART will be most useful to researchers who want to take a closer look at specific genes of interest after assembling a transcriptome, for example to tackle real-world applications such as crop yield, and stress tolerance for plant studies.

The tool could also be useful to investigate potential novel genes discovered during *de novo* transcriptome assembly. Because *de novo* transcriptome assembly is not limited to sequences that match a reference, the possibility of gene discovery is an exciting prospect. GVART would allow the researcher to see the sequences that build up potential new genes, either determining if they are mis-assembled from nodes that belong to other known genes or verifying them as potential novel genes.

While the tool is functional and feature complete in its current form, there are some issues with the app and development is ongoing. For the assemblies generated in this study, there are some cases of Trinity TRs with a large quantity of branching nodes that result in long complex paths. These are problematic to view in the current application, and there is no current solution to divide the graph into segments. For the vast majority of Trinity TRs though, the display is adequate to provide a means for comparing transcripts.

7.5 Future Work

There are some limitations to the functionality of the current softwares for *de novo* assembly of transcriptomes. Currently, there is a limit put on Trinity at 31 bp k-mers. It would be interesting to unlock this constraint and allow Trinity to generate assemblies using longer k-mers to be more comparable to the other two assemblers. This would require a major modification of the software to rewrite the dependencies that use a 64-bit integer with 2-bit encoding to store k-mer sequences. This may result in twofold increase in required memory, using a 128-bit representation of the k-mer sequences vs the 64-bit method currently used. As assembly quality for certain metrics increased up to the maximum k-mer length, improvements may continue beyond this barrier. Additionally, due to the strange quality drops for certain k-mers in the Oases assembly, allowing a merge of a non-continuous step through k-mer lengths of the single k assemblies would be interesting. Currently, Oases does not allow the merging of custom sets of single k assemblies (e.g. 21–31 and 41–51, excluding poor performing assemblies 33, 35, 37 and 39).

An often proposed means for improving assemblies is to merge or cluster multiple single k-mer assemblies [68, 69]. In general, the Oases merge function did not prove useful in the duplicate *B. napus* transcriptome. Oases assemblies resulted in an excessive number of transcripts that proved to be representative of the same proportion of the expressed transcripts observed in the other assemblies. The clustering method, which is to

combine multiple assemblies and remove redundant sequences, captures the worst of both extremes of the k spectrum. From short k assemblies, the cluster retains long transcripts with many variants originating from all homeologs, while long k assemblies provide short fragments of highly precise sequence. Instead, multiple k strategies could use the highly precise sequences from long k assemblies to resolve ambiguous pairings of sequence variants. These fragments could instead be mapped to clusters of similar short k transcripts to determine the chimers from the real transcripts.

Further investigation towards the ability of each assembler to differentiate between homeologous genes could be done via a SNP-based approach. The PolyCat [70] software makes use of read data from both progenitor species, compared to a reference of one of the progenitors, to determine single nucleotide polymorphisms that are representative of a specific homeolog of a pair. This software has been utilized in wheat, another polyploid crop, to use the SNPs identified to assign RNA-seq reads to one of the sub-genomes. When aligning the reads back to the transcripts, it is then possible to determine whether only reads derived from one progenitor make up a transcript, or if the transcript contains sequences from both (chimeric). In *B. napus*, RNA-seq data from *Brassica rapa* and *Brassica oleracea* could be used to perform this analysis. Additionally, with the read-data split into categories of belonging to progenitor *B. rapa*, *B. oleracea* and undetermined, two assembly processes could be undertaken in tandem using the assigned reads for each progenitor combined with the undetermined origin reads. In a final step, these two assemblies combined may produce a better quality overall assembly than what is possible when the assembler is responsible for resolving the read similarities caused by homeologous genes.

In conjunction, these studies would further enable analysis of the *B. napus* transcriptome. With better identification of chimeric assembled transcripts using homeolog specific SNPs, segments of each of the assembled transcripts could be assigned to one of the progenitors and visualized by colour coding nodes, or node segments in the GVART tool. This would be a useful method to visualize chimerism of homeologs within transcriptome assembly and allow for better identification of the origins of assembled transcripts. Downstream analysis of the transcripts, and their relation to important plant traits are facilitated by new methods for exploration and examination of *de novo* transcriptome assemblies. The potential benefits of these studies, extend to gene trait identification which can assist plant breeders in identifying the genetic elements that correspond to certain traits of interest such as yield, and plant robustness to abiotic or biotic stresses. The global rise in demand for canola and other rapeseed production is driving competition to make the best use of agricultural land space. Increasing yields, and developing crops that can withstand the conditions of previously un-tapped land will allow Canada to meet these demands, and remain a global agricultural powerhouse.

References

[1] Food and Agriculture Organization of the United Nations. Worldwide production of rapeseed. December 2017.

[2] Statistics Canada. Principal field crop areas, March 2017.

[3] Manfred G Grabherr, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn A Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, Zehua Chen, Evan Mauceli, Nir Hacohen, Andreas Gnirke, Nicholas Rhind, Federica di Palma, Bruce W Birren, Chad Nusbaum, Kerstin Lindblad-Toh, Nir Friedman, and Aviv Regev. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature Biotechnology*, 29(7):644–652, 2011.

[4] Marcel H Schulz, Daniel R Zerbino, Martin Vingron, and Ewan Birney. Oases: robust *de novo* RNA-seq assembly across the dynamic range of expression values. *Bioinformatics*, 28(8):1086–1092, 2012.

[5] Yinlong Xie, Gengxiong Wu, Jingbo Tang, Ruibang Luo, Jordan Patterson, Shanlin Liu, Weihua Huang, Guangzhu He, Shengchang Gu, Shengkang Li, Xin Zhou, Tak-Wah Lam, Yingrui Li, Xun Xu, Gane Ka-Shu Wong, and Jun Wang. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics*, 30(12):1660–1666, 2014.

[6] National Human Genome Research Institute. The human genome project completion: Frequently asked questions. *National Institue of Health*, 2010.

[7] K A Wetterstrand. DNA sequencing costs: data from the NHGRI genome sequencing program(GSP), May 2016.

[8] H Tettelin, V Masignani, M J Ciesiewicz, et al. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc Natl Acad Sci USA*, 102(39):13950–5, 2005.

[9] Canola Council of Canada. What is canola? Available at online at www.canolacouncil.org/oil-and-meal/what-is-canola, October 2016.

[10] Veronique Barthet. Canola. *The Canadian Encyclopedia*, 2013.

[11] Boulos Chalhoub, France Denoeud, Shengyi Liu, Isobel A. P. Parkin, Haibao Tang, Xiyin Wang, Julien Chiquet, Harry Belcram, Chaobo Tong, Birgit Samans, Margot Corréa, Corinne Da Silva, Jérémy Just, Cyril Falentin, Chu Shin Koh, Isabelle Le Clainche, Maria Bernard, Pascal Bento, Benjamin Noel, Karine Labadie, Adriana Alberti, Mathieu Charles, Dominique Arnaud, Hui Guo, Christian Daviaud, Salman Alamery, Kamel Jabbari, Meixia Zhao, Patrick P. Edger, Houda Chelaifa, David Tack, Gilles Lassalle, Imen Mestiri, Nicolas Schnel, Marie-Christine Le Paslier, Guangyi Fan, Victor Renault, Philipp E. Bayer, Agnieszka A. Golicz, Sahana Manoli, Tae-Ho Lee, Vinh Ha Dinh Thi, Smahane Chalabi, Qiong Hu, Chuchuan Fan, Reece Tollenaere, Yunhai Lu, Christophe Battail, Jinxiong Shen, Christine H. D. Sidebottom, Xinfa Wang, Aurélie Canaguier, Aurélie Chauveau, Aurélie Bérard, Gwenaëlle Deniot, Mei Guan, Zhongsong Liu, Fengming Sun, Yong Pyo Lim, Eric Lyons, Christopher D. Town, Ian Bancroft, Xiaowu Wang, Jinling Meng, Jianxin Ma, J. Chris Pires, Graham J. King, Dominique Brunel, Régine Delourme, Michel Renard, Jean-Marc Aury, Keith L. Adams, Jacqueline Batley, Rod J. Snowdon, Jorg Tost, David Edwards, Yongming Zhou, Wei Hua, Andrew G. Sharpe, Andrew H. Paterson, Chunyun Guan, and Patrick Wincker. Early allopolyploid evolution in the post-neolithic *Brassica napus* oilseed genome. *Science*, 345(6199):950–953, 2014.

[12] Yves Van de Peer. The flowering world: a tale of duplications. *Trends plant science*, 14(12):680–688, 2009.

[13] Justin Ramsey and Douglas W. Schemske. Pathways, mechanisms, and rates of polyploid formation in flowering plants. *Annu. Rev. Ecol. Syst.*, 29:467–501, 1998.

[14] Nagaharu U. Genome analysis in *brassica* with special reference to the experimental formation of *b. napus* and peculiar mode of fertilization. *Jpn. J. Bot.*, 7:389–452, 1935.

[15] E W Myers, G G Sutton, A L Delcher, I M Dew, D P Fasulo, M J Flanigan, S A Kravitz, C M Mobarry, K H Reinert, K A Remington, E L Anson, R A Bolanos, H H Chou, C M Jordan, A L Halpern, S Lonardi, E M Beasley, R C Brandon, L Chen, P J Dunn, Z Lai, Y Liang, D R Nusskern, M Zhan, Q Zhang, X Zheng, G M Rubin, M D Adams, and J C Venter. A whole-genome assembly of *Drosophila*. *Science*, 287(5461):2196–2204, 2000.

[16] Philip Campbell (Ed.). The yeast genome directory. *Nature*, 387(6632S):3–103, 1997.

[17] The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 408:796–815, 2000.

[18] D L Nelson, A L Lehninger, and M M Cox. *Lehninger Principles of Biochemistry*. Macmillan, 2008.

[19] S F Altschul, W Gish, W Miller, E W Myers, and D J Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, 1990.

[20] R L Tatusov, E V Koonin, and D J Lipman. A genomic perspective on protein families. *Science*, 278(5338):631–637, 1997.

[21] Jun Wang, Feng Tao, Nicholas C Marowsky, and Chuanzhu Fan. Evolutionary fates and dynamic functionalization of young duplicate genes in arabidopsis genomes. *Plant Physiology*, 2016.

[22] Shengyi Liu, Yumei Liu, Xinhua Yang, Chaobo Tong, David Edwards, Isobel A. P. Parkin, Meixia Zhao, Jianxin Ma, Jingyin Yu, Shunmou Huang, Xiyin Wang, Junyi Wang, Kun Lu, Zhiyuan Fang, Ian Bancroft, Tae-Jin Yang, Qiong Hu, Xinfa Wang, Zhen Yue, Haojie Li, Linfeng Yang, Jian Wu, Qing Zhou, Wanxin Wang, Graham J King, J. Chris Pires, Changxin Lu, Zhangyan Wu, Perumal Sampath, Zhuo Wang, Hui Guo, Shengkai Pan, Limei Yang, Jiumeng Min, Dong Zhang, Dianchuan Jin, Wanshun Li, Harry Belcram, Jinxing Tu, Mei Guan, Cunkou Qi, Dezhi Du, Jiana Li, Liangcai Jiang, Jacqueline Batley, Andrew G Sharpe, Beom-Seok Park, Pradeep Ruperao, Feng Cheng, Nomar Espinosa Waminal, Yin Huang, Caihua Dong, Li Wang, Jingping Li, Zhiyong Hu, Mu Zhuang, Yi Huang, Junyan Huang, Jiaqin Shi, Desheng Mei, Jing Liu, Tae-Ho Lee, Jinpeng Wang, Huizhe Jin, Zaiyun Li, Xun Li, Jiefu Zhang, Lu Xiao, Yongming Zhou, Zhongsong Liu, Xuequn Liu, Rui Qin, Xu Tang, Wenbin Liu, Yupeng Wang, Yangyong Zhang, Jonghoon Lee, Hyun Hee Kim, France Denoeud, Xun Xu, Xinming Liang, Wei Hua, Xiaowu Wang, Jun Wang, Boulos Chalhoub, and Andrew H Paterson. The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. 5:3930 EP –, 2014.

[23] Walter M. Fitch. Homology. *Trends in Genetics*, 16(5):227–231, 2000.

[24] S Nicklen F Sanger and A R Coulson. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA*, 74(12):5463–5467, 1977.

[25] Eric S Lander et al. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[26] J. Craig Venter, Mark D. Adams, Eugene W. Myers, Peter W. Li, Richard J. Mural, Granger G. Sutton, Hamilton O. Smith, Mark Yandell, Cheryl A. Evans, Robert A. Holt, Jeannine D. Gocayne, Peter Amanatides, Richard M. Ballew, Daniel H. Huson, Jennifer Russo Wortman, Qing Zhang, Chinnappa D. Kodira, Xiangqun H. Zheng, Lin Chen, Marian Skupski, Gangadharan Subramanian, Paul D. Thomas, Jinghui Zhang, George L. Gabor Miklos, Catherine Nelson, Samuel Broder, Andrew G. Clark, Joe Nadeau, Victor A. McKusick, Norton Zinder, Arnold J. Levine, Richard J. Roberts, Mel Simon, Carolyn

Slayman, Michael Hunkapiller, Randall Bolanos, Arthur Delcher, Ian Dew, Daniel Fasulo, Michael Flanigan, Liliana Florea, Aaron Halpern, Sridhar Hannenhalli, Saul Kravitz, Samuel Levy, Clark Mobarry, Knut Reinert, Karin Remington, Jane Abu-Threideh, Ellen Beasley, Kendra Biddick, Vivien Bonazzi, Rhonda Brandon, Michele Cargill, Ishwar Chandramouliswaran, Rosane Charlab, Kabir Chaturvedi, Zuoming Deng, Valentina Di Francesco, Patrick Dunn, Karen Eilbeck, Carlos Evangelista, Andrei E. Gabrielian, Weiniu Gan, Wangmao Ge, Fangcheng Gong, Zhiping Gu, Ping Guan, Thomas J. Heiman, Maureen E. Higgins, Rui-Ru Ji, Zhaoxi Ke, Karen A. Ketchum, Zhongwu Lai, Yiding Lei, Zhenya Li, Jiayin Li, Yong Liang, Xiaoying Lin, Fu Lu, Gennady V. Merkulov, Natalia Milshina, Helen M. Moore, Ashwinikumar K Naik, Vaibhav A. Narayan, Beena Neelam, Deborah Nusskern, Douglas B. Rusch, Steven Salzberg, Wei Shao, Bixiong Shue, Jingtao Sun, Zhen Yuan Wang, Aihui Wang, Xin Wang, Jian Wang, Ming-Hui Wei, Ron Wides, Chunlin Xiao, Chunhua Yan, Alison Yao, Jane Ye, Ming Zhan, Weiqing Zhang, Hongyu Zhang, Qi Zhao, Liansheng Zheng, Fei Zhong, Wenyan Zhong, Shiaoping C. Zhu, Shaying Zhao, Dennis Gilbert, Suzanna Baumhueter, Gene Spier, Christine Carter, Anibal Cravchik, Trevor Woodage, Feroze Ali, Huijin An, Aderonke Awe, Danita Baldwin, Holly Baden, Mary Barnstead, Ian Barrow, Karen Beeson, Dana Busam, Amy Carver, Angela Center, Ming Lai Cheng, Liz Curry, Steve Danaher, Lionel Davenport, Raymond Desilets, Susanne Dietz, Kristina Dodson, Lisa Doup, Steven Ferriera, Neha Garg, Andres Gluecksmann, Brit Hart, Jason Haynes, Charles Haynes, Cheryl Heiner, Suzanne Hladun, Damon Hostin, Jarrett Houck, Timothy Howland, Chinyere Ibegwam, Jeffery Johnson, Francis Kalush, Lesley Kline, Shashi Koduru, Amy Love, Felecia Mann, David May, Steven McCawley, Tina McIntosh, Ivy McMullen, Mee Moy, Linda Moy, Brian Murphy, Keith Nelson, Cynthia Pfannkoch, Eric Pratts, Vinita Puri, Hina Qureshi, Matthew Reardon, Robert Rodriguez, Yu-Hui Rogers, Deanna Romblad, Bob Ruhfel, Richard Scott, Cynthia Sitter, Michelle Smallwood, Erin Stewart, Renee Strong, Ellen Suh, Reginald Thomas, Ni Ni Tint, Sukyee Tse, Claire Vech, Gary Wang, Jeremy Wetter, Sherita Williams, Monica Williams, Sandra Windsor, Emily Winn-Deen, Keriellen Wolfe, Jayshree Zaveri, Karena Zaveri, Josep F. Abril, Roderic Guigó, Michael J. Campbell, Kimmen V. Sjolander, Brian Karlak, Anish Kejariwal, Huaiyu Mi, Betty Lazareva, Thomas Hatton, Apurva Narechania, Karen Diemer, Anushya Muruganujan, Nan Guo, Shinji Sato, Vineet Bafna, Sorin Istrail, Ross Lippert, Russell Schwartz, Brian Walenz, Shibu Yooseph, David Allen, Anand Basu, James Baxendale, Louis Blick, Marcelo Caminha, John Carnes-Stine, Parris Caulk, Yen-Hui Chiang, My Coyne, Carl Dahlke, Anne Deslattes Mays, Maria Dombroski, Michael Donnelly, Dale Ely, Shiva Esparham, Carl Fosler, Harold Gire, Stephen Glanowski, Kenneth Glasser, Anna Glodek, Mark Gorokhov, Ken Graham, Barry Gropman, Michael Harris, Jeremy Heil, Scott Henderson, Jeffrey Hoover, Donald Jennings, Catherine Jordan, James Jordan, John Kasha, Leonid Kagan, Cheryl Kraft, Alexander Levitsky, Mark Lewis, Xiangjun Liu, John Lopez, Daniel Ma, William Majoros, Joe McDaniel, Sean Murphy, Matthew Newman, Trung Nguyen, Ngoc Nguyen, Marc Nodell, Sue Pan, Jim Peck, Marshall Peterson, William Rowe, Robert Sanders, John Scott, Michael Simpson, Thomas Smith, Arlan Sprague, Timothy Stockwell, Russell Turner, Eli Venter, Mei Wang, Meiyuan Wen, David Wu, Mitchell Wu, Ashley Xia, Ali Zandieh, and Xiaohong Zhu. The sequence of the human genome. *Science*, 291(5507):1304, 2001.

[27] Michael L Metzker. Sequencing technologies — the next generation. *Nature Rev*, 11:31–46, 2010.

[28] David R. Bentley, Shankar Balasubramanian, Harold P. Swerdlow, Geoffrey P. Smith, John Milton, Clive G. Brown, Kevin P. Hall, Dirk J. Evers, Colin L. Barnes, Helen R. Bignell, Jonathan M. Boutell, Jason Bryant, Richard J. Carter, R. Keira Cheetham, Anthony J. Cox, Darren J. Ellis, Michael R. Flatbush, Niall A. Gormley, Sean J. Humphray, Leslie J. Irving, Mirian S. Karbelashvili, Scott M. Kirk, Heng Li, Xiaohai Liu, Klaus S. Maisinger, Lisa J. Murray, Bojan Obradovic, Tobias Ost, Michael L. Parkinson, Mark R. Pratt, Isabelle M. J. Rasolonjatovo, Mark T. Reed, Roberto Rigatti, Chiara Rodighiero, Mark T. Ross, Andrea Sabot, Subramanian V. Sankar, Aylwyn Scally, Gary P. Schroth, Mark E. Smith, Vincent P. Smith, Anastassia Spiridou, Peta E. Torrance, Svilen S. Tzonev, Eric H. Vermaas, Klaudia Walter, Xiaolin Wu, Lu Zhang, Mohammed D. Alam, Carole Anastasi, Ify C. Aniebo, David M. D. Bailey, Iain R. Bancarz, Saibal Banerjee, Selena G. Barbour, Primo A. Baybayan, Vincent A. Benoit, Kevin F. Benson, Claire Bevis, Phillip J. Black, Asha Boodhun, Joe S. Brennan, John A. Bridgham, Rob C. Brown, Andrew A. Brown, Dale H. Buermann, Abass A. Bundu, James C. Burrows, Nigel P. Carter, Nestor Castillo, Maria Chiara E. Catenazzi, Simon Chang, R. Neil Cooley, Natasha R. Crake, Olubunmi O. Dada, Konstantinos D. Diakoumakos, Belen Dominguez-Fernandez, David J. Earnshaw,

Ugonna C. Egbujor, David W. Elmore, Sergey S. Etchin, Mark R. Ewan, Milan Fedurco, Louise J. Fraser, Karin V. Fuentes Fajardo, W. Scott Furey, David George, Kimberley J. Gietzen, Colin P. Goddard, George S. Golda, Philip A. Granieri, David E. Green, David L. Gustafson, Nancy F. Hansen, Kevin Harnish, Christian D. Haudenschild, Narinder I. Heyer, Matthew M. Hims, Johnny T. Ho, Adrian M. Horgan, Katya Hoschler, Steve Hurwitz, Denis V. Ivanov, Maria Q. Johnson, Terena James, T. A. Huw Jones, Gyoung-Dong Kang, Tzvetana H. Kerelska, Alan D. Kersey, Irina Khrebtukova, Alex P. Kindwall, Zoya Kingsbury, Paula I. Kokko-Gonzales, Anil Kumar, Marc A. Laurent, Cynthia T. Lawley, Sarah E. Lee, Xavier Lee, Arnold K. Liao, Jennifer A. Loch, Mitch Lok, Shujun Luo, Radhika M. Mammen, John W. Martin, Patrick G. McCauley, Paul McNitt, Parul Mehta, Keith W. Moon, Joe W. Mullens, Taksina Newington, Zemin Ning, Bee Ling Ng, Sonia M. Novo, Michael J. O/'Neill, Mark A. Osborne, Andrew Osnowski, Omead Ostadan, Lambros L. Paraschos, Lea Pickering, Andrew C. Pike, Alger C. Pike, D. Chris Pinkard, Daniel P. Pliskin, Joe Podhasky, Victor J. Quijano, Come Raczy, Vicki H. Rae, Stephen R. Rawlings, Ana Chiva Rodriguez, Phyllida M. Roe, John Rogers, Maria C. Rogert Bacigalupo, Nikolai Romanov, Anthony Romieu, Rithy K. Roth, Natalie J. Rourke, Silke T. Ruediger, Eli Rusman, Raquel M. Sanches-Kuiper, Martin R. Schenker, Josefina M. Seoane, Richard J. Shaw, Mitch K. Shiver, Steven W. Short, Ning L. Sizto, Johannes P. Sluis, Melanie A. Smith, Jean Ernest Sohna Sohna, Eric J. Spence, Kim Stevens, Neil Sutton, Lukasz Szajkowski, Carolyn L. Tregidgo, Gerardo Turcatti, Stephanie vandeVondele, Yuli Verhovsky, Selene M. Virk, Suzanne Wakelin, Gregory C. Walcott, Jingwen Wang, Graham J. Worsley, Juying Yan, Ling Yau, Mike Zuerlein, Jane Rogers, James C. Mullikin, Matthew E. Hurles, Nick J. McCooke, John S. West, Frank L. Oaks, Peter L. Lundberg, David Klenerman, Richard Durbin, and Anthony J. Smith. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, 2008.

[29] Lin Liu, Yinhu Li, Siliang Li, Ni Hu, Yimin He, Ray Pong, Danni Lin, Lihua Lu, and Maggie Law. Comparison of next-generation sequencing systems. *J Biomed Biotechnol*, 2012(251364):11 pages, 2012.

[30] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.

[31] Gregory Hannon et al. FASTX-Toolkit: FASTQ/A short-reads pre-processing tools, 2009.

[32] Illumina, 9885 Towne Centre Drive, San Diego, CA. *Data Sheet: Sequencing*, Jan 2009.

[33] H. P. J. Buermans and J. T. den Dunnen. Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1842(10):1932–1941, 2014.

[34] Travis C. Glenn. Field guide to next-generation DNA sequencers. *Mol Ecol Resour*, 11(5):759–769, 2011.

[35] John Eid, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, David Rank, Primo Baybayan, Brad Bettman, Arkadiusz Bibillo, Keith Bjornson, Bidhan Chaudhuri, Frederick Christians, Ronald Cicero, Sonya Clark, Ravindra Dalal, Alex Dewinter, John Dixon, Mathieu Foquet, Alfred Gaertner, Paul Hardenbol, Cheryl Heiner, Kevin Hester, David Holden, Gregory Kearns, Xiangxu Kong, Ronald Kuse, Yves Lacroix, Steven Lin, Paul Lundquist, Congcong Ma, Patrick Marks, Mark Maxham, Devon Murphy, Insil Park, Thang Pham, Michael Phillips, Joy Roy, Robert Sebra, Gene Shen, Jon Sorenson, Austin Tomaney, Kevin Travers, Mark Trulson, John Vieceli, Jeffrey Wegener, Dawn Wu, Alicia Yang, Denis Zaccarin, Peter Zhao, Frank Zhong, Jonas Korlach, and Stephen Turner. Real-time DNA sequencing from single polymerase molecules. *Science*, 323(5910):133–138, 2009.

[36] Miten Jain, High E Olsen, Benedict Paten, and Mark Akeson. The Oxford nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol*, 17(239), 2016.

[37] Granger G Sutton, Owen White, Mark D Adams, and Anthony R Kerlavage. TIGR assembler: A new tool for assembling large shotgun sequencing projects. *Genome Science and Technology*, 1(1):9–19, 1995.

[38] R D Fleischmann, M D Adams, O White, R A Clayton, E F Kirkness, A R Kerlavage, C J Bult, J F Tomb, B A Dougherty, and J M Merrick. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, 269(5223):496–512, 1995.

[39] C M Fraser, J D Gocayne, O White, M D Adams, R A Clayton, R D Fleischmann, C J Bult, A R Kerlavage, G Sutton, J M Kelley, R D Fritchman, J F Weidman, K V Small, M Sandusky, J Fuhrmann, D Nguyen, T R Utterback, D M Saudek, C A Phillips, J M Merrick, J F Tomb, B A Dougherty, K F Bott, P C Hu, T S Lucier, S N Peterson, H O Smith, C A 3rd Hutchison, and J C Venter. The minimal gene complement of *Mycoplasma genitalium*. *Science*, 270(5235):397–403, 1995.

[40] Philip E C Compeau, Pavel A Pevzner, and Glenn Tesler. How to apply de Bruijn graphs to genome assembly. *Nat. Biotech.*, 29(11):987–991, 2011.

[41] Veli Mäkinen, Leena Salmela, and Johannes Ylinen. Normalized N50 assembly metric using gap-restricted co-linear chaining. *BMC Bioinformatics*, 13(1):255, 2012.

[42] Dent Earl, Keith Bradnam, John St. John, Aaron Darling, Dawei Lin, Joseph Fass, Hung On Ken Yu, Vince Buffalo, Daniel R. Zerbino, Mark Diekhans, Ngan Nguyen, Pramila Nuwantha Ariyaratne, Wing-Kin Sung, Zemin Ning, Matthias Haimel, Jared T. Simpson, Nuno A. Fonseca, İnanç Birol, T. Roderick Docking, Isaac Y. Ho, Daniel S. Rokhsar, Rayan Chikhi, Dominique Lavenier, Guillaume Chapuis, Delphine Naquin, Nicolas Maillet, Michael C. Schatz, David R. Kelley, Adam M. Phillippy, Sergey Koren, Shiaw-Pyng Yang, Wei Wu, Wen-Chi Chou, Anuj Srivastava, Timothy I. Shaw, J. Graham Ruby, Peter Skewes-Cox, Miguel Betegon, Michelle T. Dimon, Victor Solovyev, Igor Seledtsov, Petr Kosarev, Denis Vorobyev, Ricardo Ramirez-Gonzalez, Richard Leggett, Dan MacLean, Fangfang Xia, Ruibang Luo, Zhenyu Li, Yinlong Xie, Binghang Liu, Sante Gnerre, Iain MacCallum, Dariusz Przybylski, Filipe J. Ribeiro, Shuangye Yin, Ted Sharpe, Giles Hall, Paul J. Kersey, Richard Durbin, Shaun D. Jackman, Jarrod A. Chapman, Xiaoqiu Huang, Joseph L. DeRisi, Mario Caccamo, Yingrui Li, David B. Jaffe, Richard E. Green, David Haussler, Ian Korf, and Benedict Paten. Assemblathon 1: A competitive assessment of de novo short read assembly methods. *Genome Research*, 21(12):2224–2241, 2011.

[43] Elin Videvall. N50 for transcriptome assemblies.

[44] Richard Smith-Unna, Chris Bournsnell, Rob Patro, Julian M. Hibberd, and Steven Kelly. TransRate: reference-free quality assessment of denovo transcriptome assemblies. *Genome Res*, 26(8):1134–1144, 2016.

[45] W J Kent. BLAT- the BLAST-like alignment tool. *Genome Res.*, 12(4):656–664, 2002.

[46] Bo Li, Nathanael Fillmore, Yongsheng Bai, Mike Collins, James A. Thomson, Ron Stewart, and Colin N. Dewey. Evaluation of de novo transcriptome assemblies from RNA-Seq data. *Genome Biology*, 15(12):553, 2014.

[47] Bo Li and Colin Dewey. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMS Bioinformatics*, 12(323), 2011.

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

[49] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[50] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009.

[51] Ruiqiang Li, Hongmei Zhu, Jue Ruan, Wubin Qian, Xiaodong Fang, Zhongbin Shi, Yingrui Li, Shengting Li, Gao Shan, Karsten Kristiansen, Songgang Li, Huanming Yang, Jian Wang, and Jun Wang. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research*, 2009.

[52] Brandi L Cantarel, Ian Korf, Sofia M C Robb, Genis Parra, Eric Ross, Barry Moore, Carson Holt, Alejandro Sanchez Alvarado, and Mark Yandell. Maker: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res*, 18(1):188–196, 2008.

[53] Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn, and Lior Pachter. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.*, 7:562–578, 2012.

[54] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nat Meth*, 9(4):357–359, 2012.

[55] Joseph Fass. count_fasta.pl, November 2010.

[56] S. Andrews. FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc, 2010.

[57] Peer Bork, Thomas Dandekar, Yolande Diaz-Lazcoz, Frank Eisenhaber, Martijn Huynen, and Yanping Yuan. Predicting function: from genes to genomes and back11edited by p. e. wright. *Journal of Molecular Biology*, 283(4):707–725, 1998.

[58] Ryan R. Wick, Mark B. Schultz, Justin Zobel, and Kathryn E. Holt. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics*, 31(20):3350–3352, 2015.

[59] Daniel R. Zerbino and Ewan Birney. Velvet: algorithms for *de novo* short read assembly using de bruijn graphs. *Genome Research*, 18(5):821–829, 2008.

[60] Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey A Gurevich, Mikhail Dvorkin, Alexander S Kulikov, Valery M Lesin, Sergey I Nikolenko, Son Pham, Andrey D Prjibelski, Alexey V Pyshkin, Alexander V Sirotkin, Nikolay Vyahhi, Glenn Tesler, Max A Alekseyev, and Pavel A Pevzner. Spades: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, 19(5):455–477, 2012.

[61] Cydney B Nielsen, Shaun D Jackman, Inanc Birol, and Steven J M Jones. Abyss-explorer: visualizing genome sequence assemblies. *IEEE Trans Vis Comput Graph*, 15(6):881–888, 2009.

[62] Emden R. Gansner and Stephen C. North. An open graph visualization system and its applications to software engineering. *Softw Pract Exp*, 30(11):1203–1233, 2000.

[63] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D3: Data-driven documents. *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)*, 2011.

[64] Michael Leibman. SlickGrid. github, July 2010.

[65] John G. Carvajal, Leyta Castro, and Sebastian Wilzbach. biojs-vis-sequence. github, 2014.

[66] Isobel A P Parkin, Wayne E Clarke, Christine Sidebottom, Wentao Zhang, Stephen J Robinson, Matthew G Links, Steve Karcz, Erin E Higgins, Pierre Fobert, and Andrew G Sharpe. Towards unambiguous transcript mapping in the allotetraploid brassica napus. *Genome*, 53(11):929–938, 2010.

[67] Satshil B. Rana, Frank J. Zadlock, IV, Ziping Zhang, Wyatt R. Murphy, and Carolyn S. Bentivegna. Comparison of de novo transcriptome assemblers and k-mer strategies using the killifish, fundulus heteroclitus. *PLOS ONE*, 11(4):1–16, 2016.

[68] Dilip A Durai and Marcel H Schulz. Informed kmer selection for de novo transcriptome assembly. *Bioinformatics*, 32(11):1670–1677, 2016.

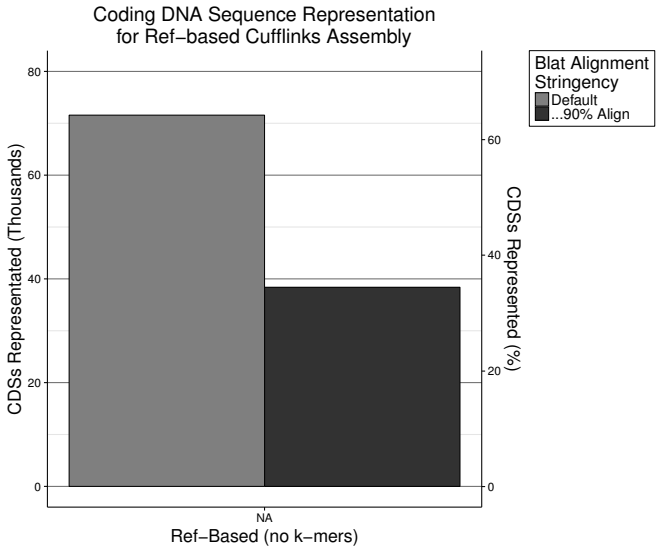
[69] Berat Z Haznedaroglu, Darryl Reeves, Hamid Rismani-Yazdi, and Jordan Peccia. Optimization of de novo transcriptome assembly from high-throughput short read sequencing data improves functional annotation for non-model organisms. *BMC Bioinformatics*, 13:170, Jul 2012.

[70] Justin T Page, Alan R Gingle, and Joshua A Udall. PolyCat: a resource for genome categorization of sequencing reads from allopolyploid organisms. *G3 (Bethesda)*, 3(3):517–525, 2013.

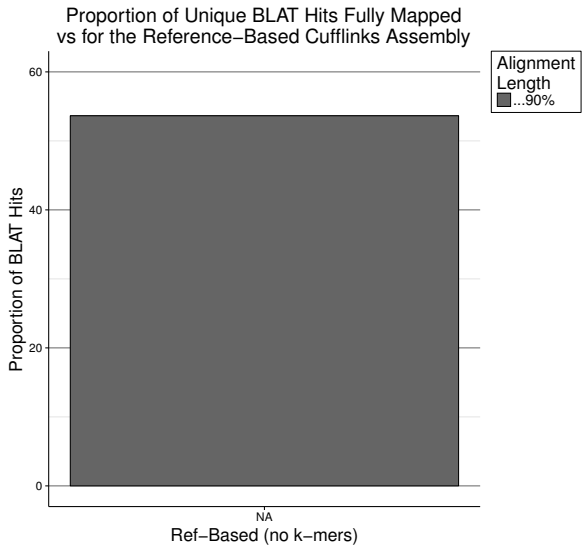
Appendix A

CDS Representation of the Reference-Based Cufflinks Assembly

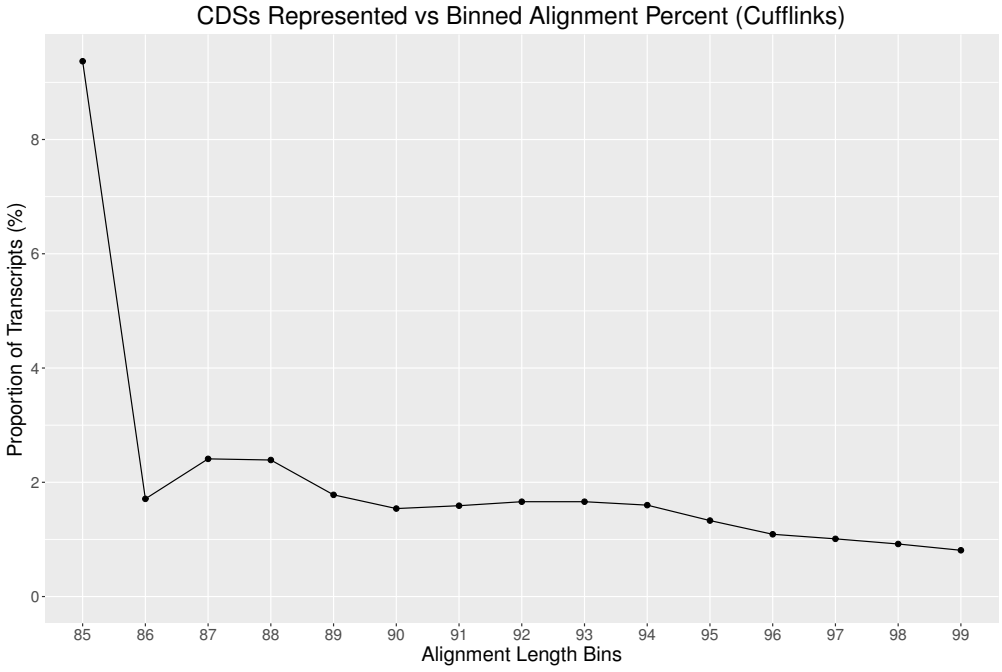
Because only one Cufflinks assembly was produced, figures displaying the CDS representation statistics for the single assembly were not shown in the Results chapter. The plots corresponding to those shown for the three *de novo* assemblers are presented here for continuity.



(a) CDS Representation



(b) Propotion of Assembly Full Length



(c) Binned Alignment Length

Figure A.1: Representation of coding DNA sequences for the Cufflinks reference-based assembly. (a) Quantity and percentage of the total CDSs represented by assembled transcripts with two BLAT stringencies: the default parameter set (gray) and full length BLAT alignments where the contig covered at least 90% of the CDS bases with matches (black). (b) CDSs represented at full length as a proportion of the CDSs represented in total by each assembly. (c) Proportion of the assembled transcripts aligning to CDSs at sequence identities in 1% bins.

Appendix B

Interpretation of Illumina RNA-seq Data Quality

The Illumina platform used has a known trend of per-base quality where a drop in phred score is observed for 100 bp reads near the 40-50 bp range. This was considered as a possible explanation for assembly quality drops for *de novo* de Bruijn graph assemblies using k-mer lengths of 40-50 bps. It is possibility that these k-mers originating from either end of a read would have a single erroneous base at either end, which could potentially result in an extension from another k-mer from a similar read originating from a different gene locus. Although each read would still contain the path of k-mers from end to end, these erroneous k-mers at the center could lead to some extra branching in the de Bruijn graph that is not biologically true. Additionally, using k-mers shorter than this length would also result in k-mers with the potentially error prone 50th bp at each end. The only difference is that longer k-mers would have the problematic 50th base pair, as well as the potentially error prone ends of the sequences, resulting in both ends of the k-mer being more error prone than shorter k-mer lengths.

B.1 RNA-seq Data per-base Read Quality

The RNA-seq dataset used in this thesis was examined for per-base read quality using the FastQC tool [56]. A drop in quality was observed around the 50 bp region, resulting in a minimum phred score of 30 corresponding to an error rate of 0.01% (99.9% base accuracy). It is unclear if such a drop is enough to result in sufficient errors in de Bruijn graph traversal to explain assembly quality drops for assemblies using k-mers near 50 bps. k-mer lengths greater than 50 bps would go beyond the 50th base pair on both sides, which may explain the sudden increase in assembly quality observed for the 51 bp k-mer SOAPdenovo-Trans assembly over the 41 bp assembly.

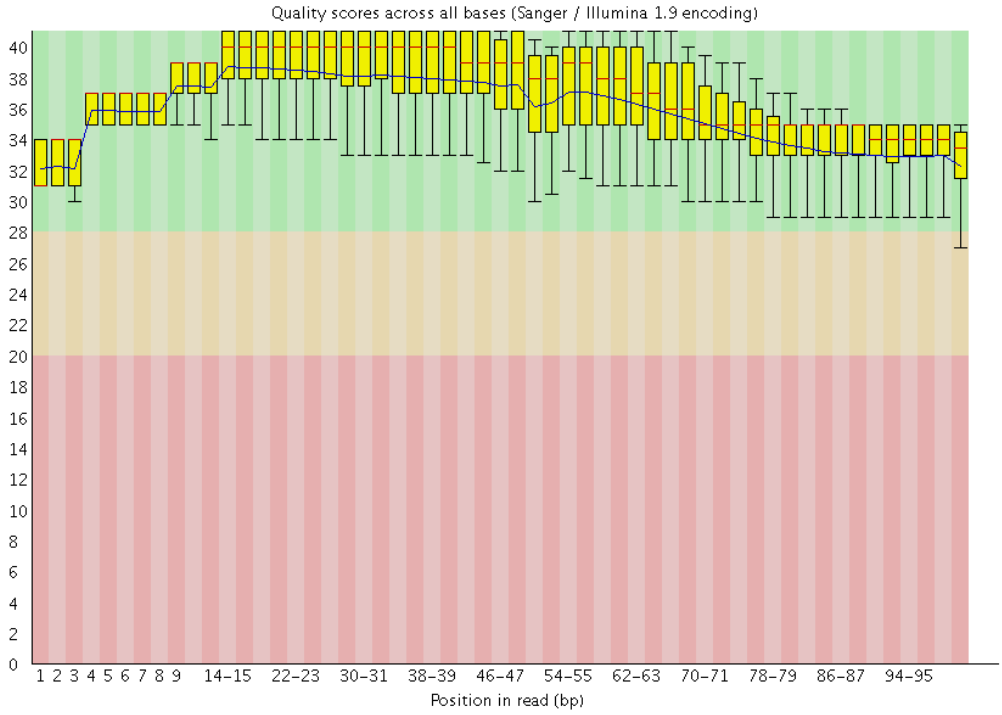


Figure B.1: Per-base quality scores of the Illumina RNA-seq data used for assembly (R1, forward reads).

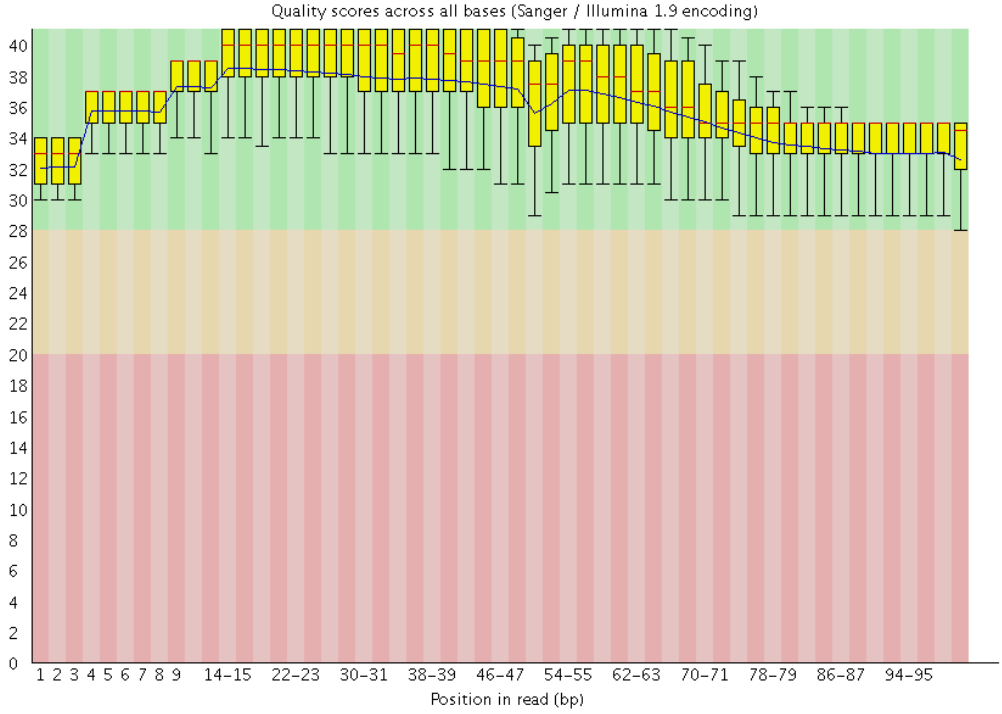


Figure B.2: Per-base quality scores of the Illumina RNA-seq data used for assembly (R2, reverse reads).

Appendix C

k-mer Repetition Analysis

We examined the repetition of k-mers of various lengths corresponding to the k-mer lengths used during *de novo* transcriptome assembly. In particular, we hoped to find an explanation for the drop in quality observed for assemblies using k-mer lengths between 31 and 41 bps, corresponding to the merged 31-41 bp Oases assembly which displayed abnormally poor assembly quality relative to the other two assemblies. One hypothesis for this poor quality was the existence of repetitive sequences within the read data corresponding to the k-mer lengths used for assembly, resulting in an increase in ambiguity for de Bruijn graphs.

C.1 k-mer Repetition Histograms for Odd k-mer Lengths of 21 bp up to 51 bp.

Unique k-mer occurrences were counted by the Jellyfish [48] software. The output is a set of ordered pairs of the repetition counts from 1 to 10,000 and 10,00+ repeats, as well as the number of unique k-mers that repeat the given number of times in the input read data. This data was plotted in R [49] using the ggplot2 [50] package (Figures C.1, C.2, C.3 and C.4)

C.2 Sums of Unique k-mers for Repeat Thresholds

The sum of the unique k-mers repeating for two thresholds, > 5 repeats (Figure C.5) and $> 5,000$ repeats (Figure C.6) These graphs show the distribution of the number of k-mers that repeat more than a given threshold of times. We expect to see fewer k-mers with large number of repeats as k increases, as probability dictates longer k-mers are less likely to be repeated. None of the k-mer lengths exhibited an obvious surfeit of repetitive k-mers in comparison that would indicate a systematic problem at that k-mer length for de Bruijn graph assembly.

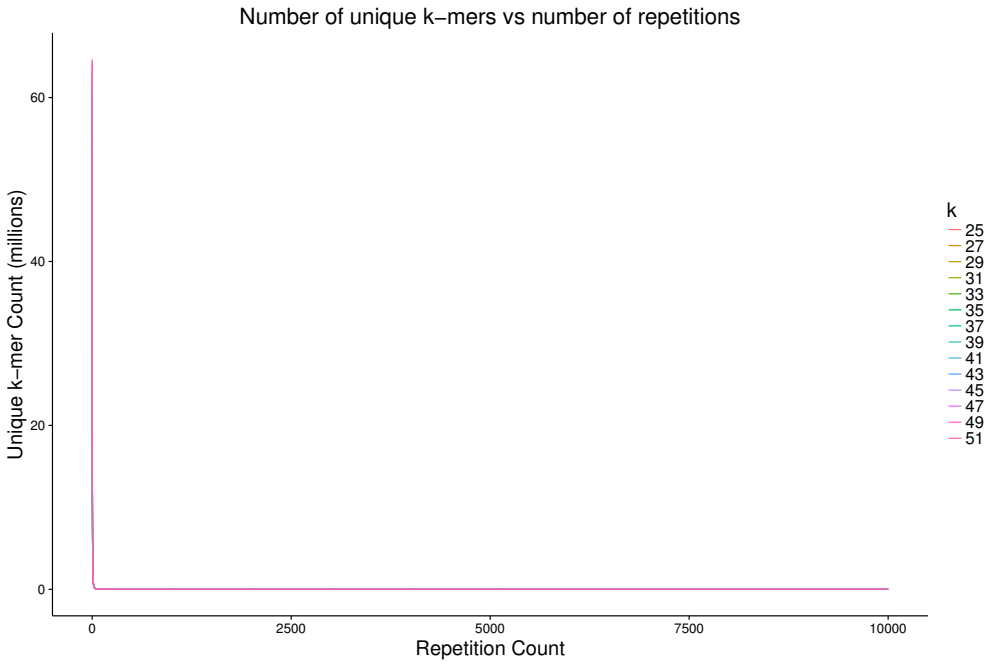


Figure C.1: Repetition counts of unique k-mers for odd k values from 21 bp up to 51 bp. Relatively long k-mer lengths result in a large quantity of possible unique k-mers (unique k-mers = 4^k) which causes the extreme graph shape.

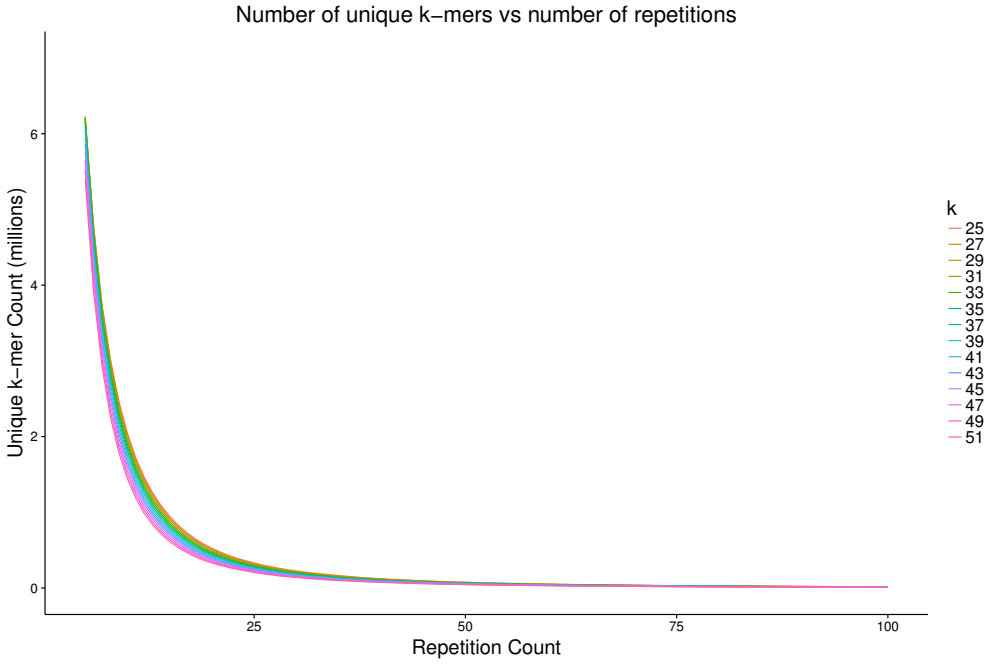


Figure C.2: Repetition counts of unique k-mers for odd k values from 21 bp up to 51 bp between 5 and 100 repeats. Limiting the graph to relatively low repetition k-mers (less than 100 repeats) shows some stratification of the various k-mers.

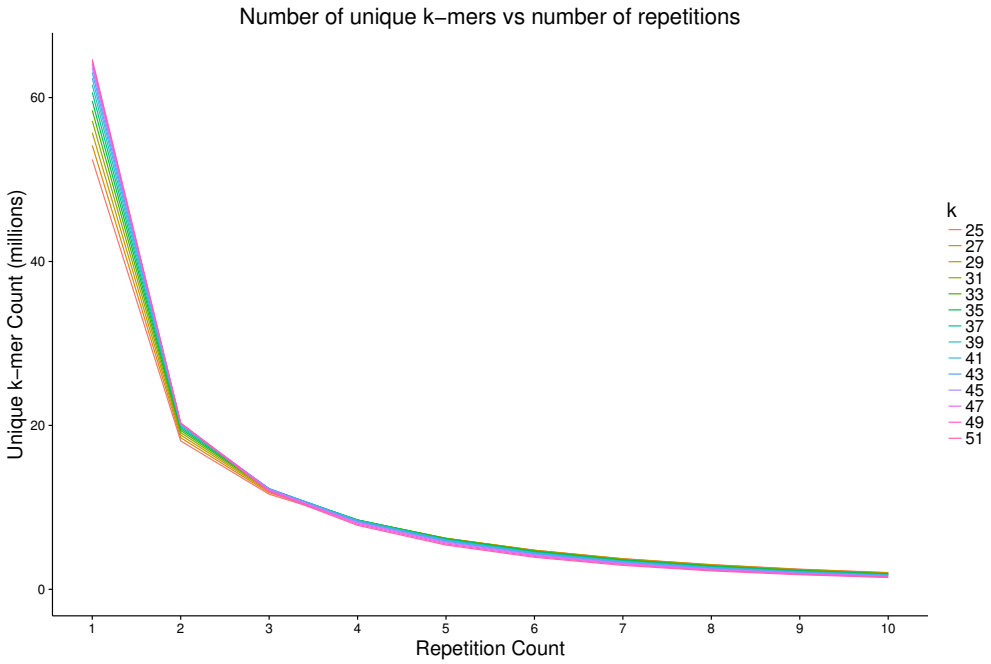


Figure C.3: Repetition counts of unique k-mers for odd k values from 21 bp up to 51 bp between 1 and 10 repeats. Further limiting to less than 10 repeats shows the inversion of the trends. Longer k-mers have a larger quantity of unique and low repeat k-mers (1, 2 and 3 repeats) while the trend inverts around the 4 repeat mark.

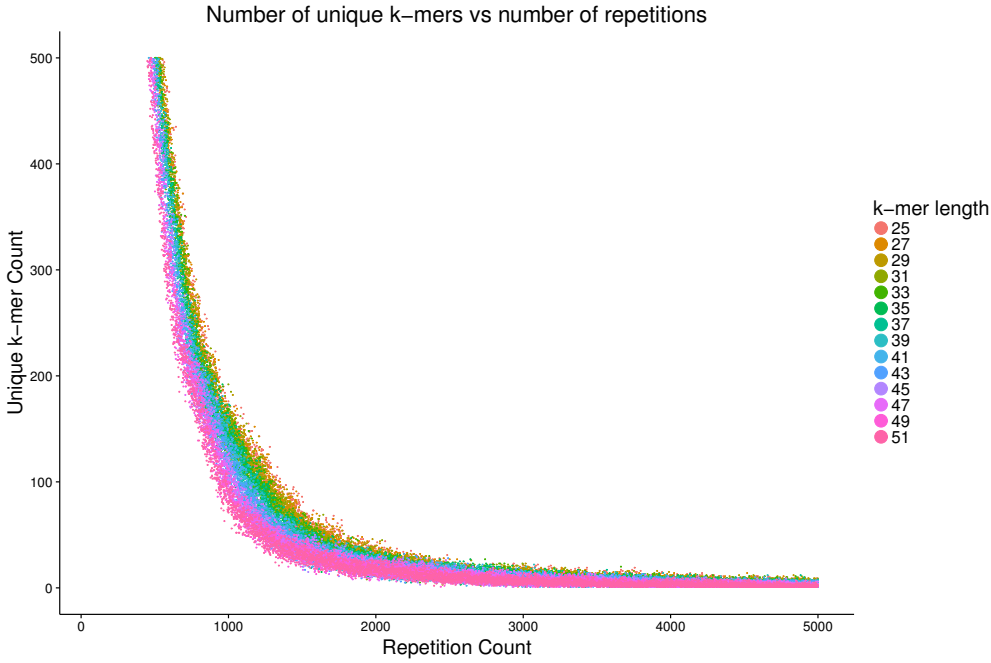


Figure C.4: Repetition counts of unique k-mers for odd k values from 21 bp up to 51 bp between 500 and 5,000 repeats. Points are used here instead of lines due to the more variable y values.

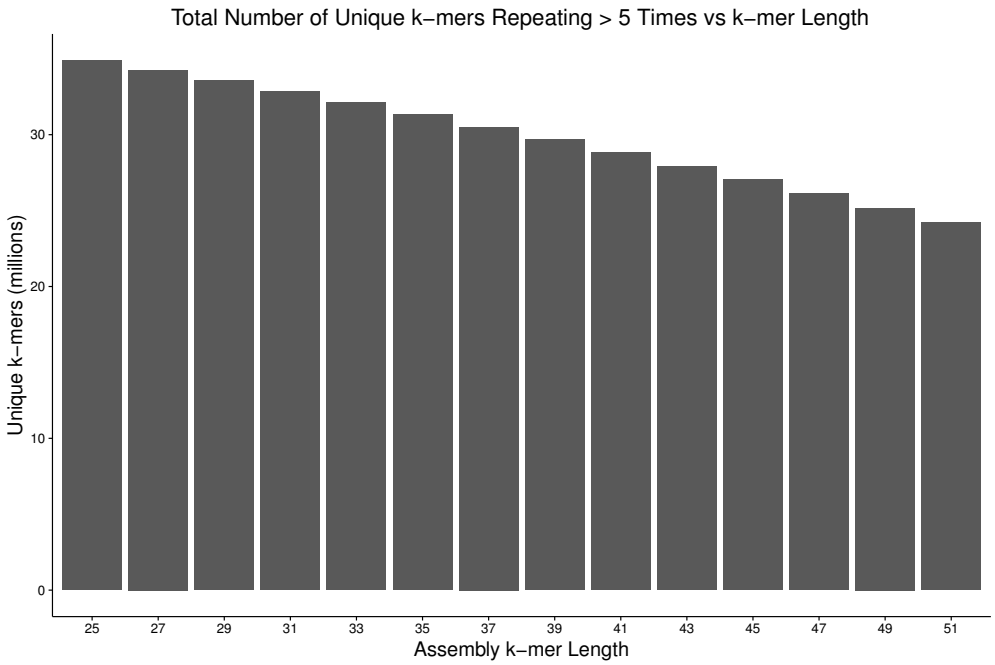


Figure C.5: Sum of unique k-mers with repetition counts greater than 5 for odd k values from 21 bp up to 51 bp.

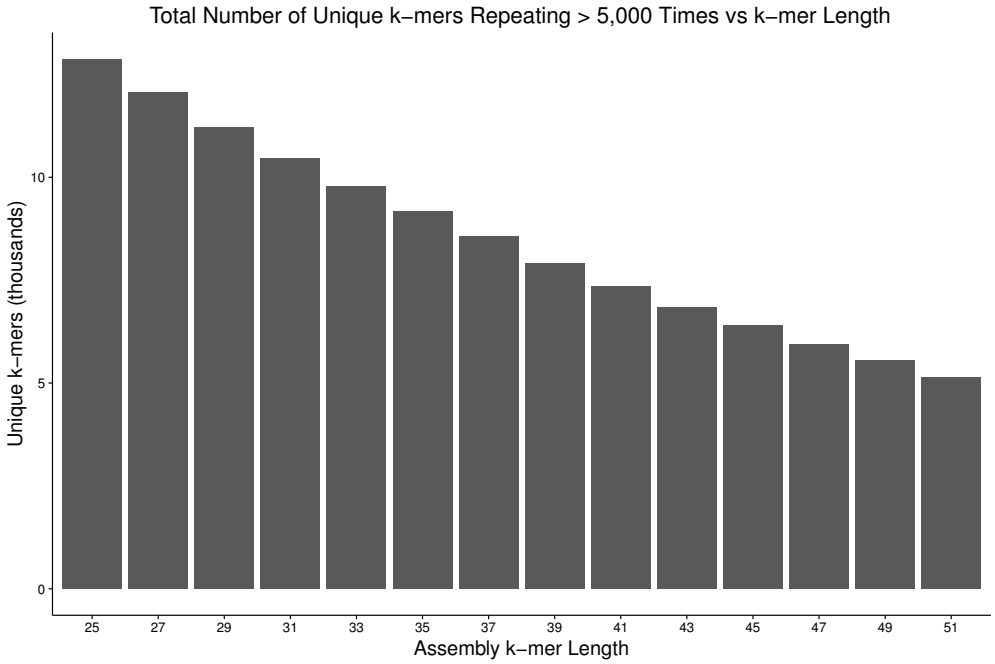


Figure C.6: Sum of unique k-mers with repetition counts greater than 5,000 for odd k values from 21 bp up to 51 bp.