DEVELOPMENT OF COMPUTATIONAL TOOLS FOR AFRICAN OIL PALM GENOME AND GENE EXPRESSION ANALYSES

JOEL LOW ZI-BIN

FACULTY OF SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR

2019

DEVELOPMENT OF COMPUTATIONAL TOOLS FOR AFRICAN OIL PALM GENOME AND GENE EXPRESSION ANALYSES

JOEL LOW ZI-BIN

THESIS SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

INSTITUTE OF BIOLOGICAL SCIENCES
FACULTY OF SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR

2019

UNIVERSITY OF MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: **JOEL LOW ZI-BIN** [redacted]

Registration/Matric No.: **SHC130095**

Name of Degree: **DOCTOR OF PHILOSOPHY**

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

DEVELOPMENT OF COMPUTATIONAL TOOLS FOR AFRICAN OIL PALM GENOME AND GENE EXPRESSION ANALYSES

Field of Study: **BIOINFORMATICS**

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date:

Subscribed and solemnly declared before,

Witness's Signature Date:

Name:
Designation:

DEVELOPMENT OF COMPUTATIONAL TOOLS FOR AFRICAN OIL PALM GENOME AND GENE EXPRESSION ANALYSES

ABSTRACT

Continuing improvements in the yield of oil palm requires knowledge of the genes and mechanisms that regulate oil accumulation. This is believed attainable with the sequencing of the oil palm genome. However, Sime Darby's oil palm genome assembly is far from complete. Here I look at ways to improve the genome assembly quality by means of computational tools developed to build exome contigs (GenSeed Pipeline Suite), detection of potential regions of misassembly due to repeats (BridgeReader), and the use of molecular markers as a means to arrange scaffolds into a physical map (MarkMyMap). I show that with the use of these tools, the most recent assembly version, OPg3, had improved over the first version in capturing the gene space (39% more mappable transcripts) and molecular marker representation (3% increase in mappable SSRs and DArTs). Furthermore, the constructed physical map representing the oil palm's 16 chromosomes had improved genome coverage from Sime Darby's previous version by 79%. The improvements to the genome draft in this work will assist future Genome-Wide Association Studies and functional studies of genes. Besides that, I have developed a fast Bayesian method to overcome analytical bottlenecks in RNA-Seq experiments with limited number of replicates and low sequencing coverage, such as those found for oil palm studies. I incorporated a previously unused sequencing coverage parameter determined from the concentration of an RNA sample into a procedure to make differentially expressed gene calls. This method had better or comparable performance with NOISeq and GFOLD, according to the results from simulations and experiments with real unreplicated data. The method is called CORNAS

(Coverage-dependent RNA-Seq), and I show that robust differentially expressed gene calls can be made in an RNA-Seq study of oil palm inflorescences using CORNAS.

Keywords: Oil Palm, Genome, Transcriptome, Bioinformatics.

PEMAJUAN PERALATAN PENGKOMPUTERAN UNTUK ANALISIS GENOM DAN EKSPRESI GEN KELAPA SAWIT AFRIKA

ABSTRAK

Peningkatan berterusan dalam hasil kelapa sawit memerlukan pengetahuan tentang gen dan mekanisme yang mengawal pengumpulan minyak dalam buah kelapa sawit. Hal ini dipercayai mampu dicapai oleh penjujukan genom kelapa sawit. Namun, penyusunan genom sawit Sime Darby masih tidak lengkap. Di sini saya mencari cara-cara bagi meningkatkan kualiti penyusunan genom dengan memajukan kaedah-kaedah pengkomputeran untuk membina kontig exom (*GenSeed Pipeline Suite*), pengesanan rantau yang mempunyai masalah penyusunan akibat jujukan ulangan (*BridgeReader*), dan penggunaan penanda molekul sebagai cara untuk mengatur perancah ke dalam peta fizikal (*MarkMyMap*). Saya menunjukkan bahawa dengan menggunakan kaedah-kaedah ini, versi penyusunan terbaru, OPg3, telah bertambahbaik berbanding dengan versi pertama dari segi liputan ruang gen (39% lebih banyak transkrip yang dapat dipetakan) dan perwakilan penanda molekul (peningkatan sebanyak 3% SSRs dan DArT yang dapat dipetakan). Selain itu, peta fizikal yang dibina untuk mewakili 16 kromosom kelapa sawit telah meningkatkan liputan genom sebanyak 79% berbanding dengan versi Sime Darby sebelumnya. Penambahbaikan draf genom dalam penyelidikan ini akan membantu pengajian *genome-wide association* masa depan dan kajian fungsi-fungsi gen. Di samping itu, saya telah mencipta kaedah Bayesian yang cepat untuk mengatasi sekatan analisis akibat bilangan replikasi dan liputan jujukan yang rendah dalam eksperimen RNA-Seq, seperti yang didapati untuk kajian kelapa sawit. Saya memasukkan parameter liputan jujukan yang belum pernah digunakan sebelum ini yang ditentukan daripada kepekatan sampel RNA ke dalam prosedur untuk membuat

panggilan gen yang diekspresikan secara berbeza. Kaedah ini mempunyai prestasi yang lebih baik atau setanding berbanding dengan NOISeq dan GFOLD, menurut keputusan dari simulasi dan eksperimen dengan data nyata tanpa replikasi. Kaedah ini dipanggil CORNAS (*Coverage-dependent RNA-Seq*), dan saya menunjukkan bahawa panggilan gen yang diekspresikan secara berbeza yang teguh dapat dibuat dalam kajian RNA-Seq perbungaan kelapa sawit dengan menggunakan CORNAS.

Kata kunci: Kelapa sawit, Genom, Transkriptom, Bioinformatik.

ACKNOWLEDGEMENTS

This body of work would not have been possible without the following people:

My supervisors, Prof. Dr. Jennifer Ann Harikrishna and Dr. Khang Tsung Fei, whose guidance I am most grateful for. I appreciate the time you had dedicated to review my work and provide critical feedback that kept me on the right track.

I am grateful to Dr. Martti Tapani Tammi for the frontiers you have opened up to me, the great work we did together, and for your mentoring. I would like to also thank Ho Hui Li and Ranganath Gudemella, whom I had the pleasure of working closely with.

Thank you to Dr. Harikrishna Kulaveerasingam, Dr. David Ross Appleton, Scott Su and Sime Darby Plantation for seeing my potential in the field of bioinformatics, providing me the right environment to grow as a scientist and the opportunity to work on such a challenging project.

I dedicate this work to my girls: Theresa, Jaynmae and Tesxia, my source of joy and my home. Finally, to God for all His blessings He has shown to make this all possible and the reason to seek the truth in all things: “And ye shall know the truth, and the truth shall make you free” - John 8:32.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

AFLP	:	amplified fragment length polymorphism.
bp	:	base pair.
cDNA	:	complementary DNA.
DArT	:	diversity array technology.
DBG	:	De Bruijn graph.
DEG	:	differentially expressed gene.
DNA	:	deoxyribonucleic acid.
GP	:	generalised Poisson.
GWAS	:	genome-wide association studies.
HMM	:	Hidden Markov Model.
LG	:	linkage group.
LINE	:	long interspersed nuclear element.
LM	:	linkage map.
LTR	:	long terminal repeat.
MPOB	:	Malaysian Palm Oil Board.
mRNA	:	messenger RNA.
nt	:	nucleotide.
OLC	:	Overlap-Layout-Consensus.
PCR	:	polymerase chain reaction.
PM	:	physical map.
qPCR	:	quantitative polymerase chain reaction.
RAM	:	Random Access Memory.
RAPD	:	random amplified polymorphic DNA.
RFLP	:	restriction fragment length polymorphism.

RNA : ribonucleic acid.

RNA-Seq : ribonucleic acid sequencing.

rRNA : ribosomal RNA.

SINE : short interspersed nuclear element.

SNP : single nucleotide polymorphism.

SSR : simple sequence repeat.

LIST OF APPENDICES

CHAPTER 1: INTRODUCTION

1.1 The oil palm

The monocot native to Africa, *Elaeis guineensis*, is one of two species of oil palms that flourish along the tropical belt of the world, the other being *Elaeis oleifera*, found mainly in Central America. It was brought to Malaysia in the 1900s and the oil from its fleshy mesocarp tissue is among the top export commodity of the country. Palm oil is used as an ingredient for cooking, as emulsifier in food products and as a source for biofuel (Corley & Tinker, 2008; Malaysian Palm Oil Council, 2019).

Global demands on food, healthcare and energy will continue to increase as the world's population grows. The requirement of oil for food production and transportation is quickly depleting the world's fossil fuel reserves. One way to meet the increase is to improve the production of biological renewable energy sources, such as oil palm.

The common African oil palm (hereafter referred to as the oil palm) varieties are mainly determined by the difference in kernel shell thickness found in the fruitlets (Hardon et al., 1985; Corley & Tinker, 2008). The *dura* oil palm variety produces fruitlets with a thick kernel shell (2 - 8 mm) and a low mesocarp to fruit ratio (M/F) of 35 - 55%; whereas the *pisifera* oil palm does not have a kernel shell at all in its thick mesocarp fruitlets (95% M/F). Just basing on the high M/F ratio, the *pisifera* variety would be considered highly productive if not for the fact that *pisiferas* are female-sterile with bunches seldom developing to maturity. The hybrid offspring of a *dura* mother and a *pisifera* father is a *tenera* oil palm that would grow to produce fruitlets that have thin kernel shells (0.5 - 4 mm) and a high M/F of 60 - 96%. It is this *tenera* oil palm that is the most widely planted variety in commercial fields.

In comparison to other oil crops, oil palm is already the most efficient oil producer

per unit area, producing 6.3 times more oil-yield than the next highest oil producer, rapeseed (MPOB, 2010). Oil palm contributes the largest portion (30%) of the oils and fats produced worldwide (Chandran, 2010). The perennial oil palm is not only resilient, but also continuously productive over 20 years. In comparison, annual crops, such as soybean, have much shorter lifetime productivity.

While productive, oil palm cultivation requires a lot of space and can only be cultivated in the limited landmass around the equator. Generally, an oil palm grows up to 20 metres high and has a crown area of about three metres in diameter; a size that is considerably larger than other oil-producing plants such as the two-metre tall soybean. Large areas of rainforests are frequently converted into agricultural land for oil palm cultivation in tropical countries. This practice has contributed to the negative perception that oil palm is environmentally destructive (Vijay et al., 2016). If oil yield per hectare can be improved substantially, demand for rainforest land for cultivation purpose may drop, thus mitigating the negative impact to the environment whilst meeting the increasing demand for oil resource in the world.

To improve the yield of oil palm requires integration of phenotypic and genetic knowledge to understand the mechanisms that regulate oil accumulation. Scientists believe this is attainable with the availability of an oil palm genome reference.

1.2 The rise of genomics, transcriptomics and bioinformatics

A genome is the complete set of deoxyribonucleic acid (DNA) sequences of an organism. The term, genome, was initially coined by Hans Winkler in 1920, at a time when its structure was not known. Geneticists of the time only depended on good phenotype and pedigree data to infer trait associations and heredity. These “discrete units of inheritance”, first suggested by Gregor Mendel, were later called genes by Wilhelm Johannsen in 1905. We have since found that phenotypes are the product of gene expression (Kærn et al., 2005;

Harper, 2008). So it became clear that understanding what genes are made of will allow greater resolution in genetics studies.

The discovery of the molecular structure of DNA by Watson and Crick (1953) and the ability to determine the sequence of DNA (Wu, 1972) had flung the doorway to genomics wide open. Mankind can finally "read" genes and reconstruct the instruction manual for the building blocks of life. The instructions are made out of only four letters: A, C, G and T. These letters represent the sugar bases of DNA: adenine (A), cytosine (C), guanine (G) and thymine (T). If a protein is a product of reading a sentence in the manual, then the words are single amino acid molecules that can be represented by 3-letter long codes made out of any of the four possible letters arranged in a specific sequence. A whole book containing hundred thousands of sentences can be quite complex, since the complexity generated by the varied combinations possible in a sequence increases exponentially as a function of sequence length. A DNA sequence that is 100 letters long will already have $4^{100}(1.61 \times 10^{60})$ possible combinations. The prokaryote *Escherichia coli* genome, which contains about five million letters, already codes for 4,288 proteins (Blattner et al., 1997). How can we interpret this biological instruction manual, especially for eukaryotic genomes, which are substantially larger and contain billions of letters, like the oil palm (Lynch & Conery, 2003)?

According to the central dogma of molecular biology, ribonucleic acid (RNA) molecules are the mediators of gene expression. It is this molecule that gets transcribed from DNA sequences in the cell's nucleus, that then travels to the cytoplasm to be part of the protein synthesis process (Simmons et al., 2006). Transcription is performed by RNA polymerases on the template strand of the double-stranded DNA, producing a single-stranded RNA complementary coding transcript. Gene expression for protein-coding genes in eukaryotes is more complex than in prokaryotes, because the transcripts undergo post-transcriptional

modifications before they are translated into proteins. The modifications include 5' capping, 3' cleavage and polyadenylation, and RNA splicing. The 5' and 3' modifications aid transport of the RNA molecule to the cytoplasm, and protect it from degradation by enzymes. RNA splicing allows a single gene to code for a variety of transcript isoforms by removing or retaining portions of coding sequences (exons) or non-coding sequences (introns) before translation (Cech & Rio, 1979; Black, 2003). Consequently, the alternative splicing mechanism enables eukaryotes to have greater protein diversity compared to prokaryotes. Studying the transcriptome, with its spliced variants and RNA species abundances, would further advance our understanding of how genotypes effect phenotypes. Since RNA uses the same nucleobases as DNA, with the only difference being the replacement of T by U (uracil), this makes the molecule amenable to sequencing, thus creating the field of transcriptomics.

Current sequencing technologies now provide biologists with a deluge of data to process and interpret (Schatz & Langmead, 2013). This would have been overwhelming if it were not for the concurrent exponential growth of computer hardware according to Moore's law (Schaller, 1997). Now biologists have an avenue to address complex biological questions with large multivariate datasets, but they need to learn how to manipulate the data with computers (Stevens, 2013). The experiments also increasingly need to incorporate a statistical framework for data analysis and interpretation. In order to perform new statistical methodologies, new efficient software are needed to run computations on the large amount of data, and new infrastructure are required to manage the storage, flow and efficiency of handling the data. The interdisciplinary field of bioinformatics was thus born through this melding of biology, computer science and statistics (Ouzounis & Valencia, 2003).

While it may be tempting to think that we can automate all analysis pipelines on computers, in general, improvisations or even new methods may need to be developed on

a case by case basis, depending on the peculiarities of a particular set of biological data. Thus, there is no one-size-fits-all analysis that will work completely for every organism or experiment. A bioinformatician is needed to identify the right computational methods to address the technical challenges faced in such data-rich analyses. Furthermore, the bioinformatician will need to know how to navigate and program computers to fully deploy these computational methods effectively (Chang, 2015).

1.3 Oil palm whole genome assembly improvement

With new massively parallelised sequencing methods in pyrosequencing (454), sequencing by ligation (ABI SOLiD) and reversible dye-terminators on slides (Illumina) technologies, eukaryotic genome drafts can now be completed in about two years instead of a decade. These methods solved, to some degree, the time and cost problems, but created new challenges with regards to assembling the enormous amounts of data generated. The challenge stems from two major sources: short sequenced fragments, and errors in signal detection and base calling (Kircher & Kelso, 2010).

In the case of Sime Darby's oil palm genome that was completed in 2009 (unpublished data), the 454 FLX sequencing technique provided relatively long and good quality reads ($\approx$ 300 base pair (bp)) compared to Illumina sequencing ($\approx$ 30 bp) at the time. The first draft assembly took 12 months to complete, with about 1.7 Gbp assembled to cover 93.8% of the genome, at 30 times depth of coverage. Even so, Sime Darby's oil palm genome assembly is not finished.

The *de novo* assembly of a genome is akin to an attempt to build a finished product without knowing what it actually looks like. Many of the current sequencing technologies use methods that first select the DNA fragments according to size before sequencing. Therefore the entire length of an organism's DNA is not read in a single run. Even for established long read third generation sequencers such as the PacBio RS II system, the

average read is about 10,000 bp (Buermans & Den Dunnen, 2014) (see Chapter 2 for details), while the smallest bacterial genome currently known is still 160,000 bp long (Nakabachi et al., 2006).

Current sequencing processes require the fragmentation of the genome for sequencers to read in a process called shotgun sequencing (Messing et al., 1981; Bankier, 2001). This process generates random fragments of variable lengths that will contain redundant overlaps from numerous cellular DNA in a tissue sample. A collection of fragments is called a library and it is usually categorised according to its fragment lengths (e.g. 20 Kbp, 3 Kbp or 100 bp libraries). In addition, libraries can also be categorised according to the methods used to sequence the fragments. Taking the Illumina sequencing platform as an example, a genome is typically fragmented into 300 bp long pieces that are inserted between adaptors. The sequencing of the insert from the two ends of the fragments creates a pair of reads. This is known as a 300 bp paired-end (PE) read library. The sequenced reads are shorter than the entire fragment (e.g. 75 bp). Singletons, also known as single-end (SE) or orphaned reads, consist of reads that are sequenced only from one end of the DNA insert. There is another pairing technique that is used to create libraries that span even greater distances of between 1 Kbp to 150 Kbp. These are called mate-pair libraries and are achieved by circularising the large inserts with the ends marked and joined together to be sequenced (Edwards & Caskey, 1991; Roach et al., 1995).

Once the reads are generated from these libraries and trimmed to retain bases with reliable quality, software called assemblers are used to assemble them - a process that is similar to solving a jigzaw puzzle. The end results are scaffolds or contigs. A scaffold is a portion of the genome reconstructed from contigs and contains gaps, while a contig is a contiguous length of genomic sequence in which the order of bases is known to a high confidence level. Gaps occur where information between contigs are unavailable.

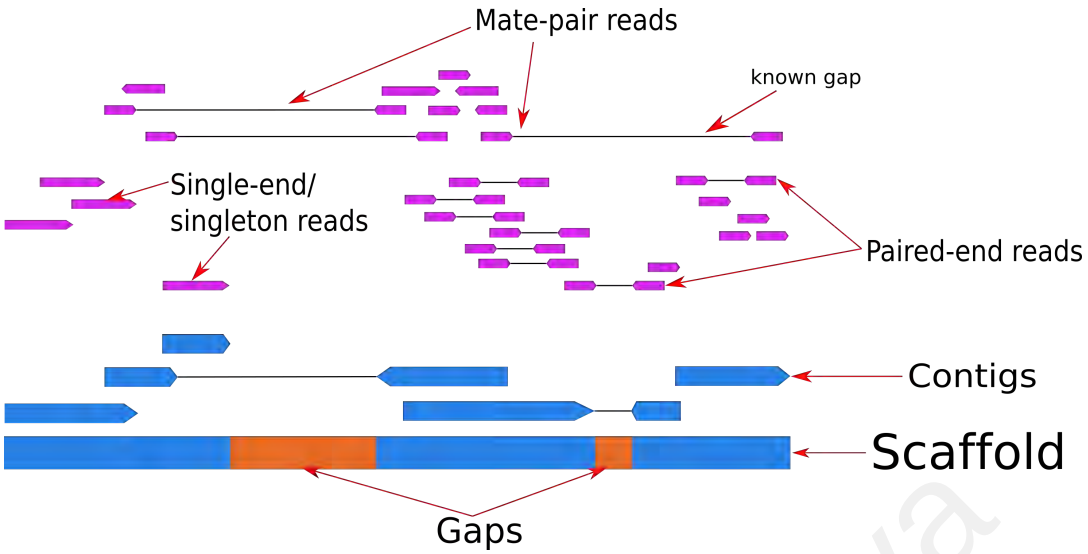


Figure 1.1: An illustration of assembly of singleton, paired-end and mate-pair reads to form a scaffold

Possible causes for gaps are sequence repeats or unsequenced regions of the genome. The unsequenced parts of the insert for PE and mate-pair reads become gaps in a scaffold when no overlapping sequence can be found for the region. Figure 1.1 shows an overview of the pieces in genome assembly.

The puzzle is rarely complete and is harder to solve if the genome in question is complex. In general, the complexity of a genome increases with genome size, number of chromosomes, higher GC content, and multiple repeats (Lynch & Conery, 2003). With an estimated size of 1.8 Gbp covering 16 chromosomes, a GC content of about 40% and approximately 60% repetitive elements, the oil palm genome is, without doubt, complex (Castilho et al., 2000; Singh et al., 2013).

The key challenge in genome assembly is the ubiquity of repetitive sequence elements (Treangen & Salzberg, 2012). Repeated sequences may range from large genomic duplications of several kilo bases long to short monomeric repeats. The repeat motifs may be structured in tandem, inverted and are not necessarily identical. It is also important to note that the assembly of a genome is haploid, and the diploid nature of oil palm, with its polymorphisms, further complicate the task of resolving contigs as alleles may be

incorrectly assembled due to dissimilarities in the sequences. Thus, sequence assemblers are rarely successful in correctly assembling repeated sequences, due to the many possible combinations caused by ambiguity in matching sequences (Pop, 2009).

One way around the problem of repeats is to remove the repeats out of the equation. Though programs such as Repeat Masker (Tarailo-Graovac & Chen, 2009) are available to mask repeats, they are reliant on preexisting repeat databases like Repbase (Jurka et al., 2005). In my study, I try to compartmentalise the repeats *ab initio*, by way of analysing the regions where unique sites meet repeat sites alone (see Chapter 3 for details). Identification of such regions allows us to remove the regions that are repeated and assemble non-repeated regions first. By identification of unique repeat/non-repeat sections, known as bridges, these can be added to the assembly, with the repeats themselves inserted between the bridge regions. This way, we can correctly link the non-repeated portions of the genome prior to adding the repeats. In this thesis, I will look at a post-hoc implementation, where knowledge of bridge reads is used to correct post-assembled scaffolds.

Compared to genomic sequences, mRNAs (messenger ribonucleic acids) tend to be less repeated. Transcripts are derived from the coding parts of the genome that are conserved in order to preserve gene functions. These sequences could thus help in resolving assembly issues in a genome assembly, by at least ensuring genomic portions that contain mRNA sequences are represented in the genome. In this study, I suggest the use of a directed assembly using a method of "fishing" for genome fragments with mRNA transcripts to improve the reliability of the assembly of the coding regions. This process will generate new contigs that give sufficient representation of the exons, with the potential to cover the introns and regulatory elements of the gene, in what we call exome contigs.

An important determinant of the completeness of a genome is the proportion that one is able to sequence. Sufficient reads are needed to cover all the gaps in an assembly. To do

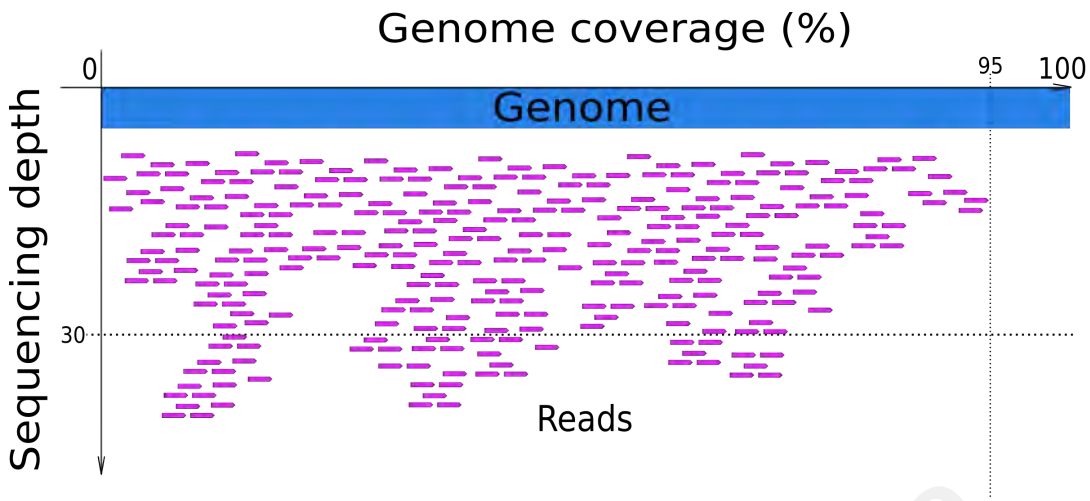


Figure 1.2: Sequencing depth versus genome coverage

that, it may be required to sequence deeply, that is, to sequence the genome many times over to increase the chances of capturing all possible read overlaps. The sequencing depth, or depth of coverage, is the number of times a sequence is covered by the total length of all reads. Sometimes the term coverage is used interchangeably with depth (Sims et al., 2014) but genome coverage only specifically means the percentage of the target genome size that is actually captured by the sequences. For example, a genome with an average sequencing depth of 30X may only have a genome coverage of 95% (Figure 1.2).

Even with the genome coverage metrics, how do we ensure the genome is complete enough for practical use? An evolutionary basis can be used to ascertain the quality of the genome draft by identifying the presence and completeness of basic genes all organism of a clade share. Programs like CEGMA (Parra et al., 2007) and BUSCO (Simão et al., 2015) evaluate completeness of genome drafts based on this evolutionary approach. CEGMA (Core Eukaryotic Genes Mapping Approach) defines a set of conserved protein families that occur in a wide range of eukaryotes. This set of Core Eukaryotic Genes (CEGs) was built from the euKaryotic clusters of Orthologous Groups database of NCBI (National Center for Biotechnology Information) (Tatusov et al., 2003). BUSCO, which is a successor to CEGMA, uses the biological basis of universal single-copy orthologs to

benchmark genome quality. However, at the time of writing, BUSCO did not have a plant database that could be used for evaluating the oil palm genome.

So how do we improve the genome draft? The simple answer is to do more sequencing. To improve the completeness of the genome, we could sequence deeper in order to capture more parts of the genome that were missed by chance. However, this option would not address the issues in regions of the genome that are repetitive. A better alternative would be to generate sequences with longer reads, for example, using PacBio. Longer reads that are in the ten of thousand bases will bridge most short tandem repeats, but the large repetitive inserts will still be missed. Until sequencing technology is able to sequence a single chromosome continuously, we will continue to struggle with the genome puzzle. Arguably, if we do not bother to accurately capture the lengths of the repetitive elements, the best current method to complete large scaffolds is to create large insert mate pairs, such as Bacteria Artifical Chromosomes (BAC) end sequencing (Shizuya et al., 1992), which can potentially be used to complete the chromosomal representation of the scaffolds.

For the Sime Darby oil palm genome, to do more sequencing would not be cost effective. I believe that the current version can yet be improved with the existing data, because I believe there is still untapped information in them that can be used to repair and improve an assembly. The refinement work described in the preceding paragraphs is part of the finishing process for the oil palm genome, a process that is arguably the most time-consuming and relatively expensive step in assembling a genome. It includes closing the gaps, contig rearrangements, sequencing error repair and physical map building. The challenge then for me is to complete the Sime Darby oil palm genome up to the point of having a draft physical map of the 16 chromosomes available, using information from the annotation of repeats, available transcriptome sequences and experimentally validated molecular markers such as SSR (Simple Sequence Repeats) and SNPs (Single Nucleotide

Polymorphisms).

1.4 Identifying differentially expressed genes from oil palm RNA-Seq data

A systems understanding of the observed variation between two different biological states typically begins with differentially expressed gene (DEG) studies. RNA-Sequencing (RNA-Seq) is the most recent high throughput technology in studying gene expression. The method has several advantages over hybridization-based approaches like microarrays: (i) a wider dynamic detection range; (ii) single nucleotide resolution; (iii) no prior dependency on a reference genome (Wang et al., 2009). Additionally, in principle, RNA-Seq can detect novel transcripts, thus allowing finer biological processes such as alternative splicing and RNA editing to be studied. These advancements naturally spurred concurrent development of data processing and analysis methods to extract biological meaning from RNA-Seq data.

At the basic level, making DEG calls is as simple as comparing the observed counts of RNA species generated from RNA-Seq data (Wilhelm & Landry, 2009). However, the observed counts are not actually the true counts found in the sample. The disparity is due to effects of the library preparation and sequencing, thus requiring replication to cover the distribution inherent in the sampling process. An ideal experiment would have sufficient sample replication to have statistical confidence in making the DEG calls. However, many RNA-Seq experiments are far from ideal due to expensive sequencing costs and the limited availability of tissue samples.

There are practical difficulties in obtaining sufficient tissue samples for studying the male and female immature inflorescences of the oil palm. The oil palm produces separate male and female inflorescences in alternate cycles throughout a year (Corley & Tinker, 2008). An inflorescence primordium is formed in the axil of each leaf during leaf inititation with the potential to develop into either a male or a female inflorescence (Verheye, 2010). The sex of the inflorescence is not distinguishable from the ground until it blooms, 30

months after inflorescence meristem initiation, because it will be covered by the prophyll (Adam et al., 2005). In the study conducted by Ho et al. (2016), immature inflorescences are classified to be 3-4cm long, corresponding to leaf +6 stage (the number after "+" corresponds to the number of their axillary leaf on the palm) (Adam et al., 2005). Many healthy palms at the end of their productive life need to be felled in order to harvest the immature inflorescences due to the uncertainty of the sex of the inflorescence. Such destructive sampling precludes the possibility of obtaining replicates from the same palm. The alternative of harvesting from productive standing palms is not feasible. Such a sampling process is dangerous to the sample collector, who must work at a height of no less than 10 metres. For the palm, accidental over-sampling of the inflorescence can kill it. Finally, very often, tissues from different palms had to be pooled in order to obtain sufficient RNA for sequencing.

At the time of writing, the cost of sequencing is still prohibitive, with each sequencing run costing tens of thousands of ringgit. Ho et al. (2016) made two sequencing runs, one male inflorescence pool and one female inflorescence pool with 454 sequencing. With sequencing lengths of 500 bp, 454 sequencing was ideal in obtaining reads for transcriptomes without a reference. The ability to identify unique mRNA sequences and assemble them was more important than obtaining the abundances of the mRNA. The abundances can be validated with more standard methods, such as quantitative polymerase chain reaction (qPCR), once sufficiently long contigs are available for primers to be designed. With a lack of replication in this experiment, a new robust method is needed to make confident DEG calls for qPCR validation from the observed counts in RNA-Seq.

To this end, it is necessary to evaluate the statistical interpretability of observed count data, leading to a procedure in which identification of DEGs between samples is done by first determining the true counts. This procedure utilises the knowledge of the sequencing

coverage to model the possible true counts from which an observed count originates from. Unlike current methods in which sequencing coverage for RNA-Seq refers to the mapped read depth over a gene model of an organism, I propose that the true coverage should be determined from the RNA concentration of a sample. To this end, I developed a method that is useful in calling reliable DEGs in unreplicated experiments, such as those in oil palm research, that are still prevalent due to cost and sample size limitations.

1.5 Aims and Objectives

I have two aims; The first is improvement of Sime Darby oil palm genome assembly through the use of transcriptome data, reduction of number of erroneously assembled regions caused by repeat sequences, repair of misassembled contigs, and scaffold consolidation into chromosomes using molecular markers. The second aim is to detect differential gene expression in the unreplicated RNA-Seq datasets of oil palm.

To achieve my aims, my objectives in this thesis are to design software tools and methods to: (i) improve the gene-centric information in the genome using transcriptome data; (ii) identify the repeated parts of the assembly and use the information to re-assemble the genome; (iii) re-arrange the scaffolds to form a physical map using molecular markers; and (iv) conduct differential gene expression analysis of RNA-Seq data generated from unreplicated experiments.

CHAPTER 2: LITERATURE REVIEW

2.1 Sequencing technology

The first attempt to sequence and construct a plant genome was successfully completed for *Arabidopsis thaliana* by a consortium of research teams from across the globe (Arabidopsis Genome Initiative, 2000). At the time, a large research body was needed to sequence a genome that was only 120 Mbp large and spans just 5 chromosomes. The completion, and subsequent availability of the Arabidopsis reference genome for the research community, catalyzed progress in plant genomics (Bevan & Walsh, 2005).

The success of the Arabidopsis genome project spurred many research groups to find means to generate their own reference genome for their plant of interest, including the oil palm (Singh et al., 2013). Furthermore with domesticated plants generating multiple crop varieties, research groups had to generate specific reference genomes for their variety of interests, like rice (Goff et al., 2002; Yu et al., 2002). The last clear census conducted in 2014 showed that there had been 95 plant genomes sequenced and published, and the trend seems to be growing exponentially (CoGe, 2015). This explosion of available reference genomes is a consequence of the availability of higher throughput and more cost effective sequencers in the instruments market at the turn of the twenty-first century.

The first technique that gained widespread use for DNA sequencing was the chain-termination method developed by Frederick Sanger in 1977 (Sanger et al., 1977). As the name suggests, the method employs dideoxynucleotide triphosphates (ddNTPs) which terminate DNA strand elongation selectively during DNA replication of the single-stranded DNA template strand. Subsequent innovation produced dye-tagged ddNTPs, thus enabling the chromatogram representation of the entire sequence elongation process. Sanger sequencing is capable of generating 1,000 bp long sequences. However, the quality of the

base calls after 700 bases from the 5' end deteriorates, as it becomes more difficult to resolve DNA length comparisons of only one nucleotide. Furthermore, the technique is time-consuming and expensive. It took two decades before second generation sequencers were introduced that further increased the throughput by several orders of magnitude and reduced the cost of DNA sequencing by a hundred fold (Liu et al., 2012).

454 sequencing is a pyrosequencing technique developed by 454 Life Sciences, that was later acquired by Roche (Margulies et al., 2005; Liu et al., 2012). The method depends on polymerase chain reaction (PCR) amplification of DNA fragments fixed onto beads. The amplification process generates inorganic pyrophosphates (PPi), which are released during DNA synthesis, resulting in a conversion cascade from ATP to oxyluciferin and light in the presence of sulfurylase and luciferase (King & Scott-Horton, 2007). 454 sequencing can generate reads with lengths between 400 bp to 800 bp. The output lengths can vary in a run due to the randomness of the nebulization method of fragmentation. 454 sequencing is known to make homopolymer errors because it becomes difficult to distinguish the light intensity difference once the homopolymer length is more than six bases (Huse et al., 2007; Liu et al., 2012).

Applied Biosystems' (ABI) Sequencing by Oligonucleotide Ligation and Detection (SOLiD) sequencing method does sequencing by ligation of oligonucleotide probes that encode two bases (Valouev et al., 2008; Liu et al., 2012). Matching probes at the two-base 3' end to the template strand are ligated together through a series of ligation, detection and cleavage cycles. The cleavage of the fluorescent labelled 5' end of the probe provides the signal for the color-space coding in the system. The reads generated are about 25-50 bp long, and are limited by the number of ligation cycles. The technique has the advantage of reading each base twice, thus reducing the error of SNP miscalls significantly (McKernan et al., 2009). However, it has been reported that palindromic sequences cannot be sequenced

reliably with this technique (Huang et al., 2012).

The Illumina sequencing method was developed by Shankar Balasubramanian and David Klenerman, the founders of Solexa, and was subsequently acquired by Illumina (Liu et al., 2012). This sequencing-by-synthesis method uses reversible dye-terminators to identify single bases as they are introduced into DNA strands during amplification (Ju et al., 2006). Illumina reads are generally shorter than 454 sequencing, from 60 bp to 150 bp (Liu et al., 2012). A fragment size selection during the library preparation phase ensures the output lengths are fixed.

The second generation sequencers still have limits in the read lengths generated. To overcome this limit, two new promising methods were developed: Pacific Biosciences's Single-Molecule Real-Time (SMRT) sequencing and Oxford Nanopore Technologies (ONT) sequencing. SMRT Sequencing works by sequentially detecting the four different fluorescent signals released when phospholinked nucleotides are incorporated onto a single DNA molecule template as it undergoes DNA strand elongation (Eid et al., 2009). The detection is done with zero-mode waveguides (ZMWs) with the DNA polymerase/template complex immobilised at the bottom of the well-like structure of the ZMWs (Levene et al., 2003). The ZMWs structure allows light to only illuminate a tiny volume sufficient for observing a single nucleotide addition at the bottom of the well. The technique is capable of generating reads with lengths exceeding 5,000 bp. Furthermore, DNA polymerase activity information can indicate whether a base is methylated since the data stream has a time factor. Base call errors were reported to be high and distributed randomly, so deeper sequencing is required to overcome this limitation (Korlach, 2013).

In DNA nanopore sequencing, a protein with a nanoscale hole (nanopore) is set in an electrically resistant polymer membrane. Detection is done by passing the DNA template through the nanopore and measuring the disruption in current for different bases. Currently,

bacterial porins are used as these nanopores (Clarke et al., 2009; Manrao et al., 2011). The MinION sequencer developed by ONT is the first commercially available product using nanopore technology. It is a relatively affordable and very portable device that has been deployed for real-time diagnostics applications (Greninger et al., 2015; Lu et al., 2016). The ONT sequencing method can generate reads with lengths similar to SMRT sequencing, albeit at high error rate (30%), which is mainly caused by indels instead of randomly distributed base call errors (Mikheyev & Tin, 2014).

As of May 2019, short read sequencing is relatively cheaper compared to long read sequencing technologies. Researchers benefit from the massively parellel capabilities of short read sequencers to generate a large amount of data in a short time. The technology is useful for sequencing fragmented DNA samples that do not benefit from long read sequencing, and the high sequencing depth of the sequencers allows base call error rates to be reduced statistically (Diouf et al., 2018; Wittig et al., 2018). Long read sequencing, on the other hand, allows researchers to read across repeated regions of the genome, thus overcoming one of the key challenges in genome assembly. However, long read sequencers currently require greater run time to increase accuracy of base calls by repeated measurements. They can also be expensive because of the lower throughput compared to short read sequencers (Travers et al., 2010). Most sequencing projects now combine short and long read sequencing to overcome each technology's limitations, where the long reads improve contiguity, and the short reads are used to "polish" the base call quality (Fuselli et al., 2018; Wang et al., 2018).

2.2 Sequence assemblers

Current sequencing technologies generate reads that need to be assembled in order to infer the original sequence of the genome. Assemblers are software that uses special algorithms to achieve this. There are two major algorithms:

(a) ***Overlap-Layout-Consensus (OLC)***

The classical method introduced by Staden in 1979 to assemble reads generated by Sanger sequencing technology is known as the overlap-layout-consensus (OLC) algorithm (Staden, 1979). The OLC algorithm first identifies pairs of reads that overlap (O) sufficiently well, normally allowing mismatches of around 10%. Then, the layout (L) of the reads is organised into a graph containing a node for every read and an edge between any pair of reads that overlap each other. Contigs are generated as a consensus (C) by inferences from information of all edges in the possible path. The heuristic is greedy, in which the assembly progresses when a local optimal configuration is found. This configuration is solved by adding reads sequentially based on the best overlap score found between the neighbouring nodes. Currently, the algorithm powers most assemblers that are optimised for long reads (> 300 bp), such as CAP3 (Huang & Madan, 1999), Celera Assembler (Myers et al., 2000), Newbler (Margulies et al., 2005) and PHRAP (de la Bastide & McCombie, 2007).

(b) ***De Bruijn graph (DBG)***

With the newer massively parallelised sequencing methods that generate short reads, like Illumina sequencing, a new algorithm that uses De Bruijn graph theory (DBG) was introduced by Idury and Waterman in 1995 to assemble the data efficiently (Idury & Waterman, 1995). First, all reads are sectioned into parts of fixed-lengths strings, called k-mers. The DBG algorithm models the relationship between exact k-mers from the reads with a directed graph. The nodes in the graph represent k-mers, that must be shorter than the reads, and the edges represent the overlap of adjacent k-mers by k-1 letters. Assembly is done by tracing the path with the most consistency through the graph. Thus, the consensus is built indirectly by solving for the global optimal configuration. Some assemblers that are based on DBG include EULER (Pevzner et al., 2001), Velvet (Zerbino & Birney, 2008), SOAPdenovo (Li et al., 2010), IDBA (Peng et al., 2010) and AllPath-LG (Gnerre et al.,

2011).

While both algorithms seek to address the assembly of reads, they are quite different in how effective they handle the type of datasets generated by sequencers (Li et al., 2012). The OLC assemblers approach assembly as finding a path in an overlap graph and visiting each node only once, as in a Hamiltonian Path Problem (Pevzner et al., 2001). This is a non-deterministic polynomial-time complete (NP-complete) problem, which is a decision problem with no known efficient way to solve (Garey & Johnson, 1979). On the other hand, DBG assemblers approach assembly as finding a path in the graph that visits every edge only once, and therefore changes it into an Eulerian Path Problem, making it solvable in linear-time (Fleischner & Fleischner, 1990). Therefore, the OLC assemblers would be better suited at assembling long reads with low-depth coverage data sets, while DBG assemblers assemble short reads with high-depth coverage data sets. The effectiveness of both algorithms in successfully finding correct matches deteriorates when they encounter sequencing errors, heterozygozity due to polyploid genomes, and repeats. Since overlap detection can be made to allow mismatches, OLC assemblers will perform better than DBG assemblers when it comes to consolidating base differences. DBG assemblers run on the assumption of no sequencing errors to work effectively. Many attempts to address the compounding effects of sequencing errors have been made over the years with error correction steps such as techniques devised by Tammi et al. (Tammi et al., 2003, 2004; Arner et al., 2006). However, the effectiveness of the error correction step depends on the assumption that there is sufficient statistical power in the alignments for errors to be detected. Under uneven sequence coverage and polyploid heterozygosity, this assumption is unlikely to hold (Nagarajan & Pop, 2013).

2.3 RNA-Seq and exome contigs

While the sequencing technologies detailed above were mainly developed for DNA sequencing, the same sequencing technologies can be adapted in ribonucleic acid sequencing (RNA-Seq). The difference mainly lies in the library preparation stage. Essentially, RNA is converted to double stranded complementary DNA (cDNA) through reverse transcription, followed by the same procedures in DNA sequencing. The bulk of RNA molecules expressed in cells are ribosomal RNA (rRNA), and are usually removed during sequencing library preparation to enrich for the messenger RNA (mRNA) (Zhao et al., 2014).

Before the availability of high-throughput RNA-Seq library techniques, scientists could only generate truncated cDNA sequences to represent the original mRNA molecule. These Expressed Sequence Tags (ESTs) are the sequenced ends of cloned cDNA fragments, usually sequenced with Sanger or 454 sequencers (Adams et al., 1991). ESTs are useful as a means to count mRNA abundances, but the incompleteness of the information provided by such relatively short reads (200 - 800 bp) hampers unambiguous gene assignments (Nagaraj et al., 2006).

Current RNA-Seq library preparations that mimic the shotgun sequencing approach allow us to assemble a sequence reference of an organism's transcriptome to characterise these intermediate messenger molecules, or transcripts, *en masse* (Wang et al., 2009). The *de novo* assembly of the transcriptome, usually only the mRNA, has the advantage of serving as a reference to enable transcript-counting in the absence of a reference genome. Such an assembly is relatively cheaper and easier to build than a genome reference (Martin & Wang, 2011). Transcriptome assemblies have the advantage of revealing novel protein isoforms due to alternative splicing events that would otherwise lay hidden in a genome draft. Further benefits in doing RNA-Seq includes the ability to detect post-transcriptional modifications, gene fusion, mutations and changes in gene expression over time, or between

different treatments (Wang et al., 2009).

While transcriptome assembly does not encounter challenges due to repetitive elements found in genomes, challenges in resolving ambiguity caused by genes with spliced isoforms and minor variations within a gene family need to be addressed. The basic algorithms used for genomic assembly are still used, with slight modification to account for wide variation in sequencing depths in a transcriptome due to different gene expression abundances, and strand specificity inherent for mRNA. Therefore, modified and/or special assemblers dedicated to overcoming this challenge were developed, such as SOAPdenovo-Trans (Xie et al., 2014), Velvet (Zerbino & Birney, 2008), ABySS (Birol et al., 2009) and Trinity (Grabherr et al., 2011). Long read sequencing in RNA-Seq experiments has also been used to overcome the challenges of true isoform identification (Byrne et al., 2019). Long reads, such as those produced by PacBio circular consensus sequencing, have very low base call errors and do not require assembly (Cheng et al., 2017). Recently, new library preparation methods that use unique molecular identifiers (UMIs) as barcodes to differentiate between PCR duplicates and transcripts in short read sequencers enable accurate isoform identification and abundance counts in human tissues (Wu & Ben-Yehezkel, 2019). The method, called LoopSeq Synthetic Long Read Sequencing, can either be used for reconstructing long RNA transcripts or for counting transcripts at low coverage.

Since a transcriptome assembly contains the expression units of the genome of the same organism, it can be used to estimate the completeness of a genome assembly. Of course, the caveat is that the transcriptome assembly should be built from a heterogenous community of cells at various times on the life cycle of the organism in question to obtain a sufficiently comprehensive set of expressed genes. This is because transcript expression is dynamic, and RNA-Seq only captures a snapshot of expression at a point of time and

place. Moreover, transcript expression is also tissue-specific. Many RNA-Seq studies on plants have substantially improved our understanding of gene regulation and expression (Martin et al., 2013).

Besides using transcriptomes for assessment of the genome assembly quality, transcriptome contigs can be used to direct improvement of the genome assemblies. A seed-driven iterative assembly approach could be used to assemble missing or incomplete fragments of the genome. GenSeed (Sobreira & Gruber, 2008) was the first tool designed for such a task, albeit without the actual use of transcriptome contigs as the starting material. Since its release, only a few approaches based on the same concept were proposed for the assembly of viral sequences from metagenomic data (Smits et al., 2015). The same team that developed GenSeed had recently published an updated version of their algorithm using Hidden Markov Model (HMM) profiles as starting seeds for target-driven reconstruction called GenSeed-HMM (Alves et al., 2016). The algorithm for GenSeed is straighforward - given a starting sequence of sufficient length for conducting local alignment, it will identify similar contigs/reads from a pool, and proceed to construct a consensus with these sequences. The ends of the completed contig are then used to repeat the search and assembly process with additional similar sequences from the pool, until no further similarity can be found.

2.4 The repeat problem

One of the biggest challenges in sequencing and the cause of many gaps in an assembly is the presence of many repeats in the genome. Repeats in the genome can be categorised into two broad categories: tandem repeats and interspersed repeats.

Tandem repeats are DNA element patterns of one or more nucleotides repeated consecutively. Tandem repeats that occur in isolated islands in the genome with 1-5 nucleotide (nt) motifs repeated up to 50 times are known as microsatellites (Richard et al.,

2008). DNA microsatellites, or alternatively known as simple sequence repeats (SSRs), are used widely in genetic diversity and population studies because of their high heritability and measureable mutation rates (Roewer et al., 1992; Jarne & Lagoda, 1996; Ellegren, 2004). SSRs are known to be highly variable in the number of repetition of its motif from organism to organism, allowing its use as molecular markers in genetic profiling. Tandem repeats that have longer motifs of up to 100 bp are known as minisatellites (Vergnaud & Denoeud, 2000). Minisatellites are prominent in the centromeres and telomeres of chromosomes (Tran et al., 2015).

Interspersed repeats have motifs that are dispersed throughout the genome and do not occur in tandem. This type of repeat occurs when DNA sequences known as transposable elements (TE), produce a duplication error during transposition. When DNA sequences are duplicated via the mediation of transposase enzymes, they are known as Class II type TE. However, the majority of interspersed repeats are caused by retrotransposons, or Class I type TEs, which are mediated with an RNA intermediate. Retrotransposons can be further grouped into three types (Xiong & Eickbush, 1990; Schmidt, 1999): Long terminal repeats (LTRs), which encode reverse transcriptase, similar to retroviruses; long interspersed nuclear elements (LINEs), which encode reverse transcriptase but lack LTRs, and are transcribed by RNA polymerase II; and short interspersed nuclear elements (SINEs), which do not encode reverse transcriptase and are transcribed by RNA polymerase III.

Current established methods that detect repeats require an assembled draft genome. If the query species' genome is available, repeats are identified with a pattern signature database specific to the genome. The most popular software for identifying repeats in DNA sequences is Repeat Masker (Tarailo-Graovac & Chen, 2009). It performs an efficient implementation of the Smith-Waterman-Gotoh algorithm, developed by Phil Green, called Cross Match (Tarailo-Graovac & Chen, 2009). Once identified, the repeated sequence is

usually replaced by Ns, hence the masking effect implied. Repeat Masker depends on a manually curated repeat consensus version of the RepBase database (Jurka et al., 2005).

If a repeat database is not available for the query species' genome, one can either use a closely-related species to predict repeats based on homology with Repeat Masker, or compile a new repeat library using *de novo* methods. RECON identifies repeats by conducting multiple sequence alignments of reads against a genome, and uses a heuristic process to determine boundaries of repeats (Bao & Eddy, 2002). RepeatScout uses a similar approach with the improvement in accuracy and computational speed with a greedy seeding protocol to identify repeats from the multiple sequence alignments (Price et al., 2005). REPuter (Kurtz et al., 2001) and Repseek (Achaz et al., 2006) both adopt a seed-and-extend paradigm to identify identical and degenerate repetitive sequence. P-clouds (de Koning et al., 2011) determines repetitive motifs by clustering similar but divergent sequences together.

The above methods still require an assembled genome for repeat detection. The software, ReAS, on the other hand, generates repeat libraries directly from sequenced reads rather than assembled contigs (Li et al., 2005). The method requires reads longer than 100 bp for the seed size. Another method, Tallymer (Kurtz et al., 2008), was made for plant genomes that does k-mer counting and indexing using enhanced suffix arrays on assembled genomes. Sequence reads can be used to generate the k-mer indices for searches in a draft genome. This method is not restricted to repeat identification, and is memory efficient (Manekar & Sathe, 2018). Another software called RepARK uses reads to build a *de novo* repeat library (Koch et al., 2014). RepARK identifies k-mers that occur more than once genome-wide and proceeds to build consensuses from the reads. RepARK determines the occurrence threshold of a k-mer with a linear function fitting the k-mer index frequencies.

Another way to find repeats is with assemblers during genome assembly, such as the

Celera Assembler (Myers et al., 2000; Denisov et al., 2008). It is possible to detect high coverage regions, which may be caused by collapsed repeats. However, the random selection process of shotgun sequencing fragments results in the coverage being distributed according to a Poisson distribution; which means that the coverage can vary widely, and high coverage regions may simply be caused by the random selection process, and not because it is a repeat region.

2.5 Completing the genome map

A genome is considered truly complete if an organism's genome can be represented unambiguously by a contiguous sequence. This is achievable for bacterial genomes, but not for genomes of multicellular organisms. Our current sequencing technology has yet to allow us to overcome computational complexities arising from the genomic architecture of eukaryotes, such as the subdivision of genomes into multiple chromosomes, and the repetitive nature of centromeres and telomeres. Still, one can argue that a genome draft is complete when it becomes useful in genetic studies (Mardis et al., 2002).

A genetic map, or linkage map (LM), is made by identifying the locus of genetic markers and their relative distances by means of an inheritance study. The first such map was developed by Alfred Sturtevant for *Drosophila* (Sturtevant, 1913). The nearer two genetic markers are on a chromosome, the more likely they are to be inherited together in the progenies. This is because the close proximity of two markers, which are said to be linked, reduces the chance of recombination happening between them during the meiosis phase of sexual reproduction. Genetic linkage is measured in centimorgan (cM). When two markers are said to be 1cM apart, it means that the markers are at a distance that has the potential rate of recombination to occur on average once per 100 meioses. A good quality genetic map would have a large amount of genetic markers and a large mapping population (a population of controlled crosses).

Any sequence feature that can be faithfully distinguished from the parents can be used as a genetic marker. Historically, traits were used as genetic markers. Now, we use DNA sequences that we call molecular genetic markers (molecular markers). The most popular types of molecular markers used to date are: (i) restriction fragment length polymorphism (RFLP) markers; (ii) random amplified polymorphic DNA (RAPD) markers; (iii) amplified fragment length polymorphism (AFLP) markers; (iv) diversity array technology (DArT) markers; (v) SSR markers; and (vi) single nucleotide polymorphism (SNP) markers.

RFLP markers are detected by identifying different sequence lengths generated by restriction enzyme digestion on DNA (Botstein et al., 1980). RAPD markers are based on the varying result caused by sequence variation at primer binding sites and DNA length differences between primers when amplification of random DNA segments are conducted (Williams et al., 1990). AFLP markers are generated by selective Polymerase Chain Reaction (PCR) amplification with restriction-site-specific primers of digested DNA (Vos et al., 1995). DArT markers are identified using microarray hybridizations that detect the presence or absence of a DNA sequence fragments from a genomic representation (Jaccoud et al., 2001). SSR markers are repetitive sequences that are highly variable in length, and can be identified by PCR amplification using unique primers flanking the repeat sequence. (Tautz, 1989; Gulcher, 2012).

SNP is a variation in a single nucleotide that occurs at a specific position in the genome. SNPs can be used as molecular markers when the variation occurs with sufficient frequency in the population due to its heritability, especially in plants (Gupta et al., 2001; Syvänen, 2001). SNPs can occur anywhere in the genome. If a base difference causes a different amino acid to be translated in the protein synthesis process, it is known as a nonsynonymous SNP, while a synonymous SNP does not cause any difference.

The complete sequence of a genome is a physical map (PM) where all the position of

the bases are known without ambiguity. Another way to look at it is that a PM is a map of the locations of genetic markers along DNA where the distance is absolute and measured in base pairs. The ideal PM represents a chromosome from end to end. In its fragmentary form, the physical map of current eukaryote genome drafts can be built with the aid of molecular markers. In 2015, the melon genome draft was improved with the use of a SNP-based genetic map to orient and anchor the scaffolds according to the chromosomes (Argyris et al., 2015). More recently, a restriction site associated DNA (RAD)-based genetic map for bitter gourd assisted in anchoring 85.48% of the assembled genome (Cui et al., 2018). As of May 2019, only Chromonomer (Catchen & Amores, 2016) had been developed to rearrange genome scaffolds according to a genetic map. Chromonomer had been used in the assembly of the Gulf pipefish and Platyfish (Amores et al., 2014; Small et al., 2016). These physical maps of chromosome-level genome assemblies were also referred to as "chromonomes" (Braasch et al., 2015).

Another step towards a complete genome is gap-filling of the assembly. Currently, programs such as ABACAS (Assefa et al., 2009) automate the pipeline that extends the gaps with unused reads during the prior assembly process, and design primers for further walking and sequencing validation. However, this program requires a reference genome for such tasks.

Sommer et al. (2007) suggested that a simpler algorithm is needed for gap filling compared to current whole genome shotgun assemblers. To that end, his team developed the Minimus assembler which utilises a Smith-Waterman hash-overlap to compute all pair-wise alignments between input sequences in its Overlap-Layout-Consensus (OLC) method of assembly. Minimus does not utilise quality values for its assembly and assumes the input sequences have already been trimmed or filtered of poor reads.

Another means to fill gaps is by combining multiple assemblies generated from different

assembly pipelines. This approach was used in the assembly of the rhesus macaque genome (Gibbs et al., 2007). The process of mapping the three intermediate assemblies and human genome reference done by the authors was likely tedious and time-consuming, as they had to conduct three different assembly pipelines sequentially. Now, programs such as the updated Minimus, and GAM-NGS (Vicedomini et al., 2013) make this task simpler.

Minimus2 is a modified version of the Minimus pipeline meant to merge two sequence sets together. It uses a faster overlap detector than Minimus called nucmer (Delcher et al., 2002). Nucmer (NUCleotide MUMmer) is a multiple DNA sequence aligner that is part of the MUMmer suite (Delcher et al., 1999; Kurtz et al., 2004). The algorithm improves speed and memory efficiency by first finding maximal unique matches (MUMs) using suffix trees for one set of input sequence, and have the second set added via a streaming behaviour before using a modified Smith–Waterman dynamic programming algorithm to align the sequences.

Further refinement steps to achieve a completed physical map of the genome currently use two additional technologies that leverage on structural information of the genome to scaffold the sequenced contigs: Hi-C sequencing, and optical mapping. Hi-C sequencing is a relatively new approach for arranging scaffolds with a technique originally designed to study the three-dimensional structure of the genome in the nucleus of a cell (Lieberman-Aiden et al., 2009; Kaplan & Dekker, 2013; Marie-Nelly et al., 2014; Bickhart et al., 2017; Dudchenko et al., 2017; Mascher et al., 2017). The technique provides linkage information that can span tens of megabases by measuring the frequency of contact between intra- or inter-chromosomal pairs of loci. Briefly, the process involves cutting cross-linked chromatin with restriction enzymes first, then labelling the linked pieces with biotin, ligating the ends, and finally, sequencing the biotin-labeled regions.

Optical maps, on the other hand, are generated by digesting the genome with specific

restriction enzymes and adding different intercalating dyes to each fragments to make them visually discernable under a fluorescent microscope for realignments (Schwartz et al., 1993; Valouev et al., 2008). Subsequent advancement with nanochannel arrays is used to digitise these patterns, and an assembly algorithm using an OLC approach is used to assemble the restriction map contigs (Levy-Sakin & Ebenstein, 2013; Chaney et al., 2016; Tang et al., 2016; Yuan et al., 2017; Udall & Dawe, 2018). The resulting map is then used as a reference to scaffold the contigs from sequence assembly.

Typically, Hi-C sequencing is followed by optical mapping in order to verify the results, as well as assist in modifying the draft genome (Burton et al., 2013; Korbel & Lee, 2013). Hi-C sequencing is capable of only providing arbitrary size estimates of gaps, while optical mapping can provide the final accurate gap sizes for N-filling.

2.6 Current methods for calling differentially expressed genes

Since the publication of the first RNA-Seq paper (Lister et al., 2008), which was based on the 454 sequencing platform, extensive interest in RNA-Seq has resulted in the development of additional platforms such as Illumina sequencing. The bioinformatics landscape for RNA-Seq analysis is large, with many different methods to infer gene expression from the sequencing data (Soneson & Delorenzi, 2013). The methods deal mostly in normalizing and applying robust statistical analyses to identify significant differentially expressed genes (DEG) in a general RNA-Seq analysis workflow (Oshlack et al., 2010). As Bullard et al. (2010) had indicated, the choice of normalisation procedure vastly influences the outcome of a DEG analysis. Most of the bewildering number of data normalization techniques currently deal with Illumina RNA-Seq data (Dillies et al., 2013). The methods can be categorised broadly as either parametric or nonparametric.

2.6.1 Parametric methods

Parametric methods model the distribution of read count data using appropriate statistical distributions. Currently, the most popular methods for making DEG calls using RNA-Seq data are parametric methods that assume a negative binomial distribution on the count data. In the description of the methods that follow, a lane refers to the data column of a sample.

DESeq and DESeq2 (Anders & Huber, 2010; Love et al., 2014) scaling factor for a given lane is computed as the median of the ratio, for each gene, of its read count over its geometric mean across all lanes. The underlying idea is that non-DEGs should have similar read counts across samples, leading to a ratio of 1. Assuming that most genes are non-DEGs, the median of this ratio for the lane provides an estimate of the correction factor that should be applied to all read counts of this lane to fulfill the hypothesis. DESeq is very conservative in generating DEGs, with a computation time that increases with the sample size.

EdgeR (Robinson et al., 2010), on the other hand, computes the Trimmed Mean of M-values (TMM) factor for each lane, with one lane being considered as a reference sample and the others as test samples. For each test sample, TMM is computed as the weighted mean of log ratios between this test and the reference, after exclusion of the most expressed genes and the genes with the largest log ratios. An empirical Bayes procedure is used to moderate the degree of overdispersion across genes by borrowing information between genes. An exact test analogous to Fisher's exact test but adapted to overdispersed data is used to assess differential expression for each gene. EdgeR is less conservative in calling DEGs compared to DESeq with a high true positive rate, and the computation time is independent of the sample size.

BaySeq (Hardcastle & Kelly, 2010) employs a selection of scaling factors (TMM/quantile/total) for normalizing the read counts. A Bayesian approach is then used to asses

differential gene expression based on posterior probabilities. BaySeq's algorithm makes it less susceptible to outliers. Its computation time is relatively long compared to DESeq, but it offers the possibility of parallel computing.

2.6.2 Nonparametric methods

Nonparametric methods do not model the distribution of the count data explicitly, and relies on more general data permutation approaches to evaluate statistical significance. NOISeq (Tarazona et al., 2011) is a nonparametric method that first employs a selection of scaling factors (TMM/upper quartile/RPKM) for normalizing the read counts. Then, a null distribution is simulated by permutation to become the base for comparisons by contrasting the fold changes and absolute differences within a condition. NOISeq performs well when count distribution of different phenotypic conditions have varying dispersion patterns. It's computation time is dependent on the sample size.

SAMSeq (Li & Tibshirani, 2013) is another nonparametric method that normalises the read count by taking the mean read count over the null features (genes that do not correlate significantly with any condition) of the dataset. A resampling strategy is used to build a distribution for comparisons before subjecting the results to a Wilcoxon rank statistic. SAMSeq performs well even with large sample sizes, and its computation time is dependent on the sample size.

2.6.3 Assumptions in RNA-Seq

All of the current RNA-Seq analysis methods assume the statistical noise in observed count data caused by sequence sampling and read mapping steps to be negligible compared to biological variation. Most biologists prioritise biological replicates over technical ones to ensure that a RNA-Seq experiment has sufficient statistical power to detect DEG (Liu et al., 2014; Gierliński et al., 2015; Schurch et al., 2016). Furthermore, the majority of normalisation strategies assume that most genes are not differentially expressed, and that for those differentially expressed there is an approximately balanced proportion of over- and under-expression.

Current RNA-Seq methods also assume more or less homogeneous gene expression levels among single population of cells, and the average expression in RNA-Seq provides a good estimate of gene activity level in the same tissue (Fu et al., 2009). This notion had been challenged by Sanchez et al. (2013), who showed that the relative proportions of mRNA species between cells can be highly variable. Furthermore, in genetically identical yeast cells, variation of more than 800 copies of an mRNA species per cell has been observed (Marguerat et al., 2012). This insight has resulted in growing numbers of RNA-Seq experiments that are at the single cell level (Saliba et al., 2014). Thus, in most RNA-Seq studies where multicellular samples are used, accurate sample-to-sample comparisons require reliable transcript counts in each cell.

The non-homogeneity of gene expression levels in cells affects the calculation of transcriptome sequencing coverage. Unlike genome sequencing where the genome size can be estimated quite accurately, transcriptome size is much harder to estimate, varying greatly between cells and tissue types of the same organism (Lovén et al., 2012). For accurate quantification of 95% of transcripts in a human cell line, up to 700 million reads are needed (Blencowe et al., 2009). In contrast, Genohub and ENCODE Consortium

recommend that a typical RNA-Seq experiment for quantification only needs at most 30 million reads (Genohub, 2015; ENCODE, 2011). Consequently, in most experiments, the sequenced reads that constitute the observed counts for each RNA species represent but a tiny fraction of the true count in a sample. Observed counts are therefore subject to potentially large stochastic effects, particularly if the corresponding true count is large. Compounding the problem of interpretability of count data are biases inherent in technical RNA-Seq library preparation and sequencing (Sendler et al., 2011), a problem that has since received serious attention (Lahens et al., 2014).

CHAPTER 3: METHODOLOGY

3.1 Oil palm whole genome assembly improvement

3.1.1 Initial resource

(a) ***Genome***

A Sime Darby commercial *tenera* hybrid palm (EKONA descent), identified as Palm 99, was sequenced using Roche 454 GS-FLX by an external service provider. The generated 238,151,400 reads were between 200 bp - 400 bp in length. The first oil palm draft in 2009 (OPg1) was assembled using proprietary methods that generated 37,882 scaffolds (unpublished results).

(b) ***Transcriptome***

Sime Darby also conducted RNA-Seq sequencing with Roche 454 GS-FLX sequencers for six different types of oil palm tissues: mesocarp, root, leaf, meristem, male inflorescence and female inflorescence. The mesocarp tissues were taken at three time points of fruit growth since pollination, which were week 12, week 16, and week 18. Eight transcriptome data sets with over 20 million 454 reads with an average read length of 400 bp were generated.

(c) ***Molecular markers***

I used 75 SSRs published by Billotte et al. (2005), as well as 501 polymorphic SSRs developed and experimentally validated to be polymorphic internally by Sime Darby (unpublished results). I also used 101 DArT markers that were developed internally by Sime Darby (unpublished results). Additionally, Sime Darby had developed a SNP array that consisted of 170,860 informative SNPs called OP200K (Kwong et al., 2016). Of these markers, 26,240 SNPs make up Sime Darby's Linkage Map.

(d) ***Computational resource***

The experiments were conducted in four IBM System x3650 M3 (2x6 core Xeon 5600) machines with 96 GB RAM, running on the RedHat 6 operating system. The machines were set up in a High Performance Computing (HPC) environment facilitated by Sun Grid Engine (SGE) at Sime Darby Technology Centre, Serdang, Selangor. I further deployed Rackspace cloud servers when additional bursts of computation power were required.

3.1.2 External programs used

In this study, I used various bioinformatics tools that were available at the time for my analyses. Table 3.1 summarises the tools that make up crucial parts of my computational pipeline. All result graphs were generated using R (R Core Team, 2018).

3.1.3 Sime Darby's transcriptome reference

(a) ***EGrefseq assembly***

A consensus transcriptome assembly, which I named EGrefseq, was built by combining the reads generated from male inflorescence, female inflorescence, apical meristem, mesocarp, leaf and root of oil palm. The reads generated by Roche 454 GS-FLX were used as input to the Newbler program package V2.5 (454 Life Sciences, Roche Diagnostics Corporation, Branford, CT, USA). Sequences originating from organelles and rRNA were removed. The assembly process used default settings with the addition of the '-urt' option that assembles transcripts with low-read coverage. In post-processing, I used in-house Perl scripts to remove redundant sequences and sequences smaller than 200 bp (accepted length of NCBI data repository and annotation). The resulting assembled sequences are known as isotigs or contigs. The Newbler program was used to cluster the isotigs into isogroups. The notion of a gene corresponds to an isogroup, while the splice variants were

Table 3.1: External programs used

Name	Version	Description and reference
BLAST	2.2.25	A suite of tools used to find regions of similarity between biological sequences. (Altschul et al., 1990)
BLAT	34	BLAT is an alignment tool like BLAST that search matches with indexes kept in memory. (Kent, 2002)
BWA	0.7.12	Fast and accurate short read alignment with Burrows-Wheeler transform. (Li & Durbin, 2009)
CAP3	10/15/07	A sequence assembly program utilizing OLC method mainly for Sanger sequences. (Huang & Madan, 1999)
CEGMA	2.4	A pipeline to accurately annotate core genes in eukaryotic genomes. (Parra et al., 2007)
GenSeed	1.0.22	A seed-driven progressive assembly program. (Sobreira & Gruber, 2008)
IDBA	1.1.0	An iterative De Bruijn graph *de novo* assembler for sequence assembly. (Peng et al., 2010)
Minimus2	3.1.0	A program pipeline meant to merge two sequence sets together. (Sommer et al., 2007)
PHRAP	20	A program for assembling shotgun DNA sequence data. (de la Bastide & McCombie, 2007)
SAMtools	0.1.18	A program suite that provide various utilities for manipulating alignments in the SAM format. (Li et al., 2009)
SSPACE	2011	A program for scaffolding pre-assembled contigs using NGS paired-read data. (Boetzer et al., 2011)

represented as isotigs or contigs. In the following sections, both contigs and isotigs are referred to as EGrefseq contigs.

(b) ***Functional annotation and quality assessment of EGrefseq contigs***

Using BLASTX with an E-value threshold of 1×10^{-10}, the EGrefseq sequence functions were annotated using the Swiss-Prot database (October 2013 release) (Boutet et al., 2007), the TAIR10 database (Swarbreck et al., 2007), the RGAP 7 database (Kawahara et al., 2013) and the KEGG database (Kanehisa & Goto, 2000). I identified the best match (minimum length of 200 bp) with the highest score in bits to annotate each contig. I also adapted CEGMA (see Section 3.1.4) to assess the reliability of the assembly.

3.1.4 Genome quality assesment

(a) ***Genome draft statistics***

Throughout the study, I evaluated the different iterations of the genome draft by collecting basic statistics on the total size, length and N50 for comparison.

(b) ***Evolutionarily-conserved genes evaluation***

I used an evolutionary framework to evaluate the completeness of the genome assembly by using the program pipeline CEGMA (Core Eukaryotic Genes Mapping Approach) (Parra et al., 2007). The approach evaluates the number and completeness of 458 evolutionary conserved genes, known as Core Eukaryotic Genes (CEGs), present in the assembly. I also added CEGMA analyses for six highly-cited plant genomes as comparison to the oil palm drafts: grape (Jaillon et al., 2007), two rice varieties (Goff et al., 2002; Yu et al., 2002), sorghum (Paterson et al., 2009), poplar (Tuskan et al., 2006), and Arabidopsis (Arabidopsis Genome Initiative, 2000).

Table 3.2: Publicly available oil palm ESTs from GenBank

Researchers (Year)	Number of contigs	Tissues
Ho et al. (2007)	14,537	root, shoot apical meristem, young flower, mature flower, suspension cell culture, zygotic embryo.
Bourgis et al. (2011)	41,695	mesocarp (weeks after pollination: 15, 17, 19, 21 and 23) and leaves.
Tranbarger et al. (2011)	29,034	mesocarp (day after pollination: 100, 120, 140, 160).

(c) ***Exome completeness evaluation***

EGrefseq contigs were mapped against the genome assembly using the program BLAT (Kent, 2002) with default parameters. To mitigate the effects of possible biases due to solely using in-house transcript data, ESTs (Expressed Sequence Tags) from published datasets (Ho et al., 2007; Bourgis et al., 2011; Tranbarger et al., 2011) were also mapped with the same BLAT parameters (Table 3.2).

(d) ***Molecular marker representation evaluation***

SSR primers and DArT markers that had been experimentally validated to be polymorphic were mapped by applying a BLAST search against the genome assembly, with low complexity filters turned off (Altschul et al., 1990). The location and statistics of succesfully mapped markers were identified with the MarkMyMap program I developed (Section 3.1.8).

3.1.5 Improvement by additional sequencing

The first draft, OPg1, was improved by conducting further sequencing with new pair-mated libraries. The library preparation, sequencing and assembly were conducted by an external service provider according to Illumina sequencing protocols. Briefly, new pair-mated libraries of 300 bp, 5 Kbp, 8 Kbp and 40 Kbp were sequenced from samples

derived from Palm 99. The programs IDBA (Peng et al., 2010), PHRAP (de la Bastide & McCombie, 2007) and SSPACE (Boetzer et al., 2011) were used to assemble the new reads with OPg1 scaffolds. This produced the newer iteration of the genome draft called OPg2. The additional sequencing reads were not made available by the service provider.

3.1.6 Improvement by adding exome contigs with GenSeed Pipeline Suite

I developed a program pipeline (Appendix B) expanding the capabilites of the program GenSeed (Sobreira & Gruber, 2008). The pipeline requires the following dependencies:

1. GenSeed program
2. BLAST program suite
3. CAP3 program
4. BioPerl & modules: Bio::SearchIO, Getopt::Long, GD, Number::Range

Two files were used as input for the pipeline:

1. Individual transcriptome contigs in fasta
2. BLAST database of genome reads, indexed (./formatdb -o T -p F)

A single shell script wrapper was executed to run the pipeline. I first identified EGrefseq contigs that did not map at least 60% of total length in the results of Section 3.1.4 with OPg2. These were compiled into a single multi-sequence fasta format file. Then, the program proceeds to identify genome reads that matches to each EGrefseq contig via BLAST. CAP3 was then called to assemble the matched reads. The resulting contig was further used to find additional genome reads in the same database. This process of search and assemble iterates until no further reads can be found that match well. I used BLAST with the low complexity filters turned off and the E-value threshold set at 1×10^{-3} in the matching steps. The contig assembly process was run with CAP3 using default parameters. Finally, the program compiles all generated exome contigs as output in fasta format.

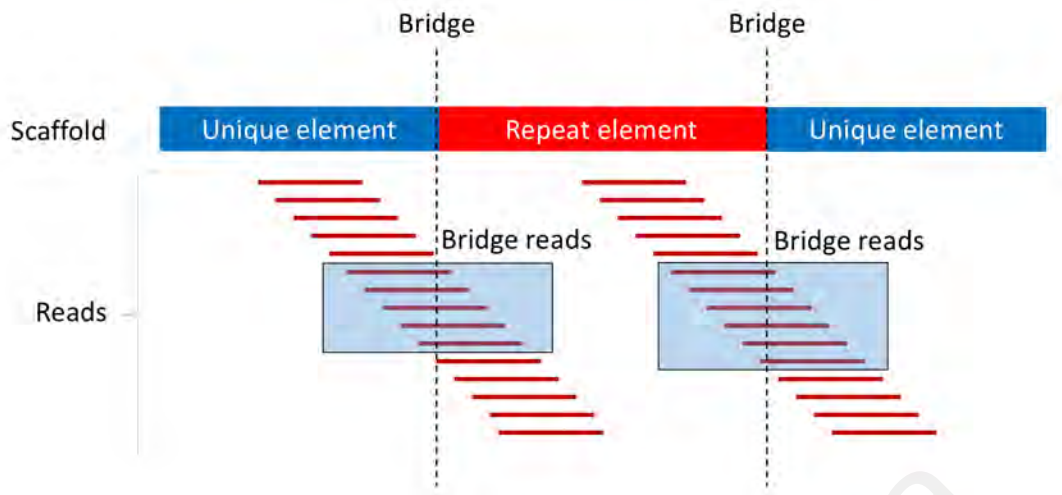


Figure 3.1: Concept of bridge reads in sequence assembly

The exome contigs and OPg2 scaffolds were then assembled together using the program SSPACE (Boetzer et al., 2011), with the use of all available pair-mated libraries from both the Illumina and 454 sequencing, to produce OPg3.

3.1.7 Bridge read detection with BridgeReader

A program called BridgeReader was developed by Martti Tammi and I (Appendix B) to detect reads that span over unique-repeat and repeat-unique regions of a genome, which we call bridge reads (Figure 3.1). Briefly, a genome read is segmented into six regions and indexed. False overlaps are detected by comparing the middle sections of one read against all flanking sections of other reads by a sliding window search. A depth of coverage profile of each read is then made to detect the exact point where unique and repeat regions end. The unique portions of these reads are then mapped back against the genome to identify the positions on the physical map.

This program is optimised for use on 454 sequenced reads, which are longer than Illumina's. The split to six regions is practical for reads that are 100 bp or longer, and provides improved sensitivity.

The following subsections give a more detailed look at the algorithm contained in the

two modules that make up BridgeReader.

3.1.7.1 iCountDBMate

This module creates a database that contains a unique index of every word of user-defined length, and in which read it appears. The output consists of text files containing a list of rows with read identities (IDs) ordered by the word index. The number of text files corresponds to the number of sections each read is divided into. For example, if a read is divided into five sections, five separate database files will be created.

Reads were trimmed to equal lengths for the partitioning to work effectively. Only 454 sequencing reads of the genome were used, and these were clipped to 100 bp size fragments. Only fragments with contiguously high-quality base calls (Phred score ≥ 30) were selected. Reverse-complemented duplicates of the reads were made, and the total number of reads was indexed in a database for searches in the next step. The database contains word indices of each read partitioned into six sections (S_n, where $n = 6$):

1. S_1 - base 1 to base 15
2. S_2 - base 16 to base 31
3. S_3 - base 32 to base 47
4. S_4 - base 48 to base 63
5. S_5 - base 64 to base 79
6. S_6 - base 80 to base 100

3.1.7.2 RepeatCandyMate

This module reads the word index database constructed using iCountDBmate and uses the information to discover bridge reads. The algorithm is based on assessment of word matches on a read. Each read in the raw dataset is analyzed by the quantity and position of matching words by a sliding window method. A pair of sliding windows is used to

detect a candidate break point within pre-set margins. If a candidate break point is found, a coverage ratio of both sides of this break point is computed with the number of matching words over the total length of a read. If a pre-set threshold for this ratio is exceeded, the read is labelled a candidate bridge.

To improve the time taken to complete this search, the reads were randomised and divided into 10 runs, each run having approximately 7 million reads. These runs were searched against the database built in step Section 3.1.7.1. The following parameters were set in the definition file (def.h):

1. Maximum and minimum word length: 100
2. Index word size: 11
3. Step size: 2
4. Begin Margin: 15
5. End Margin: 21
6. Breakpoint Sum Threshold: 50
7. Breakpoint Range: 20

The preceeding algorithm and parameters were used to identify the overlaps of reads with one another. Given two reads (read A and read B), there are four categories of overlaps possible:

1. An overlap begins at the beginning of read B and end at the end of read A.
2. An overlap begins at the beginning of read A and end at the end of read B.
3. Read A is contained within read B.
4. Read B is contained within read A.

Only true overlaps between reads are found to have the above characteristics. A false overlap on the other hand will exhibit what we call False Ends and False Begins, which

occurs when the end sections of a read do not match another read when its middle section does.

The algorithm that finds these False Ends and False Begins are as follows:

1. Read begin margin (S_1), into computer memory (RAM) from _BEGIN file.
2. Each read is sequentially read from the original fasta file as a reference read. This file also contains the reverse complements of each read.
3. Using a sliding window at each corresponding middle section (S_2 to S_5) the word indices are computed on the reference read and mapped against S_1. It is faster to re-compute the word indices than to look up reference read IDs in the database, since all the rows and IDs must be read in that case. By re-computing the indices, we only need to look at the specific rows having the matching indices.
4. Read in Section S_2 to Section S_5, one at a time and search for read IDs that end in each Section.
5. S_1 is in the RAM, therefore, it is fast to check whether a read that ends in any of the following Sections also overlaps in S_1, otherwise label as a False Begin.
6. Replace S_1 with End Margin database, S_6.
7. Read in S_2 to Section S_5, one at a time and looks for read IDs that begins in each Section.
8. S_6 is in the RAM, therefore, it is fast to check whether a read that begins in any of the following Sections also overlaps in S_6, otherwise labeled as a False End.
9. All False Begins and False Ends are stored as overlap counts of the corresponding section in the reference read, as well as overlaps that are determined to be true. Total number of overlaps is determined by the number of ends and beginnings found to overlap the reference read.
10. Any read that has the number of true overlaps that is less than a set threshold (the

Breakpoint Sum Threshold) is removed from the analysis and stored in a file for investigation. Such reads may not have additional coverage information and may be a result of contamination, sequencing errors, or sequencing artifacts.

Reads that contained the false overlaps were marked as candidate bridge reads, which were then mapped against OPg3 using BLAST.

3.1.8 Consolidating the oil palm genome physical map

3.1.8.1 Fragmenting scaffolds and arrangement with MarkMyMap

I developed a program called MarkMyMap (Appendix B), that is able to fragment and re-arrange scaffolds according to information from a linkage map with molecular markers. The algorithm for fragmenting a scaffold is as follows:

1. Index the start and end positions of bridge regions on a scaffold with bridge read candidates from Section 3.1.7.2.
2. Index molecular marker positions of scaffold.
3. Identify and index the start and end positions of any N-gap filled regions present (generated during scaffold assembly process from Section 3.1.5).
4. Sort all the position index per scaffold. Each position is a potential start and end of a fragment.
5. Splits in a scaffold are made by working through the position index, starting with 1 (scaffold start) and ending at an index position (bridge or N-gap start) after at least one molecular marker position index is accounted for. The next fragment begins with the end position of either a bridge or N-gap region. Process continues until the scaffold end position is reached.

Figure 3.2 shows a representation of the process. The split will only be made if there is at least one marker available on the generated fragments. Molecular markers that fall into

Figure 3.2: Example of the scaffold splitting process of MarkMyMap, with the top block representing the original OPg3 scaffold, while the bottom 3 blocks are the fragmented result. The red regions are bridge regions, the Ns are N-gaps, and the M1 to M5 are molecular markers.

the bridge regions can be chosen to be removed together with the split (default) or not.

The first Sime Darby oil palm genome physical map used 26,240 Sime Darby Linkage Map SNPs that had been mapped in a linkage study conducted internally by Sime Darby. This is known as the Sime Darby Physical Map. I wanted to seek improvement of the physical map with the available genome scaffolds published by Singh et al. (2013). In the publication, Singh et al.'s oil palm genome scaffolds (MPOB scaffolds) had already been arranged into their MPOB Physical Map representation with their own Linkage Groups. I first consolidated the chromosome representation between Sime Darby's and MPOB's Linkage Groups (LG). The Sime Darby Linkage Map SNPs (with their 60 nt flanks) that are in LG in Sime Darby's linkage map were mapped against the MPOB's 16 chromosome scaffolds using BLAST and the MarkMyMap program. I then grouped our LGs according to MPOB's by determining the majority of Sime Darby's LG SNPs mapping to MPOB physical map scaffolds. The LG's were subsequently referred to MPOB's original indexing for ease of future reference.

The MarkMyMap program was then used to fragment OPg3 scaffolds to generate Sime Darby fragments. I then collected MPOB scaffolds that were unlinked but contained Sime Darby LG SNPs, and Sime Darby fragments according to MPOB LGs. Subsequently, I used MarkMyMap to re-order and arrange the scaffolds, and to remove redundancies (Figure 3.3), thus generating the MarkMyMap Physical Map.

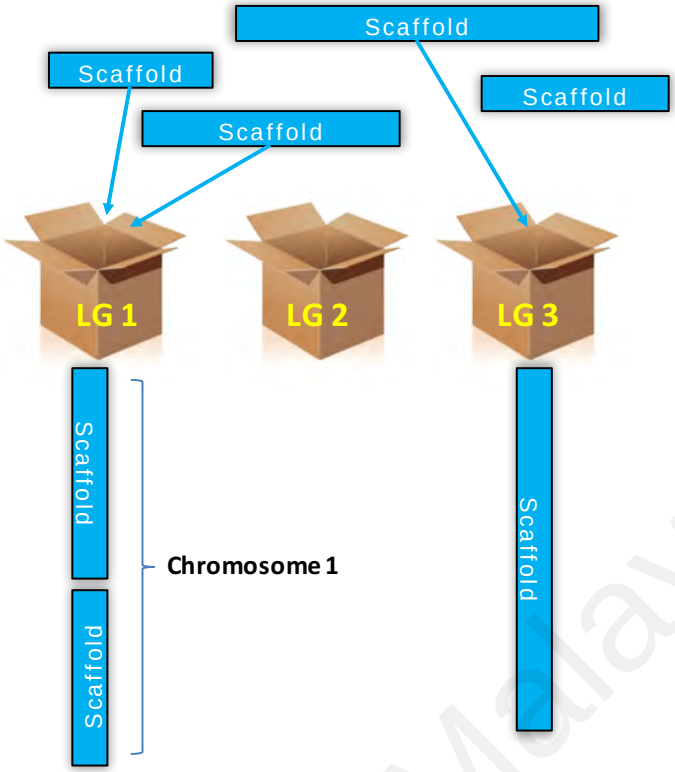


Figure 3.3: Conceptual example of scaffold fragment binning using molecular markers into collections representing available Linkage Groups (LG).

3.1.8.2 Improvement of the oil palm physical map by merging using Minimus2

I further tested a merged assembly between the Sime Darby and MPOB physical maps using Minimus2, a modified version of Minimus (Sommer et al., 2007) found in the AMOS assembly package (Treangen et al., 2011). The process of merging relies on finding overlaps between two scaffolds and fusing both together. I collected MPOB scaffolds and contigs that Sime Darby Linkage Map SNPs map in a non-redundant manner, i.e. only uniquely mapping SNPs were used, and prioritised scaffolds over contigs should a SNP matches both datasets. I then binned scaffolds from the Sime Darby oil palm genome with the rest according to LG, and run the Minimus2 program to merge the scaffolds. The base calls were optimised to prioritise Sime Darby's sequences, thus yielding the result called the Minimus2 Physical Map.

3.2 Identifying differentially expressed genes from oil palm RNA-Seq data

3.2.1 Definition of true gene count and sample coverage

I first define the true gene count as the total number of mRNA copies of a gene, in a sample prepared for a sequencing run. This definition holds for a sample containing single or multiple cells. This value cannot be known with certainty solely from the observed gene count, since the latter can, in principle, be derived from multiple different true gene counts. However, information about sample coverage can lead to more accurate estimates of the true gene count, as I shall show in this study.

3.2.2 Estimating the RNA-Seq coverage from sample concentration

3.2.2.1 Illumina's sequencing procedure

Illumina's platform is widely used for differential gene expression analyses due to its ability to sequence deeper than 454 sequencing at lower costs. Briefly, I describe six important steps during the library preparation process that contribute to variations in observed RNA-Seq gene count:

(a) ***Starting material***

For an Illumina run using TrueSeq stranded mRNA sequencing library preparation, 1 μg of total RNA is usually needed.

(b) ***mRNA isolation***

Most RNA-Seq studies are conducted on mRNA. Less than 1% of the total RNA survives mRNA isolation (poly-dT beads), including mRNA. Usually, the loss of mRNA in the wash is due to degraded mRNA, i.e. poor total RNA quality.

(c) ***Fragmentation***

This process produces approximately 500 bp long fragments. The process is followed by a size selection procedure which further increases mRNA loss.

(d) ***cDNA preparation***

The next phase is then cDNA preparation with random hexamer priming which introduces priming biases.

(e) ***PCR***

This step is needed to increase the amount of RNA. It is noted that overloading (too concentrated) the flow cell produces no results, and underloading (too diluted) can cause very skewed results. Most cases require PCR, because underloading is common in most RNA samples. Furthermore, the amount is greatly affected by the starting sample concentration, e.g. 200 ng, which is not the same for all samples. One sample may need PCR, while another does not, so doing PCR for both will introduce equal duplication events to cancel out comparison bias. However, to reduce duplication bias, the cycle is kept as low as possible, which is generally 14 cycles.

(f) ***Loading volume for sequencing***

The final product of PCR yield approximately 40 μL of 200 nM (nanoMolar). The amount then gets diluted 20,000 times to a loading amount of 120 μL for the flow cell. The 40 μL is first diluted 100 times to 2-3 nM, and then further diluted 200 times as aliquots.

3.2.2.2 What is the total mRNA found in a sample?

To identify a sample's sequencing coverage, we will need to first identify what is the size of the mRNA population to compute the sample's sequenced proportion. While it is ideal to obtain the number of total mRNA available prior to library preparation, the

biases mentioned above make it difficult, if not impossible, to allow accurate estimates of the total mRNA in the sample. I reasoned that the amount of cDNA produced at the step prior to PCR would provide us the most reliable means for computation for three reasons. Firstly, the fragmentation step causes homogeneity of the cDNA molecule sizes. Secondly, the volume and concentration after PCR is known. Finally, the number of PCR cycles is known.

3.2.2.3 What is the original amount of cDNA before PCR?

We can calculate this quantity since the cDNA molecules would have similar molecular weights after size selection ($\approx$ 500 bp). The PCR final volume of 40 μL has 200 nM (200 nmol/L) concentration of 500 bp cDNA molecules, which translates to 4.818×10^{12} cDNA molecules. Assuming complete replication efficiency, a cDNA molecule is amplified 2^{14} (16,384) times for 14 cycles of PCR. Therefore, in the ideal case where all cDNA are amplifiied, the number of cDNA before PCR is $\frac{4.818 \times 10^{12}}{2^{14}} = 294,067,382$.

3.2.3 Simulation of the fragment sampling process and the relationship between coverage and the ratio of mean to variance of observed counts

When cDNA fragments are loaded into a sequencing run, short reads are assumed to be generated randomly from the loaded cDNA fragments. Thus, a true gene count induces a probability distribution of observed gene count. To find a probabilistic model that best describes the latter, I made a series of simulations to determine the mean-variance relationship of the observed gene counts.

Consider a population of N cDNA fragments of the same length. In this study, I set $N = 300 \times 10^6$ (300M). I used the following numbers of sequenced reads (S): 150 M, 120 M, 75 M, 30 M, 3 M and 0.3 M for simulating coverages (S/N) of 0.5, 0.4, 0.25, 0.1, 0.01 and 0.001, respectively.

To simulate the process of sampling from the cDNA fragment population, I first indexed each of the N cDNA molecule from 1 to N. Next, I used the Fisher-Yates shuffle algorithm (Fisher & Yates, 1963) to shuffle the indices, creating a permutation $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_N)$. The first S elements of $\boldsymbol{\pi}$ represent the indices of sequenced fragments. For each true gene count k from 1 to 100,000, I determined the corresponding observed gene count as

$$X = \sum_{i=1}^{S} I_{(\pi_i \leq k)},$$

where I_A is the indicator function that takes value 1 when the event A is true, and 0 otherwise. A total of 2,000 iterations were made, and the mean and the variance of the observed gene counts were estimated from them.

Theoretically, the observed counts generated from this process follow a hypergeometric distribution. Thus I am able to calculate the ratio of the mean and variance (m) of the hypergeometric distribution. For a given coverage b:

$$\begin{aligned} m &= \frac{S(k/N)}{S(k/N)((N-k)/N)((N-S)/(N-1))} \\ &= \left(\frac{N}{N-k}\right)\left(\frac{N-1}{N-S}\right) \\ &\approx \frac{N}{N-S} \\ &= \frac{1}{1-b} \end{aligned} \tag{3.1}$$

given N is very large (300M), k is very much smaller than N ($\leq$ 100K) and $b = S/N$. For sufficiently small b, $m \approx 1 + b$.

3.2.4 Modelling the posterior mean and the posterior variance as functions of the coverage parameter

The posterior distribution of true count k for an observed count x was determined as in Equation 4.2. Then the relationship of both the mean and variance of the posterior distribution was identified with the coverage parameter. The mean and variance can be modelled as linear functions such that:

$$\mu = x \cdot Gm + Im; \qquad \sigma^2 = x \cdot Gs + Is,$$

where the parameters Gm and Gs are the gradients, and Im and Is are the intercepts respectively. Then, I fitted models for each of the parameters from simulations at various coverages (Figure 4.7). Equations 4.3 and 4.4 are the final approximations to model the mean and variance of the posterior distribution as a function of the observed gene count (x) and the sequencing coverage (b).

3.2.5 Evaluation of CORNAS

Based on the results from Section 3.2.3 and 3.2.4, I wrote a program called CORNAS (COverage-dependent RNA-Seq) (Appendix B) which is a Bayesian statistical test for calling differentially expressed genes (DEGs) in the case of unreplicated RNA-Seq experiments. I proceeded to evaluate how well CORNAS performed against NOISeq (Tarazona et al., 2011) and GFOLD (Feng et al., 2012). Both NOISeq and GFOLD are current methods that can be used to find DEGs from unreplicated RNA-Seq experiments.

(a) ***Program settings***

CORNAS has two parameters that can be set. The first one is α, which is used for determining the lower $(1-\alpha)/2\times 100$th percentile ($p_{(1-\alpha)/2}$) and the upper $(1+\alpha)/2\times 100$th percentile ($p_{(1+\alpha)/2}$). The second parameter is the fold-change cut-off ϕ. To make a DEG

call, we require $p^{+}_{(1-\alpha)/2}/p^{-}_{(1+\alpha)/2} \geq \phi$, where the superscript + and − indicate the posterior distribution with higher and lower mean, respectively. The default settings are $\alpha = 0.99$ and $\phi = 1.5$. These values can be changed to make CORNAS more conservative (e.g. increasing α and/or ϕ), or more liberal (e.g. lowering α and/or ϕ).

NOISeq was run with a q=0.9 cut-off. GFOLD was run with a 0.01 significance cut-off for fold changes. The expression of a gene was considered up-regulated if the GFOLD value was 1 or greater and down-regulated if the GFOLD value was -1 or smaller.

(b) ***Performance metrics***

In the context of differential gene expression analysis, a true positive (TP) is a true DEG call that is correctly flagged as being differentially expressed by a DEG method. False DEG calls are false positives (FP), while false negatives (FN) are missed true DEG calls. For a DEG call method, its positive predictive value (PPV) is the proportion of calls that are true DEG (TP/(TP+FP)); and its sensitivity is the proportion of true DEGs that are called (TP/(TP+FN)). The sensitivity and PPV of each method were jointly considered for Tests 2, 3 and 4. The F-score, which is the harmonic mean of sensitivity and PPV, was calculated for each comparison as $2 \times$ (sensitivity×PPV)/(sensitivity+PPV). The mean F-score for each method was reported.

For Test 1, the false positive rate (FPR) is determined from the no-fold change scenario as the true negatives (TN) are explicitly known (FP/(FP+TN)), while the sensitivity is calculated similarly as that in Tests 2, 3 and 4 for the weak and strong effect scenarios.

(c) ***Test 1: Detection of differentially expressed genes in simulated true gene count data***

For this simulation, I tested CORNAS using four coverages: 0.5, 0.25, 0.1 and 0.01, and three scenarios of biological effects were considered: no fold change (no effect), 1.5-fold change (weak effect), 2-fold change (strong effect). The maximum true counts considered

under these three scenarios were 10,000, 6,666 and 5,000, respectively. Each true gene count is assumed to be expressed by a gene, so that the set of all true gene counts under all three scenarios corresponded to a total of 21,666 genes. The observed counts for each gene was generated following the procedure described in the simulation of the fragment sampling process (Section 3.2.3). A total of 100 iterations were made to account for sampling variability in observed gene counts. Where gene length information is required for a particular method, we set it at 1,000 bases.

(d) ***Test 2: compcodeR simulation***

I generated the simulated data set B_625_625 according to the example provided in Soneson (2014) to create a control-treatment comparison (five replicates in each group), with 624 up-regulated genes and 625 down-regulated genes in the control group for a simulated transcriptome of 12,498 genes. From this data matrix, a total of 25 unreplicated data sets were constructed. Gene lengths were assumed to be equal and set at 1,000 bases. For CORNAS, I evaluated the outcome of two different coverages on the sample comparisons; one estimated at 10 times less than compcodeR coverage (CORNAS_10xless), and another at 100 times less (CORNAS_100xless). I made two separate NOISeq runs, one without length normalization (NOISeq_nln), and another using the trimmed mean of M-values normalization (NOISeq_tmmnl).

(e) ***Test 3: Human sex-specific gene expression***

For the Pickrell (2010) study consisting of 29 females and 25 males from Nigeria, I used the number of total sequenced reads from the published paper. The RNA-Seq count data was obtained from the ReCount database (Frazee et al., 2011). The sequencing coverage for each sample was calculated as the number of total reads reported divided by the standard 300M cDNA fragment size. For samples with more than one sequencing run, I took the

average of the total reads generated. The differentially expressed genes were identified as 19 genes with Y chromosome-related expression (Khang & Lau, 2015). Genes that are not differentially expressed on biological grounds include 61 X-inactivated (XiE) genes (Carrel & Willard, 2005; Esnaola et al., 2013) and 11 housekeeping genes (Eisenberg & Levanon, 2013).

(f) ***Test 4: Coverage effects in tissue-specific gene expression data***

In the Marioni (2008) data set, the same human liver and kidney samples were sequenced in seven lanes each, with five lanes loaded at an RNA concentration of 3 pM, and another two with 1.5 pM. The 14 lanes were sequenced in two separate runs. To reduce technical variation, I used only data from run 2, where loadings with different concentrations were run under the same conditions and time. I estimated the number of cDNA fragments representing the sample's transcriptome as the product of the loading concentration, the loading volume (assumed as standard 120 μL), and the Avogadro constant $6.022 \times 10^{23}\text{mol}^{-1}$. The set of true DEGs used was identified based on curated information extracted from the TISSUES database (Santos et al., 2015) on the 14th of June 2016. I selected 737 human kidney genes and 4,126 human liver genes that have supporting experimental validation results and are identifiable with Ensembl gene ID.

(g) ***Effect of PCR amplification efficiency on sensitivity***

The evaluation was conducted with the same dataset used for Test 1. To simulate the effect of PCR amplification efficiency in the study, I recalculated the sequencing coverages for each CORNAS run by reducing the assumed total number of fragments prior to PCR caused by different PCR amplification efficiencies (70%, 49%, 34%, 23% of total fragments for 95%, 90%, 85%, 80% amplification efficiency respectively).

Let X be a random variable that represents the proportion of DNA fragments unamplified

during PCR. Suppose we model X as a beta random variable with mean $\alpha/(\alpha+\beta)$. We can use X to model the deviation from perfect amplification by considering the random variable $2-X$. Let k be the number of PCR cycles, and N_0 the initial number of DNA fragments. Assuming perfect amplification, the number of fragments after k cycles of amplification is

$$S_p = N_0 2^k.$$

If we assume amplification efficacy in each cycle is independent of one another, then the actual number of fragments after k cycles is

$$S_a = N_0 \prod_{i=1}^{k} (2 - X_i).$$

Thus, the expected relative effect of variation in amplification efficiency is given by

$$\begin{aligned}
\mathbb{E}\left(\frac{S_a}{S_p}\right) &= \frac{1}{2^k} \prod_{i=1}^{k} \mathbb{E}(2 - X_i) \\
&= \frac{1}{2^k} [\mathbb{E}(2 - X_1)]^k \\
&= \left(1 - \frac{\alpha}{2(\alpha+\beta)}\right)^k.
\end{aligned}$$

Table 3.3: The expected proportion (mean = $\mathbb{E}(S_a/S_p)$, SD = Standard Deviation) of DNA fragments amplified by PCR under a beta model with mean $\alpha/(\alpha+\beta)$ relative to perfect amplification

α	β	$\mathbb{E}(S_a/S_p)$	$\mathbb{E}(S_a/S_p) \pm 2\mathbf{SD}$
5	95	0.70	0.64 - 0.76
10	90	0.49	0.43 - 0.55
15	85	0.34	0.29 - 0.38
20	80	0.23	0.19 - 0.27

For the variance of S_a/S_p, we have

$$\begin{aligned}
\mathrm{Var}\left(\frac{1}{2^k}\prod_{i=1}^{k}(2-X_i)\right) &= \mathrm{Var}\left[\prod_{i=1}^{k}\left(1-\frac{X_i}{2}\right)\right] \\
&= \mathbb{E}\left[\prod_{i=1}^{k}\left(1-\frac{X_i}{2}\right)\right]^2 - \left[\prod_{i=1}^{k}\mathbb{E}\left(1-\frac{X_i}{2}\right)\right]^2 \\
&= \left\{\mathrm{Var}\left(1-\frac{X_1}{2}\right)+\left[\mathbb{E}\left(1-\frac{X_1}{2}\right)\right]^2\right\}^k - \left(1-\frac{\alpha}{2(\alpha+\beta)}\right)^{2k} \\
&= \left[\frac{\alpha\beta}{4(\alpha+\beta)^2(\alpha+\beta+1)}+\frac{\beta^2}{4(\alpha+\beta)^2}\right]^k - \left(1-\frac{\alpha}{2(\alpha+\beta)}\right)^{2k} \\
&= \frac{[\beta(\alpha/(\alpha+\beta+1)+\beta)]^k - (\alpha+2\beta)^{2k}}{[4(\alpha+\beta)^2]^k}.
\end{aligned}$$

Table 3.3 gives the expected proportion of fragments under a beta model of amplification variation relative to perfect amplification. As an example, a sample that had perfect amplification but had a sequencing coverage of 0.25 would have 300M fragments prior to PCR and 75M reads produced. Supposed the reads produced remain unchanged, but the PCR amplification efficiency is now 95%, the sequencing coverage estimated will then be 0.36 (75M / (300M $\times$ 0.7)). The new coverage is then used in the 0.25 coverage CORNAS run with 95% PCR amplification efficiency. For each coverage, the FPR was calculated from the number of DEG called in the no effect scenario, and sensitivity was calculated from the DEG called from the strong effect scenario. I generated the Receiver Operating Characteristic (ROC) curves using the ROCR R package (Sing et al., 2005). The cut-offs for making a differential expression call were obtained by fixing $\alpha = 0.99$ and then varying ϕ from 1.5 to 0.75, and by fixing $\phi = 0.75$ and then varying α from 0.99 to 0.01.

3.2.6 CORNAS on oil palm male and female inflorescence unreplicated samples

(a) ***RNA extraction from samples***

Oil palm inflorescences (male and female) were sampled from *Elaeis guineensis tenera* hybrid palms (GH500 series *dura* $\times$ *pisifera*). Each tissue sample was collected on the same day from six different 20-year-old oil palms from the same estate in Carey Island, Selangor, Malaysia. These were healthy palms that were at the end of their production cycle and marked for culling. The inflorescences of each palm were sampled by removing the fronds from the felled palms followed by dissection of the stem to reach the tissues, which were then immediately placed in liquid nitrogen. The length of male and female inflorescences ranged between 3–4 cm corresponds to leaf +6 stage (Adam et al., 2005). Total RNA from these tissues were extracted and pooled as detailed in Ho et al. (2016).

(b) ***Sequencing and making counts***

The cDNA libraries were sequenced using a 454 GS-FLX Sequencing System (Roche Molecular Diagnostics, Indianapolis, IN, USA). The generated reads were mapped to EGrefseq using *bwasw* of the BWA package (Li & Durbin, 2010). Uniquely mapped reads were counted from the reference alignment using SAMtools (Li et al., 2009) in each sample dataset.

(c) ***Validation with nCounter method***

The nCounter analysis system (NanoString Technologies, Seattle, WA, USA) was used to determine the expression level of 16 sex-specific transcripts in male inflorescence and female inflorescence of oil palm. These 16 candidates were chosen by considering whether the transcript had at least 10 observed counts in the male inflorescence RNA-Seq sample and no observed count in the female inflorescence RNA-Seq sample as male-specific, while the reverse was considered as female-specific (Ho et al., 2016). In order to obtain sufficient RNA for analysis, pooled RNA samples were used (the RNA for each tissue type was extracted from six different palms and was then combined in equal amounts). A transcript was considered sex-specific when it showed a sex-predominant expression pattern (log_2-transformed fold-change ratios ≥ 0.50) between male and female inflorescences. Pre-mRNA splicing factor SLU7 and glutaredoxin genes that had been shown to stably express in different oil palm tissues (Yeap et al., 2014) were used as the housekeeping gene controls. Standard negative and positive controls were spiked into the samples according to the manufacturer's protocol. Four technical replicates for each tissue type were used. The raw counts were normalised using the geometric mean of the positive controls and the two housekeeping genes in the nSolver Analysis Software provided by NanoString Technologies (Seattle, WA, USA).

(d) ***CORNAS parameter setup***

To apply CORNAS, I work on the assumption that the coverage could never achieve > 0.01 and is therefore Poisson distributed (Section 4.2.2). This is because 454 sequencing generates far fewer reads per sample run, which therefore leads to lower expected coverages than Illumina sequencing. This lower coverage is the consequence of the length/depth limitation: longer lengths are generated at the expense of sequencing depth. Transcript expression profiles of oil palm male (Mi) and female (Fi) inflorescences were compared using CORNAS with $\alpha = 0.99$ and $\phi = 1$.

(e) ***Comparison with NOISeq and GFOLD***

I also ran NOISeq and GFOLD with the same configuration in Section 3.2.5 to benchmark against the CORNAS results.

CHAPTER 4: RESULTS

4.1 Oil palm whole genome assembly improvement

4.1.1 Sime Darby's transcriptome reference

The consensus transcriptome assembly built from various oil palm tissues, called EGrefseq, was instrumental for the rest of my analyses throughout the study. EGrefseq comprises of 60,210 non-redundant contigs (Table 4.1) which have been deposited at DDBJ/EMBL/GenBank under the accession GCKD00000000 (Ho et al., 2016).

About 45% of EGrefseq transcripts are longer than 1,000 bp. I found 37,737 and 40,162 transcripts to have significant matches against TAIR10 and RGAP 7 databases respectively. A total of 30,192 transcripts had BLAST matches in the Swiss-Prot database and 40,208 transcripts had significant matches in the KEGG database. Transcripts for all 458 CEGs were represented in EGrefseq, with 453 out of the total (98.9%) having alignments with lengths exceeding 60% of either the CEG or the EGrefseq sequence. These statistics validate the EGrefseq as likely to have good representation of the genome and as a comprehensive reference set.

4.1.2 Comparison of Sime Darby oil palm genome assemblies

In this study, I sought to improve the first draft of Sime Darby's oil palm genome (OPg1). This assembly was built using 454 sequencing reads. Additional sequencing from

Table 4.1: Overview of EGrefseq consensus assembly

EGrefseq attributes	Statistic
Number of contigs	60,210
Number of genes	38,981
Total transcriptome size (bp)	70,422,832
Longest contig length (bp)	11,413
Shortest contig length (bp)	200
Average contig length (bp)	1,169
N50 (bp)	1,652

Table 4.2: N-statistics on OPg1, OPg2 and OPg3

Genome statistics	Genome version		
	OPg1	**OPg2**	**OPg3**
Total draft size (bp)	1,680,286,271	1,819,755,685	1,779,709,065
Number of scaffolds	37,882	246,587	117,574
Largest scaffold size (bp)	21,372,121	71,844,751	71,844,751
Percentage of N bases	9	11	10
N50 (bp)	134,844	91,160	95,157

the Illumina platform was implemented in the second iteration of the draft (OPg2), while the third OPg3 version was the culmination of the technique to add exome contigs to the assembly. There is an overall improvement of the genome draft sizes with each iteration as it approach closer to the estimated DNA size of 1.8 Gbp (Table 4.2). OPg1 seemed to be under-representing the size, while OPg2 was larger than the other two. The number of scaffolds generated increased 5.5 times more from OPg1 to OPg2 with additional Illumina sequencing, but saw a reduction from OPg2 to OPg3 once the sequences were consolidated using exome contigs. With the inclusion of new paired-end sequences to consolidate the contigs, the largest scaffold size was increased to 72 Mbp in OPg2, compared to OPg1, which was just about 21 Mbp. The longest scaffold recorded for OPg2 remained the same in the OPg3 iteration. The percentage of unknown bases in the three draft genomes were only marginally different (between 9 to 11 %). The N50 of OPg1 indicates it had the best contiguous genome representation (135 Kbp), followed by OPg3 (95 Kbp) and OPg2 (91 Kbp). Interestingly, the N50 statistic fell by 30% in OPg2.

I further compared the oil palm genome drafts to evaluate their degree of completeness of their gene content. Out of 458 Core Eukaryotic Genes (CEGs), OPg1 had 182 CEGs missing (39.7%), which indicates considerably low gene representation (Figure 4.1, Supplementary Table A.3). With additional sequencing, the draft improved with a reduction in missing CEGs to 81 in OPg2 (17.7%). The addition of the exome contigs further in the assembly

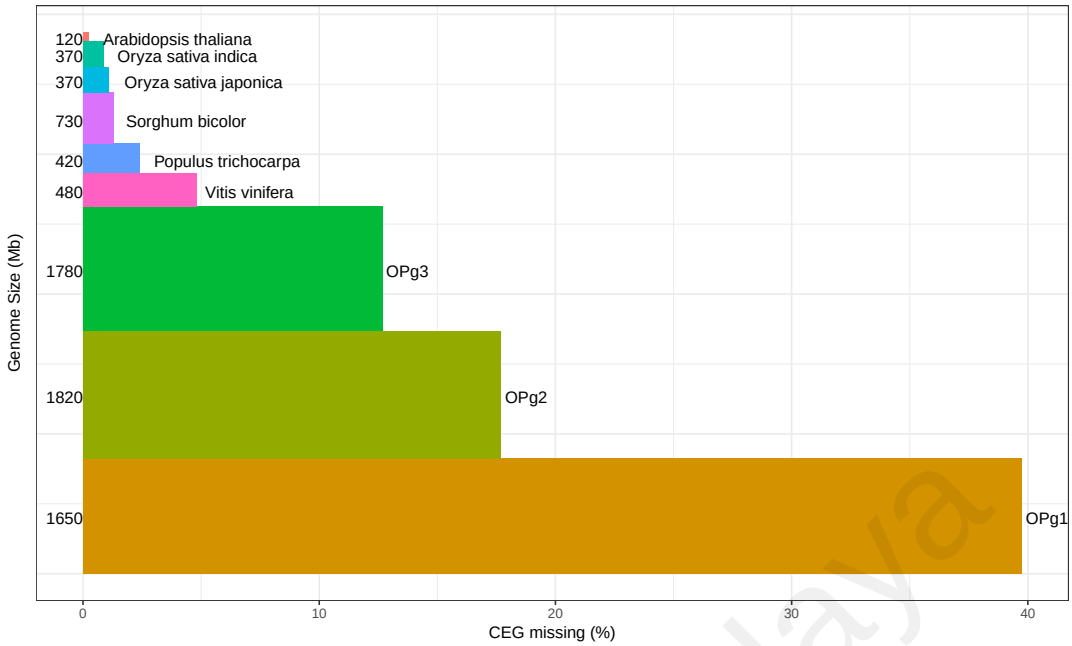


Figure 4.1: Percentage of missing Core Eukaryotic Genes (CEGs) across various genome assemblies. There are 458 CEGs in total.

decreased the missing CEGs to 58 in OPg3 (12.7%). The improvements seen in OPg3 approaches, but does not quite reach, the completeness of other plant genome references that had been published. The next plant ranked with poor CEG representation is the grape genome, *Vitis vinifera*, which has about 8% less missing CEGs compared to OPg3.

Since the EGrefseq resource is available, I was able to further determine the extent of the expression landscape covered in the oil palm genome drafts. This dataset does not limit us with only 458 highly-conservative genes. We see that EGrefseq representation improved with each iteration, from OPg1 to OPg3 (Figure 4.2); with 39% improvement from OPg1 to OPg3 of transcripts mappable at 80% of their lengths. Since publicly available ESTs were also available, I compared them to the OPg3, which exhibits the best gene representation. With the exception of the public ESTs from Ho et al. (2007), the rest of the ESTs mapped to OPg3 with comparable standards as EGrefseq (Figure 4.3).

I further assessed if the genome assemblies contain sufficient molecular marker information for genetic studies. I observed an overall 3% increase of total markers mapping

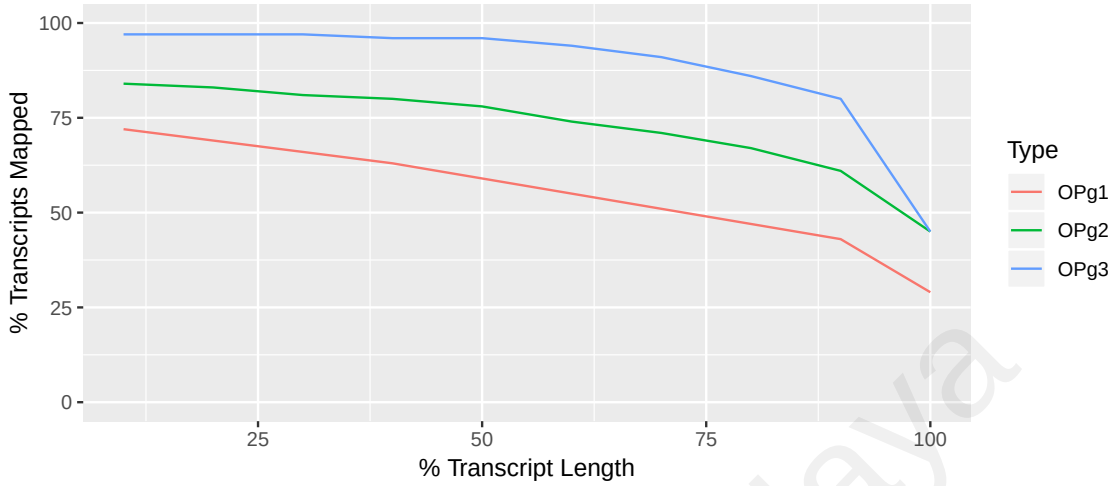


Figure 4.2: Coverage of 60,210 EGrefseq contigs against OPg1, OPg2 and OPg3

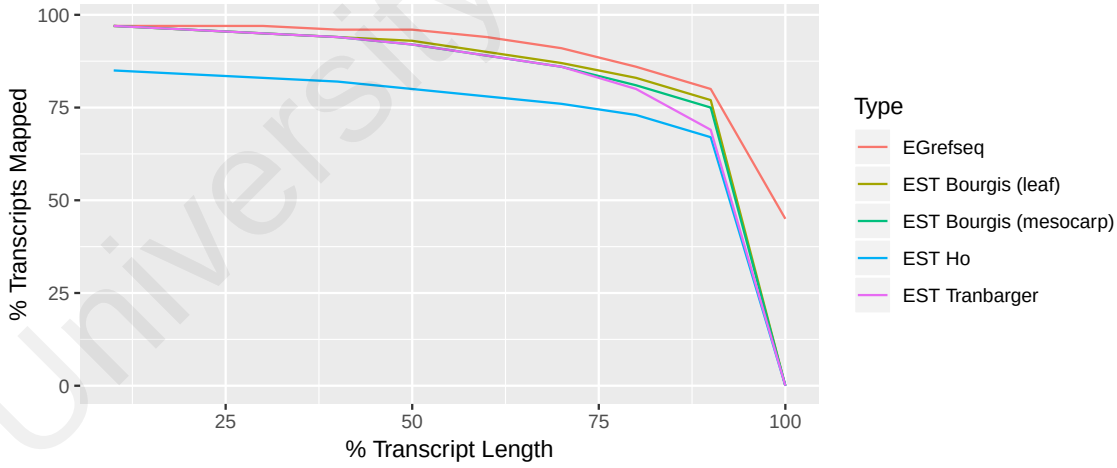


Figure 4.3: Coverage of various published oil palm transcriptome sequences over OPg3. EGrefseq consists of 60,210 consensus contigs generated from various tissues; EST Ho contained 14,537 contigs from root, shoot apical meristem, young flower, mature flower, suspension cell culture and zygotic embryo; EST Tranbarger contained 29,034 contigs from mesocarp; EST Bourgis (mesocarp) contained 33,841 contigs from mesocarp and EST Bourgis (leaf) contained 7,854 contigs from leaves.

Table 4.3: Statistics of mappable SSR and DArT markers against OPg1, OPg2 and OPg3. The number in brackets next to marker types denote the total markers used for evaluation.

Molecular Marker statistics	**Genome version**		
	OPg1	**OPg2**	**OPg3**
Billotte et al. SSRs (75)	44	58	58
Sime Darby SSRs (501)	467	428	450
Sime Darby DArT (101)	48	72	72
Total markers	559	558	580
Total % (over 677 markers)	82.6	82.4	85.7
Total uniquely mapped scaffolds	330	301	309
Total length of uniquely mapped scaffolds	70,293,744	350,954,987	352,118,750

to the genome after improvements, from 82.6% in OPg1 to 85.7% in OPg3 (Table 4.3). Except for a reduction of 22 SSRs developed by Sime Darby found in OPg2, the rest of the mapped markers were the same between OPg2 and OPg3. There was a slight reduction of 12 markers from OPg1 to OPg3. The number of uniquely mapped scaffolds had also reduced but the total length of those scaffolds became about 5 times larger in OPg3 (352 Mbp) compared to OPg1 (70 Mbp), an indication that the scaffolds became joined in the new assembly.

4.1.3 Improvement by adding exome contigs with GenSeed Pipeline Suite

I developed a program pipeline, written in Perl, (Section 3.1.6) that expands the capabilites of the program GenSeed (Sobreira & Gruber, 2008). With my program, I found 13,582 EGrefseq contigs that did not match OPg2 with identities more than 60%. Of these contigs, 95.4% (12,958 contigs) were able to seed a total of 22,652 exome contigs (Table 4.4). The general lengths of the exome contigs were 84% shorter on average compared to the largest exome contig of 18 Kbp. Also, EGrefseq contigs seemed to seed two exome contigs on average.

Table 4.4: Statistics of completed exome contigs built from EGrefseq with less than 60% match identity to OPg2

Exome contigs attributes	Statistic
Total contigs	22,652
Most contig per EGrefseq	16
Least contig per EGrefseq	1
Average contig per EGrefseq	2
Longest contig (bp)	17,770
Shortest contig (bp)	91
Average contig length (bp)	2,836

4.1.4 Bridge read detection with BridgeReader

BridgeReader was developed to detect bridge reads (Section 3.1.7). The program was written in C++ (for optimal computation) and Perl. BridgeReader identified a total of 10,152,549 candidate bridge reads in the 454 sequencing reads data set. With these annotated against OPg3 genome, 27,375 potential break sites were identified on OPg3 scaffolds. Figure 4.4 shows a visual result of annotating the genome with bridge reads. The 'NGS Reads' track consists of raw reads from Illumina sequencing. The bridge reads track, 'greads101_aln', indicates a gap in the 'NGS Reads' mapping. The two differences in coverage seen on the left and right of the bridge reads are an indication of a possible misassembled region.

4.1.5 Consolidating the oil palm genome physical map

With the latest iteration of Sime Darby's oil palm genome (OPg3), an in-house linkage map based on 26,240 SNPs, and the annotation of bridge reads, I proceeded to build Sime Darby's oil palm genome physical map. Taking advantage of the release of the MPOB oil palm genome (Singh et al., 2013), I further used their physical map to assist in improving the contiguity and completeness of genetic marker content of Sime Darby's physical map.

During the linkage group (LG) consolidation exercise, the Sime Darby's LG was able to map to MPOB's LGs with more than 90% confidence (Table 4.5). This confidence metric

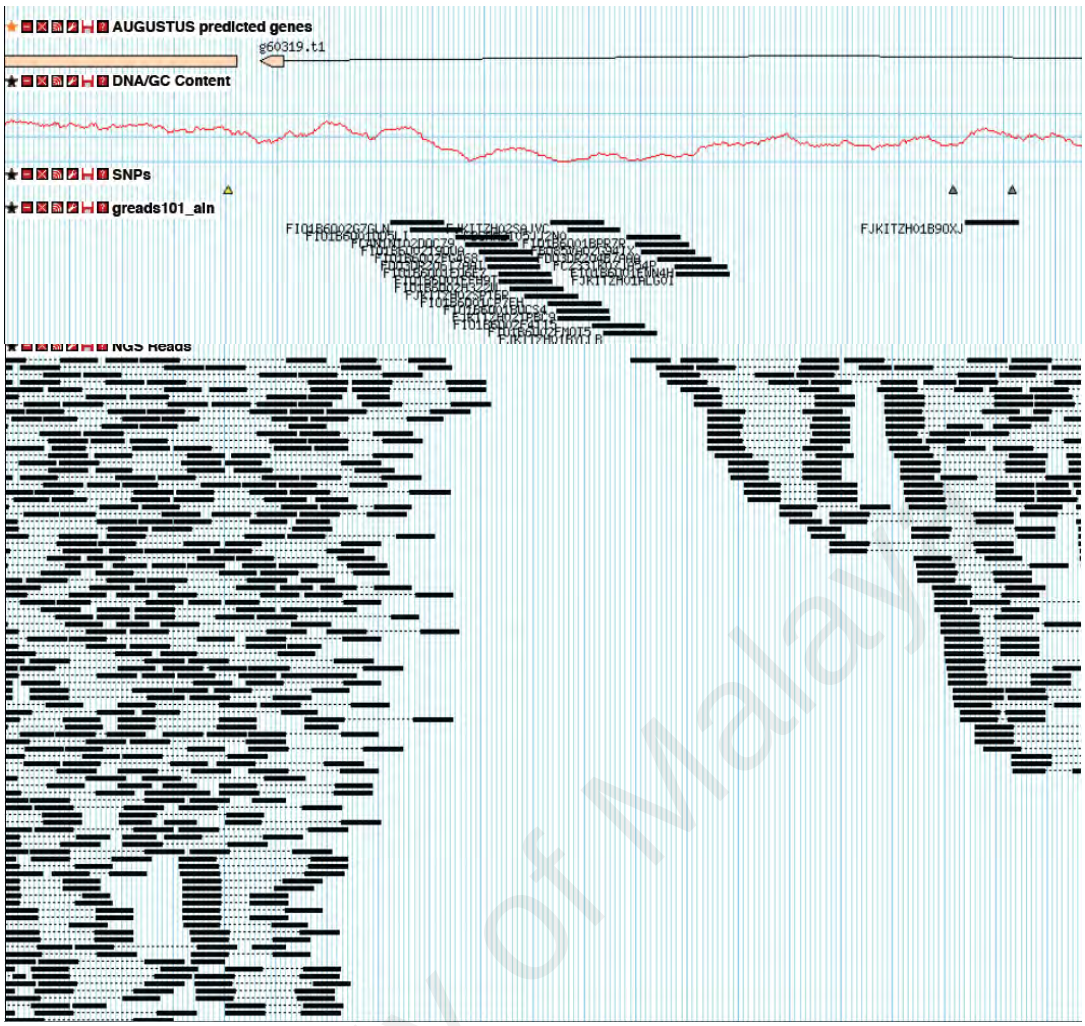


Figure 4.4: Example of bridge read annotation on OPg3

is derived from the number of SNPs I found matching between the mapped LGs, giving us the degree of similarity between Sime Darby's LG to MPOB's LGs.

I developed a Perl program called MarkMyMap (Section 3.1.8), that is able to fragment and re-arrange scaffolds according to information from a linkage map with molecular markers. The results from the MarkMyMap program yielded the largest physical map (PM) size, nearly doubling the similarly sized Sime Darby and MPOB physical maps from about 658 Mbp to 1.2Gbp (79% increase). The Minimus2 PM reduced the number of scaffolds and increased the mean scaffold lengths per LG compared to Sime Darby PM, with median improvements of 30% and 46% respectively (Supplementary Table A.2), with only two Sime Darby Linkage Map SNP markers missing (Table 4.6). However, the Minimus2 PM

Table 4.5: Linkage group (LG) comparison between MPOB and Sime Darby physical maps (PM). Matched SNPs: Number of Sime Darby LG SNPs found in MPOB PM.

MPOB PM ID	MPOB LG	Sime Darby LG	Matched SNPs	% Matched SNPs
gb|CM002081.1|	LG1	LG08	1782	99.26
gb|CM002082.1|	LG2	LG04	1547	99.42
gb|CM002083.1|	LG3	LG01	1960	98.20
gb|CM002084.1|	LG4	LG11	381	96.21
gb|CM002085.1|	LG5	LG12	705	92.40
gb|CM002086.1|	LG6	LG10	1086	99.91
gb|CM002087.1|	LG7	LG06	969	100.00
gb|CM002088.1|	LG8	LG02	1282	90.41
gb|CM002089.1|	LG9	LG07	896	92.66
gb|CM002090.1|	LG10	LG15	1428	97.81
gb|CM002091.1|	LG11	LG14	555	99.64
gb|CM002092.1|	LG12	LG13	769	98.72
gb|CM002093.1|	LG13	LG09	427	96.83
gb|CM002094.1|	LG14	LG03	766	98.21
gb|CM002095.1|	LG15	LG16	545	99.82
gb|CM002096.1|	LG16	LG05	188	100.00

Table 4.6: Comparisons between physical maps generated. Note that LG 17 is just a placeholder for unlinked scaffolds, and does not represent an actual chromosome in oil palm. There are 26,240 LM SNPs (Sime Darby linkage map) and 170,860 OP200K SNPs.

Physical map statistics	Physical map version			
	Sime Darby	MPOB	Minimus2	MarkMyMap
Total size (bp)	657,211,498	657,968,836	733,454,586	1,177,588,601
Number of LG	17	16	17	16
Number of scaffolds	25,285	16	16,930	4,383
Mean size (bp)	25,992	41,123,052	43,322	268,671
Number of LM SNPs	26,240	16,330	26,238	22,641
Number of OP200K SNPs	91,562	82,854	88,176	113,210

only had relatively smaller difference in the number of CEGs represented, from 43.2% in Sime Darby PM to 32.3% in Minimus2 PM (Supplementary Table A.3), and I have found evidence of assembly errors (Figure 4.5). The MarkMyMap PM, on the other hand, had a reduction of 3,599 Sime Darby linkage map SNP markers (LM SNPs) compared to Sime Darby PM, but had the highest recovery of Sime Darby's OP200K SNPs (Kwong et al., 2016) with 21,648 additional informative SNPs (Table 4.6). Furthermore, MarkMyMap PM had the lowest amount of CEGs missing at 9.2% (Supplementary Table A.3).

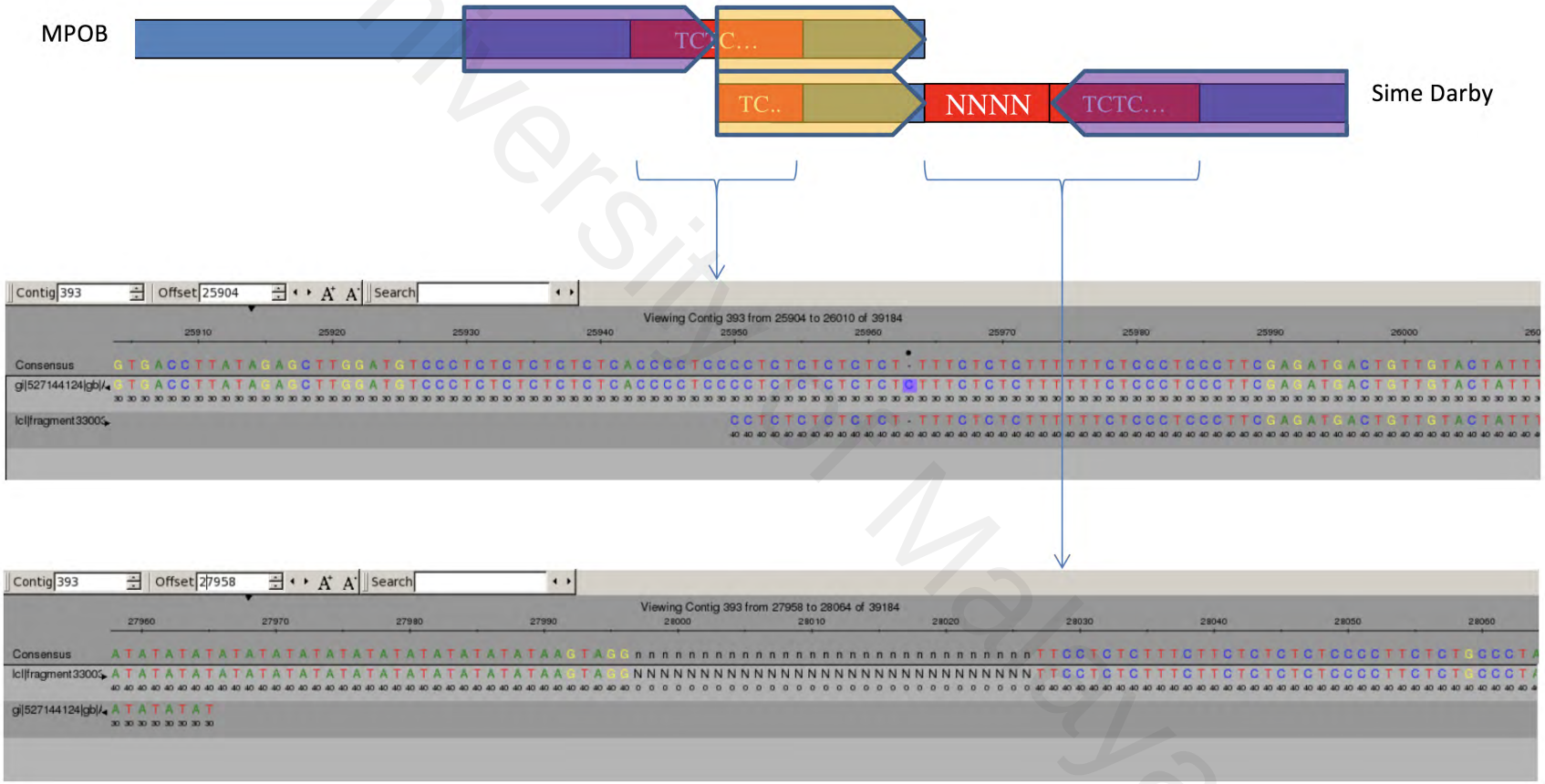


Figure 4.5: Example of a Minimus2 physical map merge error between MPOB scaffolds and Sime Darby scaffolds

4.2 Identifying differentially expressed genes from oil palm RNA-Seq data

4.2.1 The true coverage of RNA-Seq experiments

The coverage of a sample (b) is defined as the number of cDNA fragments sequenced (S) divided by the total cDNA fragment population size (N). Single-end sequencing produces one read to represent one cDNA sequenced, while paired-end sequencing produces two reads to represent one cDNA sequenced. The calculation of sample coverage in the context of the Illumina sequencing protocol can be based on mRNA sample concentration. The amount of cDNA produced at the step prior to PCR provides the key to a reasonable estimate of sample coverage because: 1) the fragmentation step during sample library preparation causes homogeneity of the cDNA molecule sizes (500bp); 2) the volume and concentration after PCR is known (40 μL of 200nM cDNA) and; 3) the number of PCR cycles is known (14 cycles). The cDNA fragments undergo PCR to improve the chance of getting at least a sequencing coverage of one. Assuming perfect amplification efficiency, each cDNA fragment is amplified 2^{14} times during PCR. Thus, I calculated the number of cDNA fragments prior to PCR as $\approx$ 300 M. I used this quantity as the estimated N to determine coverage, since it most closely resembles the mRNA amount we expect to start off with.

4.2.2 Chance mechanism generating a Generalised Poisson distribution for observed gene counts

I simulated the sampling effect that occurs in the sequencing run to find a probabilistic model that best describes the probability distribution of observed counts. For each coverage, I generated an empirical distribution of the observed counts for true count values ranging from 1 to 100,000.

The simulation results provided three important observations: the mean of observed counts is proportional to the coverage, underdispersion occurs (i.e. variance less than mean)

with increasing coverage (Figure 4.6), and the relationship between the mean-variance ratio and coverage can be described adequately with a linear model (Eq. 3.1). The sampling process naturally leads to a hypergeometric distribution of the observed counts because N is finite. However, N is large and unknown in practice, hence the need for an approximating distribution that does not have an upper bound. These results suggest that the generalised Poisson (GP) distribution (Consul & Jain, 1973) is suitable for modelling the distribution of observed gene counts (X) given a true gene count (T). The probability mass function of the GP with parameters λ_1 and λ_2 is given by

$$P(X = x|T = k) = \frac{\lambda_1(\lambda_1 + x\lambda_2)^{x-1}e^{-(\lambda_1+x\lambda_2)}}{x!}, \tag{4.1}$$

where $x = 0, 1, 2, \ldots$, $\lambda_1 > 0$, and $|\lambda_2| < 1$. Its mean and its variance are given by

$$\begin{aligned} \mathbb{E}(X|T) &= \lambda_1/(1-\lambda_2), \\ \mathrm{Var}(X|T) &= \mathbb{E}(X|T)/(1-\lambda_2)^2, \end{aligned}$$

implying that $\lambda_2 = 1 - \sqrt{m}$, where m is the mean-variance ratio ($0 < m < 4$). The mean of the observed gene count given the true gene count is proportional to the product of the coverage b and the true count k, giving $\lambda_1 = bk\sqrt{m}$. The Poisson distribution with mean λ_1 is a special case of the GP when $m = 1$.

4.2.3 A Bayesian model for estimating true gene counts given observed gene counts and sequencing coverage

The importance of the GP model in Eq. 4.1 stems from the fact that reverse conditioning enables us to consider the probability distribution of the true gene count (T) given an observed gene count and sequencing coverage (i.e. the posterior distribution of the true gene count). Let us assume a uniform prior distribution for T over values of $1, 2, \ldots$.

Application of Bayes Theorem yields:

$$P(T = k|X = x) = \frac{P(X = x|T = k)}{\sum_{j=x}^{\infty} P(X = x|T = j)}$$
$$= \frac{k(bk\sqrt{m} + x(1 - \sqrt{m}))^{x-1} e^{-bk\sqrt{m}}}{\sum_{j=x}^{\infty} j(bj\sqrt{m} + x(1 - \sqrt{m}))^{x-1} e^{-bj\sqrt{m}}}, \quad (4.2)$$

where $k \geq x$. Note that although an improper prior was used, the resulting posterior distribution is proper. Interestingly, the gamma distribution provides a good approximation to Eq. 4.2 (see Khang (2016) for mathematical proof). Here, I found that the approximation is excellent if the mean μ and the variance σ^2 of the gamma distribution relates to the coverage b and the observed gene count x as (Figure 4.7):

$$\mu \approx \frac{x+1}{b} - \left(1 + \frac{1}{2b}\right)^{-1}; \quad (4.3)$$

$$\sigma^2 \approx \frac{x+1}{[b(b+1)]^2}. \quad (4.4)$$

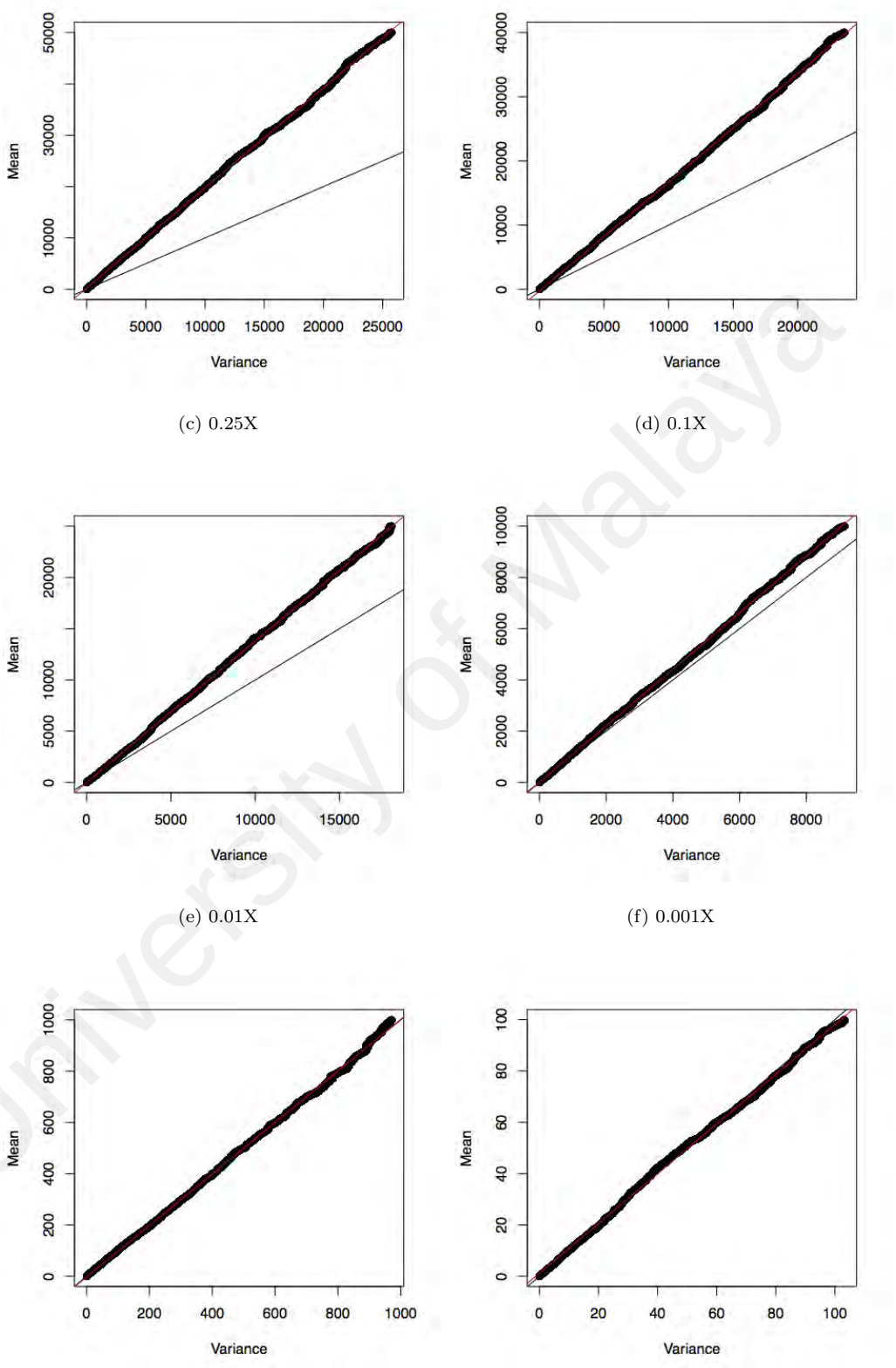
Thus the probability density of the approximating gamma distribution is given by

$$f(k|x) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} k^{\alpha-1} e^{-k/\beta}, \quad (4.5)$$

where $k \geq 0$, $\alpha = \mu^2/\sigma^2$ and $\beta = \sigma^2/\mu$. The approximation provides a computationally efficient means to calculate the cumulative distribution function of the posterior distribution of the true gene count.

(a) 0.5X

(b) 0.4X



(c) 0.25X

(d) 0.1X

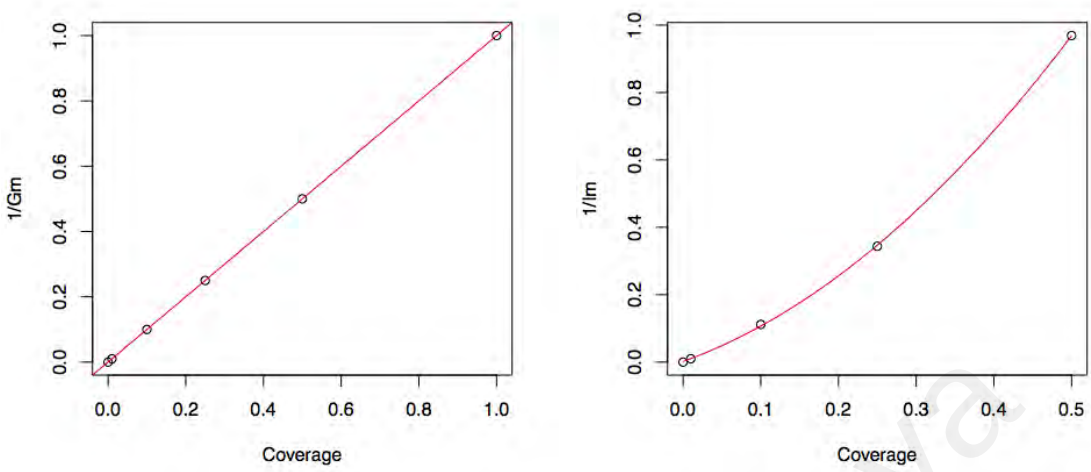
(e) 0.01X

(f) 0.001X

Figure 4.6: Mean vs variance of observed counts in 2,000 replicates for the following coverages (a): 0.5X, (b): 0.4X, (c): 0.25X, (d): 0.1X, (e): 0.01X, (f): 0.001X. The black line is where mean is equal to variance. The red line is the fitted linear model.

(a) Gm

(b) Im



(c) Gs

(d) Is

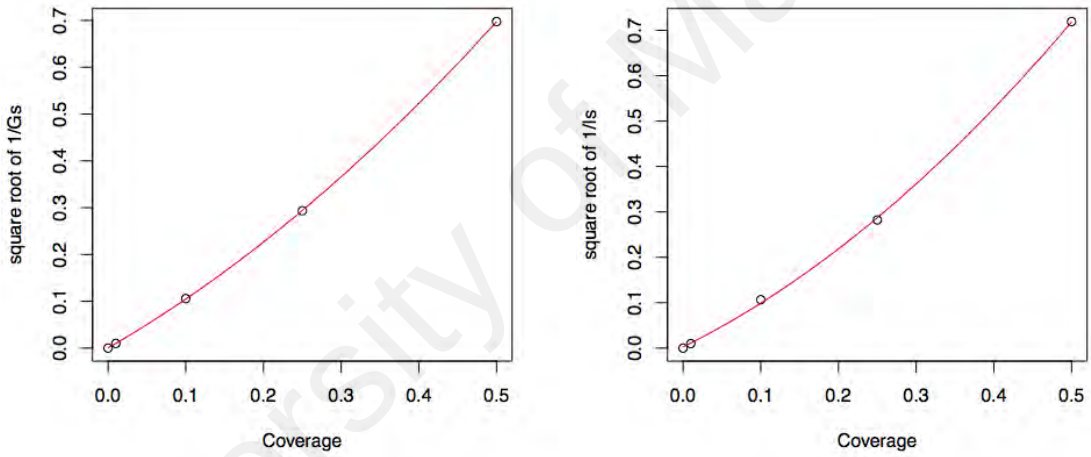


Figure 4.7: The relationship of the sequencing coverage with the slope and intercept parameters of linear models of the posterior mean and posterior variance; where (a): Gm, (b): Im, (c): Gs, (d): Is are respectively modelled in the equations of Section 3.2.4 and Section 4.2.3. The open circles represent the simulated data used to estimate the model.

(a) A DEG (b) Not a DEG (c) Not a DEG

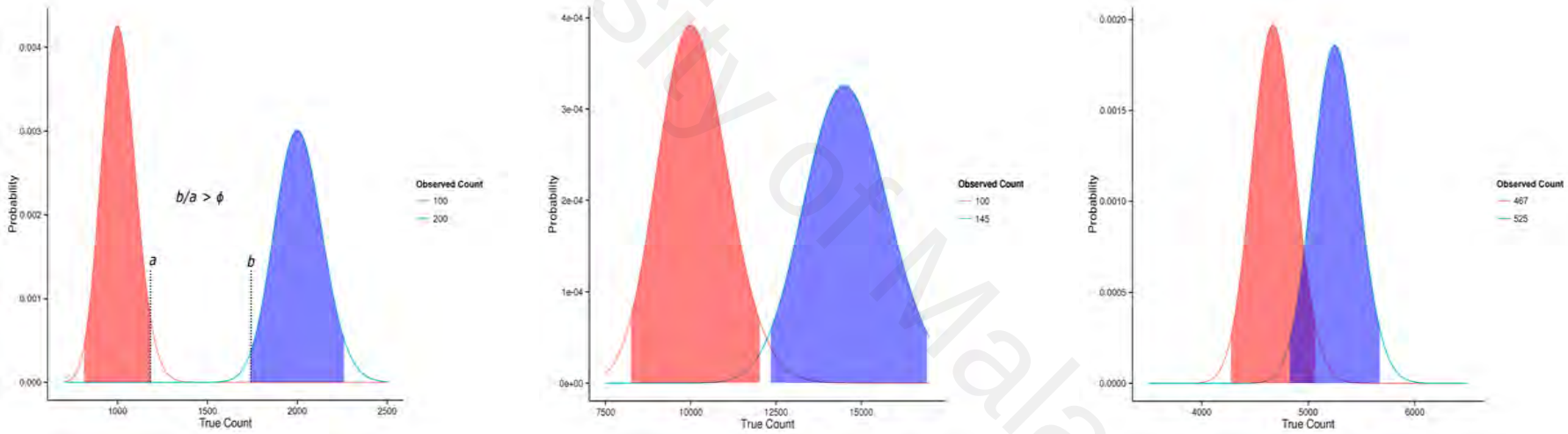


Figure 4.8: Illustration of how DEG calls are made in CORNAS. By default, the fold-change (ϕ) is 1.5.

4.2.4 COverage-dependent RNA-Seq (CORNAS)

A statistical test for calling differentially expressed genes in the case of unreplicated RNA-Seq experiments can be based on the posterior distribution of the true gene count (Eq. 4.2) as follows: For a single control and a single treatment sample, if we have information about sequencing coverage for the control sample (b_0) and the treatment sample (b_1), then, given the observed gene count for the control (x_0) and the treatment (x_1) group, the posterior distribution of their true gene count is approximately gamma (Eq. 4.5). A gene is declared to be differentially up-regulated in the treatment group if the latter has a larger posterior mean, and its 0.5th percentile is at least 1.5 fold (default) larger than the 99.5th percentile of the control group. Conversely, a gene is differentially down-regulated in the treatment group if the latter has a smaller posterior mean, and its 99.5th percentile is at least 1.5 fold (default) smaller than the 0.5th percentile of the control group (Figure 4.8). This procedure is fast because the percentiles of the gamma distribution are easily computed. Furthermore, declaring genes to be differentially expressed using this procedure implies there is a $0.995^2 \approx 0.99$ probability that the true gene count in the two samples differ by at least 1.5 fold. I call this algorithm CORNAS (COverage-dependent RNA-Seq) and wrote the program to perform CORNAS analysis in R (Low, Khang, & Tammi, 2017).

4.2.5 Performance evaluation of CORNAS

I conducted a series of tests comparing the performance of CORNAS against NOISeq (Tarazona et al., 2011) and GFOLD (Feng et al., 2012) using both simulated and real data sets. I chose GFOLD and NOISeq, because both have been reported to return relatively small number of false positives among the genes flagged as differentially expressed when applied to unreplicated RNA-Seq data sets compared to other popular methods such as DESeq2 and edgeR (Khang & Lau, 2015).

Table 4.7: The mean F-score calculated for each method for Test 2, Test 3 and Test 4 cases

Method	F-score
Test 2	
GFOLD	0.31
NOISeq_tmmnl	0.30
CORNAS_100xless	0.30
CORNAS_10xless	0.30
NOISeq_nln	0.28
Test 3	
GFOLD	0.51
CORNAS	0.45
NOISeq	0.22
Test 4	
CORNAS	0.36
GFOLD	0.31
NOISeq	0.19

(a) ***Test 1: Detection of differentially expressed genes in simulated true gene count data***

I tested CORNAS, NOISeq and GFOLD on simulated true gene counts for the scenario of no-fold change (no effect), 1.5-fold change (weak effect) and 2-fold change (strong effect) between control and treatment. The false positive rate (FPR) was estimated as the DEG call rate in the scenario of no-fold change. The true positive rate (TPR), or sensitivity, is the DEG call rate in the weak and strong effect scenarios.

In general, a decreased false positives and increased DEG call rates with increasing coverage and increasing number of true gene counts were observed (Figure 4.9). Compared to GFOLD and CORNAS default, NOISeq produced the largest FPR when true gene counts are low. NOISeq's sensitivity is generally good except at low coverage of 0.01; its DEG call rate begins to fall when true counts are over 1,000. GFOLD showed very low sensitivity, which is consistent with its conservative behaviour reported in (Khang & Lau, 2015). CORNAS showed excellent control of FPR and a dependence on the fold change threshold for detecting DEG under weak and strong signal scenarios. For example,

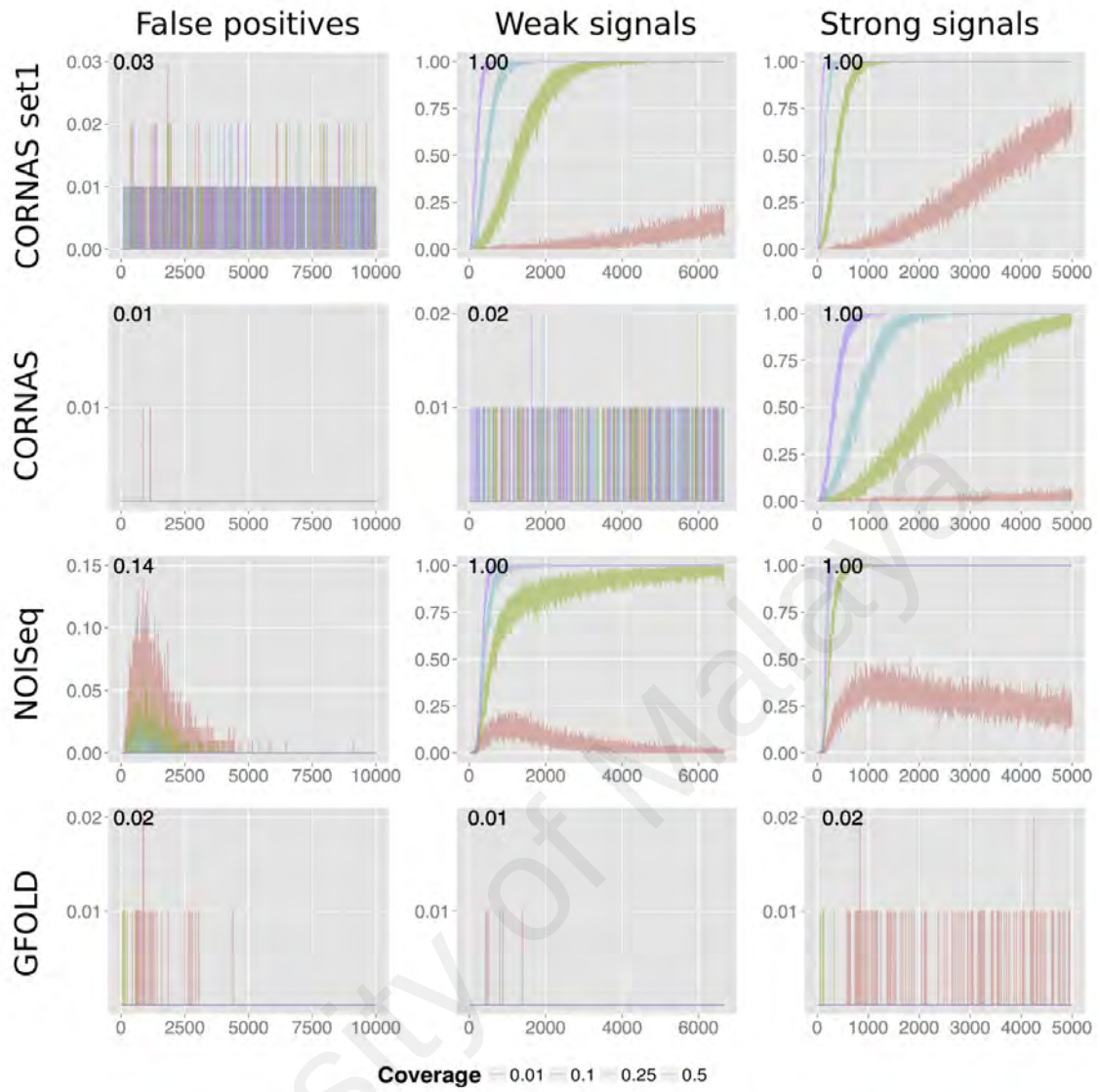


Figure 4.9: DEG detection using simulated true count data. The Y-axis is the proportion of DEG called in 100 replicates. The X-axis is the true count of Sample 1. Comparison is made against Sample 2, which either has the same (False positives), 1.5 times more (Weak signals), or 2 times more (Strong signals) true counts. The numbers at the top left of each plot denotes the Y-axis maximum. The maximum true counts for false positive, weak signal and strong signal conditions are 10,000, 6,666 and 5,000 respectively. CORNAS set1 refers to CORNAS with $\phi = 1$, while CORNAS refers to the default $\phi = 1.5$.

CORNAS default ($\phi = 1.5$) performed very poorly under the weak signal scenario, so that if the detection of such genes are of interest, then ϕ should be adjusted to a lower value such as 1 (CORNAS set1). In general, the sensitivity of CORNAS increases with larger true count, and converges to 1 quickly for coverage values of 0.1 or more.

(b) ***Test 2: compcodeR simulation***

The distribution of observed gene counts is popularly modelled using the negative binomial distribution, and the compcodeR R package (Soneson, 2014) provides a simulator for simulating RNA-Seq count data based on this distribution. I generated a dataset of 12,498 genes 10% DEGs to be tested.

Positive predictive value (PPV) and sensitivity were low for all methods; nonetheless, CORNAS showed relatively greater sensitivity than the other methods, whereas GFOLD had relatively better PPV (Figure 4.10a). The F-scores for all methods were very similar (Table 4.7). CORNAS called a larger DEG set size compared to other methods. Unlike NOISeq_nln, the larger DEG set size called by CORNAS did not substantially reduce its PPV. Both CORNAS_100xless and CORNAS_10xless showed similar performance.

Average runtimes for the comparisons were about three minutes for NOISeq_nln and NOISeq_tmmnl, one minute for GFOLD, and three seconds for CORNAS_10xless and CORNAS_100xless.

(c) ***Test 3: Human sex-specific gene expression***

The evaluation of the applicability of CORNAS on real data is based on the human lymphoblastoid cell RNA-Seq data set from Pickrell's study (Pickrell et al., 2010). In this data set, male and female gender constitute the two phenotype classes, so the true DEG can be determined purely using biological reasoning using sex-specific genes.

I randomly chose 100 single female-single male pairs from a total of 725 possible pairs (29 females, 25 males), and compared the performance of GFOLD, NOISeq and CORNAS. The results indicated that NOISeq performed poorly compared to CORNAS and GFOLD, while GFOLD performed slightly better than CORNAS (Figure 4.10b, Table 4.7). However, similar to the compcodeR simulation result, CORNAS called larger DEG sets.

Average runtimes were about two minutes for NOISeq, thirty seconds for GFOLD and ten seconds for CORNAS.

(d) ***Test 4: Coverage effects in tissue-specific gene expression data***

The Marioni data set (Marioni et al., 2008) consists of RNA-Seq data from human liver and kidney sequenced at two different loading concentrations, 3 pM (high) and 1.5 pM (low). I used a set of 4,863 genes identified to be uniquely expressed in either human liver or kidney tissues catalogued in the tissue expression database, TISSUES (Santos et al., 2015) as DEGs.

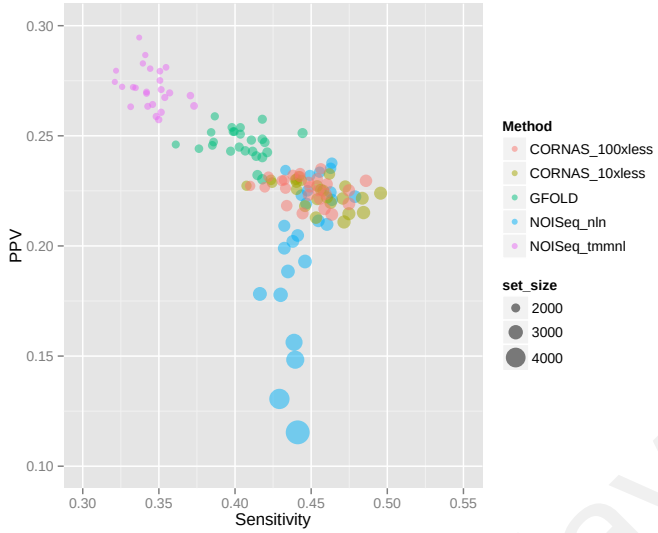
I investigated whether CORNAS would be misled into making DEG calls simply on the basis of differing concentration, when both samples are taken from the same tissue. False positive rates were low in CORNAS, with no DEG calls made for comparisons within the same tissue samples with equal concentrations (Table 4.8). However, for samples with different concentrations, GFOLD showed fewer false positives than CORNAS. In all instances, NOISeq returned the highest FPR.

For DEG evaluation, NOISeq again performed poorly compared to CORNAS and GFOLD, while CORNAS performed the best (Figure 4.10c, Table 4.7). For all 12 comparisons between different tissue types, the largest DEG sets were called by CORNAS, and the smallest ones by NOISeq.

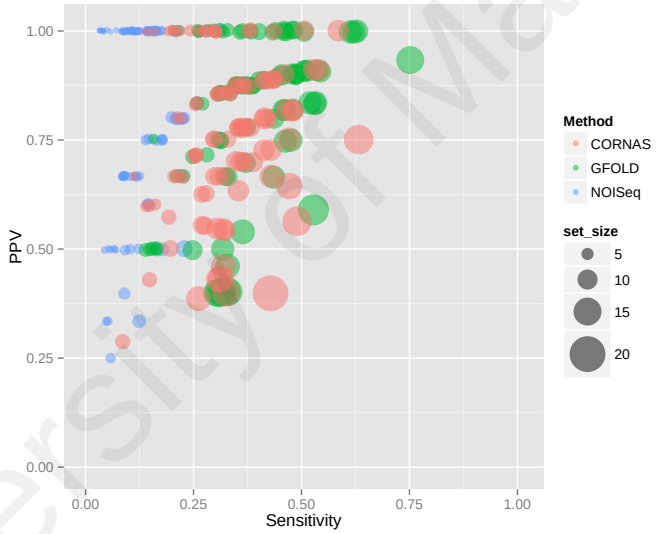
Generally for different tissue types, the DEG sets called by NOISeq and GFOLD showed poor overlap, compared to overlaps between GFOLD and CORNAS, and between NOISeq and CORNAS (Figure 4.11). CORNAS indicated more unique DEG calls for different tissue types. At the same time, a large percentage of DEG calls from GFOLD or NOISeq were also called by CORNAS.

Average runtimes were about five minutes for NOISeq, thirty seconds for GFOLD, and five seconds for CORNAS.

(a) compcodeR simulation



(b) Human sex-specific gene expression



(c) Human tissue-specific gene expression

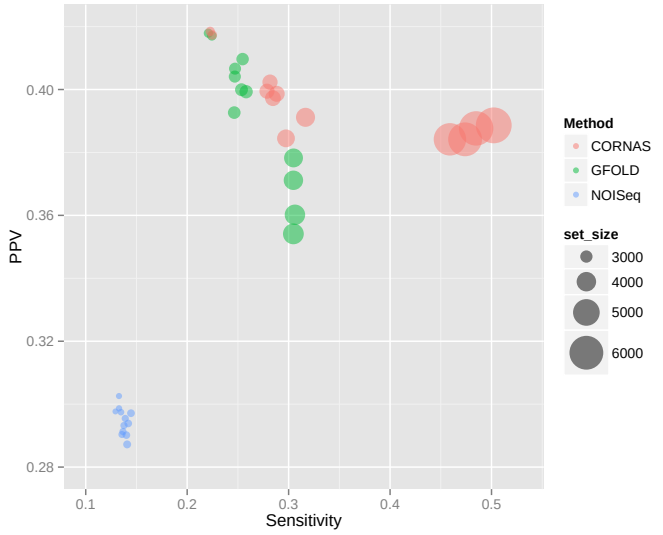


Figure 4.10: Scatterplots of PPV against sensitivity. The size of each dot is proportional to the DEG set size.

Table 4.8: DEG calls made by NOISeq, GFOLD and CORNAS between two samples from Marioni's data. The sample combinations consisted of two human tissue types (Liver and Kidney) with two loading concentrations, 3 pM (high) and 1.5 pM (low).

Concentration	Type	Sample A	Sample B	NOISeq	GFOLD	CORNAS
low vs low	same tissue	R2L4Kidney	R2L8Kidney	275	0	0
high vs high	same tissue	R2L2Kidney	R2L6Kidney	333	1	0
low vs high	same tissue	R2L4Kidney	R2L2Kidney	329	0	42
low vs high	same tissue	R2L8Kidney	R2L2Kidney	335	0	29
low vs high	same tissue	R2L4Kidney	R2L6Kidney	356	1	124
low vs high	same tissue	R2L8Kidney	R2L6Kidney	325	0	82
low vs high	same tissue	R2L1Liver	R2L3Liver	324	0	105
low vs high	same tissue	R2L7Liver	R2L3Liver	308	1	46
low vs low	same tissue	R2L1Liver	R2L7Liver	307	1	0
low vs high	different tissue	R2L4Kidney	R2L3Liver	2347	2616	2588
low vs high	different tissue	R2L8Kidney	R2L3Liver	2288	2570	2619
high vs high	different tissue	R2L3Liver	R2L2Kidney	2366	2972	3761
high vs high	different tissue	R2L3Liver	R2L6Kidney	2348	3051	3937
low vs low	different tissue	R2L1Liver	R2L4Kidney	2113	3143	3484
low vs low	different tissue	R2L1Liver	R2L8Kidney	2135	3083	3517
low vs high	different tissue	R2L1Liver	R2L2Kidney	2285	4185	6000
low vs high	different tissue	R2L1Liver	R2L6Kidney	2273	4134	6284
low vs low	different tissue	R2L7Liver	R2L4Kidney	2202	2956	3392
low vs low	different tissue	R2L7Liver	R2L8Kidney	2163	3022	3405
low vs high	different tissue	R2L7Liver	R2L2Kidney	2283	3993	5810
low vs high	different tissue	R2L7Liver	R2L6Kidney	2385	3918	6083

(a) CORNAS vs GFOLD

(b) CORNAS vs NOISeq

(c) GFOLD vs NOISeq

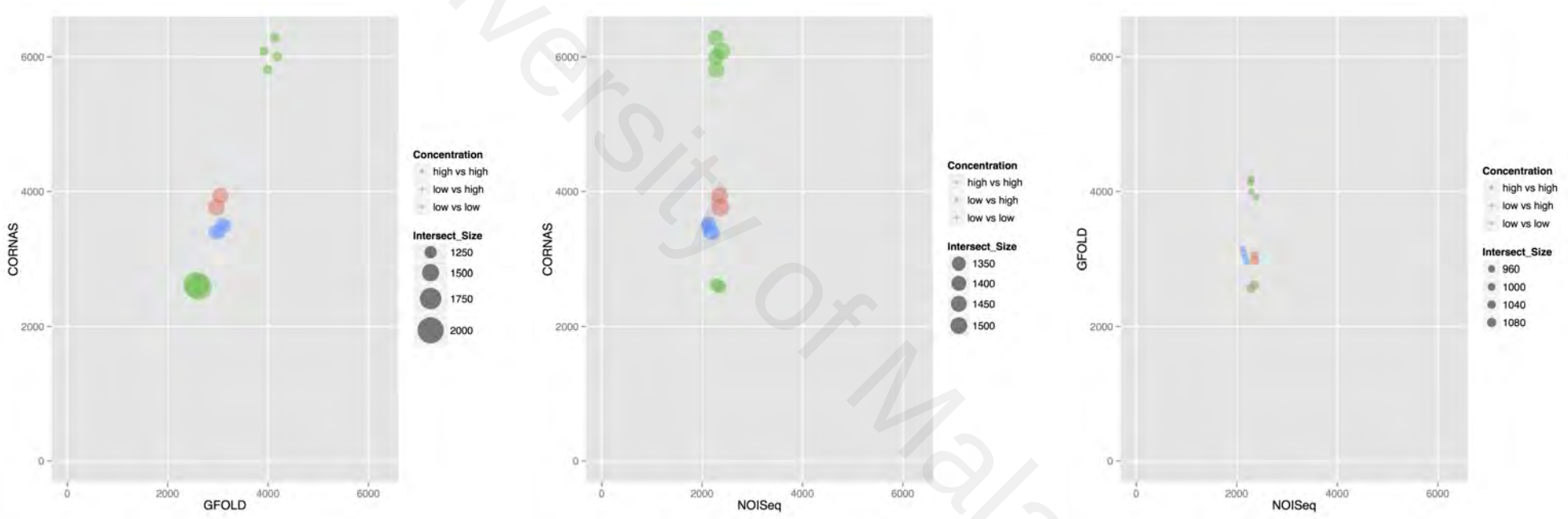


Figure 4.11: DEG set agreement between methods in analysing 12 comparisons between three human liver and four kidney samples. The axes represents the number of DEG called for each method, while the circle size approximates the intersect size. Two types of sample loading concentrations were used, 3 pM (high) and 1.5 pM (low). Details can be found in Table 4.8.

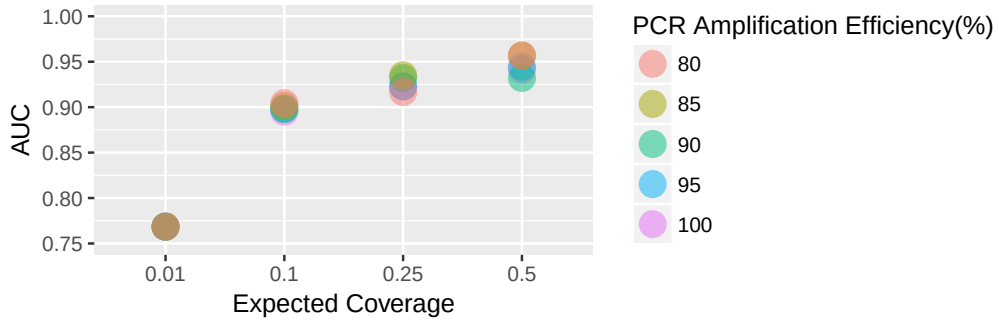


Figure 4.12: The area under the curve (AUC) of Receiver Operating Characteristic (ROC) analysis for CORNAS runs on data simulated to have 100% 95%, 90%, 85% and 80% PCR amplification efficiencies. The Expected Coverages are the original coverage estimate at 100% PCR amplification efficiency (0.5, 0.25, 0.1 and 0.01).

(e) ***Effect of PCR amplification efficiency on sensitivity***

While I assumed perfect PCR amplification efficiency in building the model, the possible effects of 95%, 90%, 85% and 80% PCR efficiencies on the sensitivity and FPR of CORNAS were still evaluated. CORNAS appeared to be robust to small violation of perfect PCR amplification efficiency, as no substantial changes to sensitivity and FPR, even at 80% PCR efficiency, were found. The area under the curve (AUC) of the Receiver Operating Characteristic (ROC) graphs of all four tested expected coverages had less than 5% difference (Figure 4.12 and 4.13).

4.2.6 Application of CORNAS in the analysis of unreplicated transcriptomes of male and female oil palm inflorescences

I proceeded to use CORNAS to detect DEGs in the oil palm RNA-Seq data set which consists of 454 platform sequenced reads of oil palm male and female inflorescence tissues mapped against EGrefseq. CORNAS made 1,218 DEG calls for comparison of male and female inflorescence transcriptomes.

To validate the candidate DEGs obtained using CORNAS, a probe-based method, nCounter analysis system (NanoString Technologies, Seattle, WA, USA), was used to validate the expression levels of 16 sex-specific transcripts in male inflorescence and

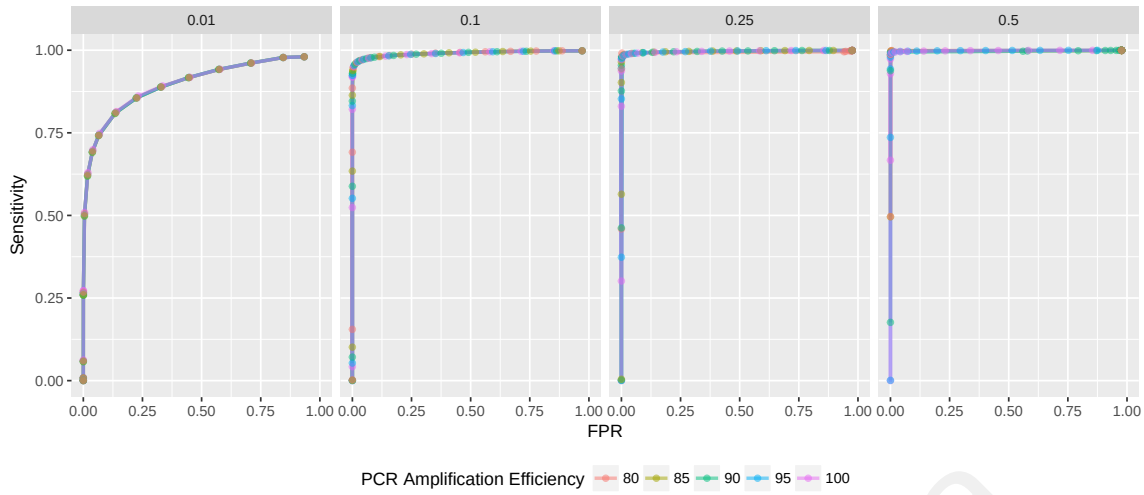


Figure 4.13: CORNAS sensitivity against false positive rates (FPR) for data simulated to have 100% 95%, 90%, 85% and 80% PCR amplification efficiencies, facetted according to the expected coverage estimates at 100% PCR amplification efficiency (0.5, 0.25, 0.1 and 0.01).

female inflorescence of oil palm found in the CORNAS results (Section 3.2.6). The validation results showed that four transcripts had significantly higher mean abundance in the female inflorescence, four were more abundant in the male inflorescence and eight did not show significant difference between male and female inflorescences. Among the 16 transcripts, isotig40710 (putative DEFICIENS), isotig53408 (putative acid phosphatase), isotig59228 (unannotated) and isotig67634 (unannotated), were more highly expressed in male inflorescence in comparison with female inflorescence; whereas isotig23091 (putative TASSELSEED1), isotig28587 (unannotated), isotig40414 (putative bZIP transcription factor) and isotig54309 (unannotated), were more highly expressed in female inflorescence in comparison to male inflorescence (Figure 4.14).

With the validated expression values of the 16 sex-specific transcripts, I evaluated the PPV and sensitivity of CORNAS, GFOLD and NOISeq. I found that CORNAS had the highest F-score (0.3), followed by NOISeq (0.12) and then GFOLD (0.1), which respectively had PPV/sensitivity scores of 0.5/1, 0.22/0.25 and 1/0.1. Table 4.9 shows the DEG call rates for the 16 sex-specific transcripts; GFOLD made only one significant

call, while NOISeq had nine significant calls of which only two matched the nCounter validation results. CORNAS identified all 16 transcripts as DEG.

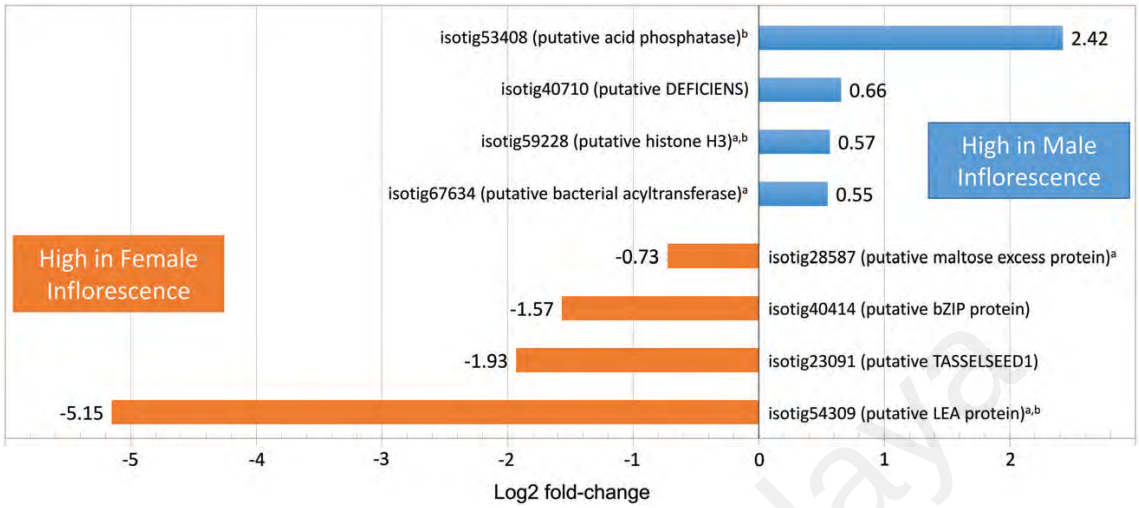


Figure 4.14: Differential expression levels of sex-specific transcripts in male and female inflorescences of oil palm. Values represent log2-transformed fold-change ratios of relative expression between male and female inflorescences. Only the transcripts that have log_2-transformed fold-change ratios ≥ 0.50 were shown. ^{a} Indicates putative function based on ORF prediction. ^{b} Indicates that the transcript also showed inflorescence-specific expression.

Table 4.9: DEG calls comparison for 16 sex-specific transcripts. CORNAS, GFOLD and NOISeq called DEG high either in the female inflorescence (Fi) or male inflorescence (Mi) similarly (High in). While all 16 were called significant DEG by CORNAS (Y), only the ones noted with a "*" is significant in GFOLD or NOISeq (see Section 3.2.5) and in our nCounter validation (nCounter high).

Gene Name	High in	CORNAS DEG	GFOLD value	NOISeq q	nCounter high
isotig23091	Fi	Y	0.3	0.5	Fi*
isotig40414	Fi	Y	0.9	0.8	Fi*
isotig54309	Fi	Y	0.6	0.8	Fi*
isotig28587	Fi	Y	0.5	0.7	Fi*
isotig63768	Fi	Y	0.9	1.0*	Fi
isotig69051	Fi	Y	0.5	1.0*	Mi
isotig58939	Mi	Y	-0.7	0.9*	Mi
isotig40710	Mi	Y	-0.4	0.7	Mi*
isotig53408	Mi	Y	-0.4	0.8	Mi*
isotig67634	Mi	Y	-0.7	1.0*	Mi*
isotig59228	Mi	Y	-0.4	0.9*	Mi*
isotig49719	Mi	Y	-0.5	0.7	Mi
isotig70480	Mi	Y	1.0*	1.0*	Fi
isotig70872	Mi	Y	-0.7	1.0*	Fi
isotig42197	Mi	Y	-0.5	0.9*	Fi
isotig70874	Mi	Y	-0.5	1.0*	Fi

CHAPTER 5: DISCUSSION

5.1 Oil palm whole genome assembly improvement

5.1.1 Sime Darby oil palm genome assemblies comparison

The measure of scaffold sizes is a common statistic used by many researchers to evaluate the quality of a genome assembly (Bradnam et al., 2013). The standard statistic used is typically the N50, which is the minimum scaffold length in a draft assembly needed to cover 50% of the total genome length, which also means 50% of the genome is made out of scaffolds with N50 lengths or longer (International Human Genome Sequencing Consortium, 2001). Consequently, it is believed that the larger the N50, the better the genome assembly is likely to be. It is expected that as more information is available to assist in long-range contiguous sequence construction, such as large-insert paired-end libraries, the improved sizes of the scaffolds generated should increase the N50. However, this is not necessarily so as the N50 is only the median of scaffold lengths. The results in Section 4.1.2 show that the largest scaffold of OPg1 is smaller than OPg2 and OPg3, yet it has the highest N50 score. This is likely due to OPg2 and OPg3 having only a few large scaffolds generated, so that their scaffold length distribution is skewed to the right.

Different sequencing platforms have specific biases and limitations. Having reads generated from another different sequencing technology on the same organism allows one to overcome the shortcomings of another technology. In OPg2, Illumina paired-end reads with large inserts was used to improve the assembly because it was more cost-effective than 454 sequencing. However, the short reads generated by the Illumina sequencing technique produce ambiguity in matching corresponding contigs for scaffolding (Pop & Salzberg, 2008). This may account for the few scaffold size improvements in OPg2 that did not push the N50 higher than anticipated.

Therefore ideally, we should employ additional sequencing technologies that produce longer reads such as PacBio and Nanopore to further improve the oil palm assembly (Levene et al., 2003; Branton et al., 2008; Bleidorn, 2016). Additional large insert mate pairs, such as BAC (Bacterial Artifical Chromosome) end sequencing could potentially further improve the scaffold sizes. Such techniques have been deployed for improving the rice (Kawahara et al., 2013), medicago (Tang et al., 2014) and even the MPOB oil palm genome (Singh et al., 2013) (Supplementary Table A.1). However, conducting such projects is costly, and potentially leads to marginal improvements as this study seemed to indicate with each iteration of the Sime Darby oil palm genome.

While the N50 may be useful for measuring the contiguity of the assembly, this statistic is insufficient in informing useful completeness of the genome assembly. A genome draft may have a high N50, but could exhibit incomplete genomic content, e.g. missing or incomplete genes, or even be incorrectly represented in misassemblies caused by overly aggressive joining of contigs (Salzberg & Yorke, 2005). Therefore, it is important to evaluate genome assembly quality by an evolutionary-based metric with the identification of highly conserved genes expected to occur in the organism under study. This is the case for the Sime Darby oil palm genome assemblies, where OPg1 exhibited a higher N50 than the other versions, but has very poor gene representation.

I used CEGMA (Parra et al., 2007) with its reference gene set of 458 genes (CEGs) that are highly conserved in six eukaryotic species (*Arabidopsis thaliana, Homo sapiens, Drosophila melanogaster, Caenorhabditis elegans, Saccharomyces cerevisiae, and Schizosaccharomyces pombe*) to evaluate their presence and completeness in the oil palm genome assembly. This gene set is conserved across the abovementioned species because these genes most likely comprise the bulk of housekeeping genes known. In comparison to CEGMA, BUSCO (Simão et al., 2015) defined reference gene sets separately by

phylogenetic clades. Each gene set per clade consists of genes that are expected to be found as a single-copy in the genome; 3,023 genes for vertebrates, 2,675 for arthropods, 843 for metazoans, 1,438 for fungi and 429 for eukaryotes. At the time this study was conducted, the plant database was still under development. Once released, the gene set could provide a better assessment of assembly quality for the oil palm genome with greater sensitivity compared to CEGs. I believe plant polyploidy may be a reason it is difficult to generate a robust BUSCO gene set and apply it. Identifying single-copy genes when the actual genome consists of heterogeneous alleles can be difficult with our current sequencing technologies to differentiate true single-copy genes from chimeric, mis-read or pseudo-gene results (Cai et al., 2012). Thus, using CEGMA to ascertain genome assembly quality of plant genomes such as wheat (hexaploid) may be a better alternative. However, this may not be an evident disadvantage for the diploid oil palm genome, since the ploidy level is not high. In any case, the CEGMA authors had conceded that future evolutionary-based evaluation of genome assembly quality should use BUSCO, as they had discontinued CEGMA code maintenance due to the lack of funding. Since 2017, the latest version of BUSCO (version 3) now has plant databases, but are split into seven lineages: Embryophyta, Viridiplantae, Chlorophyta, Embryophyta, Liliopsida, Eudicotyledons and Solanaceae (Waterhouse et al., 2017).

While we see an improvement of the quality of the genome assembly by an evolutionary-based criteria from OPg1 to OPg3, OPg3 is not yet complete, with about 12% of the CEGs still missing. Understandably, the genome statistics show that OPg3 still has 10% unknown bases in the genome, which translates to roughly 180,000,000 bases unaccounted for. The information available in the 454 and Illumina sequencing may have been exhausted with the application of the exome contig process (Section 3.1.6). Therefore, it is likely that we will need to use a combination of PacBio sequencing to get long sequences, and

polish the results with deep Illumina sequencing reads to further improve the assembly. Such a technique was recently applied to the grass, *Oropetium thomaeum*, to achieve a chromosome-scale assembly (VanBuren et al., 2018).

Besides using CEGMA, I also evaluated the gene space with published oil palm expression data. Ideally, a complete reference genome should contain all the codes for expressed mRNA. Therefore, a reference transcriptome, such as EGrefseq, can be used to identify gaps in the gene content of the genome assembly. Yet, the reference transcriptome may not cover exactly every gene possible, since RNA-Seq captures the mRNA expression as snapshots in space and time; which means that different tissues (space) have different gene sets expressed at different development stages (time). I sought to reduce this limitation by ensuring many different tissue types were represented in EGrefseq (mesocarp, root, leaf, meristem, male inflorescence and female inflorescence), with some associated with different time points (mesocarp). Besides that, I further utilised published ESTs in my evaluations. I had used EGrefseq contigs to improve OPg3 in Section 3.1.6, which comprises of RNA-Seq results from multiple mature oil palm tissues, so we would expect the results of EGrefseq mapping to OPg3 to be good. With at least 73% mapping 80% of their lengths, Ho et al.'s ESTs (Ho et al., 2007) had the least transcripts mapped to OPg3 compared to the other EST data. This result is not surprising as their ESTs were generated from embryonic callus and thus, would have very different genes expressed compared to the other tissues. The regions where these genes are in the genome may not have been captured during the exome contig assembly process in Section 3.1.6. Ho et al.'s EST dataset also contained many unannotated sequences and more strings of unidentified bases compared to the other public ESTs, which also probably contributed to the poorer mapping results.

Besides the improvement in gene content for the latest iteration of Sime Darby's oil palm

genome draft (OPg3), we also saw an improvement in the molecular marker representation available. There was a slight decrease in the Sime Darby SSR markers in OPg3 compared to OPg1, and this is likely due to these primers being specifically designed from the OPg1 draft. The new assembly parameters used in Section 3.1.5 for OPg2 may have collapsed these SSRs, since they are technically tandem repeats. This assembly step also reduced the number of uniquely mapped scaffolds but increased the total length of those scaffolds, an indication of scaffolds joined in the new assembly. The slight improvement of Sime Darby SSR recovery from OPg2 to OPg3 may be attributed to the addition of the exome contig step in Section 3.1.6.

The metrics I used to evaluate the improvements in the genome assembly are not the only ones. Competitions such as the Assemblathon 1 and Assemblathon 2 had set the benchmarks in evaluating the results of assemblers (Earl et al., 2011; Bradnam et al., 2013). In Assemblethon 2, real world datasets from various sequencing platforms were used to evaluate the capabilities of various assembly techniques. Twenty one teams took part in the challenge to assemble a bird (*Melopsittacus undulatus*), a snake (*Boa constrictor constrictor*) and a fish (*Maylandia zebra*) genome. There were ten key metrics that the authors of Assemblathon 2 had used to evaluate the results, of which four had been covered in this study. The other metrics require additional information that I do not have, such as validated fosmid regions (VFRs) and optical maps. VFRs are 1 to 40 Kbp sized contigs assembled from genome fragment inserts of bacterial F-plasmid cloning vectors, while optical maps provide genome structural information (Dimalanta et al., 2004; Valouev et al., 2006). The authors generated VFRs and optical maps for each of the three species as references to compare against the assemblers' results. These methods do have their caveats however, with VFRs relying on the fosmid sequencing and assembly to be accurate, and optical maps limiting evaluation to scaffolds larger than 300 Kbp. Another metric, called

the REAPR analysis, measures errors in assembly by evaluating how paired-end reads map against the genome drafts (Hunt et al., 2013). I was unable to use this measure, due to restrictions in the data access for the paired-end reads generated at the time.

5.1.2 Improvement by adding exome contigs with GenSeed Pipeline Suite

The GenSeed Pipeline Suite I developed expands the capabilities of the program GenSeed for use with this study's unique dataset. It automated the progressive assembly of missing genome contigs that are arguably vital, since these are parts of the genome that are transcribed to RNA. With this program, I was also able to speed up the process by introducing parallelisation capabilities over SGE, as GenSeed can take quite long to complete due to the iterative nature of the seed-driven assembly. Interestingly, not all the EGrefseq contigs seeded assemblies with genome reads. This may yet be due to unsequenced regions of the genome, as we do see an inadequate amount of polishing possible with the current dataset. We also cannot rule out the possibility of transcriptome misassemblies in the EGrefseq, which can happen due to splicing variants that complicate the assembly process. One can imagine this seed-driven assembly process building large contigs, but once the contig reaches a region containing repeats, the seeding process will stall (Figure 5.1). The limitation of length of the short genome reads ($\approx$ 300 bp) makes it difficult for the programs used for finding matches (BLAST) and assembly (CAP3) to complete accurately. This may have led to the generation of more contigs than seeded on average in our results in Section 4.1.3. Alves et al. (2016) released an updated version of GenSeed in 2016 that uses HMM profiles in the seeding process, as well supports a greater number of assemblers for the task. This update improved the overall speed and sensitivity of the program, and can be replaced into my pipeline suite with minimal modifications.

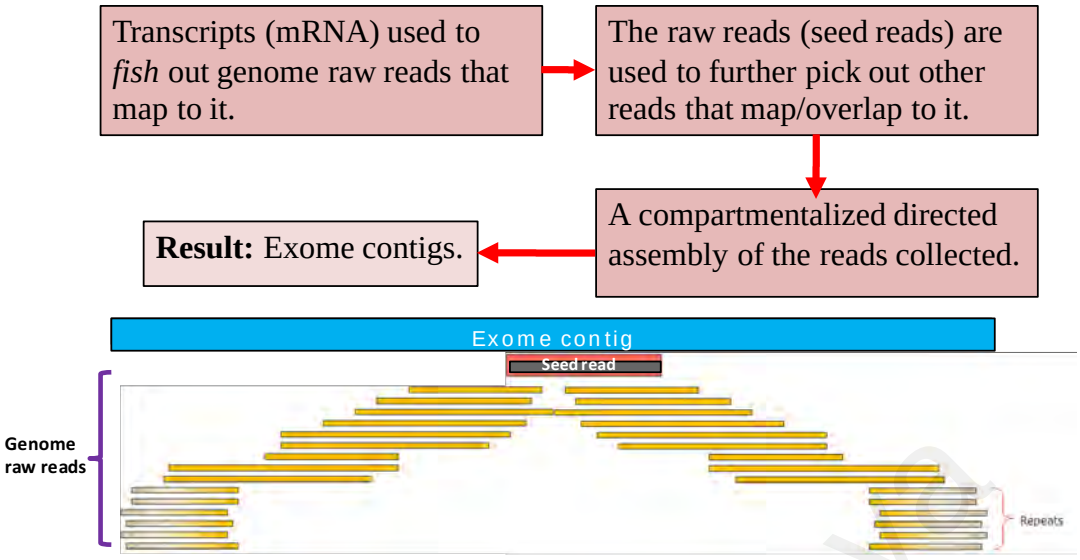


Figure 5.1: An exome contig built with the GenSeed pipeline

5.1.3 Bridge read detection with BridgeReader

The annotation of bridge reads allows the detection of possible misassemblies due to repeats. This allows us to use these regions as candidate breakpoints for rearranging the scaffold correctly with the aid of loci-based information, such as molecular markers.

In order to analyse overlaps, each read needs to be aligned to every other read. Due to the huge number of reads generated by the whole genome shotgun project, the alignment step is expected to take longer time or more resources than desired. One rapid way is to compare the sequences by using a k-mer sliding window method, where a unique integer for each k-mer is assigned and all overlaps are determined by solely using k-mer indices. As the preceding process is sequential, the data can be read and written to a hard disk after each read is processed, thus requiring a minimal amount of computer memory (RAM). Ideally the position of each k-mer is stored, but this potentially puts pressure on computer storage space. Nevertheless, stored on the hard disk, the database of k-mer positions would significantly increase the computing time in the next step due to slower read speeds of hard disks compared to RAM. Therefore, each and every position is not stored, but the read is

partitioned into a number of sections and only each k-mer is assigned into an appropriate section given by its position on a read. This way, a lot of computing time is saved in subsequent modules because the data does not contain positional information for each k-mer; instead the database can now be partitioned into corresponding sections. While this approach is not as accurate as base-by-base indexing, it attains a good tradeoff between speed and accuracy.

The problem with non-identical repeats is the false collapsing or expanding of the overlap region, which leads to incorrectly assembled regions. A sliding window approach allows the search to account for sequencing error in those determined to have false overlaps. The sensitivity and specificity of the approach depend on: (i) the sequencing error rate; (ii) the k-mer length and the resolution of the sectioning; and (iii) the step size of the sliding window. In my study, only the 454 sequencing reads were used, as they provided longer reads for the partitioning compared to the Illumina paired-end dataset. This partitioning method increases the speed in which we can run the analysis by parallelisation, but does have the potential drawback of increased false positive rates. The program parameters offer some control over false positive rates, but some prior knowledge on the actual genome sequencing coverage and the length of the repeats will be needed to assist in controlling the method's sensitivity.

While BridgeReader was being developed and used in this study, a great deal of work on other methods to study repeats had been published. BridgeReader is primarily designed to identify bridge reads, which are reads that contain a unique part and a repeat part, to be removed from assemblies. With that goal in mind, a consensus repeat contig was not implemented in the pipeline, and therefore differs from *de novo* methods such as RECON and RepeatScout. RECON and RepeatScout principally tackles the problem of defining repeat boundaries in sequences but take pairwise alignments as input, thus requiring

additional data preparation while BridgeReader uses only reads directly.

ReAS uses reads directly to identify repeats and even uses a sliding window with a fixed k-mer length (default used is 17 mer). In BridgeReader, a similar approach is used, but I further segmented the 100 bp fragment into smaller fragments in order to narrow down the identification of repeat boundaries. This segmentation of reads also means BridgeReader is not limited to requiring reads of more than 500 bp to identify repeats, unlike ReAS.

For the Sime Darby oil palm genome, BridgeReader was used to determine bridge reads in order to map and annotate the draft with potential sites of missasemblies. Consequently, I did not annotate these repeats in this study.

5.1.4 Consolidating the oil palm genome physical map

During the time this study was conducted, Singh et al. (2013) had sequenced and published the MPOB oil palm reference genome. MPOB's published oil palm physical map consists of only 45% of the genome successfully scaffolded into 16 chromosomes. With the aid of this public dataset and the MarkMyMap program, I successfully increased Sime Darby's oil palm physical map scaffold representation to 67%. Although Singh et al. used similar sequencing technologies as Sime Darby, the one notable addition in the MPOB genome sequencing project was that they conducted BAC-end sequencing. With it, they were able to recover larger scaffolds than Sime Darby's oil palm genome, with an N50 that is 100x larger. The Sime Darby physical map is twice as large as the MPOB physical map, but 100 times more fragmentary. As the results seem to indicate, the Sime Darby oil palm genome would benefit in having BAC end sequencing conducted to improve the PM representation.

Minimus2 was used as the program of choice as it was able to work with large scaffolds, implement low-level optimisation and does not require an additional reference genome to merge between two drafts. However, I found that the merging process using

Minimus2 yielded less optimal results than the non-redundant combination method with a total assembly size that was not substantially larger than either Sime Darby's or MPOB's physical map. Merging also did not produce more favourable informative SNP representation compared to the MarkMyMap results. This is probably due to the sequences originating from different palms.

The MPOB oil palm genome is of the highly inbred *pisifera* variety, while the Sime Darby oil palm genome is a *tenera*; the latter being a hybrid cross between a *pisifera* and a *dura* (Noh et al., 2012). This means the Sime Darby oil palm genome is likely to have greater heterogeneity in content compared to the MPOB oil palm genome, and therefore contain greater complexity for the assemblers to deal with. As a genome reference, the Sime Darby oil palm genome may contain a more comprehensive set for use in genetic studies, as well as having more relevance to commercial application since the *tenera* type is the productive variety of palms. However, the MPOB genome may yet exhibit less errors caused by assemblies, and therefore have greater reliability.

The poor Minimus2 result may be due to large-scale variations that confound the merging algorithm. I found that the simpler non-redundant combination employed in my MarkMyMap program worked better. The results showed an increase in the physical map total size and representation of informative Sime Darby SNPs. While MarkMyMap consolidated scaffolds based on non-redundancy of Sime Darby molecular markers and careful fragmentation based on annotated bridge reads, we may still be retaining gene content redundancies that were erroneous. Even with the fragmentation steps undertaken during the process, I have yet to ascertain if the fragments generated have high degree of homology with other fragments. The program assumed that the assemblies did not produce these redundancies, which can be a wrong. Conducting a more comprehensive genome comparison of the physical maps would give us better insight, but for this study, it is

sufficient to observe that the MarkMyMap program can be used to consolidate assembled scaffolds effectively with molecular markers.

Chromonomer (Catchen & Amores, 2016) is a C++ program that was designed with similar objectives with MarkMyMap. It was released for public use at a time MarkMyMap had already been developed for the work in this study. Chromonomer currently implements the latest formats dealing with marker alignments, and is capable of producing visualisations for the results generated.

5.2 Identifying differentially expressed genes from oil palm RNA-Seq data

5.2.1 CORNAS as a framework for estimating the true gene count

Currently, the mapped read depth over a gene model of an organism is used to estimate coverage in RNA-Seq experiments. This is in fact not true, as the true coverage should be based on the actual amount of mRNA in a sample, and therefore related directly with sample concentration. This therefore leads us to dispel the assumption that transcriptome sizes can be similar from one sample to the next, and that similar sequencing depth can be applied for a comparative study. We know that the total amount of mRNA in a sample is not captured in Illumina sequencers, which have a fixed finite saturation amount that can over- or under-represent sample concentrations. The coverage is generally accepted as an under-representation, a limitation that is usually thought to be rectifiable by deep sequencing, which is used to detect genes that have very low mRNA expression (Blencowe et al., 2009; Haas et al., 2012; Liu et al., 2014). The coverage parameter (between 0 and 1) in CORNAS should cover most practical cases where deep sequencing is not done. Outside of this range, the usefulness of CORNAS is unclear.

A completely efficient PCR amplification process is one of the several simplifying assumptions used in constructing the model. PCR is required to improve the chance of one cDNA to be picked for sequencing by having it copied ten thousand times. The chance of

picking the original amount for each cDNA species prior to PCR should be very high if the dilutions are perfectly homogenous after PCR. If we work with the assumptions above as the ideal case, we are able to calculate the coverage of sequencing, based on the number of sequenced reads obtained over the number of cDNAs available before PCR. The effect of PCR amplification efficiency I simulated does indicate that the sensitivity and FPR increases when I over-estimate the coverages, but the difference is not adversely significant.

The assumption of ideal random cDNA fragment sampling in the current work was made in order to keep the observed count model (hence the posterior distribution) sufficiently simple for us to study the effect of introducing the coverage parameter into the DEG call procedure. Since real RNA-Seq experiments contain library preparation biases, the effect of such biases may be better explored by full sequencing process simulators such as rlsim (Sipos et al., 2013).

The GP model is being increasingly studied as an alternative to the negative binomial distribution in RNA-Seq count data modelling (Srivastava & Chen, 2010; Li & Jiang, 2012; Zhang et al., 2014; Wang et al., 2015). In my simulations, I found the GP as a suitable model for observed gene count data. By relating the parameters of GP to the true gene count and sequencing coverage using RNA sample concentration, I was able to determine the posterior distribution of the true gene count. This distribution forms the basis for making DEG calls in unreplicated RNA-Seq experiments.

A potential source of variation in the observed gene count that was not explicitly handled in my simulation concerns the way different algorithms map the short reads to a reference genome (e.g. using BWA (Li & Durbin, 2009), OSA (Hu et al., 2012), TopHat (Trapnell et al., 2012) and Bowtie (Langmead & Salzberg, 2012)), and how such mapped reads are quantified (e.g. using HTSeq (Anders et al., 2015), and Cufflinks (Trapnell et al., 2012)). I suggest that variation in the observed gene count due to this source of variation

is relatively unimportant, and hence does not severely affect the posterior distribution of the true gene count. Firstly, algorithms that improve the quality of read alignment (Le et al., 2013), and thus minimise counting errors, are available. Furthermore, combinations of read-mapper and gene count quantification have been empirically studied, and optimal recommendations are available to obtain the most reliable observed gene count (e.g. OSA + HTSeq as suggested by Fonseca et al. (2014)).

5.2.2 Robustness of CORNAS

CORNAS showed comparable performance as GFOLD and NOISeq in the compcodeR simulation, despite being based on a different data model for the observed gene counts (i.e. Generalised Poisson vs. Negative Binomial). This finding provides confidence in integrating the CORNAS framework into current RNA-Seq data analysis protocols. Furthermore, despite the fact that the coverages were estimated, and thus subject to errors, both CORNAS settings (10xless and 100xless) showed similar performance on average. CORNAS struck a good compromise between sensitivity, PPV and DEG set size compared to GFOLD and NOISeq. In real world experiments, CORNAS can outperform competing methods when coverage is more reliably ascertained, such as from the Marioni dataset in Test 4.

Without incorporating information from the coverage parameter, traditional methods such as GFOLD and NOISeq for analysing unreplicated RNA-Seq count data are either too conservative, making very few calls but most of which are true positives (GFOLD), or making relatively more false positive calls (NOISeq) under very low coverage scenario (e.g. $b = 0.01$) (Figure 4.9). On the other hand, I showed that CORNAS controlled the FPR well and had high TPR when coverages are not too small (e.g. $b \geq 0.1$). Furthermore, if detection of weak fold change difference is of interest, then the fold-change parameter (ϕ) can be reduced from 1.5 to, say, 1.0 (details in the Methods Section). The TPR profiles

of CORNAS at fold-change parameter of 1.0 becomes similar to that of NOISeq for weak and strong signals, except when coverage is very low. With increasing true gene count, CORNAS continued to show a general increase in TPR, whereas NOISeq showed decline.

At present, most RNA-Seq experiments do not report an estimate of the actual amount of RNA in the starting material prior to sequencing. As a result, I could only study the effect of correcting the observed gene count using the posterior mean by simulations. Given the encouraging results, researchers may wish to collect information about the coverage parameter in the future to take advantage of CORNAS in the analysis of real RNA-Seq data sets.

A major problem in analysing unreplicated RNA-Seq count data is the lack of effective normalisation methods in the absence of biological replicates. The Bayesian framework on which CORNAS is based on in this study avoids the normalisation problem by working with the posterior distribution of the gene's true count. As a result, transcript length information is not required. This makes CORNAS suitable for organisms with incomplete or evolving transcriptome reference data, as new transcript information will not change how true counts are estimated over time.

My results suggest that CORNAS can be used to overcome analytical bottlenecks in experiments with limited replicates and low sequencing coverage by enabling the detection of DEGs with better prospects of downstream validation, using platforms such as quantitative PCR and NanoString nCounter (Kulkarni, 2011). However, good experimental design requires replication to derive robust biological interpretations. If cost is prohibitive and there are no limitations to obtaining samples, prioritising more biological replicates is more beneficial than increasing sequencing depth per sample. Liu et al. (2014) showed that increasing the number of replicates rather than sequencing depth is more effective at increasing the statistical power for detecting DEGs. Another good practice in experimental

design is the addition of controls in the assays. Synthetic spike-in standards, such as the External RNA Controls Consortium (ERCC) synthetic RNAs, can be added before library preparation to assist in measuring the performance of the sequencing protocol, and to normalise the count data (ENCODE, 2011; Jiang et al., 2011).

5.2.3 Application to oil palm samples

My simulation results indicated that as sequencing coverage approaches one, the variance of observed counts began to reduce dramatically (Section 4.2.2). This is logical, as the probability to pick what is intended during a random sampling will drop as the population size increases. Since the variance becomes the same as the mean when sample coverage is less than 0.01, this would mean that a gene with high mRNA copies in the sample will have a higher observed count variance across samples compared to a gene with lower mRNA copies. Therefore it becomes less likely to ascertain a good distinction between two samples if their coverages are lower than 0.01. In fact, in cases of low coverage samples, it would be more desirable to study differential expression on genes with low observed count numbers. This seems to be the case if we wish to apply CORNAS on RNA-Seq results from 454 sequencing.

RNA-Seq with 454 sequencing is known to produce generally less sequenced reads compared to Illumina sequencing, with about 2 M reads in one run. The library preparation is more straightforward and does not contain a PCR step that may bias random sampling events prior to loading (454LifeSciences, 2014). However, I chose to build the model used by CORNAS on the Illumina sequencing platform because the Illumina sequencing library preparation includes size selection. The 454 sequencing library preparation does not have this size selection step, which does not allow us to calculate concentration reliably. Figure 5.2 briefly summarises the library preparation process. While the library preparation may be different, the same coverage principal in Illumina sequencing can be applied to

these runs, albeit with coverages lower than 0.01. Therefore, my results indicated that 454 sequencing can produce reliable results when comparing very low counts in a differential gene expression analysis (Section 4.2.6).

The results from the nCounter validation of 16 DEGs indicated that CORNAS is sensitive to 454 sequencing data, but is unclear as to how specific in can be with such high false positives. Whether this is acceptable is still in question, as there has yet been sufficient published results that show RNA-Seq DEG calls are 100% accurate. Arguably I did not test the whole CORNAS DEG list, but selected the nCounter candidates based on other criteria, such as biological significance, potential novelty, relation to future/current projects, availability of samples and cost to validate.

It may seem logical that we should look at a consensus approach to making a DEG list that would be more specific, such as filtering significant DEGs that overlap in all, or the majority, of multiple DEG call programs, i.e. in CORNAS, GFOLD and NOISeq. However, my results seem to indicate that this will not work if the nCounter was used to validate the results: Only one transcript, isotig70480, was flagged as significant in all three programs, but not significant in nCounter; Only two transcripts, isotig67634 and isotig59228, were flagged as statistically significant in nCounter, NOISeq and CORNAS (Table 4.9).

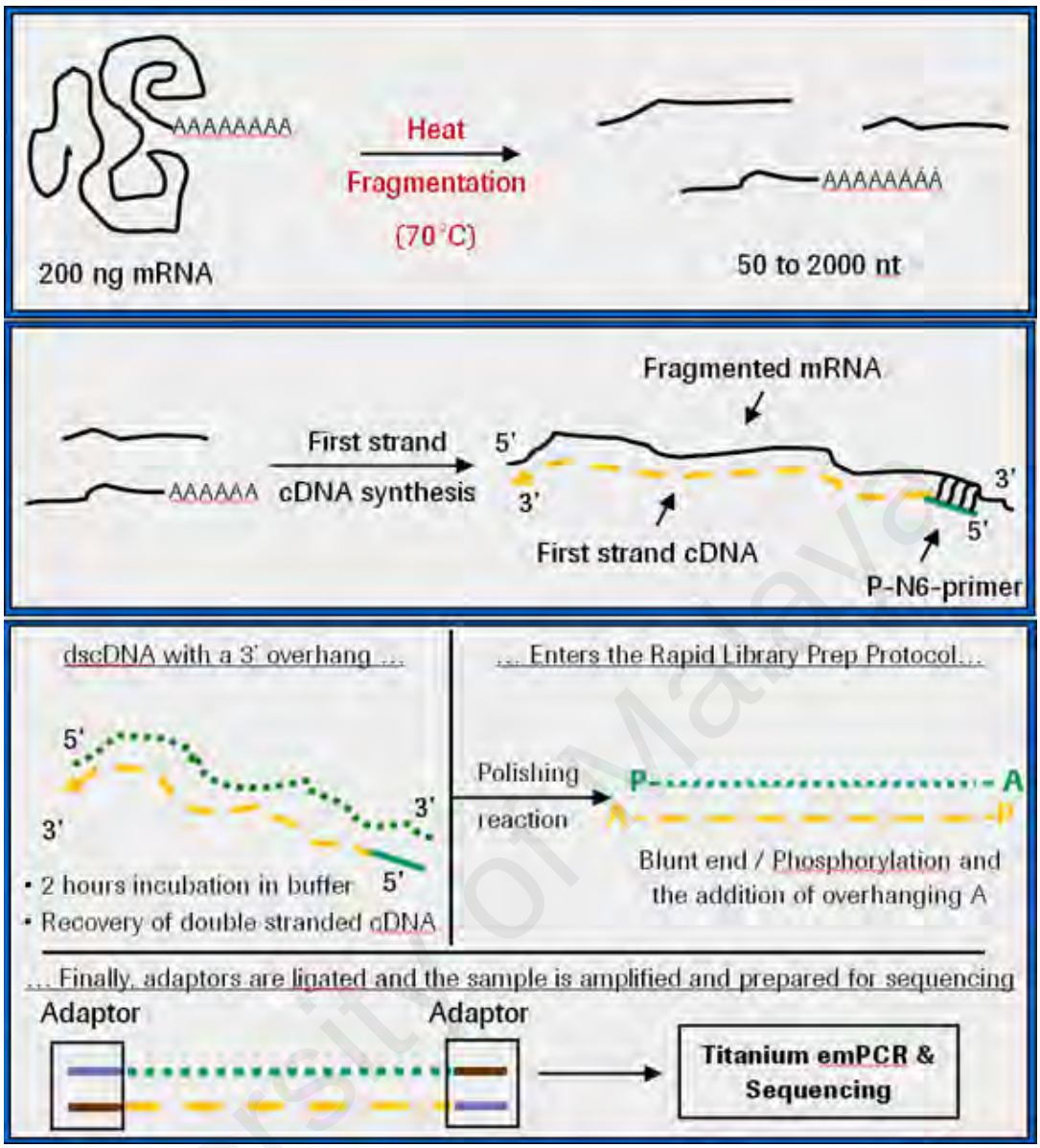


Figure 5.2: 454 transcriptome library preparation briefly explained (454LifeSciences, 2014)

CHAPTER 6: CONCLUSIONS

The Sime Darby oil palm genome assembly had made major improvements over the course of this study. The latest genome assembly, OPg3, has improved gene space representation by 39% over the first version. This was achieved by improving the gene-centric information in the genome using transcriptome data to build back parts of the genome that was incomplete or missing. Furthermore, Sime Darby's physical map coverage of the oil palm's 16 chromosomes was able to improve by 79%, with the method of using molecular markers and annotation of bridge reads, to aid fragmentation and rearrangement of both OPg3 and MPOB oil palm's physical map scaffolds. I believe OPg3 is of sufficiently quality to be deployed for genome-wide association studies (GWAS) and fuctional genomic studies. In the process of improving the oil palm genome, three new computational methods were developed to overcome the challenges in assembly. These are (i) GenSeed Pipeline Suite; (ii) BridgeReader; and (iii) MarkMyMap.

While sufficient for genetics studies, OPg3 is not complete. To make progress, I suggest to employ the newer sequencing technologies that generate longer read lengths to overcome the challenges in contiguity and repeats. It may not be necessary to sequence deep, but to use these long reads as scaffolds. Then, the aforementioned computational methods can be used to further polish the genome assembly. This time however, the BridgeReader program can be deployed to identify and remove bridge reads prior to assembly to improve the non-repetitive regions in the genome.

Finally, I have developed CORNAS (COverage-dependent RNA-Seq), a fast Bayesian method that incorporates a novel coverage parameter to estimate the posterior distribution of the true gene count. Under the CORNAS framework, orthogonal information from sequence coverage that is determined from the concentration of an RNA sample can

be used to improve the accuracy of calling DEG. Through simulations and analyses of real data sets, the performance of CORNAS was shown to be comparable or superior to GFOLD and NOIseq in the case of unreplicated RNA-Seq experiments. While CORNAS was developed based on Illumina sequencing, I have shown that CORNAS can provide reliable results on 454 sequenced unreplicated comparison of oil palm male inflorescences and female inflorescences.

REFERENCES

454LifeSciences. (2014). *454 Transcriptome Sequencing*. Retrieved 3 March 2014 from http://454.com/applications/transcriptome-sequencing/.

Achaz, G., Boyer, F., Rocha, E. P., Viari, A., Coissac, E. (2006). Repseek, a tool to retrieve approximate repeats from large DNA sequences. *Bioinformatics*, *23*(1), 119–121.

Adam, H., Jouannic, S., Escoute, J., Duval, Y., Verdeil, J.-L., Tregear, J. W. (2005). Reproductive developmental complexity in the African oil palm (*Elaeis guineensis*, Arecaceae). *American Journal of Botany*, *92*(11), 1836–1852.

Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., ... Moreno, R. F. (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, *252*(5013), 1651–1656.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410.

Alves, J. M., de Oliveira, A. L., Sandberg, T. O., Moreno-Gallego, J. L., de Toledo, M. A., de Moura, E. M., ... Reyes, A. (2016). GenSeed-HMM: A tool for progressive assembly using profile HMMs as seeds and its application in Alpavirinae viral discovery from metagenomic data. *Frontiers in Microbiology*, *7*, Article#269.

Amores, A., Catchen, J., Nanda, I., Warren, W., Walter, R., Schartl, M., Postlethwait, J. H. (2014). A RAD-tag genetic map for the platyfish (*Xiphophorus maculatus*) reveals mechanisms of karyotype evolution among teleost fish. *Genetics*, *197*(2), 625–641.

Anders, S., Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*(10), Article#R106.

Anders, S., Pyl, P. T., Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, *31*(2), 166–169.

Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, *408*(6814), 796–815.

Argyris, J. M., Ruiz-Herrera, A., Madriz-Masis, P., Sanseverino, W., Morata, J., Pujol, M., . . . Garcia-Mas, J. (2015). Use of targeted SNP selection for an improved anchoring of the melon (*Cucumis melo* l.) scaffold genome assembly. *BMC Genomics*, *16*(1), Article#4.

Arner, E., Tammi, M. T., Tran, A.-N., Kindlund, E., Andersson, B. (2006). DNPTrapper: an assembly editing tool for finishing and analysis of complex repeat regions. *BMC Bioinformatics*, *7*(1), Article#155.

Assefa, S., Keane, T. M., Otto, T. D., Newbold, C., Berriman, M. (2009). ABACAS: algorithm-based automatic contiguation of assembled sequences. *Bioinformatics*, *25*(15), 1968–1969.

Bankier, A. T. (2001). Shotgun DNA sequencing. In *DNA Sequencing Protocols* (pp. 89–100). New York: Springer.

Bao, Z., Eddy, S. R. (2002). Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Research*, *12*(8), 1269–1276.

Bevan, M., Walsh, S. (2005). The Arabidopsis genome: a foundation for plant research. *Genome Research*, *15*(12), 1632–1642.

Bickhart, D. M., Rosen, B. D., Koren, S., Sayre, B. L., Hastie, A. R., Chan, S., . . . Smith, T. (2017). Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nature Genetics*, *49*(4), Article#643.

Billotte, N., Marseillac, N., Risterucci, A.-M., Adon, B., Brottier, P., Baurens, F.-C., . . . Amblard, P. (2005). Microsatellite-based high density linkage map in oil palm (*Elaeis guineensis* jacq.). *Theoretical and Applied Genetics*, *110*(4), 754–765.

Birol, I., Jackman, S. D., Nielsen, C. B., Qian, J. Q., Varhol, R., Stazyk, G., . . . Horsman, D. E. (2009). De novo transcriptome assembly with ABySS. *Bioinformatics*, *25*(21), 2872–2877.

Black, D. L. (2003). Mechanisms of alternative pre-messenger RNA splicing. *Annual Review of Biochemistry*, *72*(1), 291–336.

Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., . . . Gregor,

J. (1997). The complete genome sequence of *Escherichia coli* K-12. *Science*, *277*(5331), 1453–1462.

Bleidorn, C. (2016). Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. *Systematics and Biodiversity*, *14*(1), 1–8.

Blencowe, B. J., Ahmad, S., Lee, L. J. (2009). Current-generation high-throughput sequencing: deepening insights into mammalian transcriptomes. *Genes & Development*, *23*(12), 1379–1386.

Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D., Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics*, *27*(4), 578–579.

Botstein, D., White, R. L., Skolnick, M., Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, *32*(3), 314–331.

Bourgis, F., Kilaru, A., Cao, X., Ngando-Ebongue, G.-F., Drira, N., Ohlrogge, J. B., Arondel, V. (2011). Comparative transcriptome and metabolite analysis of oil palm and date palm mesocarp that differ dramatically in carbon partitioning. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(30), 12527–12532.

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., Bairoch, A. (2007). Uniprotkb/swiss-prot. In *Plant Bioinformatics* (pp. 89–112). New York: Springer.

Braasch, I., Peterson, S. M., Desvignes, T., McCluskey, B. M., Batzel, P., Postlethwait, J. H. (2015). A new model army: Emerging fish models to study the genomics of vertebrate Evo-Devo. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, *324*(4), 316–341.

Bradnam, K. R., Fass, J. N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., . . . Chitsaz, H. (2013). Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience*, *2*(1), Article#10.

Branton, D., Deamer, D. W., Marziali, A., Bayley, H., Benner, S. A., Butler, T., . . . Jovanovich, S. B. (2008). The potential and challenges of nanopore sequencing. *Nature Biotechnology*, *26*(10), 1146–1153.

Buermans, H., Den Dunnen, J. (2014). Next generation sequencing technology: advances and applications. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, *1842*(10), 1932–1941.

Bullard, J. H., Purdom, E., Hansen, K. D., Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, *11*(1), Article#94.

Burton, J. N., Adey, A., Patwardhan, R. P., Qiu, R., Kitzman, J. O., Shendure, J. (2013). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nature Biotechnology*, *31*(12), Article#1119.

Byrne, A., Cole, C., Volden, R., Vollmers, C. (2019). Realizing the potential of full-length transcriptome sequencing. *Philosophical Transactions of the Royal Society B*, *374*(1786), Article#20190097.

Cai, D., Rodríguez, F., Teng, Y., Ané, C., Bonierbale, M., Mueller, L. A., Spooner, D. M. (2012). Single copy nuclear gene analysis of polyploidy in wild potatoes (*Solanum* section *Petota*). *BMC Evolutionary Biology*, *12*(1), Article#70.

Carrel, L., Willard, H. F. (2005). X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature*, *434*(7031), 400–404.

Castilho, A., Vershinin, A., Heslop-Harrison, J. (2000). Repetitive DNA and the chromosomes in the genome of oil palm (*Elaeis guineensis*). *Annals of Botany*, *85*(6), 837–844.

Catchen, J., Amores, A. (2016). *Chromonomer*. Software available online at http://catchenlab.life.illinois.edu/chromonomer/.

Cech, T. R., Rio, D. C. (1979). Localization of transcribed regions on extrachromosomal ribosomal RNA genes of *Tetrahymena thermophila* by R-loop mapping. *Proceedings of the National Academy of Sciences of the United States of America*, *76*(10), 5051–5055.

Chandran, M. R. (2010). Malaysian palm oil industry performance 2009. *Global Oils & Fats Business Magazine*, *7*(1), Article#11.

Chaney, L., Sharp, A. R., Evans, C. R., Udall, J. A. (2016). Genome mapping in plant

comparative genomics. *Trends in Plant Science*, *21*(9), 770–780.

Chang, J. (2015). Core services: reward bioinformaticians. *Nature News*, *520*(7546), Article#151.

Cheng, B., Furtado, A., Henry, R. J. (2017). Long-read sequencing of the coffee bean transcriptome reveals the diversity of full-length transcripts. *Gigascience*, *6*(11), Article#gix086.

Clarke, J., Wu, H.-C., Jayasinghe, L., Patel, A., Reid, S., Bayley, H. (2009). Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology*, *4*(4), 265–270.

CoGe. (2015). *Sequenced plant genomes*. Retrieved 5 July 2015 from http://genomevolution.org/wiki/index.php/Sequenced_plant_genomes.

Consul, P. C., Jain, G. C. (1973). A generalization of the Poisson distribution. *Technometrics*, *15*(4), 791–799.

Corley, R. H. V., Tinker, P. B. (2008). *The oil palm*. Hoboken: John Wiley & Sons.

Cui, J., Luo, S., Niu, Y., Huang, R., Wen, Q., Su, J., ... Hu, K. (2018). A RAD-based genetic map for anchoring scaffold sequences and identifying QTLs in bitter gourd (*Momordica charantia*). *Frontiers in Plant Science*, *9*, Article#477.

de Koning, A. J., Gu, W., Castoe, T. A., Batzer, M. A., Pollock, D. D. (2011). Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genetics*, *7*(12), Article#e1002384.

de la Bastide, M., McCombie, W. R. (2007). Assembling genomic DNA sequences with PHRAP. *Current Protocols in Bioinformatics*, *17*(1), 11–4.

Delcher, A. L., Kasif, S., Fleischmann, R. D., Peterson, J., White, O., Salzberg, S. L. (1999). Alignment of whole genomes. *Nucleic Acids Research*, *27*(11), 2369–2376.

Delcher, A. L., Phillippy, A., Carlton, J., Salzberg, S. L. (2002). Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Research*, *30*(11), 2478–2483.

Denisov, G., Walenz, B., Halpern, A. L., Miller, J., Axelrod, N., Levy, S., Sutton, G. (2008). Consensus generation and variant detection by Celera Assembler. *Bioinformatics*, *24*(8), 1035–1040.

Dillies, M.-A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., ... Guernec, G. (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics*, *14*(6), 671–683.

Dimalanta, E. T., Lim, A., Runnheim, R., Lamers, C., Churas, C., Forrest, D. K., ... Schwartz, D. C. (2004). A microfluidic system for large DNA molecule arrays. *Analytical Chemistry*, *76*(18), 5293–5301.

Diouf, L., Magwanga, R., Gong, W., He, S., Pan, Z., Jia, Y., ... Du, X. (2018). QTL mapping of fiber quality and yield-related traits in an intra-specific upland cotton using genotype by sequencing (GBS). *International Journal of Molecular Sciences*, *19*(2), Article#441.

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., ... Aiden, E. (2017). De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*, *356*(6333), 92–95.

Earl, D., Bradnam, K., John, J. S., Darling, A., Lin, D., Fass, J., ... Nguyen, N. (2011). Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome Research*, *21*(12), 2224–2241.

Edwards, A., Caskey, C. T. (1991). Closure strategies for random DNA sequencing. *Methods*, *3*(1), 41–47.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., ... Bibillo, A. (2009). Real-time DNA sequencing from single polymerase molecules. *Science*, *323*(5910), 133–138.

Eisenberg, E., Levanon, E. Y. (2013). Human housekeeping genes, revisited. *Trends in Genetics*, *29*(10), 569–574.

Ellegren, H. (2004). Microsatellites: simple sequences with complex evolution. *Nature Reviews Genetics*, *5*(6), 435–445.

ENCODE. (2011). *Standards, Guidelines and Best Practices for RNA-Seq*. Retrieved 19

January 2016 from https://genome.ucsc.edu/ENCODE/protocols/dataStandards/.

Esnaola, M., Puig, P., Gonzalez, D., Castelo, R., Gonzalez, J. R. (2013). A flexible count data model to fit the wide diversity of expression profiles arising from extensively replicated RNA-seq experiments. *BMC Bioinformatics*, *14*(1), Article#254.

Feng, J., Meyer, C. A., Wang, Q., Liu, J. S., Liu, X. S., Zhang, Y. (2012). GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data. *Bioinformatics*, *28*(21), 2782–2788.

Fisher, R. A., Yates, F. (1963). *Statistical tables for biological, agricultural and medical research* (6th ed.). London: Oliver & Boyd Ltd.

Fleischner, J., Fleischner, H. (1990). *Eulerian graphs and related topics* (Vol. 1). Amsterdam: Elsevier.

Fonseca, N. A., Marioni, J., Brazma, A. (2014). RNA-seq gene profiling-a systematic empirical comparison. *PLoS ONE*, *9*(9), Article#e107026.

Frazee, A. C., Langmead, B., Leek, J. T. (2011). ReCount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets. *BMC Bioinformatics*, *12*(1), Article#449.

Fu, X., Fu, N., Guo, S., Yan, Z., Xu, Y., Hu, H., . . . Khaitovich, P. (2009). Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC Genomics*, *10*(1), Article#161.

Fuselli, S., Baptista, R., Panziera, A., Magi, A., Guglielmi, S., Tonin, R., . . . Bertorelle, G. (2018). A new hybrid approach for MHC genotyping: high-throughput NGS and long read MinION nanopore sequencing, with application to the non-model vertebrate Alpine chamois (*Rupicapra rupicapra*). *Heredity*, *121*(4), Article#293.

Garey, M. R., Johnson, D. S. (1979). *Computers and intractability: A Guide to the Theory of NP-Completeness*. New York: Freeman.

Genohub. (2015). *Designing your Next Generation Sequencing Run*. Retrieved 19 January 2016 from https://genohub.com/next-generation-sequencing-guide/.

Gibbs, R. A., Rogers, J., Katze, M. G., Bumgarner, R., Weinstock, G. M., Mardis, E. R., ... Batzer, M. A. (2007). Evolutionary and biomedical insights from the rhesus macaque genome. *Science*, *316*(5822), 222–234.

Gierliński, M., Cole, C., Schofield, P., Schurch, N. J., Sherstnev, A., Singh, V., ... Barton, G. (2015). Statistical models for RNA-seq data derived from a two-condition 48-replicate experiment. *Bioinformatics*, *31*(22), 3625–3630.

Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, F. J., Burton, J. N., Walker, B. J., ... Berlin, A. M. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(4), 1513–1518.

Goff, S. A., Ricke, D., Lan, T.-H., Presting, G., Wang, R., Dunn, M., ... Hadley, D. (2002). A draft sequence of the rice genome (*Oryza sativa* l. ssp. japonica). *Science*, *296*(5565), 92–100.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., ... Chen, Z. (2011). Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature Biotechnology*, *29*(7), Article#644.

Greninger, A. L., Naccache, S. N., Federman, S., Yu, G., Mbala, P., Bres, V., ... Dodd, R. (2015). Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. *Genome Medicine*, *7*(1), Article#99.

Gulcher, J. (2012). Microsatellite markers for linkage and association studies. *Cold Spring Harbor Protocols*, *2012*(4), 425–432.

Gupta, P., Roy, J., Prasad, M. (2001). Single nucleotide polymorphisms: a new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants. *Current Science*, *80*(4), 524–535.

Haas, B. J., Chin, M., Nusbaum, C., Birren, B. W., Livny, J. (2012). How deep is deep enough for RNA-Seq profiling of bacterial transcriptomes? *BMC Genomics*, *13*(1), Article#734.

Hardcastle, T. J., Kelly, K. A. (2010). baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, *11*(1), Article#422.

Hardon, J., Rao, V., Rajanaidu, N. (1985). Progress in Plant Breeding. In G.E. Russell (Ed.), *A review of oil-palm breeding*. (pp. 139-163). London: Butterworths.

Harper, P. S. (2008). *A short history of medical genetics* (No. 57). Oxford: Oxford University Press.

Ho, C.-L., Kwan, Y.-Y., Choi, M.-C., Tee, S.-S., Ng, W.-H., Lim, K.-A., . . . Tan, S. H. (2007). Analysis and functional annotation of expressed sequence tags (ESTs) from multiple tissues of oil palm (*Elaeis guineensis* Jacq.). *BMC Genomics*, *8*(1), Article#381.

Hu, J., Ge, H., Newman, M., Liu, K. (2012). OSA: a fast and accurate alignment tool for RNA-Seq. *Bioinformatics*, *28*(14), 1933–1934.

Huang, X., Madan, A. (1999). CAP3: A DNA sequence assembly program. *Genome Research*, *9*(9), 868–877.

Huang, Y.-F., Chen, S.-C., Chiang, Y.-S., Chen, T.-H., Chiu, K.-P. (2012). Palindromic sequence impedes sequencing-by-ligation mechanism. *BMC Systems Biology*, *6*(2), 2–10.

Hunt, M., Kikuchi, T., Sanders, M., Newbold, C., Berriman, M., Otto, T. D. (2013). REAPR: a universal tool for genome assembly evaluation. *Genome Biology*, *14*(5), Article#R47.

Huse, S. M., Huber, J. A., Morrison, H. G., Sogin, M. L., Welch, D. M. (2007). Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biology*, *8*(7), Article#R143.

Idury, R. M., Waterman, M. S. (1995). A new algorithm for DNA sequence assembly. *Journal of Computational Biology*, *2*(2), 291–306.

International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921.

Jaccoud, D., Peng, K., Feinstein, D., Kilian, A. (2001). Diversity arrays: a solid state technology for sequence information independent genotyping. *Nucleic Acids Research*, *29*(4), e25–e25.

Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., . . . Vezzi, A. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, *449*(7161), 463–467.

Jarne, P., Lagoda, P. J. (1996). Microsatellites, from molecules to populations and back. *Trends in Ecology & Evolution*, *11*(10), 424–429.

Jiang, L., Schlesinger, F., Davis, C. A., Zhang, Y., Li, R., Salit, M., . . . Oliver, B. (2011). Synthetic spike-in standards for RNA-seq experiments. *Genome Research*, *21*(9), 1543–1551.

Ju, J., Kim, D. H., Bi, L., Meng, Q., Bai, X., Li, Z., . . . Edwards, J. R. (2006). Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(52), 19635–19640.

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., Walichiewicz, J. (2005). Repbase update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research*, *110*(1-4), 462–467.

Kærn, M., Elston, T. C., Blake, W. J., Collins, J. J. (2005). Stochasticity in gene expression: from theories to phenotypes. *Nature Reviews Genetics*, *6*(6), Article#451.

Kanehisa, M., Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, *28*(1), 27–30.

Kaplan, N., Dekker, J. (2013). High-throughput genome scaffolding from in vivo DNA interaction frequency. *Nature Biotechnology*, *31*(12), Article#1143.

Kawahara, Y., de la Bastide, M., Hamilton, J. P., Kanamori, H., McCombie, W. R., Ouyang, S., . . . Childs, K. L. (2013). Improvement of the *Oryza sativa* nipponbare reference genome using next generation sequence and optical map data. *Rice*, *6*(1), Article#4.

Kent, W. J. (2002). BLAT—the BLAST-like alignment tool. *Genome Research*, *12*(4), 656–664.

Khang, T. F. (2016). A gamma approximation to the Bayesian posterior distribution of a discrete parameter of the Generalized Poisson model. *arXiv:1606.01749v1*.

Khang, T. F., Lau, C. Y. (2015). Getting the most out of RNA-seq data analysis. *PeerJ*, *3*, Article#e1360.

King, C. R., Scott-Horton, T. (2007). Pyrosequencing: a simple method for accurate genotyping. *Pyrosequencing Protocols*, 39–55.

Kircher, M., Kelso, J. (2010). High-throughput DNA sequencing–concepts and limitations. *Bioessays*, *32*(6), 524–536.

Koch, P., Platzer, M., Downie, B. R. (2014). RepARK—de novo creation of repeat libraries from whole-genome NGS reads. *Nucleic Acids Research*, *42*(9), Article#e80.

Korbel, J. O., Lee, C. (2013). Genome assembly and haplotyping with Hi-C. *Nature Biotechnology*, *31*(12), Article#1099.

Korlach, J. (2013). *Understanding accuracy in SMRT sequencing*. Retrieved 3 September 2017 from https://www.mscience.com.au/upload/pages/pacbioaccuracy/perspective-understanding-accuracy-in-smrt-sequencing.pdf.

Kulkarni, M. M. (2011). Digital multiplexed gene expression analysis using the NanoString nCounter system. *Current Protocols in Molecular Biology*, *25*(B10), 1–17.

Kurtz, S., Choudhuri, J. V., Ohlebusch, E., Schleiermacher, C., Stoye, J., Giegerich, R. (2001). REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Research*, *29*(22), 4633–4642.

Kurtz, S., Narechania, A., Stein, J. C., Ware, D. (2008). A new method to compute k-mer frequencies and its application to annotate large repetitive plant genomes. *BMC Genomics*, *9*(1), Article#517.

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome Biology*, *5*(2), Article#R12.

Lahens, N. F., Kavakli, I. H., Zhang, R., Hayer, K., Black, M. B., Dueck, H., . . . Grant, G. R. (2014). IVT-seq reveals extreme bias in RNA sequencing. *Genome Biology*, *15*(6), Article#R86.

Langmead, B., Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359.

Le, H.-S., Schulz, M. H., McCauley, B. M., Hinman, V. F., Bar-Joseph, Z. (2013). Probabilistic error correction for RNA sequencing. *Nucleic Acids Research*, *41*(10), Article#e109.

Levene, M. J., Korlach, J., Turner, S. W., Foquet, M., Craighead, H. G., Webb, W. W. (2003). Zero-mode waveguides for single-molecule analysis at high concentrations. *Science*, *299*(5607), 682–686.

Levy-Sakin, M., Ebenstein, Y. (2013). Beyond sequencing: optical mapping of DNA in the age of nanotechnology and nanoscopy. *Current Opinion in Biotechnology*, *24*(4), 690–698.

Li, H., Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760.

Li, H., Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, *26*(5), 589–595.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079.

Li, J., Tibshirani, R. (2013). Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data. *Statistical Methods in Medical Research*, *22*(5), 519–536.

Li, R., Ye, J., Li, S., Wang, J., Han, Y., Ye, C., ... Wang, J. (2005). ReAS: Recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun. *PLoS Computational Biology*, *1*(4), Article#e43.

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., ... Li, S. (2010). De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research*, *20*(2), 265–272.

Li, W., Jiang, T. (2012). Transcriptome assembly and isoform expression level estimation from biased RNA-Seq reads. *Bioinformatics*, *28*(22), 2914–2921.

Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., . . . Yang, B. (2012). Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph. *Briefings in Functional Genomics*, *11*(1), 25–37.

Lieberman-Aiden, E., Van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., . . . Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, *326*(5950), 289–293.

Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H., Ecker, J. R. (2008). Highly integrated single-base resolution maps of the epigenome in arabidopsis. *Cell*, *133*(3), 523–536.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., . . . Law, M. (2012). Comparison of next-generation sequencing systems. *BioMed Research International*, *2012*(251364), 1–11.

Liu, Y., Zhou, J., White, K. P. (2014). RNA-seq differential expression studies: more sequence or more replication? *Bioinformatics*, *30*(3), 301–304.

Love, M. I., Huber, W., Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), Article#550.

Lovén, J., Orlando, D. A., Sigova, A. A., Lin, C. Y., Rahl, P. B., Burge, C. B., . . . Young, R. A. (2012). Revisiting global gene expression analysis. *Cell*, *151*(3), 476–482.

Lu, H., Giordano, F., Ning, Z. (2016). Oxford Nanopore MinION sequencing and genome assembly. *Genomics, Proteomics & Bioinformatics*, *14*(5), 265–279.

Lynch, M., Conery, J. S. (2003). The origins of genome complexity. *Science*, *302*(5649), 1401–1404.

Malaysian Palm Oil Council. (2019). *History and Origin*. Retrieved 14 June 2019 from https://theoilpalm.org/history-and-origin/.

Manekar, S. C., Sathe, S. R. (2018). A benchmark study of k-mer counting methods for high-throughput sequencing. *GigaScience*, *7*(12), 1–13.

Manrao, E. A., Derrington, I. M., Pavlenok, M., Niederweis, M., Gundlach, J. H. (2011).

Nucleotide discrimination with DNA immobilized in the MspA nanopore. *PLoS ONE*, *6*(10), Article#e25723.

Mardis, E., McPherson, J., Martienssen, R., Wilson, R. K., McCombie, W. R. (2002). What is finished, and why does it matter. *Genome Research*, *12*(5), 669–671.

Marguerat, S., Schmidt, A., Codlin, S., Chen, W., Aebersold, R., Bähler, J. (2012). Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells. *Cell*, *151*(3), 671–683.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., . . . Dewell, S. B. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, *437*(7057), Article#376.

Marie-Nelly, H., Marbouty, M., Cournac, A., Flot, J.-F., Liti, G., Parodi, D. P., . . . Koszul, R. (2014). High-quality genome (re) assembly using chromosomal contact data. *Nature Communications*, *5*, Article#5695.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, *18*(9), 1509–1517.

Martin, J. A., Wang, Z. (2011). Next-generation transcriptome assembly. *Nature Reviews Genetics*, *12*(10), 671–682.

Martin, L., Fei, Z., Giovannoni, J., Rose, J. K. C. (2013). Catalyzing plant science research with RNA-seq. *Frontiers in Plant Science*, *4*, Article#66.

Mascher, M., Gundlach, H., Himmelbach, A., Beier, S., Twardziok, S. O., Wicker, T., . . . Stein, N. (2017). A chromosome conformation capture ordered sequence of the barley genome. *Nature*, *544*(7651), Article#427.

McKernan, K. J., Peckham, H. E., Costa, G. L., McLaughlin, S. F., Fu, Y., Tsung, E. F., . . . Zhang, Z. (2009). Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Research*, *19*(9), 1527–1541.

Messing, J., Crea, R., Seeburg, P. H. (1981). A system for shotgun DNA sequencing. *Nucleic Acids Research*, *9*(2), 309–321.

Mikheyev, A. S., Tin, M. M. (2014). A first look at the Oxford Nanopore MinION sequencer. *Molecular Ecology Resources*, *14*(6), 1097–1102.

MPOB. (2010). *Oil Palm & The Environment*. Retrieved 5 August 2011 from http://www.mpob.gov.my/en/palm-info/environment/520-achievements.

Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J., . . . Anson, E. L. (2000). A whole-genome assembly of drosophila. *Science*, *287*(5461), 2196–2204.

Nagaraj, S. H., Gasser, R. B., Ranganathan, S. (2006). A hitchhiker's guide to expressed sequence tag (EST) analysis. *Briefings in Bioinformatics*, *8*(1), 6–21.

Nagarajan, N., Pop, M. (2013). Sequence assembly demystified. *Nature Reviews Genetics*, *14*(3), Article#157.

Nakabachi, A., Yamashita, A., Toh, H., Ishikawa, H., Dunbar, H. E., Moran, N. A., Hattori, M. (2006). The 160-kilobase genome of the bacterial endosymbiont carsonella. *Science*, *314*(5797), 267–267.

Noh, A., Rafii, M., Saleh, G., Kushairi, A., Latif, M. (2012). Genetic performance and general combining ability of oil palm Deli dura x AVROS pisifera tested on inland soils. *The Scientific World Journal*, *2012*, Article#792601.

Oshlack, A., Robinson, M. D., Young, M. D. (2010). From RNA-seq reads to differential expression results. *Genome Biology*, *11*(12), Article#220.

Ouzounis, C. A., Valencia, A. (2003). Early bioinformatics: the birth of a discipline—a personal view. *Bioinformatics*, *19*(17), 2176–2190.

Parra, G., Bradnam, K., Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, *23*(9), 1061–1067.

Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., . . . Schmutz, J. (2009). The sorghum bicolor genome and the diversification of grasses. *Nature*, *457*(7229), Article#551.

Peng, Y., Leung, H. C., Yiu, S.-M., Chin, F. Y. (2010). IDBA–a practical iterative de

bruijn graph de novo assembler. In *Annual international conference on research in computational molecular biology* (pp. 426–440). Berlin: Springer.

Pevzner, P. A., Tang, H., Waterman, M. S. (2001). An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(17), 9748–9753.

Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., ... Pritchard, J. K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, *464*(7289), 768–772.

Pop, M. (2009). Genome assembly reborn: recent computational challenges. *Briefings in Bioinformatics*, *10*(4), 354–366.

Pop, M., Salzberg, S. L. (2008). Bioinformatics challenges of new sequencing technology. *Trends in Genetics*, *24*(3), 142–149.

Price, A. L., Jones, N. C., Pevzner, P. A. (2005). De novo identification of repeat families in large genomes. *Bioinformatics*, *21*(*Suppl.*1), i351–i358.

R Core Team. (2018). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Richard, G.-F., Kerrest, A., Dujon, B. (2008). Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiology and Molecular Biology Reviews*, *72*(4), 686–727.

Roach, J. C., Boysen, C., Wang, K., Hood, L. (1995). Pairwise end sequencing: a unified approach to genomic mapping and sequencing. *Genomics*, *26*(2), 345–353.

Robinson, M. D., McCarthy, D. J., Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140.

Roewer, L., Amemann, J., Spurr, N., Grzeschik, K.-H., Epplen, J. (1992). Simple repeat sequences on the human y chromosome are equally polymorphic as their autosomal counterparts. *Human Genetics*, *89*(4), 389–394.

Saliba, A.-E., Westermann, A. J., Gorski, S. A., Vogel, J. (2014). Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Research*, *42*(14), 8845–8860.

Salzberg, S. L., Yorke, J. A. (2005). Beware of mis-assembled genomes. *Bioinformatics*, *21*(24), 4320–4321.

Sanchez, A., Golding, I. (2013). Genetic determinants and cellular constraints in noisy gene expression. *Science*, *342*(6163), 1188–1193.

Sanger, F., Nicklen, S., Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(12), 5463–5467.

Santos, A., Tsafou, K., Stolte, C., Pletscher-Frankild, S., O'Donoghue, S. I., Jensen, L. J. (2015). Comprehensive comparison of large-scale tissue expression datasets. *PeerJ*, *3*, Article#e1054.

Schaller, R. R. (1997). Moore's law: past, present and future. *IEEE Spectrum*, *34*(6), 52–59.

Schatz, M. C., Langmead, B. (2013). The DNA data deluge: fast, efficient genome sequencing machines are spewing out more data than geneticists can analyze. *IEEE Spectrum*, *50*(7), 26–33.

Schmidt, T. (1999). LINEs, SINEs and repetitive DNA: non-LTR retrotransposons in plant genomes. *Plant Molecular Biology*, *40*(6), 903–910.

Schurch, N. J., Schofield, P., Gierliński, M., Cole, C., Sherstnev, A., Singh, V., . . . Barton, G. (2016). How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA*, *22*(6), 839–851.

Schwartz, D. C., Li, X., Hernandez, L. I., Ramnarain, S. P., Huff, E. J., Wang, Y.-K. (1993). Ordered restriction maps of *Saccharomyces cerevisiae* chromosomes constructed by optical mapping. *Science*, *262*(5130), 110–114.

Sendler, E., Johnson, G. D., Krawetz, S. A. (2011). Local and global factors affecting RNA sequencing analysis. *Analytical Biochemistry*, *419*(2), 317–322.

Shizuya, H., Birren, B., Kim, U.-J., Mancino, V., Slepak, T., Tachiiri, Y., Simon, M. (1992). Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(18), 8794–8797.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212.

Simmons, M. J., Snustad, D. P., et al. (2006). *Principles of genetics*. John Wiley & Sons.

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, *15*(2), 121–132.

Sing, T., Sander, O., Beerenwinkel, N., Lengauer, T. (2005). ROCR: visualizing classifier performance in R. *Bioinformatics*, *21*(20), 3940–3941.

Singh, R., Ong-Abdullah, M., Low, E.-T. L., Manaf, M. A. A., Rosli, R., Nookiah, R., ... Azizi, N. (2013). Oil palm genome sequence reveals divergence of interfertile species in old and new worlds. *Nature*, *500*(7462), 335–339.

Sipos, B., Slodkowicz, G., Massingham, T., Goldman, N. (2013). Realistic simulations reveal extensive sample-specificity of RNA-seq biases. *arXiv preprint arXiv:1308.3172*.

Small, C., Bassham, S., Catchen, J., Amores, A., Fuiten, A., Brown, R., ... Cresko, W. (2016). The genome of the gulf pipefish enables understanding of evolutionary innovations. *Genome Biology*, *17*(1), Article#258.

Smits, S. L., Bodewes, R., Ruiz-González, A., Baumgärtner, W., Koopmans, M. P., Osterhaus, A. D., Schürch, A. C. (2015). Recovering full-length viral genomes from metagenomes. *Frontiers in Microbiology*, *6*, Article#1069.

Sobreira, T. J., Gruber, A. (2008). Sequence-specific reconstruction from fragmentary databases using seed sequences: implementation and validation on sage, proteome and generic sequencing data. *Bioinformatics*, *24*(15), 1676–1680.

Sommer, D. D., Delcher, A. L., Salzberg, S. L., Pop, M. (2007). Minimus: a fast,

lightweight genome assembler. *BMC Bioinformatics*, *8*(1), Article#1.

Soneson, C. (2014). compcodeR - an R package for benchmarking differential expression methods for RNA-seq data. *Bioinformatics*, *30*(17), 2517–2518.

Soneson, C., Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, *14*(1), Article#91.

Srivastava, S., Chen, L. (2010). A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. *Nucleic Acids Research*, *38*(17), Article#e170.

Staden, R. (1979). A strategy of DNA sequencing employing computer programs. *Nucleic Acids Research*, *6*(7), 2601–2610.

Stevens, H. (2013). *Life out of sequence: a data-driven history of bioinformatics*. Chicago: University of Chicago Press.

Sturtevant, A. H. (1913). The linear arrangement of six sex-linked factors in drosophila, as shown by their mode of association. *Journal of Experimental Zoology*, *14*(1), 43–59.

Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T. Z., Garcia-Hernandez, M., Foerster, H., . . . Radenbaugh, A. (2007). The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Research*, *36*(*Suppl.*1), D1009–D1014.

Syvänen, A.-C. (2001). Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nature Reviews Genetics*, *2*(12), 930–942.

Tammi, M. T., Arner, E., Kindlund, E., Andersson, B. (2003). Correcting errors in shotgun sequences. *Nucleic Acids Research*, *31*(15), 4663–4672.

Tammi, M. T., Arner, E., Kindlund, E., Andersson, B. (2004). ReDiT: Repeat Discrepancy Tagger—a shotgun assembly finishing aid. *Bioinformatics*, *20*(5), 803–804.

Tang, C., Yang, M., Fang, Y., Luo, Y., Gao, S., Xiao, X., . . . Huang, H. (2016). The rubber tree genome reveals new insights into rubber production and species adaptation. *Nature Plants*, *2*(6), Article#16073.

Tang, H., Krishnakumar, V., Bidwell, S., Rosen, B., Chan, A., Zhou, S., ... Mayer, K. F. (2014). An improved genome release (version Mt4. 0) for the model legume *Medicago truncatula*. *BMC Genomics*, *15*(1), Article#312.

Tarailo-Graovac, M., Chen, N. (2009). Using repeatmasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics*, *25*(1), 4–10.

Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A., Conesa, A. (2011). Differential expression in RNA-seq: a matter of depth. *Genome Research*, *21*(12), 2213–2223.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., ... Rao, B. S. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, *4*(1), Article#41.

Tautz, D. (1989). Hypervariabflity of simple sequences as a general source for polymorphic DNA markers. *Nucleic Acids Research*, *17*(16), 6463–6471.

Tran, T. D., Cao, H. X., Jovtchev, G., Neumann, P., Novák, P., Fojtová, M., ... Fuchs, J. (2015). Centromere and telomere sequence alterations reflect the rapid genome evolution within the carnivorous plant genus *Genlisea*. *The Plant Journal*, *84*(6), 1087–1099.

Tranbarger, T. J., Dussert, S., Joët, T., Argout, X., Summo, M., Champion, A., ... Morcillo, F. (2011). Regulatory mechanisms underlying oil palm fruit mesocarp maturation, ripening, and functional specialization in lipid and carotenoid metabolism. *Plant Physiology*, *156*(2), 564–584.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., ... Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, *7*(3), 562–578.

Travers, K. J., Chin, C.-S., Rank, D. R., Eid, J. S., Turner, S. W. (2010). A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Research*, *38*(15), e159–e159.

Treangen, T. J., Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, *13*(1), 36–46.

Treangen, T. J., Sommer, D. D., Angly, F. E., Koren, S., Pop, M. (2011). Next generation

sequence assembly with AMOS. *Current Protocols in Bioinformatics*, *33*(1), 11–8.

Tuskan, G. A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., . . . Schein, J. (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*, *313*(5793), 1596–1604.

Udall, J. A., Dawe, R. K. (2018). Is it ordered correctly? validating genome assemblies by optical mapping. *The Plant Cell*, *30*(1), 7–14.

Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H., . . . Sidow, A. (2008). A high-resolution, nucleosome position map of *c. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Research*, *18*(7), 1051–1063.

Valouev, A., Li, L., Liu, Y.-C., Schwartz, D. C., Yang, Y., Zhang, Y., Waterman, M. S. (2006). Alignment of optical maps. *Journal of Computational Biology*, *13*(2), 442–462.

VanBuren, R., Wai, C. M., Keilwagen, J., Pardo, J. (2018). A chromosome-scale assembly of the model desiccation tolerant grass *Oropetium thomaeum*. *Plant Direct*, *2*(11), Article#e00096.

Vergnaud, G., Denoeud, F. (2000). Minisatellites: mutability and genome architecture. *Genome Research*, *10*(7), 899–907.

Verheye, W. (2010). Growth and production of oil palm. In *Land use, land cover and soil sciences.* Oxford: UNESCO-EOLSS Publishers.

Vicedomini, R., Vezzi, F., Scalabrin, S., Arvestad, L., Policriti, A. (2013). GAM-NGS: genomic assemblies merger for next generation sequencing. *BMC Bioinformatics*, *14*(7), Article#S6.

Vijay, V., Pimm, S. L., Jenkins, C. N., Smith, S. J. (2016). The impacts of oil palm on recent deforestation and biodiversity loss. *PLoS ONE*, *11*(7), Article#e0159668.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., Lee, T. v. d., Hornes, M., . . . Zabeau, M. (1995). AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Research*, *23*(21), 4407–4414.

Wang, W., Schalamun, M., Morales-Suarez, A., Kainer, D., Schwessinger, B., Lanfear, R. (2018). Assembly of chloroplast genomes with long-and short-read data: a comparison of approaches using *Eucalyptus pauciflora* as a test case. *BMC Genomics*, *19*(1), Article#977.

Wang, Z., Gerstein, M., Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*(1), 57–63.

Wang, Z., Wang, J., Wu, C., Deng, M. (2015). Estimation of isoform expression in RNA-seq data using a hierarchical Bayesian model. *Journal of Bioinformatics and Computational Biology*, *13*(06), Article#1542001.

Waterhouse, R. M., Seppey, M., Simão, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., ... Zdobnov, E. M. (2017). BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular Biology and Evolution*, *35*(3), 543–548.

Watson, J. D., Crick, F. H. (1953). The structure of DNA. *Nature*, *171*, 737–738.

Wilhelm, B. T., Landry, J.-R. (2009). RNA-Seq—quantitative measurement of expression through massively parallel RNA-sequencing. *Methods*, *48*(3), 249–257.

Williams, J. G., Kubelik, A. R., Livak, K. J., Rafalski, J. A., Tingey, S. V. (1990). DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Research*, *18*(22), 6531–6535.

Wittig, M., Juzenas, S., Vollstedt, M., Franke, A. (2018). High-Resolution HLA-Typing by Next-Generation Sequencing of Randomly Fragmented Target DNA. In *HLA Typing* (pp. 63–88). Springer.

Wu, I., Ben-Yehezkel, T. (2019). A single-molecule long-read survey of human transcriptomes using LoopSeq synthetic long read sequencing. *bioRxiv*, Article#532135.

Wu, R. (1972). Nucleotide sequence analysis of DNA. *Nature New Biology*, *236*(68), Article#198.

Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., ... Zhou, X. (2014). SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics*, *30*(12), 1660–1666.

Xiong, Y., Eickbush, T. H. (1990). Origin and evolution of retroelements based upon their reverse transcriptase sequences. *The EMBO Journal*, *9*(10), 3353–3362.

Yeap, W.-C., Loo, J. M., Wong, Y. C., Kulaveerasingam, H. (2014). Evaluation of suitable reference genes for qRT-PCR gene expression normalization in reproductive, vegetative tissues and during fruit development in oil palm. *Plant Cell, Tissue and Organ Culture*, *116*(1), 55–66.

Yu, J., Hu, S., Wang, J., Wong, G. K.-S., Li, S., Liu, B., ... Cao, M. (2002). A draft sequence of the rice genome (*Oryza sativa* l. ssp. indica). *Science*, *296*(5565), 79–92.

Yuan, Y., Bayer, P. E., Batley, J., Edwards, D. (2017). Improvements in genomic technologies: application to crop genomics. *Trends in Biotechnology*, *35*(6), 547–558.

Zerbino, D. R., Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome Research*, *18*(5), 821–829.

Zhang, J., Kuo, C.-C. J., Chen, L. (2014). WemIQ: an accurate and robust isoform quantification method for RNA-seq data. *Bioinformatics*, *31*(6), 878–885.

Zhao, W., He, X., Hoadley, K. A., Parker, J. S., Hayes, D. N., Perou, C. M. (2014). Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. *BMC Genomics*, *15*(1), Article#419.

LIST OF PUBLICATIONS AND PAPERS PRESENTED

PUBLICATIONS

Ho, H., **Low, J. Z.**, Gudimella, R., Tammi, M., Harikrishna, J. (2016). Expression patterns of inflorescence-and sex-specific transcripts in male and female inflorescences of african oil palm (*Elaeis guineensis*). *Annals of Applied Biology*, *168*(2), 274–289.

Kwong, Q. B., Teh, C. K., Ong, A. L., Heng, H. Y., Lee, H. L., Mohamed, M., **Low, J. Z.**, ... Appleton, D. R. (2016). Development and validation of a high-density SNP genotyping array for African oil palm. *Molecular Plant*, *9*(8), 1132–1141.

Low, J. Z., Khang, T. F., Tammi, M. T. (2017). CORNAS: coverage-dependent RNA-Seq analysis of gene expression data without biological replicates. *BMC Bioinformatics*, *18*(16), Article#575.

Low, J. Z., Tammi, M. T. (2017). *De novo* assembly of a genome. In L.W.Y. Low & M.T. Tammi (Eds.), *Bioinformatics: A Practical Handbook of Next Generation Sequencing and Its Applications* (pp. 107-125). Singapore: World Scientific.

PAPERS PRESENTED

Low, J. Z., Khang, T. F., Tammi, M. T. (2017). *CORNAS: coverage-dependent RNA-Seq analysis of gene expression data without biological replicates.* Paper presented at the 16th International Conference on Bioinformatics, 20-22 September 2017, Shenzhen, China.