Graduate School ETD Form 9
(Revised 01/14)

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By ANNE A. ANDERE

Entitled
DE NOVO GENOME ASSEMBLY OF THE BLOW FLY PHORMIA REGINA (DIPTERA: CALLIPHORIDAE)

For the degree of Master of Science

Is approved by the final examining committee:

DR. CHRISTINE PICARD

DR. STEPHEN RANDALL

DR. YUNLONG LIU

To the best of my knowledge and as understood by the student in the *Thesis/Dissertation Agreement. Publication Delay, and Certification/Disclaimer (Graduate School Form 32)*, this thesis/dissertation adheres to the provisions of Purdue University's "Policy on Integrity in Research" and the use of copyrighted material.

Dr. Christine Picard

Approved by Major Professor(s): _____

Approved by: Dr. Simon Atkinson 06/27/2014

Head of the Department Graduate Program Date

DE NOVO GENOME ASSEMBLY OF THE BLOW FLY PHORMIA REGINA (DIPTERA: CALLIPHORIDAE)

A Thesis

Submitted to the Faculty

of

Purdue University

by

Anne A. Andere

In Partial Fulfillment of the

Requirements for the Degree

of

Master of Science

August 2014

Purdue University

Indianapolis, Indiana

ACKNOWLEDGEMENTS

I would like to take this opportunity to express a heartfelt gratitude to my mentor and advisor Dr. Christine Picard for all the supportive criticism and encouragement that she continuously gave me in the past two years. I am extremely grateful to her for entrusting me to work on this exciting project, and I am looking forward to continue broadening my knowledge in this field in the years to come. I would also like to acknowledge my MS Thesis committee members Dr. Stephen Randall and Dr. Yunlong Liu for all their guidance regarding the project. To the Picard Lab members - past and present (Gina, Kevin, John, Kelsie, Abeer, Charity & all the undergrad students), thank you from the bottom of my heart for all the encouragement and the fun times during my journey as a masters student. You all made coming to lab every day something to look forward to. To my family, of whom I thank God to be a part of, thank you for always being by my side. Despite the many days that I locked myself reading and working on this project, you were always there. I will forever be in your debt for the constant support. Last but not least, I am grateful to the IUPUI School of Science start-up funds for enabling this project to be possible.

TABLE OF CONTENTS

Page

Page

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AFLP	amplified fragment length polymorphism
BAC	bacterial artificial chromosome
BLAST	basic local alignment search tool
bp	base pair
CEG	core eukaryotic genes
CEGMA	core eukaryotic genes mapping approach
CLC-GWB	CLC genomic workbench
COI	cytochrome oxidase subunit I
COII	cytochrome oxidase subunit II
DNA	deoxyribonucleic acid
GO	gene ontology
Gbp	giga base pairs
Kb	kilo bases
Mbp	million base pairs
mtDNA	mitochondrial DNA
NADH	nicotinamide adenine dinucleotide
ND4	NADH dehydrogenase subunit 4
ND4L	NADH dehydrogenase subunit 4L
NGS	next-generation sequencing
NCBI	national center for biotechnology information
NR	non-redundant
OLC	overlap/layout consensus
PCR	polymerase chain reaction
PE	paired-end

PMI	postmortem interval
RAPD	random amplified polymorphic DNA
RNA	ribonucleic acid
rRNA	ribosomonal RNA
SE	single-end
SNP	single nucleotide polymorphism
SSR	simple sequence repeats
TE	transposable elements
WGS	whole-genome sequencing

ABSTRACT

Andere, Anne A. M.S., Purdue University, August 2014. *De novo* Genome Assembly of the Blow Fly *Phormia regina* (Diptera: Calliphoridae). Major Professor: Christine J. Picard.

Phormia regina (Meigen), commonly known as the black blow fly is a dipteran that belongs to the family Calliphoridae. Calliphorids play an important role in various research fields including ecology, medical studies, veterinary and forensic sciences. *P. regina*, a non-model organism, is one of the most common forensically relevant insects in North America and is typically used to assist in estimating postmortem intervals (PMI). To better understand the roles *P. regina* plays in the numerous research fields, we re-constructed its genome using next generation sequencing technologies. The focus was on generating a reference genome through *de novo* assembly of high-throughput short read sequences. Following assembly, genetic markers were identified in the form of microsatellites and single nucleotide polymorphisms (SNPs) to aid in future population genetic surveys of *P. regina.*

A total 530 million 100 bp paired-end reads were obtained from five pooled male and female *P. regina* flies using the Illumina HiSeq2000 sequencing platform. A 524 Mbp draft genome was assembled using both sexes with 11,037 predicted genes.

The draft reference genome assembled from this study provides an important resource for investigating the genetic diversity that exists between and among blow fly

species; and empowers the understanding of their genetic basis in terms of adaptations, population structure and evolution. The genomic tools will facilitate the analysis of genome-wide studies using modern genomic techniques to boost a refined understanding of the evolutionary processes underlying genomic evolution between blow flies and other insect species.

CHAPTER 1. INTRODUCTION

Insects are one of the most diverse organisms and have adapted to a broad range of habitats. They perform a vast number of significant functions important to agriculture, human health, natural resources and the economy [1-3]. In particular, they are important in activities ranging from the provision of food for man and wildlife [4, 5], acting as predators, parasites and parasitoids [5, 6]; aiding in the pollination of crops and other flowering plants [7, 8], in the decomposition of organic matter, fertilization of soils [2, 9]; and even in the production of commercial products such as silk and honey [10, 11]. Insects have therefore been used in a variety of landmark studies including climate change, developmental biology, ecology, evolution, genetics, medicine and forensic sciences [1-3, 12-15]. The order Diptera, which is also known as the “true flies”, is one of the largest insect orders and contains a large number of diverse species worldwide [2, 12]. Calliphoridae are a family of insects in Diptera that are important scavengers as they feed and remove decomposing plant and animal material from the environment, and also act as pests and parasites in livestock [2].

1.1 Blow Flies (Diptera: Calliphoridae)

Calliphoridae, commonly known as blow flies, are a family of flies belonging to the Diptera. There are over a thousand species of Calliphoridae with a worldwide distribution

[2], making them a valuable model of study in various scientific research studies. Researchers in fields ranging from veterinary, medical, forensic, ecology, genetics and many other disciplines have had an interest in studying various species from this family due to the diverse roles they play in different scopes of studies [2, 13, 16]. Calliphoridae are one of the most dominant insects which contribute to the decomposition process [17] and are one of the important groups of insects that are typically the first to arrive at the scene of a dead body [2, 18]. They are therefore an important group of insects in forensic sciences, as forensic entomologists commonly use them to assist in estimating the time since death, generally referred to as postmortem interval (PMI). Insects are especially useful when decomposition has progressed and traditional medical and physiological estimates of time since death are no longer accurate. Calliphoridae are not only limited to decomposition, but they also play important roles in other disciplines [19]. The larvae of certain species of blow flies such as *Lucilia sericata* (Meigen) are utilized in wound care, specifically maggot debridement therapy (MDT) [13, 20, 21]. In Australia, the larvae of another closely related species, *Lucilia cuprina* (Wiedemann), are responsible for parasitizing sheep and causing flystrike, leading to major economic losses [16].

1.2 *Phormia regina* (Black Blow Fly)

Phormia regina (Meigen, 1826), a member of the sub family Chrysomyinae [22] is the predominant forensically relevant species of blow fly in Northern America [2]. The adult flies have a characteristic dark green to olive color on both the thorax and abdomen and they typically range in length from 7 to 9 mm (Figure 1.1) [2].

P. regina is one of the most common blow fly species across North America and is present throughout the year in different seasons. It is a Holarctic blow fly species, meaning it is mostly found in non-tropical regions in the northern continents of the world. It is typically considered a cold weather fly as it is mostly abundant during the cooler spring and fall temperatures. In warmer areas it will be found in higher altitudes [23]. It is dominant in Southern USA during the winter months and in the Northern USA and Canada during summer months [2, 24]. It is not only found in North America, but is also present in Western Europe [25, 26]. Aside from forensic studies, *P. regina* has also been used as a model of study in a number of various scientific studies varying from genetic, fertility, behavioral, medical and gene expression studies [13, 14, 21, 27-31].

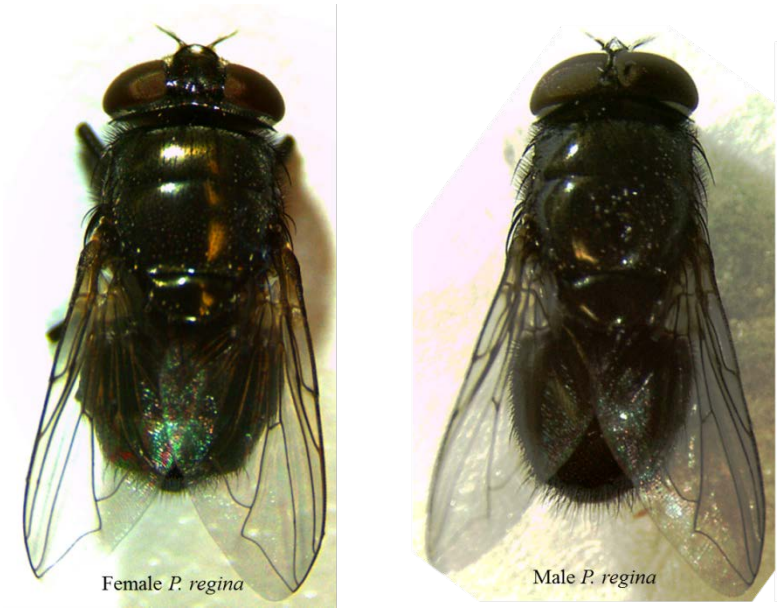


Figure 1.1 Dorsal view of the adult female and male. (Photo courtesy J. Whale)

1.2.1 Life Cycle

Blow flies use a visual search coupled with assistance from powerful olfactory receptors on their antennae to detect the location of human and animal remains [2]. Once detected, female flies locate wound openings and natural orifices such as the mouth and eyes to lay their eggs. After a period of time, the eggs hatch into a first instar larvae which voraciously feed on the carrion and eventually molts into two additional instar stages before crawling away from the corpse to pupate. Following pupation, the adult fly emerges [2, 24]. Depending on a number of variables, (e.g. the availability of suitable food resources, other blow fly species, ambient temperature and climatic conditions), the duration of *P. regina's* development from egg to adult is typically 10 - 14 days (Figure 1.2) [24].

The life cycle of *P. regina* is important in forensic investigations as their developmental rate can be used to estimate the time since death. Larvae present on the body are collected and their approximate age determined using published temperature-dependent developmental data, which is extrapolated backwards to estimate minimum PMI [2].

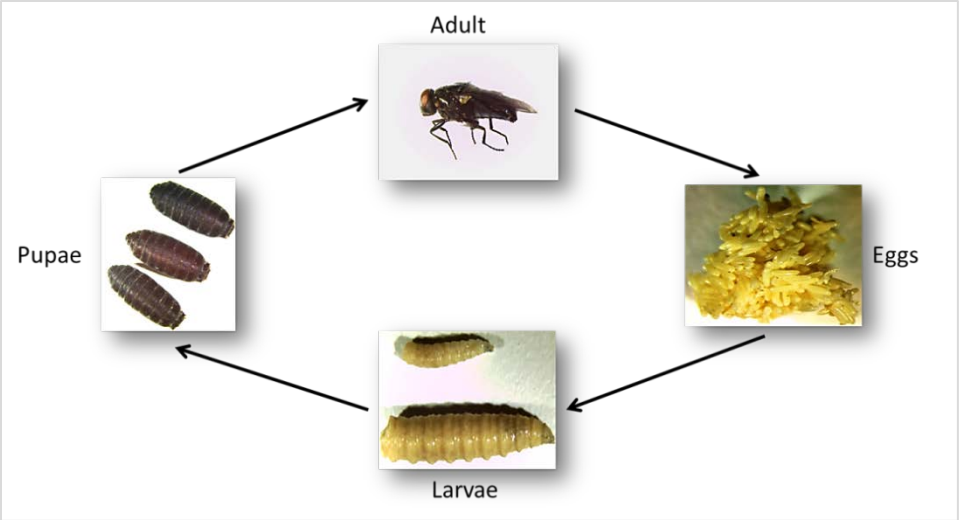


Figure 1.2 The life cycle of *Phormia regina.* The development of blow flies follows a predictable cycle that typically has four distinct life stages, namely egg, larvae, pupae and adult, that takes approximately 10 – 14 days [24]. (Photo courtesy J. Whale)

1.3 The Genetics of Blow Flies

1.3.1 Molecular Identification and Population Genetics

A majority of insects are morphologically and phenotypically similar, especially in the early stages of development, the larval and pupal stages [32, 33]. Because morphologically similar species may differ in development time, it is important to correctly identify the species of fly when conducting forensic investigations. Since larvae are collected at different stages of development, it is necessary to obtain molecular markers that will assist in the correct identification of the species, sex, and if possible, the geographic location of origin [2, 33, 34]. Molecular markers present in different developmental stages will hopefully be able to categorize the larvae into reliable age estimates, which in conjunction with published temperature-dependent developmental data will aid to estimate the time of insect activity and eventually assist in estimating PMI [26, 35].

Advances in molecular biology and genetic identification techniques have increased the use of DNA for molecular identification purposes [2]. The most common molecular method for identification is the use of mitochondrial DNA (mtDNA) sequencing. The mitochondrial gene cytochrome oxidase subunit I (COI) has been shown to serve as the core of a global bio-identification system in animals [36] and its profiles have been used to correctly analyze taxa to their appropriate phylum or order [36]. The COI gene has also been used in DNA typing of blow flies due to high interspecific nucleotide variation, making it a valuable tool for determining markers that can be used for the molecular identification of blow flies [26, 37]. The combination of COI and other three mitochondrial genes, cytochrome oxidase subunit II (COII), NADH dehydrogenase subunit 4 (ND4) and NADH dehydrogenase subunit 4L (ND4L) have been used in molecular phylogenetic analysis to test evolutionary hypotheses of different blow fly species in Australia, and results showed that they were able to identify most of the species reliably [38]. DNA sequences from ribosomal RNA (rRNA) have also been used in the molecular identification of blow flies. The large sub-unit 28S of rRNA has been used to study phylogenetic relationships among blow fly species with the intent to resolve key taxonomic relationships within the Calliphoridae family [39, 40]. Sequencing mtDNA and rRNA correctly identifies most blow fly species [26, 41]; however, these molecular markers fail for closely related species [26, 42]. This is due to the fact that low sequence divergences of sister species cause intra- and inter- specific nucleotide divergence to overlap resulting in similar haplotypes causing misidentification of a species [26, 41, 42]. A good example is of two closely related species, *Lucilia caesar* (Linnaeus) and *Lucilia illustris* (Meigen) whose COI sequences were unable to accurately

distinguish them due to some overlapping identical haplotypes [26, 38, 42]. Therefore, in cases where there is a failure of species separation using such molecular markers, they should not be used alone but in conjunction with additional markers to overcome these problems.

Mitochondrial DNA is more appropriate for species discrimination than for detecting population variation [2]. Thus the population genetic structure of blow flies have been studied using nuclear DNA, mostly by random amplified polymorphic DNA (RAPD), amplified fragment length polymorphism (AFLP) profiles, neither of which require any genomic information, and more rarely, microsatellites [43-45]. RAPD markers are DNA fragments produced from the amplification of random genomic DNA segments. Profiles from RAPD have been used to confirm the species integrity of *L. sericata* (Meigen) and *L. cuprina* (Wiedmann), which are two species that are similar in morphology and ecology, yet populations of each species vary in their importance as pests in different parts of the world [44]. The limitation of RAPDs is the sensitivity of the profiles to experimental conditions making it difficult to independently reproduce the same profiles [2, 46]. This variability limits its use in the characterization of distantly related groups, thus, it is recommended that a complementary technique such as DNA sequencing should be used in conjunction with RAPD [46].

AFLP uses selective PCR amplification of restriction fragment elements from genomic DNA [47] to detect restriction site polymorphisms. The number of loci produced from AFLP can be used to infer population genetic structure [2]. AFLP profiles for *P. regina* have been generated and used to find significant variation among discrete samples collected in one location over a short period of time [43].

They have also been used in conjunction with mtDNA to correct an error of misidentification between two afrotropical blow flies (*Chrysomya putoria* and *Chrysomya choloropyga* -Weidemann) [48].

Microsatellites are simple sequence repeats (SSR) that are popular genetic markers because they are highly polymorphic [2, 49, 50]. They are used in population genetics as they can assess the direction of gene flow and genetic differentiation between populations, and have a potential of providing data to be used for fine-scale phylogenetic analysis to the level of closely related species [2, 45, 51]. Microsatellite typing methods were used to develop polymorphic microsatellite loci in *L. illustris* and *L.sericata* [45]. The development of microsatellites for non-model species used to be labor intensive, however, next generation sequencing (NGS) techniques has made it easier and faster for the identification of large numbers of microsatellites at reduced costs using sequenced DNA [49, 50, 52].

1.3.2 Limited Genomic Resources in *Phormia regina*

Phormia regina is a non-model organism and has limited genomic resources available. It has variable developmental time which might have been influenced by the different geographic and climatic seasons in which it can survive [18, 24, 53]. Previous studies show that there is genetic variation within *P.regina* populations found in different geographic locations [43] and this concept can be applied to other blow fly species as well. The presence of population structure enable them to be geographically distinguished, and this can play a major role in various studies including potential forensic applications

in the determination of corpse relocation [54]. Due to its dominant status in North America, and its capability to survive in numerous geographic and climatic conditions all year round, *P. regina* is a good candidate for the assembly of its genome. This genome can be a representative reference genome for the Calliphoridae family, as there are currently no known genomes for this family. Next generation sequencing (NGS) technologies will aid in facilitating novel gene detection and the development of genomic tools on *P. regina* that can be used for further downstream analysis.

1.4 Sequencing Technologies

Over the years, there have been several revolutionary approaches to DNA sequencing, with the recent one being the introduction of NGS technologies which are non-Sanger based high throughput DNA sequencing technologies [55]. NGS has experienced fast technological advancements largely due to the commercial introduction and availability of sequencing instruments [55]. Advanced NGS technology has triggered rapid and cost-effective methods of sequencing genomic DNA with the technologies producing up to one billion bases in a single run [56]. It has been employed in numerous areas of genetics and genomics. Genomic information is now easily attainable leading to an explosion of NGS use in a wide array of applications. New areas of biological inquiry that involve genome assemblies have opened up and scientists across many fields have an opportunity to use this data and focus on a multitude of organisms, populations and ecologies [57-61].

1.4.1 History of DNA Sequencing

In the late 1970's, the first generation of DNA sequencing methods were independently developed by Allan Maxam and Walter Gilbert (Maxam-Gilbert sequencing method) [62], followed by the development of Sanger sequencing by Frederick Sanger and his colleagues [63]. At this time it was quite an arduous, expensive and time-consuming task to sequence DNA, but the reads produced were relatively long, high quality reads between 300 and 1000 nucleotides [63, 64]. As time progressed and sequencing technologies introduced, faster and more accurate automatic sequencing instruments were introduced [65]. By the 2000's, there was a rapid evolution in sequencing development that produced even more powerful instruments that have generated a substantial increase in throughput with moderate accuracy, and subsequently a reduction of the cost and manpower needed to perform sequencing [65, 66]. However, the reads produced have been shorter, ranging between 35 – 250 bp [56, 65].

The hallmark of NGS has been the increase in throughput and decrease in cost as compared to previous technologies [67], with much improvement in read length and coverage. Sequencing first began with the use of bacterial artificial chromosomes (BAC), then the development of automated pyro sequencing, advancing to the explosion of newer NGS methods such as the whole genome shotgun sequencing (WGS) techniques that produce high quality reads [63, 68, 69].

BAC-based sequencing was used in the initial stages of various genome sequencing projects including human and *Drosophila* genomes [70]. In BAC sequencing, several copies of the genome are randomly sheared into fragments of approximately 150 kb long, and each of the fragments is inserted into a bacterial artificial chromosomes. Each BAC

clone is amplified in bacterial culture, isolated, and sheared to produce size-selected pieces of approximately 2 – 3 kb, which are sub-cloned into plasmid vectors, amplified in bacterial culture and chosen for sequencing [55]. The sequences from each group of subclone are then assembled into larger genomic fragments.

In WGS sequencing, genomic DNA is sheared and size selected into distinct sizes and cloned into plasmid and fosmid vectors. The ends of the subclones are sequenced to generate sequenced reads, preferably paired-end sequenced reads which are one of the necessary requirements for generating linking information to assist in whole-genome assembly algorithms (more below) [55]. The sequencing instrument reads the DNA fragment starting from both ends of the template fragment, producing two reads that overlap or are separated by a short gap of an approximate known length, referred to as insert size [67], as illustrated in Figure 1.3.

Currently, most NGS sequencing platforms require that the template DNA is short, between 200-1000 bp and that each template contains forward and reverse primer-binding sites, introduced during library template preparation [67]. NGS instruments are capable of producing millions of DNA sequence reads in a single run, however there is induction of high fragmentation especially in highly polymorphic or highly repetitive genomes [55].

1.5 Sequencing Platforms

Sequencers vary in their sequencing mechanisms, the read lengths produced, the number of reads that can be produced per run, and the cost of sequencing [65]. There are a number of commercially available platforms, each with advantages and disadvantages.

Some of the commonly known platforms are by Illumina (www.illumina.com), Roche 454 Life Sciences (www.454.com) and ion torrent (www.lifetechnologies.com).

The 454 sequencing platform manufactured by Roche was the first platform to be introduced commercially [55, 65] and is known for producing longer reads from 400 bp up to 1000 bp long. It currently has two sequencers (GS Junior and GS FLX systems) that are used for whole genome and transcriptome sequencing, targeted resequencing and metagenomic sequencing (www.454.com). Illumina is a leading sequencing platform and offers a variety of sequencing instruments (MiSeq, NextSeq, HiSeq, and HiSeq X) that are used for a number of projects involving genome, epigenome and transcriptome sequencing. The read lengths produced by Illumina sequencers range in size from 125 bp to 300 bp (www.illumina.com). The Ion Torrent sequencing platform is a benchtop high-throughput sequencing instrument that is ideal for small sequencing projects and produces reads ranging from 200 bp – 400bp in length (www.lifetechnologies.com).

1.5.1 Whole Genome Sequencing

The sequencing process first involves shearing the DNA of an organism into fragments suitable for DNA sequencing. Most of the NGS platforms work by annealing linkers to blunt-ended fragment libraries generated from a genome or DNA of interest. The linkers used are often specific to the sequencing platform used. Adapter sequences are then ligated to the DNA fragments enabling selective amplification by polymerase chain reaction (PCR) [55]. This is in contrary to the BAC sequencing method where a bacterial cloning step is required to amplify the fragment in a bacterial intermediate.

Once sequences have been obtained, their quality must be assessed. Each base called for a given read is assigned a quality score. A quality score is an algorithm that measures the quality and accuracy of the DNA sequences produced, and is assigned to every base called [71, 72]. Quality scores enhance the ability of an assembler to discriminate correct base calls from incorrect base calls, by enabling the filtration or removal of lower quality reads during the pre-assembly process [73, 74]. Eventually, the accuracy of the input data is enhanced as the assembler is able to differentiate true DNA polymorphisms from sequencing errors and in identifying ambiguous nucleotides [71, 72].

Quality scores (Q) can be defined by the following equation:

$$Q = -10 * log_{10}(P)$$

Equation 1.1 Quality score equation. *P* is the probability that the corresponding base call is incorrect [71].

where 'P' is the estimated error probability. Therefore high quality values correspond to low error probabilities and vice versa [71, 72, 75] as shown in Table 1.1.

Table 1.1 Phred quality scores calculation. Quality scores are assigned to each nucleotide base call. A quality score of 20 assigned to a base indicates a 1 in 100 probability of an erroneous call, or alternatively is 99% accurate.

PHRED QUALITY SCORE	PROBABILITY OF INCORRECT BASE CALL	BASE CALL ACCURACY
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%

Early sequencing models were capable of only generating unpaired or single-end (SE) reads, however many newer models now have the option of producing paired-end (PE) reads [76]. PE reads (Figure 1.3) are pairs of reads generated from a single DNA fragment of a fixed size, with an orientation assigned to them by the sequencer i.e.

forward or reverse reads [77]. PE reads come from sequencing each of the DNA fragments twice, once from either end of the fragment [55, 66, 78] leaving a defined space in between each fragment that is not sequenced.

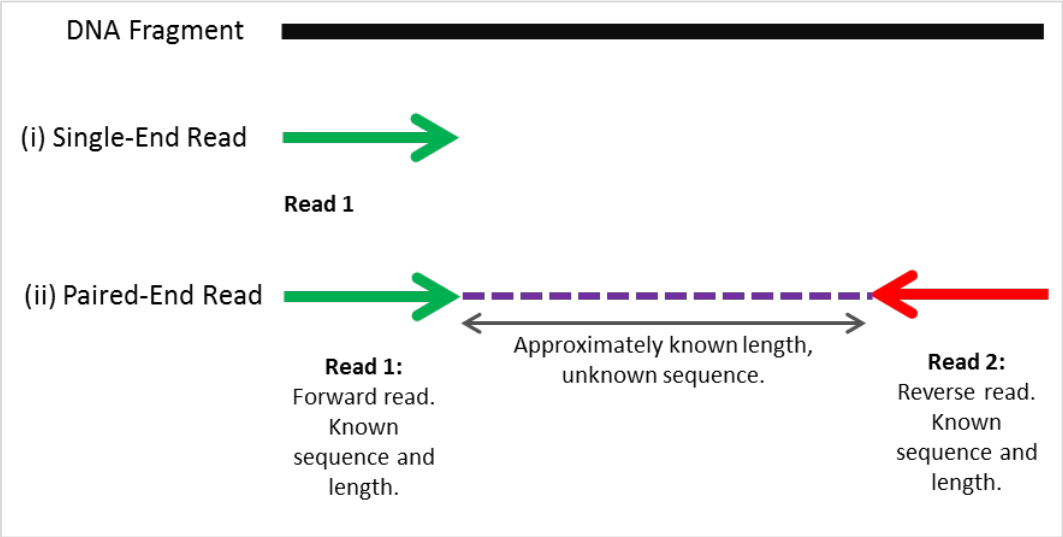


Figure 1.3 Two types of libraries and their orientation. (i) Single-end (SE) reads are short sequenced fragments that have only one end sequenced. (ii) Paired-end (PE) reads are generated from each DNA fragment sequenced both ends producing forward and reverse reads.

PE reads have a known spacing and orientation, which provides the required information to link similar sequences, improving the efficiency of sequence assembly [74]. During assembly, the assembler uses both the orientation and the expected distance between the PE reads to reconstruct a genome [77]. Contiguity in the sequences is created, facilitating the genome assembly process [55]. Depending on the library preparation technique, the distance between the sequenced ends can be as short as 200 bp or as large as several tens of kilobases [66].

PE reads can be used to estimate how far apart each read should be from its mate in the final assembly, therefore if one read is mapped to a unique position; it is possible to assign an approximate location for its partner. PE reads can also indicate the size of repetitive regions [66] and help to span a particular repeat to assist in assembling data unambiguously [77].

1.6 *De novo* Genome Assembly

De novo genome assembly refers to the reconstruction of a draft genome using a collection of randomly sampled fragments produced from sequenced DNA [64, 76, 79]. It is the process where individual sequenced reads are merged together to form long contiguous sequences (‘contigs’) that share the same nucleotide sequences as the original template of the DNA which was sequenced [80]. The assumption is that if two or more sequence reads share an overlapping substring of bases, then they are likely to have originated from the same chromosomal region in the genome [79]. The output of an assembly is typically a set of contigs created from a consensus sequence provided by multiple sequence alignments of overlapping reads (Figure 1.4) [76].

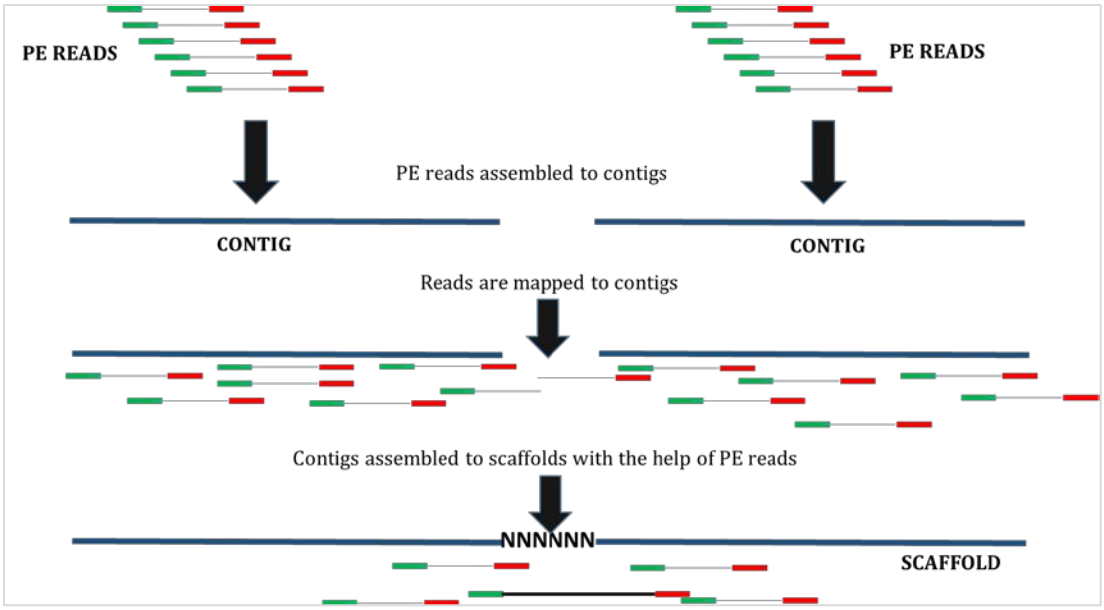


Figure 1.4 Contig and scaffold assembly. Overlapping reads are aligned and assembled to create a long contiguous sequence known as contig. Contigs are then ordered, oriented and linked with the help of pair-end reads. Gaps between the contigs are represented by the letter 'N' [76, 81].

Contigs are ordered and oriented into scaffolded sequences with the help of PE reads or additional sequencing (Figure 1.4). PE reads are used to find the approximate distances between non-repetitive contigs in the genome and connects the ones that are in the same orientation, producing scaffolds with gaps [67, 76, 81]. The gaps are often represented by a consecutive number of the letter 'N' which denotes regions of uncertainty [64]. The length of the N's may represent the gap length based on the mates of the paired-reads spanning both contigs [76].

Eukaryotic genomes vary by several orders of magnitude in terms of size. Therefore assembled draft genomes differ in size, ranging from hundreds to thousands to millions of bases. The human genome is ~3000 Mbp [70] while the fruit fly, *Drosophila melanogaster*'s genome is ~130 Mbp [59, 82].

For that reason, data storage for assemblies is an important concept to keep in mind during the sequencing and assembly process. It is worth noting that the size of the genome does not correlate with the number of genes present or the complexity of the organism [82].

The widely accepted data file format for saving an assembly is known as the FASTA format (Figure 1.5) [76]. FASTA files are text based files which represent the sequence of the elements being studied, in this case, nucleotide sequences. The nucleotide sequences are represented by a string of characters that describe DNA, namely adenine, guanine, cytosine and thymine (‘A’,’G’,’C’,’T’). Unknown characters or ambiguous nucleotides that are not read properly by the assembler will be represented by the letter ‘N’ in the assembly [79].

```
> Contig 1
GAGCNATATTAAGCGAGATATTTAAGATTAAAGAATTTTATATAAAATATCGATTTCCAG
TGAGAAATAAAGGTGATCAGAAATAAAGCTTATTTTTCCATTTGGAGCCATATTAAGCGA
GATATTTAAGATTAAAGAATTTTATATAAAATAT
> Contig 2
TAAAAATCGATTATTTAATGAAAAACATAAAAATCGACTTTAAAGCTAAACCAGCAGAGA
TAGAGCCCAAAGGACGTTTATCTGTGATCACCCATTTTTTTTCTAAAAATCGATTATTTA
ATGAAAAACATAAAAATCGACTTTAAAGCTAAACCAGCAGAGAT
```

Figure 1.5 FASTA file format. The first line is the description line which is distinguished from the sequence by the symbol ‘>’. The description is followed by the nucleotide or peptide sequences.

Sequence reads can either be saved in the FASTA or FASTQ format [75]. FASTQ files (Figure 1.6) are just an extension of the FASTA format, but are more informative in that they store a numeric quality score associated with each nucleotide in a sequence [75].

```
@HWI-ST1234:70:C1T4PACXX:1:1101:1191:2028/1 1:N:0:AGTTCC
TTTGTGTGTTAAACTATAAATAAATTTGACAAATATTGAAATATTGAATTCATTTCAATTCAATTCCTTTTACATAAATCAATATTTGCTTT
+
=DDFFFHHHHGJJJJIJJJIJJJIJJJIJJJJJJJJJIJJIIJJJJJJIJIJJJJJJJJJHJJIIHHIJJJIIHHIJJJIJIJJHHHHHHHF
@HWI-ST1234:70:C1T4PACXX:1:1101:1191:2028/2 2:N:0:AGTTCC
ATTAGAAAATATAATATTTGTTCTGCTTTAGGTAACTAAATGTATCCAAACTAAAGACTCGTAATATTAATGTCGATTTACTTGTTTACTTT
+
CCCFFFFFHHHHHJJJJJJJJJJJJJJJJJJJCHIJJJIJJJHHIJIIFIHIIJJJJJJJJIJJIJJJJGIJJJJHIJJJJHHHHHHFFFFF
```

Figure 1.6 FASTQ file format has four lines per sequence. Line 1: Starts with '@' and is a sequence identifier with an optional description. Line 2: Raw sequence with nucleotides. Line 3: starts with '+'. It signals the end of sequence lines and beginning of quality strings. Line 4: Quality values. They are indicated in form of ASCII characters (ie the numbers 0-9, the letters a-z A-Z, and punctuation symbols).

1.6.1 Importance of *De novo* Genome Assembly

De novo genome assembly facilitates the discovery of variable polymorphic regions within and among different species, especially in the absence of a reference genome of the same species or a closely related species. It also enables the discovery of novel genes and facilitates the creation of a catalog of genes from draft genomes [83, 84].

1.6.2 Challenges of *De novo* Assembly

There are a number of challenges associated with the genome assembly. Ideally, an assembled genome should comprise of one contig for every chromosome of the organism being sequenced; however, in most cases many contigs are created due to a combination of factors. They emanate from the high volume of short reads produced by the NGS platforms, the presence of repetitive sequences, sequencing errors, the absence of a reference genome for comparison purposes, time and memory used by the assembly program, uneven coverage of some genomic regions, and storage resources for the input and output data. [79]. The greatest challenge that affects the assembly process leading to mis-assemblies is the repetitiveness of the sequence data.

During sequencing, NGS covers each base position a number of predetermined times producing a large volume of data made up of short read lengths [59, 76]. Short reads have less information per read which makes the process of assembly computationally difficult [76] as the assembly would require a higher coverage of the shorter reads to make up for the less information delivered. Higher coverage subsequently introduces complications to the assembly process as it increases the computational demands that usually accompany large data sets [76]. Fragmented assemblies are commonly created when short reads are used, since the short reads make it difficult for the assembler to resolve repeat regions [56].

Assembly errors can arise when there are repetitive elements longer than the read length being used in the assembly, including non-unique elements from gene duplication or transposable elements (TE) [80]. Generally larger genomes tend to have more repetitive DNA than smaller genomes, represented by TE's, simple sequence repeats (SSR), and duplicated genes. Repetitive DNA are sequences that are similar or identical to sequences elsewhere in the genome [77]. They align to multiple positions in the reference genome and create ambiguity as to which location was the true source of the read [67, 77]. Repeats can range in size from1 - 2 bases (i.e. mono- or dinucleotides), to millions of bases [77]. There is a substantial amount of repetitive sequences in eukaryotic genomes, as evidenced in the human genome, as nearly half is composed of repetitive DNA sequences [77]. These repeats are a problem in alignment and assembly, and are one of the main causes which contribute to the complexity in the assembly process [59, 80].

They ‘confuse’ the assembly process because reads originating from distinct copies of the repeat appear identical to the assembler making it difficult to differentiate sequencing error from polymorphism between repeat copies, thus resulting in incorrect placement of the reads leading to poor quality assemblies [56, 77, 78].

Resolving repetitive sequences is difficult, especially when the length of the repeated unit is longer than the sequence reads (Figure 1.7) [80]. An important component of dealing with repetitive sequences is to have PE reads [74]. When the repeats are longer than the read length, gaps are created in the assembly [77].

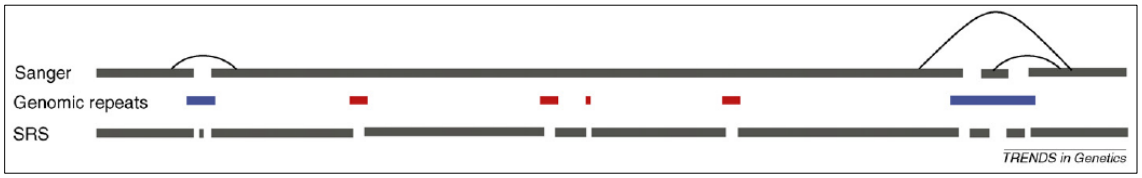


Figure 1.7. The grey thick lines with spaces represent different sized contigs from one region, created from longer reads (Sanger) and short reads (SRS).The boxes in the middle represent repeats (Blue = repeats longer than 800 bp. Red = repeats shorter than 800 bp.). An assembly of the long reads (top line) would correctly resolve the short repeats, breaking only at the boundaries of long repeats. PE reads (shown by the thin lines) would significantly help in connecting across the repeats. An assembly created from short reads (bottom line) is more fragmented as it breaks at all repeat boundaries. PE reads would considerably assist in connecting the contigs across the repeat regions [56]. Reproduced with permission from Elsevier Publishing Company.

PE reads are therefore capable to span repeats longer than the individual reads and thus are usually essential when assembling genomes [76]. An alternative to dealing with repetitive reads is sequencing longer reads (Figure 1.7), which would help in connecting gaps produced by the repeats in a sequence alignment [56] using other sequencing technologies as part of an integrative approach. Since some repeat instances are longer than the read lengths, shorter reads have a low capability of resolving these genomic repeats (Figure 1.7) [76].

1.7 *De novo* Genome Assemblers

Sequencing is not error-proof and thus each position has an associated probability of producing sequencing errors. The software used in different assemblers therefore needs to be programmed to be flexible to allow and tolerate imperfect read alignments. Due to the influx of data from the rapid evolution and availability of NGS technologies, a surge of assembler packages each with an aim of producing high quality assemblies have emerged. There are numerous assembler tools from freely available open source packages as well as commercial packages [85-88]. The open source packages are available to the public for use and modification from the original design if need be.

Choosing an assembler depends on the programs usability and the quality and composition of the assembled product [80, 89]. Usability comprises of factors such as hardware and software requirements, ease of installation, execution and speed of the whole process [80]. The user may choose an assembler through the process of elimination. This entails testing different assemblers and eventually choosing the one that generates the desired or adequate assembly as determined by a set of specific assembly metrics.

As previously mentioned, eukaryotic genomes are composed of repetitive DNA, which are a challenge to assemblers because genomic regions share perfect repeats which are indistinguishable even if originating from different genomic locations [76]. For repeats with slight variations, applying stringent settings to the parameters of the assembly and alignment process can assist the assembler to place the reads at their correct alignment position [76]. There are two standard data structure approaches frequently used in the assembly process by most common assemblers. These are de Bruijn graph approaches and overlap layout consensus (OLC) approaches [76] and each are better

suited for different read lengths and sequencing depths [90]. Both use graph algorithms in the construction of an assembly. Typically, in computational terms, a graph includes nodes and edges where nodes are the vertices or points and edges are the lines or arcs connecting the nodes. De Bruijn graph approaches are the favored approach used by most assemblers [85-87, 91] as the assembly of repetitive regions present in short sequenced reads are better handled than by the OLC approach.

1.7.1 Overlap Layout Consensus (OLC) Approach

The OLC consists of three main stages from which it derives its name: the overlap, layout and consensus stage. The overlap stage is the first stage which begins by comparing all the reads to each other in a pairwise manner and computing overlaps between them constructing an overlap graph [76, 90]. This is followed by a layout stage where the overlap graph is analyzed and simplified by the removal of redundant reads and the graph algorithm determines a relative placement of the reads along the genome [76, 90]. Finally in the consensus stage, the assembler builds an alignment of all the reads covering the genome inferring the original sequence of the assembled genome through the consensus of the aligned reads [76, 90].

The ultimate goal of the OLC is to identify appropriate paths that pass through the constructed graphs only once, which unfortunately is computationally difficult. The overlap phase is time consuming and computationally intensive especially in NGS data that contains millions of short reads, as the overlap of every single pair of reads in a data set is determined. [76, 90]. The first genome assemblers were OLC based and they

targeted reads from Sanger sequencing. Therefore the OLC was initially designed to be used for Sanger-based sequencing technology, and thus it works quite well with longer reads produced by Sanger-based assemblers [90]. It was not programmed for the assembly of short read length (~30 bp – 100bp) that usually have a high read depth (>30 X coverage) to compensate for the short length [90]. The computational complexity of OLC has limited its use by assemblers, however, as NGS data is now on the rise, current OLC assemblers such as Newbler (http://www.454.com/products/analysis-software/) and Celera [92] have been well optimized to handle the short reads from NGS data.

1.7.2 De Bruijn Graph Approach

De Bruijn graph approach uses k-mer graphs which are useful for a large volume of data with repetitive regions and short reads [76, 86]. K-mers are a sequence of consecutive length *k* nucleotides in a DNA sequence, where *k* is any positive integer (Figure 1.8).

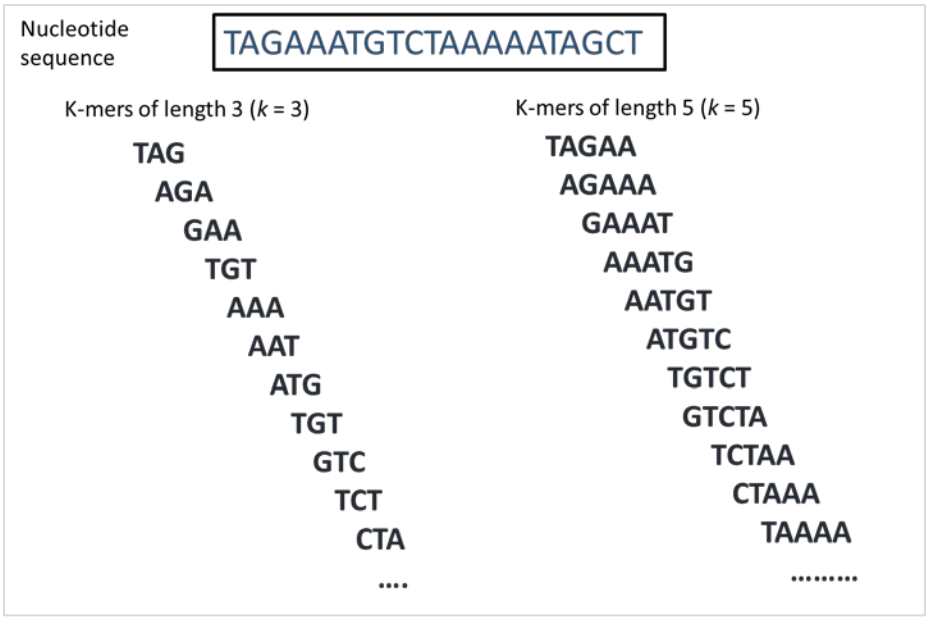


Figure 1.8 A K-mer is a substring composed of length k nucleotides in a DNA sequence, where k is a specified length of the string of nucleotides. The sequence is randomly fragmented into user-specified k-mer length. In the example shown, a 21 nucleotide sequence is fragmented into k-mers of size k = 3 and k = 5. A string of length L has (L – k + 1) k-mers. For example the read length in the figure has 19 k-mers and 17 k-mers when k = 3 and 5 respectively [91].

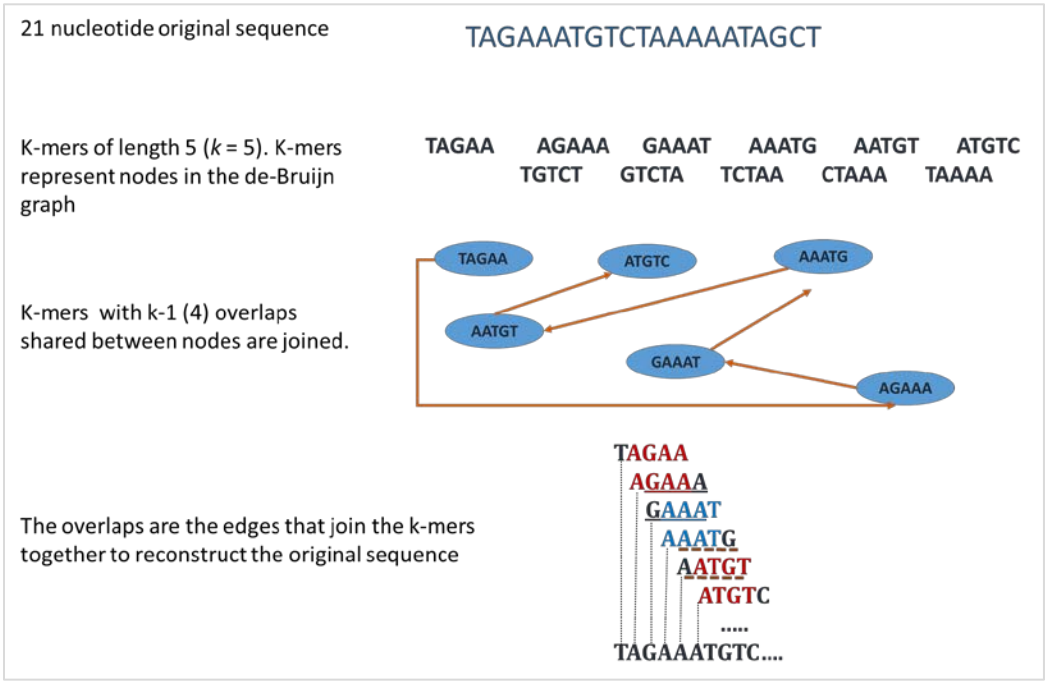


Figure 1.9 Simplified version of the de Bruijn graph construction. A 21 nucleotide genomic sequence is fragmented in k-mers of size *k*=5. K-mers represent nodes in the de Bruijn graph. The Overlaps (edges) of length k-1, in this case 4 nucleotides, are used to join the nodes (k-mers) together with the effort of reconstructing the original sequence.

The nodes are the k-mers of a specified length *k* contained within the sequencing reads [93] and the edges are the overlaps represented by the overlaps of the k-mers, and they join k-mers that overlap by *k*-1 nucleotides (Figure 1.9). If two k-mers are adjacent in at least one sequence read, they are connected and form a single path (Figure 1.9) [56, 76]. During the k-mer alignment process, non-repetitive genomic sequences would form a single path through the graph. On the other hand, repetitive genomic sequences would form bubbles in the graph (Figure 1.10) which would allow more than one possible reconstruction of a path [76].

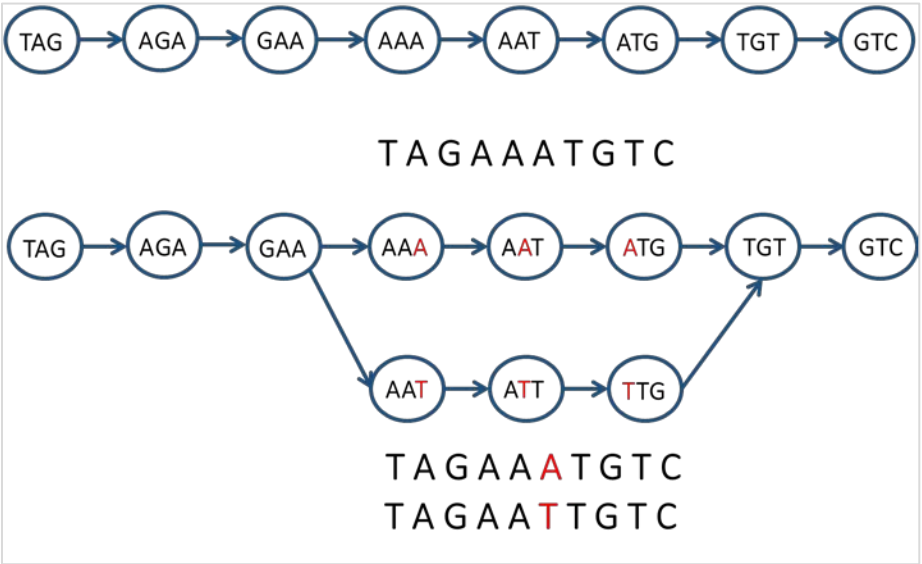


Figure 1.10 A simple 11 nucleotide sequence as a de Bruijn graph (top line). The bottom line shows the same sequence, but now with a SNP, creating a bubble in the graph.

Bubbles are diverges and converges of a path that occur due to the presence of polymorphism or sequencing errors among the k-mer sequences [86].

In summary, the de Bruijn graph begins by breaking the reads into a set of shorter segments, which are used in the construction of the graph using the segments as nodes. If

two segments are adjacent to one another in the original reads, the segments are connected [56]. The paths that pass through the graphs are the potential contigs and they are converted to sequences [76].

In eukaryotic genomes, the graphs can be very large in size as they would involve billions of k-mers [93], and are complex due to the presence of sequencing errors and repetitive regions in the reads leading to high usage of computational memory during the assembly. However, when compared to the OLC algorithm, de Bruijn approach is memory efficient as it does not have the memory intensive pairwise comparison step that is time consuming and can exhaust the memory especially if the data has a high read depth. Some of the common assemblers that use the de Bruijn graph approach are Velvet [86], ABySS [91], CLC-Genomic Workbench (www.clcbio.com), SOAPdenovo [94], and ALLPATHS [87].

Velvet and SOAPdenovo assemblers are open source software and freely available [86, 94]. Both are compiled and run on a UNIX operating system, therefore some basic knowledge in programming skills are needed in order to perform an assembly [86, 94, 95]. Velvet requires a large amount of physical memory (RAM) and so it is recommended to be downloaded on a system with as much physical memory as possible (>12 GB) [86, 95]. Alternatively, SOAPdenovo can be run in a system with a minimum of 5 GB physical memory especially for small bacterial and fungal genomes, however, for big genomes like the human genome, it would require ~150 GB RAM [94]. Both Velvet and SOAPdenovo need to be installed and run in a system that has a 64 bit environment. Velvet is designed for Illumina and SOLiD reads [86] and is recommended for use in assembling small to medium sized genomes. SOAPdenovo is designed specifically for

Illumina reads and large genomes [94]. CLC-Genomic Workbench (QIAGEN Inc., Valencia, CA, USA) is commercially based software that can be run on both UNIX and Windows operating systems. It is also recommended to be installed in a 64 bit environment with a minimum of 8 GB RAM. CLC-GWB is mainly for users that lack programming skills due to its graphic user interface and can assemble data from common sequencing platforms such as Illumina, Roche 454, SOLiD and ion torrent. CLC-GWB is relatively costly, and as a commercial product, its source code is not available and cannot be modified according to a user's specific needs (www.clcbio.com).

A number of organisms (plants, insects, animals, microbes) have had their genomes successfully assembled *de novo* [57, 58, 61, 96, 97]. For example, the giant panda's (*Ailuropoda melanoleura*) draft genome of ~2.25 Gb was assembled *de novo* using the WGS sequencing strategy. Short reads were sequenced by the Illumina sequencing platform and assembled using SOAPdenovo [58]. *De novo* assembly is not only limited to whole genomes, but can also be applied in assembling an organism's organelles, providing useful biological information for cell evolution studies. For example, the carrot's (*Daucas carota*) mitochondrial draft genome has also been assembled *de novo* using a mix of reads sequenced from the 454 Roche and Illumina platforms and assembled by the Newbler assembler [57].

1.8 General Pipeline of the Assembly Process

De novo genome assembly occurs in a multistage pipeline [64]. The general approach primarily consists of first filtering the raw reads following quality assessment, followed

by subsequent assembling steps. Filtration of raw reads is obtained by following pre-processing procedures, which aim to reduce the amount of sequencing errors within the reads. Since WGS data is error-prone, there are various types of errors that can cause problems in the construction of the assembly for different assemblers [98]. The source of errors may originate from the sequence library preparation methods, or sequencing platform errors where nucleotides are incorrectly called due to poor quality scores thus resulting in overall lower quality assemblies [80, 99]. In order to increase the accuracy and completeness of the *de novo* assembly, it is imperative that the raw reads are 'cleaned' before downstream analysis. This reduces false positives and ensures that data of high quality is used in order to have a greater confidence in the results generated. Filtering low quality reads also improves the run-time performance and reduces the amount of memory used by the assembler [99].

The cleaning or filtering process is done by performing a number of steps normally known as quality controls. These steps include trimming sequence reads with poor quality scores, undesirable read lengths, adaptor sequences, ambiguous nucleotides 'N' and homopolymer reads [80]. Once the quality control step is complete, the reads can be used in the assembly process. Studies have shown that it is difficult to produce a good assembly with just a single attempt. It is therefore wise not to trust the results of a single assembly, but construct several assemblies, if possible, with different assemblers and varying parameters with the intention of producing a better assembly after each attempt [59, 89, 100].

It is also worth noting that approaches that work well in the genome assembly of one species may not necessarily work well for another [89] due to the presence of differences in the genomic content of different organisms. Once the pre-processing procedures are complete, assembly of the genome can commence.

1.9 Assembly Evaluation

The accuracy of a newly assembled genome is difficult to measure and determine [76]. There is no easy formula to evaluate an assembly as it cannot be done by examining a single metric, therefore the most appropriate method for assessing the quality of an assembled genome remains unclear [89]. The use of a reference genome facilitates better assembly as well as the capability to assess an assembly. However, in the absence of a high-quality reference genome, new assemblies, especially for non-model organisms, are challenging to evaluate. Therefore assembly-accuracy metrics that do not depend on the presence of aligning to a reference sequence are needed in such situations [80]. The most commonly used metrics for evaluation an optimal assembly are the size of the contigs, the N50 statistic, estimated genome size, percent of reads mapped, average contig read coverage, and the completeness of the assembly [76, 80, 81].

1.9.1 Contiguity of the Assembled Genome

Once a *de novo* assembly is produced, a wide range of basic statistics can be calculated from the size of the output by analyzing the number of contigs or scaffolds, proportion of reads assembled, read depth coverage and the estimated genome size assembled [64, 66, 80, 89].

1.9.1.1 Contig and Scaffold Sizes

Contigs and scaffolds are usually evaluated by analyzing their size in terms of their lengths [76, 89]. Typically, the set of contigs produced in an assembly are not of uniform length; therefore the distribution of the lengths measures the assembly's contiguity [80]. Basic statistical calculations provides the assessment of the contigs in terms of maximum, average, median and minimum lengths in the assembled draft genome; which offer an overall picture of how fragmented or contiguous an assembly is. An ideal assembly is expected to have one contig for every chromosome of the organism's genome being sequenced. However this is usually not the case as many contigs are usually produced due to the presence of contaminants, repetitive sequences, polymorphisms and missing data introducing fragmentation in the assembly [66]. Therefore it is often preferred to have fewer longer contigs in an assembly [89]. However, these should not be the only determining factors as when assessing an assembly [66]. For example, an assembly which consists of one large contig of approximately the size of the genome being studied is not useful if incorrectly assembled.

Alternatively, an assembly consisting of many short contigs could have a very high accuracy, but the contigs may be too short to be useful for gene-annotation purposes [79].

1.9.1.2 N50 Contig

The most widely used statistic to describe the contiguity of a genome assembly is the N50 statistic [81]. The N50 contig is a weighted median statistic, which is defined as the contig length such that half of the *de novo* assembled genome, lies in blocks of this size or larger [64, 89]. It is calculated by first ordering all contigs (or scaffolds) by length from largest to smallest, then summing them beginning with the largest, until the sum just exceeds 50% of the total length of all the contigs present [64, 80, 89]. The N50 statistic provides a sense of the scale and potential contiguity of an assembly and the longer the N50 is, the more accurate the assembly is presumed to be.

1.9.2 Estimated Genome Size

It is easier to gauge the correctness of an assembly if an estimated genome size is known therefore comparisons between the draft and actual genome, in terms of size, can be made. However if a reference genome is lacking, estimates can be made by comparisons with a closely related species [101]. Draft genomes can either be larger or smaller than the expected genome size. Larger assemblies may represent errors acquired during construction but they may also infer that an assembler has successfully resolved regions of the genome with high heterozygosity into multiple scaffolds and contigs [89].

Genome assemblies which are significantly shorter than the actual genome size may be due to missing repetitive sequences or fewer reads used in the assembly [77, 100]. The total estimated genome size is calculated by summing all the contigs in the assembly. The sum is then compared to the target genome size, and the optimal assembly should be that one that closely matches the actual genome size [89].

1.9.3 Percent of Reads Mapped

The percent of reads mapped onto the assembled contigs indicate how much of the sequenced reads were used in the construction of the assembly. Ideally, all of the reads should be used in the contig construction step. Therefore an optimal assembly is one that has almost, if not all, of the reads mapped back to the contigs [66, 89].

1.9.4 Expected Coverage

The expected coverage of the contigs can be calculated by using the number of reads sequenced multiplied by their average length, and the genome size of the organism as seen in Equation 1.2:

$$\frac{Total\ no.of\ reads\ used\ in\ the\ assembly\ (R) * \ Average\ length\ of\ the\ reads\ (L)}{Expected\ genome\ size\ of\ the\ organism\ (G)}$$

Equation 1.2 Calculation of the expected coverage of the contigs.

An optimal assembly is expected to have a uniform average coverage in every contig. Regions of high or low coverage are an indicator that the assembly should be further

evaluated. Low coverage can introduce gaps in assemblies [76], while extreme differences in read depth coverage may indicate a presence of contamination such as bacteria, phage, human DNA or the presence of mitochondrial DNA whose coverage levels are typically higher than other genomic data in WGS data [102].

1.9.5 Completeness of the Assembled Genome

There are several software programs available to estimate assembly completeness of an assembly, and one of them is called CEGMA [84, 89, 103], or core eukaryotic genes mapping. CEGMA helps to map and identify a set of highly conserved eukaryotic genes (CEGs) that are believed to be present in low copy numbers in higher eukaryotes [84]. It provides a complementary means of estimating the completeness and contiguity of an assembly and assists to determine the percentage of each of the core eukaryotic genes lying on a single scaffold [81]. Additionally, it is a useful tool for predicting the orthologs of a set of core genes in newly sequenced genomes which possess little or no annotation [103]. CEGMA works by first predicting the core genes using a database of a standard set of 248 CEG's derived from eukaryotic orthologous groups (KOGS) [84, 103]. The predicted genes are aligned to a HMMER profile (a software package used for sequence analysis to identify homologous protein or nucleotide sequences [104]) built for each of the core family gene. If the fraction of alignment exceeds 70% of the protein, it is classified as a full-length CEG, if it is less than 70% it is classified as a partial gene [89, 103]. The 248 CEG's used in CEGMA are the most highly conserved eukaryotic genes and that occur in single copy genes. They are divided into four groups, based on their

degree of protein sequence conservation; the least conserved protein sequences are in group 1, and the most conserved protein sequences are in group 4 [84, 103]. An assembly is fairly complete if it is predicted to contain almost or all of the 248 CEG's [84, 89, 103]. The proportion of CEGs mapped in draft genomes provide a useful metric for describing the gene space of a genome [84].

An assembly can be termed as satisfactory when judged by one approach, but poor by another [64, 89]. Additionally, an assembler performing well in an array of metrics in one species is no guarantee that the same metrics will work as well if at all on a different species [89]. If there are computational resources available, then a greater number of reads are better for a high quality assembly, as it will help in verifying locations of high heterozygosity.

The objective of this study was to reconstruct the genome of *P. regina* through *de novo* assembly of high-throughput short read sequences using NGS technologies. A draft mitochondrial genome was also assembled in the process. The draft genomes assembled from this study provide an important resource for analyzing genetic basis of variations between and among blow fly species, which will ultimately facilitate ongoing studies in various areas of research that utilize blow flies as study models.

CHAPTER 2. MATERIALS AND METHODS

2.1 Genomic DNA Libraries and Illumina Sequencing

Genomic DNA was extracted from five male and five female flies from a lab colony of *Phormia regina* using the DNeasy Blood and Tissue DNA Extraction kit following the manufacturer's instructions (QIAGEN Inc., Valencia, CA, USA), to prepare two paired-end (PE) libraries for sequencing. Preparation of the samples and whole genome sequencing of the libraries were performed by the Purdue University Genomics Core Facility (Purdue University West Lafayette, USA). The overall protocol followed in preparing the samples was based on the TruSeq DNA sample preparation guide by Illumina (Catalog #PE-940-2001. Part # 15005180 Rev. A, November 2010). The steps followed in the preparation stage are listed in Appendix C. Each of the resulting libraries (male and female) was barcoded with adapters in order to distinguish the two sets of data. Sequencing was then done on the PE libraries using the Illumina HiSeq2000 platform (Illumina Inc, San Diego) with a read length of 2 x 100 bp. Only 1 lane was used in sequencing, using half the lane for each of the samples. In the end, genomes were assembled for the female fly, the male fly, and the combined sexes.

2.2 Data Filtering

2.2.1 Trimming of Adapter Sequences and Low Quality Reads

The male and female raw reads were pre-processed to eliminate low quality reads. Adapter sequences were removed by the sequencing core facility. The quality of the reads was analyzed using the software CLC Genomics Workbench (v6.0.5) (www.clcbio.com). A quality score limit of 0.01 was selected resulting to the removal of reads with a Phred quality score of less than 20. The CLC trimming tool follows the modified-mott trimming algorithm, which first converts the quality scores (Q) of all the bases to an error probability using Equation 1.1. Low error probability values (P_{error}) represents high quality bases and high P_{error} values represents low quality bases. For example a base with Q = 20 has a P_{error} of 0.01 while one with Q = 10 has a P_{error} of 0.1. The trimming tool then calculates a new value for every base in a sequence using a user specified probability error limit (P_{limit}), by subtracting the P_{error} from the user specified (P_{limit}). This newly calculated value is expected to be negative for low quality bases. The final step performed is the running sum of the newly calculated values for every consecutive base in a sequence. The region of the sequence that is not trimmed is the one that lies between the first positive values of the running sum and the highest value of the running sum. Everything else before and after this region is then trimmed off. The maximum number of ambiguities ‘N’ was set to 2.

2.2.2 Duplicate Removal and Merging of Overlapping Pairs of Reads

Duplicate reads were removed from all the reads using a plug-in tool of the CLC-GWB. Overlapping pairs of the duplicate-free reads were merged using a software-based merging tool (CLCbio). Mismatch cost was set to 2, thus any mismatch present was penalized with 2 points. The minimum score required for an alignment to be accepted for merging was set to 8 and the gap cost which is the cost for the introduction of an insertion or deletion was set to 3. The maximum unaligned end mismatches were set to 0 to avoid the possibility of matching poor quality reads that occur at the end of the reads especially since the quality of NGS reads often drops.

2.3 *De novo* Genome Assembly

Initial *de* novo assemblies were carried out by 3 assemblers: CLC-GWB (v6.0.5), Velvet (v1.2.03) [86] and SOAPdenovo (v1.05) [85]. Both Velvet and SOAPdenovo assemblies were run on Mason, a large memory supercomputer cluster at Indiana University (IU) courtesy of the National Center for Genome Analysis Support (NCGAS). CLC software was run on a 32 GB RAM, 64-bit local workstation. A total of more than 21 complete *de novo* assembly iterations were performed for the male, female and combined sexes; with k-mer values ranging from 24 to 60 nucleotides. The parameters described for each assembler resulted from the ideal assemblies that were selected for the combined sexes based on the optimality criteria described in the introduction. The main criteria were the total number of contigs, the N50 size and the maximum contig lengths.

2.3.1 Velvet (v1.2.03)

Optimal k-mer sizes determined for the male was 75 bp, female 75 bp and combined sexes 85 bp. Velvet is centered on two programs (velveth and velvetg) and always uses them together [86]. Velveth reads the sequence files and builds a dictionary of the user specified k-mers defining local alignments between the reads; while Velvetg reads those alignments and builds a de Bruijn graph, removing errors in the process while attempting to resolve repeats based on user specified parameters [86]. Velvet requires that paired-end FASTA and FASTQ datasets be merged into a single file before use. The reads should be identified as short, long, paired or single end as Velvet handles reads depending on their length and pairing. Typically reads longer than 200 bp would be marked as long [86]. The *P. regina* reads were therefore marked as short and paired FASTQ reads. The average insert size of the PE reads was set to 350 and the expected coverage of the reads was automatically calculated by the program.

2.3.2 SOAPdenovo (v1.0.5)

The optimal k-mer size determined for all the three assemblies was 75 bp. SOAPdenovo requires a configuration (config) file in addition to the assembly command lines [85]. A config file contains a number of parameters which informs the assembler the location of the sequence files and other relevant read information such as the length or the insert size.

In the command lines, the command ‘all’ was used which informed the assembler to carry out all the assembly subsequent steps starting with graph formation followed by

contig construction, read mapping and finally the scaffolding step. The maximum read length in the sequence reads was found to be 101 bp and reads detected that were > 101 bp were cut to this length. The average insert size of the PE reads was set to 350. Orientation of the reads was set to 0 to notify the program the PE reads are forward-reverse oriented and are generated from fragmented DNA ends with insert sizes of less than 500bp. The assembly flag parameter was set to 3 to notify the program to use the PE reads in both the contig and scaffolding process. The minimum alignment length between a read and a contig required for a reliable read location was set to 32. The file format used, in this case FASTQ, was indicated by ‘q1’ and ‘q2’ which pointed the assembler to the two forward and reverse sequence files composed of paired-end reads. The option to resolve repeats using the reads were selected and only contigs ≥ 500bp were saved

2.3.3 CLC Genomic Workbench (v6.0.5)

Optimal k-mer sizes for both male and female assembly run was 60 bp, and 45 bp for the combined sexes. Reads were mapped back to the contigs following their construction, and the option to update the contigs was selected. This meant that the contig regions needed to be supported by at least one read mapping back to them for the contig to be included in the final output. Contig regions where no reads mapped were removed. A k-mer size of 45 was elected and all contigs shorter than 500 bp were removed. A paired distance range of 180 bp – 580 bp was used in mapping the PE reads. This range describes the distance from the beginning of the forward read to the beginning of the reverse read and can be automatically detected by the assembler or specified by the user.

The average insert size of the PE reads (Figure 3.1) is kept into consideration by the assembler when detecting the paired distance range. The cost of a mismatch between the reads mapped back and the reference contig was set to 2, while the cost of having an insertion and gaps present in the read was set to 3 for both options. A minimum length fraction of 0.5 was selected, where half of the read needed to match with the reference sequence to be included in the final mapping. The similarity fraction relates to the length fraction and it was set to 0.8. In this case, at least 50% of the read must have had at least 80% identity to be included in the mapping.

All subsequent *de novo* assembly runs executed by CLC-GWB employed similar parameters, with changes occurring only on the word size (k-mer), and similarity fraction.

2.4 Mitochondrial Genome Assembly

2.4.1 Mapping Parameters

Pooled male and female *P. regina* reads were mapped onto the mitochondrial genomes of seven different blow fly species downloaded from GenBank (NCBI: www.ncbi.nlm.nih.gov). These were: *Cochliomyia hominivorax* (NC_002660), *Protophormia terraenovae* (NC_019636.1), *Chrysomya albiceps* (NC_019631.1), *Chrysomya bezziana* (NC_019632.1), *Chrysomya rufifacies* (NC_019634.1), *Chrysomya megacephala* (NC_019633.1), *Lucilia sericata* (NC_009733) and *Lucilia cuprina* (NC_019573.1). All the genomes were used as reference sequences in the mapping. Mapping was performed by the CLC read mapping tool and local alignment was used as it allows the ends of the reads to be left unaligned especially in the presence of many

mismatches between the reads and the reference. Reads that align equally well to multiple places in the contigs (non-specific reads) were ignored. The remaining parameters utilized the following values: mismatch cost = 2, insertion cost = 3, deletion cost = 3, length fraction = 0.5 and similarity fraction = 0.9.

2.4.2 Consensus Extraction

Consensus sequences from the mappings were extracted using CLCbio's extract consensus tool. Default values were used in the parameters. A low coverage threshold of 0 was selected, where low coverage is defined as no coverage. Therefore if only one read covers a region in the alignment, only that read will determine the consensus sequence. The ambiguity symbol 'N' was inserted for every base position in the low coverage regions. Conflicts (mismatches) between the reads were handled by inserting IUPAC nucleic acid codes for ambiguous nucleotides. The default value of the noise threshold used was 0.1, which means that for a base to contribute to the ambiguity code it must be in at least 10% of the reads at a given position. A minimum nucleotide count of 1 was used, which specifies the minimum number of reads required before a nucleotide is included in the consensus.

2.4.3 *De novo* Assembly Parameters

The reads that mapped to each of the seven mitochondrial genomes were extracted and assembled via CLC-GWB (v7.0.3). The parameter values were similar to the *de novo* assembly described in Section 2.3.3, with the exception of the bubble size which was detected automatically by the assembler to be 50 and the similarity fraction set to 0.9.

2.5 Contaminant Removal

A total of 1405 phage genomes were downloaded from the phage annotation tools and method website (www.phantome.org), and 595 bacterial genomes were downloaded from GenBank (NCBI: www.ncbi.nlm.nih). *P. regina* reads (male and female) were mapped to each set of genomes separately, retaining the unmapped reads to be used in subsequent steps. Reads were first mapped to the phage genomes, then to the bacterial genomes. This was conducted in CLC-GWB (v7.0.3). The same protocol as described in Section 2.4.1 was used for both mappings but in this case the references were the bacterial and phage genomes.

2.6 *De novo* Assembly of the Refined Draft Genome

The mitochondrial and contaminant free reads were used to build the contig and scaffold sequences of the refined version of the combined sexes draft genome. Approximately 313,000 additional longer *P. regina* reads of average size 344 bp from the 454 sequencing platform were used as guidance only reads to help resolve ambiguities in

the graph and assist in the scaffolding step. These reads were kindly shared to us by one of our collaborators Dr. David Ray of Mississippi State University, USA.

Contig construction and read mapping were performed in two separate steps. In the contig construction step, the minimum contig length was set to 500 bp and a k-mer size of 42 was used. Read mapping was performed using the same protocol as in Section 2.4.1 using the contigs as the reference sequences to be mapped on. An insert range of 180 bp – 600 bp was used in the mapping the PE reads.

2.7 Draft Genome Completeness Assessment

CEGMA (v2.4.010312) [103] was used to assess the completeness of the draft genome. It uses a combination of a number of programs to detect core eukaryotic genes in the draft genome. The main programs used in this version were GeneWise (v2.4.1), NCBI-BLAST+ (v2.2.28), and geneid (v1.4.4). CEGMA was installed and run on IU's primary Linux cluster Quarry.

The output includes a list of files which contains the DNA sequence of each CEGMA prediction, protein sequences of the predicted CEG's, exon details of all the predicted genes, KOG IDs for the selected proteins, and the statistical report of the findings that indicate the percent of complete and partial predicted CEG's present.

2.8 Gene Prediction

Gene prediction was conducted by the program AUGUSTUS (v2.5.5) [105], which is a software tool used for *ab initio* gene prediction in eukaryotic genomic sequences. The

user selects a species-specific parameter set to be used in the prediction process. AUGUSTUS can be trained and its model parameters (e.g. splice window sizes, coding regions, non-coding regions, exonic and intergenic regions e.t.c) estimated using sequences of already known genes from organisms that have annotated genes or genomes available. It therefore contains a database of annotated genomes of organisms whose parameters are already trained with it. It is recommended that a user selects the database which is the closest relative to their study organism when performing gene prediction on their data. Since *P. regina* is a non-model organism, *D. melanogaster* was chosen, and despite being distantly related to *P. regina,* it has a well-studied and annotated genome. Results were exported in text format consisting of exon, intron, transcript and gene boundaries in the general feature format (GFF), which is a file that is usually used for describing genes and other features of DNA, RNA and protein sequences. It includes information on the exact location of the predicted genes in the contigs that they were predicted from. Amino acid sequences of the predicted genes are also included in the same text file. Gene prediction was conducted twice, one with the option of predicting any number of genes possibly partial genes and the other one full 'complete' genes.

The predicted amino acid sequences were extracted from the GFF files using a programming script, saved in fasta format and blasted using BlastP (v2.2.28+) in GenBank via Galaxy (an open source web-based platform for NGS data analysis) [106, 107] . The database used was the non-redundant (NR) protein blast database and an E-value cutoff of >1e-3 was selected.

2.9 Gene Ontology and Functional Annotation

Functional categorization of the blasted proteins based on the NR annotation was conducted using Gene Ontology (GO) via Blast2GO (v2.7.1) [92], a tool for functional annotation and analysis. GO is a standardized functional annotation schema for gene and protein sequences, and is used in nearly all public databases [92, 93]. Blast2GO identifies similar sequences in blast results, maps the homologous sequences to GO terms [92-94] and adds a functional annotation to them. It enables GO data mining on sequence data that has no GO annotation available [92] and can assign multiple terms to the same gene, as one gene can be classified in different categories.

The resulting blast files were sorted and categorized according to their functional groups (molecular, cellular and biological processes) with an e-value threshold of less than 1e-3.

CHAPTER 3. RESULTS AND DISCUSSIONS

3.1 Sample Selection and Sequencing

Whole genome sequencing was performed on extracted DNA of five male and five female *Phormia regina* flies. Pooled sequencing was performed on each sample using Illumina HiSeq2000 sequencing platform by the Purdue University genomics core facility. Sequencing was done with the aim of producing paired-end (PE) reads with an average read length of 100 bp and 25X coverage for both sexes. The inserts of the PE reads had an average size of approximately 320 bp and 300 bp for the females and males respectively shown by the peaks of the graphs in Figure 3.1.

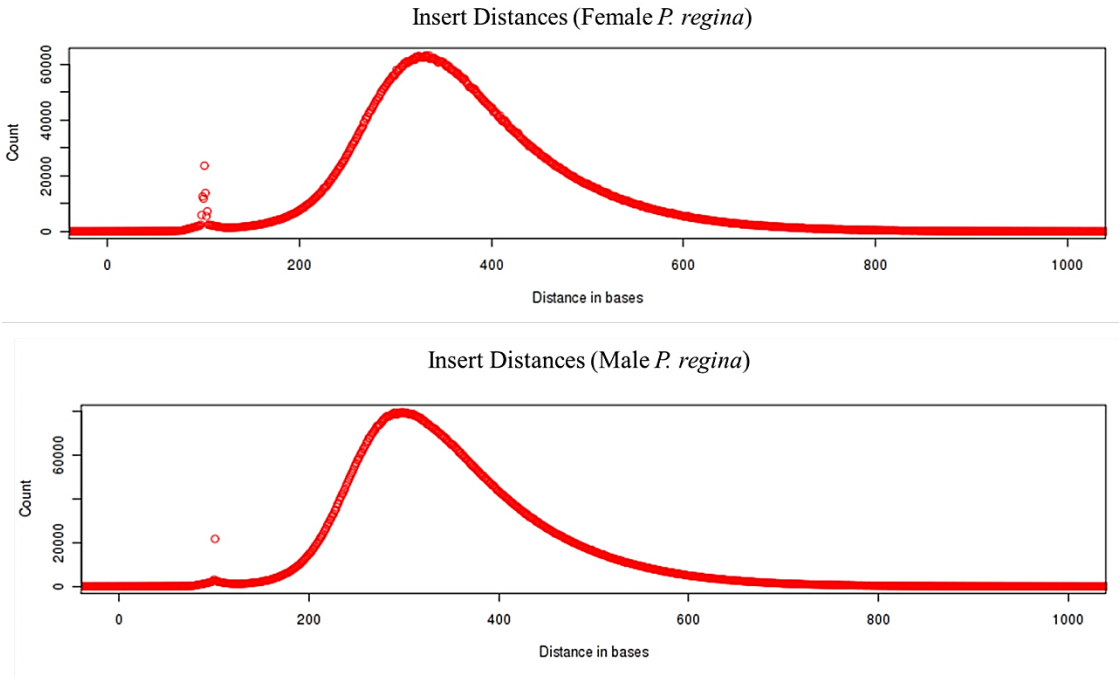


Figure 3.1 A representation of the range of insert sizes for the female and male *P. regina* sequenced reads.

A total of more than 530 million high quality paired-end raw reads from both samples were produced, the female with ~269 million raw reads, and the male with ~261 million raw reads (Table 3.1).

Table 3.1 Summary statistics of the male and female sequenced reads. Raw reads are the reads directly from the sequencer. Reads with a Phred quality score of <20 were trimmed.

SEX	**RAW READS**			**QUALITY AND ADAPTER CLIPPED READS**				
	Single Reads	Bases (bp)	Max Length (bp)	Single Reads	Bases (bp)	Min length (bp)	Mean Length (bp)	Max Length (bp)
Male	261,906,090	26,452,515,090	101	255,233,928	25,408,514,873	30	99	101
Female	269,473,502	27,216,823,702	101	262,812,488	26,172,503,447	30	99	101

3.2 Quality Control

Quality control was performed on the raw reads prior to assembly. Included was the trimming of low quality reads, merging of overlapping pairs of reads and the removal of duplicate reads. These steps need to be employed to ensure that high quality sequences are assembled [74, 108] to increase the accuracy of the assemblies. The quality controls in this study were performed on CLC-GWB (v6.0.1) (www.clcbio.com)**.** Approximately 4.5% of the female sequenced reads and 4.7% of the male reads, were discarded after the pre-processing steps, leaving a total of slightly over 260 million reads and 253 million reads for the female and male flies, respectively.

3.2.1 Low Quality Read Trimming

Reads with a mean Phred quality score of less than 20 were trimmed (Figure 3.2), which resulted in the remaining base calls to have a probability of 99% or greater of being correct (see Table 1.1). The average Phred score of the remaining reads was ~38.

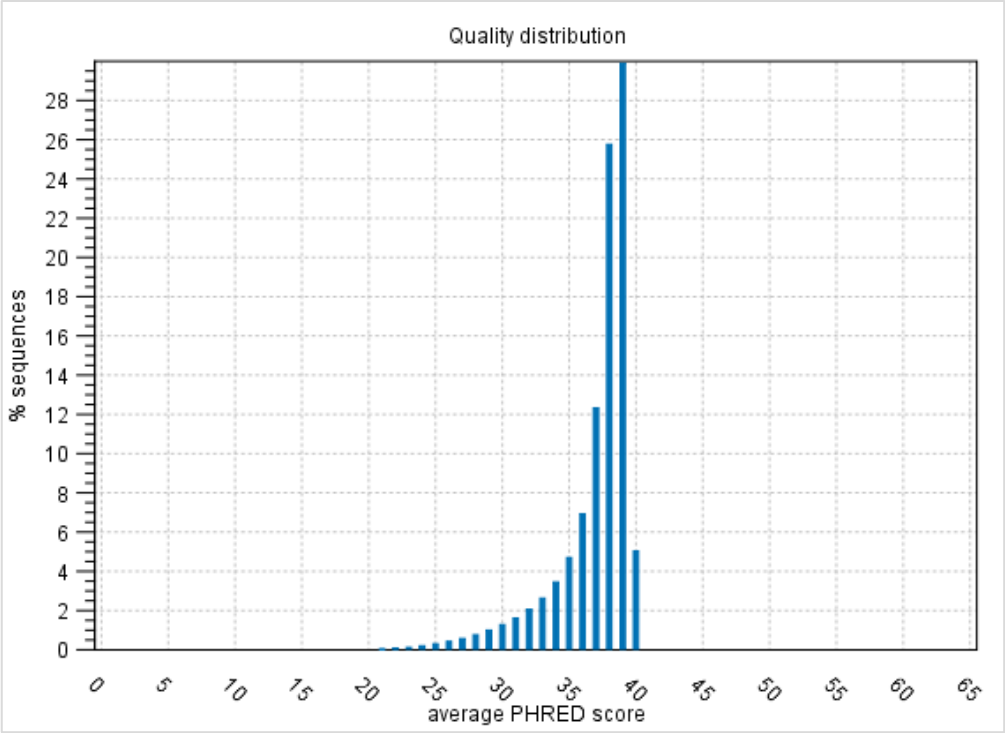


Figure 3.2 Phred Score Distribution. The graph shows the distribution of the average Phred scores following the removal of low quality sequences (Phred score < 20).

3.2.2 Duplicate Read Removal

Duplicate reads are identical reads that are present in the data more than once. Removal of duplicate reads aids in the reduction of raw data used in the assembly by including a single copy of the duplicated sequences. The software-based (CLCbio) duplicate removal tool attempts to remove only the identical reads that come from PCR amplification errors and not reads that are a result of high coverage. The software targets neighboring reads that start at the same position and are in the same orientation. Reads are marked as duplicates using the following criteria: (i) if they share a common sequence of at least 10 bases in the beginning of the read and at any of four other randomly selected regions distributed across the read and; (ii) if the rest of the read has a > 80%

alignment score. In PE reads, if both parts of the pair share the same sequence, they are marked as duplicates, and only one copy of the pair is left in the read set for assembly.

A small percentage of duplicates were detected in the sequenced reads, with the female having 0.7% of duplicates and the males 0.59% (Table 3.2).

Table 3.2 Removal of duplicate reads. Less than 1% of the reads were detected as duplicates, leaving more than 99% of the reads for subsequent assembly steps.

SEX	INPUT READS	# DUPLICATES READS	REMAINING READS	DUPLICATE READ %	REMAINING %
Female	262,812,488	1,848,922	260,963,566	0.70	99.30
Male	255,233,928	1,506,316	253,727,612	0.59	99.41

3.2.3 Merging of Overlapping Reads

Overlapping pairs of reads can be merged into a single sequence read which creates longer and higher quality reads; this is beneficial to the assembly process yielding more contiguous assemblies [80]. The software program aligned a set of two reads by using user-specified alignment scores. Since errors are expected in sequencing reads, the alignment is not expected to be perfect, and therefore, the user decides the acceptable number of errors that should be allowed. It is also critical to ensure that the length of the overlap is large enough to avoid the likelihood of matching by chance due to only a few bases overlapping.

Approximately 5% of reads from either sex were merged (Table 3.3). These resulted in more than 26 million reads longer than the original 100 bp reads, with a mean length of ~180 bp.

Table 3.3 Merged reads. Percentage of the merged pairs for male and female reads.

	FEMALE		MALE	
	No. of Reads	Percentage	No. of Reads	Percentage
Merged	13,313,916	5.10%	13,505,360	5.32%
Not Merged	247,649,650	94.90%	240,222,252	94.68%
Total	260,963,566	100%	253,727,612	100%

3.3 *De novo* Genome Assembly

3.3.1 Assembly Software Comparisons

De novo genome assembly was carried out using three different assemblers: CLC-GWB (v7.0.3) (QIAGEN Inc., Valencia, CA, USA), Velvet (v1.2.03) [86] and SOAPdenovo (v1.05) [85]. This was done to assist in the selection of the optimal assembler that would be used for the reconstruction and refinement of the draft genome later [89].

De novo assemblies were run on the male and female reads separately, and then on the pooled reads of both sexes, creating three draft genomes. All three assemblers use the de Bruijn approach in constructing the assembly, therefore, various k-mer sizes were used to conduct several runs of assemblies, ranging from 24 bp to 90 bp. Assemblies from each software program were ranked based on a select set of metrics (Table 3.4). Ideally, the optimal assembly has the fewest number contigs (generally resulting in longer contig sizes), larger N50's, and the majority of the reads are used in the draft genome assembly. Using these data, the genome size can be estimated from the assembled draft genome and the closer to the known genome size. For *P. regina*, the known genome sizes is 529 Mbp for females and 517 Mbp for males [109].

Table 3.4 Comparative draft genome assemblies. Results of the draft genome assemblies of female, male and combined sexes using CLC-GWB, Velvet and SOAPdenovo.

FEMALE	**CLC-GWB**	**VELVET**	**SOAPdenovo**
Kmer Size (bp)	60	75	75
No. of Contigs	286,936	317,916	1,953,762
N50 (bp)	2,918	1,095	451
Max Contig (bp)	70,568	49,370	49,698
% Reads Used	94.8%	47.0%	68.8%
Estimated Genome Size	~530 Mbp	~550 Mbp	~556 Mbp
MALE	**CLC-GWB**	**VELVET**	**SOAPdenovo**
Kmer Size (bp)	60	75	75
No. of Contigs	339,679	310,045	2,010,365
N50 (bp)	2,076	956	351
Max Contig (bp)	71,264	25,508	28,012
% Reads Used	92.4%	37.9%	58.6%
Estimated Genome Size	~512 Mbp	~488 Mbp	~497 Mbp
COMBINED	**CLC-GWB**	**VELVET**	**SOAPdenovo**
Kmer Size (bp)	45	85	75
No. of Contigs	385,512	266,759	4,380,801
N50 (bp)	1,841	1,658	283
Max Contig (bp)	48,059	66,334	38,640
% Reads Used	91.3%	41.1%	73.2%
Estimated Genome Size	~559 Mbp	~492 Mbp	~821 Mbp

Overall, CLC-GWB produced better assemblies than Velvet or SOAPdenovo. CLC-GWB produced assemblies with longer contigs and used the majority of the reads in the assembly. It should be noted in some cases Velvet produced a smaller number of contigs (see Table 3.4), but this is likely due to the fact that a smaller percentage of reads were used in the construction of the assembly, thus it would be expected that there would be an overall decrease in the number of contigs. Therefore, CLC-GWB was selected as the assembler of choice to produce optimized assemblies for further downstream analysis in this study.

3.4 Contaminant Removal

It is common to expect a level of contamination or a presence of extraneous DNA from sequenced genomic eukaryotic DNA, especially from bacteria [110]. Bacterial contamination is especially prevalent in de novo sequencing as all organisms on earth harbor these prokaryotes, whether on the surface, or more like, internally. In addition, though rare, contamination can be introduced during the library preparation and sequencing process. The presence of DNA samples from other organisms can affect the target assembly by adding to genome mis-assembly and the reduction of the overall quality of the assembly [110]. Since blow flies are attracted to decomposing material, and their larvae act as scavengers for carrion, it is expected to observe a percentage of bacterial DNA sequence reads.

Prior to assembly, reads were pooled and mapped to a set of complete genomes of bacteriophage and bacteria. A total of 1405 phage genomes were downloaded to create a local search database from the phage annotation tools and methods website (www.phantome.org); while a total of 595 bacterial genomes were downloaded from GenBank (NCBI: www.ncbi.nlm.nih.gov). Approximately 1.8 million reads (0.38% quality trimmed reads) mapped to the phage genomes, and 2.3 million reads (0.49% of the reads) mapped to the bacterial genomes (Table 3.5). All of these reads were removed from the pooled reads.

Table 3.5 Mapping Summary Statistics. Summary statistics for bacteria and bacteriophage genome mapping using quality-filtered reads.

MAPPING SUMMARY STATISTICS				
	Bacteriophage		**Bacteria**	
	Count	%Reads	Count	%Reads
Reference Genomes	1,405	-	595	-
Mapped Reads	1,819,343	0.38%	2,311,444	0.49%
Unmapped Reads	477,037,631	99.62%	472,413,418	99.51%

3.5 Mitochondrion Assembly

Mitochondrial reads were extracted by mapping all the reads (male and female, trimmed by quality and bacterial and bacteriophage reads removed) to published complete mitochondrion genomes of 7 different blow fly species downloaded from GenBank (NCBI: www.ncbi.nlm.nih.gov). These species were: *Cochliomyia hominivorax* (NC_002660), *Protophormia terraenovae* (NC_019636.1), *Chrysomya albiceps* (NC_019631.1), *Chrysomya bezziana* (NC_019632.1), *Chrysomya rufifacies* (NC_019634.1), *Chrysomya megacephala* (NC_019633.1), *Lucilia sericata* (NC_009733) and *Lucilia cuprina* (NC_019573.1). Of the 474 million reads, ~8 million reads mapped to the seven mitochondrial genomes (Table 3.6) with the greatest homology to *Co. hominivorax*. It is possible that not all of *P.regina*'s mitochondrial reads were extracted from the total pooled male and female reads especially if the reads are specific to *P.regina* mitochondrial genome.

Table 3.6 Mitochondrial read mapping statistics. Summary of the distribution of the read mappings to the mitochondrial genomes of 7 blow fly species

REFERENCES	#MAPPED READS
Co. hominivorax	7,437,554
Pr. Terraenovae	89,318
Ch. albiceps	165,212
L. sericata	492,813
L. cuprina	43,397
Ch. bezziana	37,038
Ch. rufifacies	27,964
Ch. megacephala	85,119
TOTAL READS MAPPED	8,378,415

The reads that mapped were then extracted and analyzed. The *P. regina* mitochondria was A-T rich (as expected, Table 3.8) at 76% of the nucleotide distributions (compared to *Co. hominivorax*, at 77%) [111].

The reads were then re-mapped to each of the mitochondrial genomes separately, in what is called assisted assembly, to assess what percentage that would map back, and extract a consensus sequence [112]. *Co. hominivorax* mitochondrion genome had the highest percentage (89%) reads mapped back (Table 3.7). The consensus sequence of the draft *P. regina's* mitochondrion genome extracted from each mapping had a range of 15,194 bp to 16,045 bp, which is within expected size of average mitochondrial genome sizes of the Animalia kingdom of ~ 16,000 bp.

Table 3.7 Extracted mitochondrial read mappings. Summary of extracted *P. regina* mitochondrial reads mapped back to the 7 mitochondrial genomes

	MAPPED READS	% READS MAPPED	% READS UNMAPPED	REFERENCE GENOME LENGTH (bp)	CONSENSUS LENGTH (bp)
Co.hominivorax	7,437,452	88.77%	11.23%	16,022	16,045
Pr.terraenovae	6,993,763	83.47%	16.53%	15,170	15,194
Ch.albiceps	7,327,873	87.46%	12.54%	15,491	15,531
L.sericata	7,390,713	88.21%	11.79%	15,945	15,981
L.cuprina	7,391,554	88.22%	11.78%	15,952	15,979
Ch.bezziana	6,766,391	80.76%	19.24%	15,236	15,234
Ch.rufifacies	6,743,364	80.48%	19.52%	15,412	15,450
Ch.megacephala	7,350,401	87.73%	12.27%	15,273	15,289

De novo assembly of the extracted mtDNA was performed on CLC-GWB (v.7.0.3) for comparison purposes with the alignment results from the reference mapping. A k-mer size of 22 bp was used in the assembly. This assembly was A-T rich with a percentage summing up to 75% (Table 3.8).

Table 3.8 Mitochondrial nucleotide distribution. The *de novo* assembled mitochondrial genome of *P.regina*

NUCLEOTIDE	COUNT (bp)	FREQUENCY
Adenine (A)	6,074	37.00%
Cytosine (C)	1,573	9.60%
Guanine (G)	2,224	13.50%
Thymine (T)	6,295	38.30%
Any Nucleotide (N)	257	1.60%

The assembly was composed of 2 scaffolded contigs, which were created from 93% of the reads (Table 3.9). The assembly had an N50 contig size of 15,795 bp long, and an estimated genome size of 16,423 bp (Table 3.9), which is within the range of the blow fly mitochondrial genomes [111]. The complete nucleotide sequence of the contigs can be found in Appendix A.

Table 3.9 Mitochondrial d*e novo* assembly summary statistics. Summary statistics of the *de novo assembly of P. regina's* mitochondrial genome

***DE NOVO* ASSEMBLY STATISTICS**	
Total Reads	8,378,415
Mapped Reads	7,813,492 (93.26%)
Unmapped Reads	564,923 (6.74%)
Contig Count	2
Contig #1	628 bp
Contig #2	15,795 bp
Estimated Genome Size	16,423 bp

The extracted sequence of *P. regina* mitochondrial genome can be used as a source of sequence information for general Diptera molecular and evolutionary approaches and it can aid in primer selection of specific mitochondrial DNA regions and phylogenetic studies to assist in understanding dipteran evolution [111, 113].

3.6 Analysis of the Optimal Combined Sexes Draft Genome

3.6.1 *De novo* Genome Assembly

The remaining reads following quality trimming, and removal of mitochondrial, bacterial and phage containing reads totaled to approximately 470 million reads. *De novo* assembly was done using CLC-GWB (v7.0.3). Out of the range of k-mer sizes initially tested (24 -60 bp), 45 bp was selected as the optimal k-mer size for this data. The genome was predominantly A-T rich, with a combined percentage at 71%, as is expected in insect genomes (Table 3.10) [111, 114-117].

Table 3.10 Nucleotide distribution of the assembled *P. regina* draft genome.

NUCLEOTIDE	COUNT (bp)	FREQUENCY
Adenine (A)	211,314,974	35.60%
Cytosine (C)	78,366,055	13.20%
Guanine (G)	78,359,402	13.20%
Thymine (T)	211,095,592	35.60%
Any Nucleotide (N)	13,613,946	2.30%

Removing mitochondrial, bacterial and bacteriophage sequence reads improved the overall assembly (compared to draft genome assembled in Table 3.4) where the N50 contig size increased from 1,841 bp to 2,488 bp, and the maximum contig size increased from 48,059 bp to 121,695 bp (Table 3.11).

Additionally, the number of contigs reduced by approximately 6% with 41,374 fewer contigs, suggesting that this assembly is more complete and contiguous [64].

Table 3.11 Comparison between the new draft assembly (assembled with the filtered reads – i.e. bacterial, phage and mitochondrial reads removed) and previous draft assembly (assembled using original unfiltered set of reads). A notable improvement in the overall contiguity is observed in the new assembly.

***DE NOVO* METRIC COMPARISON**		
Assembly Statistics	New Assembly	Previous Assembly
N75	1,192 bp	994 bp
N50	2,488 bp	1,841 bp
N25	5,304 bp	3,606 bp
Minimum	500 bp	426 bp
Maximum	121,695 bp	48,059 bp
Average	1,722 bp	1,450 bp
Estimated Genome Size	592,749,969 bp	559,135,440 bp
Contig Count	344,138 contigs	385,512 contigs

3.6.2 Assembly Evaluation and Refinement

Additional 313,000 reads (average length 344 bp) sequenced by the 454 sequencing platform, were used to aid in contig scaffolding. The reads were courteously provided to

us by one of our collaborators, Dr. David Ray, from Mississippi State Univ. The reads refined the assembly by reducing the number of ambiguous nucleotides (N), from 2.3% to 0.6%. This suggests the longer reads assisted in filling in the gaps that were present in the first set of contigs improving the overall contiguity of the assembly (Table 3.12).

Table 3.12 Nucleotide distribution of the refined assembly

NUCLEOTIDE	COUNT (bp)	FREQUENCY
Adenine (A)	189,569,676	36.2%
Cytosine (C)	70,947,935	13.5%
Guanine (G)	70,963,795	13.5%
Thymine (T)	189,366,729	36.1%
Any Nucleotide (N)	3,293,153	0.6%

The number of contigs reduced 45% from an initial count of 344,138 to 131,111 (Table 3.13), and 94% are greater than 1000 bp long. Contig sizes ranged from 500 bp – 121,969 bp (Figure 3.3), with the N50 length increasing from 2,488bp to 5,267 bp. The estimated genome size improved from 592 Mbp (new assembly, Table 3.11) to 524 Mbp, which is near the calculated average estimated genome size of ~523 Mbp (Table 3.13) [109].

Table 3.13 *De novo* assembly statistics. A summary of the statistics of the refined assembly constructed using guidance from a set of longer reads.

***DE NOVO* ASSEMBLY STATISTICS**	
N75	3,097 bp
N50	5,267 bp
N25	9,055 bp
Minimum	500 bp
Maximum	121,696 bp
Average	3,998 bp
# Contigs	131,111
Total Size	524,141,288 bp

Evaluating this new version of the draft genome in terms of contiguity by looking at the length of the contigs and the N50, we can assume that the longer reads from the 454 sequencer, assisted in improving the overall state of the draft genome.

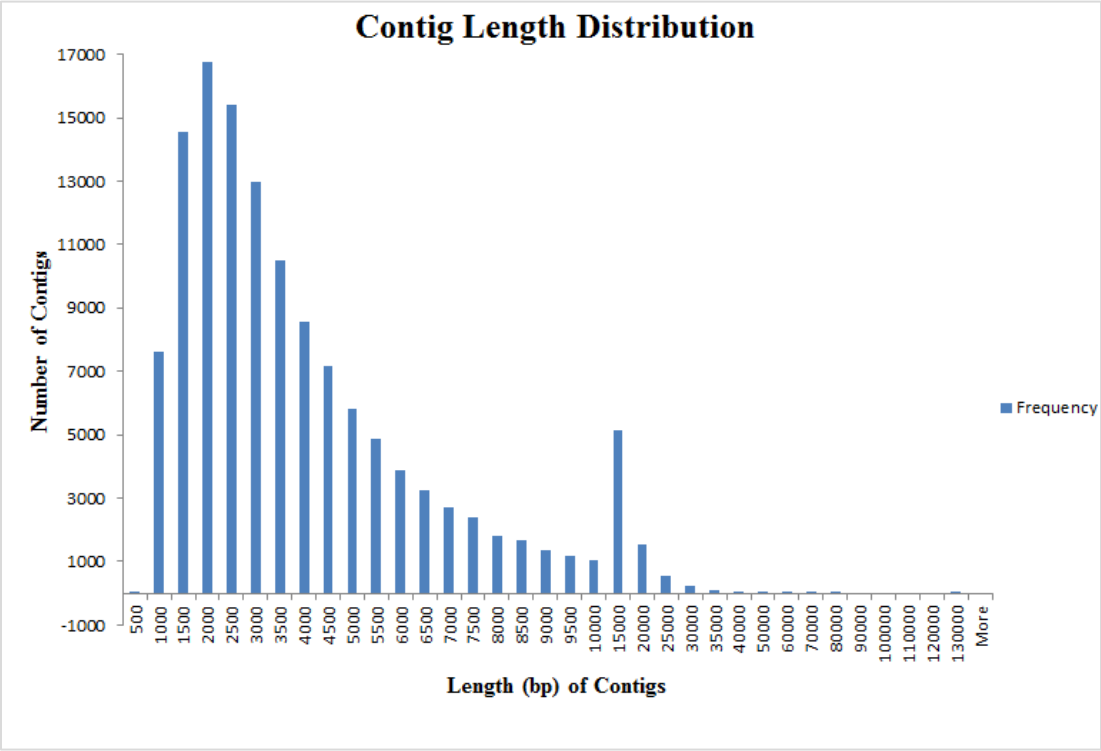


Figure 3.3 Contig length distribution of the refined *P. regina* draft genome. More than 94% of the contigs having a length greater than 1000 bp.

In order to further assess the quality of the scaffolds, the reads were mapped back to the contigs (Table 3.14).

Table 3.14. Read mapping statistics. Summary statistics of mapping the PE reads back to the assembled contigs.

MAPPING STATISTICS	
References (Contigs)	131,111
Total Reads	470,236,292 reads
Mapped Reads	391,097,987 reads (83.17%)
Unmapped Reads	79,138,985 reads (16.83%)

Approximately 83% (Table 3.14) of the PE reads mapped back with average read depth coverage per base of 76X (Figure 3.4). An ideal assembly is expected to have a uniform read coverage distribution as each contig should have roughly the same number of reads aligned to it. However in Figure 3.4, the distribution of read coverage of the draft assembly portrays a roughly log normal distribution due to the uneven distribution of the read coverage within the contigs, with a majority of the contigs having an average coverage of the expected ~76X.

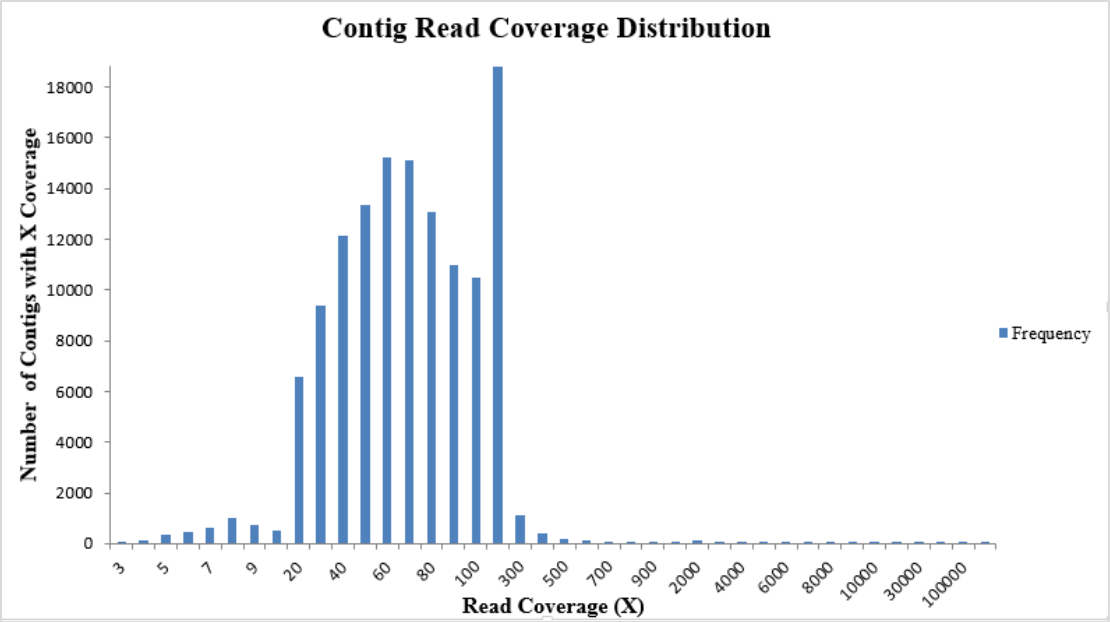


Figure 3.4 Contig coverage distribution. Average read depth coverage (X) of the assembled contigs was determined to be 76X.

3.6.3 Quality Evaluation of the Assembled Genome

The draft genome was evaluated using CEGMA (v2.4.010312), which is a useful tool for predicting orthologs of a set of core genes in newly sequenced genomes that have little or no annotation [103]. An ideal draft assembly is expected to have all of the 248 core eukaryotic genes (CEGs), therefore the greater the number of genes detected, the better [84, 89, 103]. The draft genome was found to contain ‘complete’ copies (>70% alignment length with core proteins) of 229 (92.34%) of the 248 CEG’s, with 244 (98.39%) of them partially represented (<70% alignment) (Table 3.15). The ‘complete’ predicted genes share a 77.29% orthology and the partial 87.3% with the CEG’s suggesting that there is a level of conservation between the predicted genes and other eukaryotic genes. A complete list of the eukaryotic orthologous groups (KOGs) detected in the draft genome can be found in Appendix B.

Table 3.15 Statistics of the draft genome's completeness based on 248 CEGs. Results are based on the set of genes selected by Genis Parra [103]. Prots = number of 248 ultra-conserved CEGs present in genome, %Completeness = percentage of 248 ultra-conserved CEGs present, %Ortho = percentage of detected CEGS that have more than 1 ortholog.

	#PROTS	%COMPLETENESS	%ORTHO
Complete	229	92.34	77.29
Partial	244	98.39	87.3

3.6.4 Gene Prediction

AUGUSTUS (v2.5.5) predicted 11,037 complete genes and 12,182 partial genes, where a partial gene is described as an incomplete gene which does not have all of its exons contained in the input sequence [105]. This is fewer than the number of genes predicted in other insect genomes such as *D. melanogaster* and *Anopheles gambiae*, which have 13,600 and 14,000 genes, respectively [59, 118, 119]. This could be due to the large evolutionary distance between *D. melanogaster* and *P. regina*. The longest gene in the list was 10,602 amino acids while the shortest one had 66 amino acids.

3.6.5 Gene Ontology and Annotation

The list of predicted proteins was blasted using BlastP (v2.2.28+) [106, 107] against the NR (non-redundant) protein database from GenBank (www.ncbi.nlm.nih.gov), using a minimum e-value of 1e-3. Of the 11,037 predicted genes, 87% (9,611) were successfully blasted returning a hit from the NR protein database.

The e-value distribution for the top hits show that approximately 93% of the predicted genes had an e-value of less than 1e-10 with approximately 74% having an e-value of 1e-180 (Figure 3.5). This suggests that the original genome assemblies produced viable protein sequences.

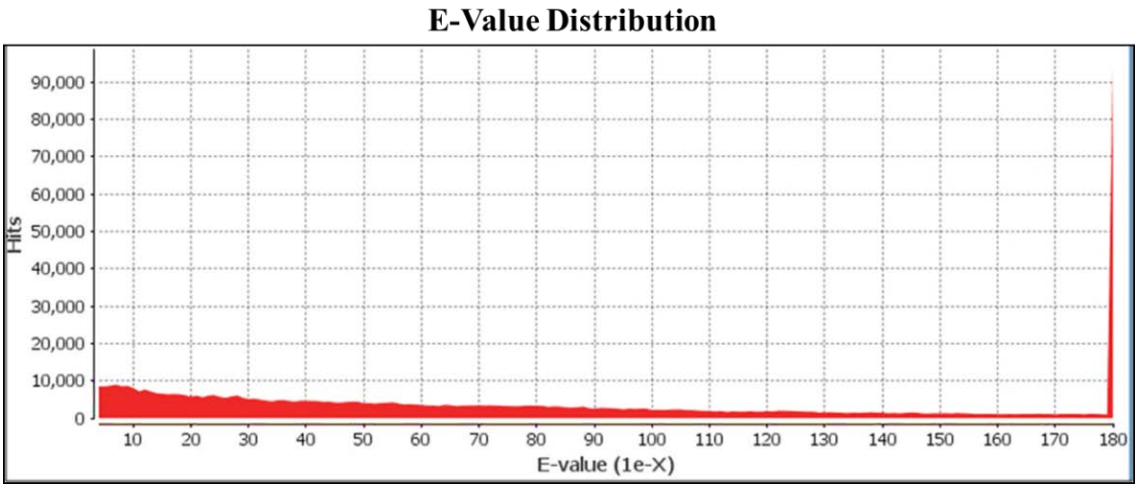


Figure 3.5 E-value distribution of the BlastP results against the NR GenBank protein database. The e-value was set at a minimum of 1e-3.

In blast results, generally, the best match (top hit) is the best estimate of the species identity. The top hit species distribution shows that more than 4,500 of the top hits had sequences similar to proteins from *Musca domestica*, also known as the common house fly, which is more closely related to *P. regina* than *D. melanogaster* (Figure 3.6) and has recently had its genome sequenced.

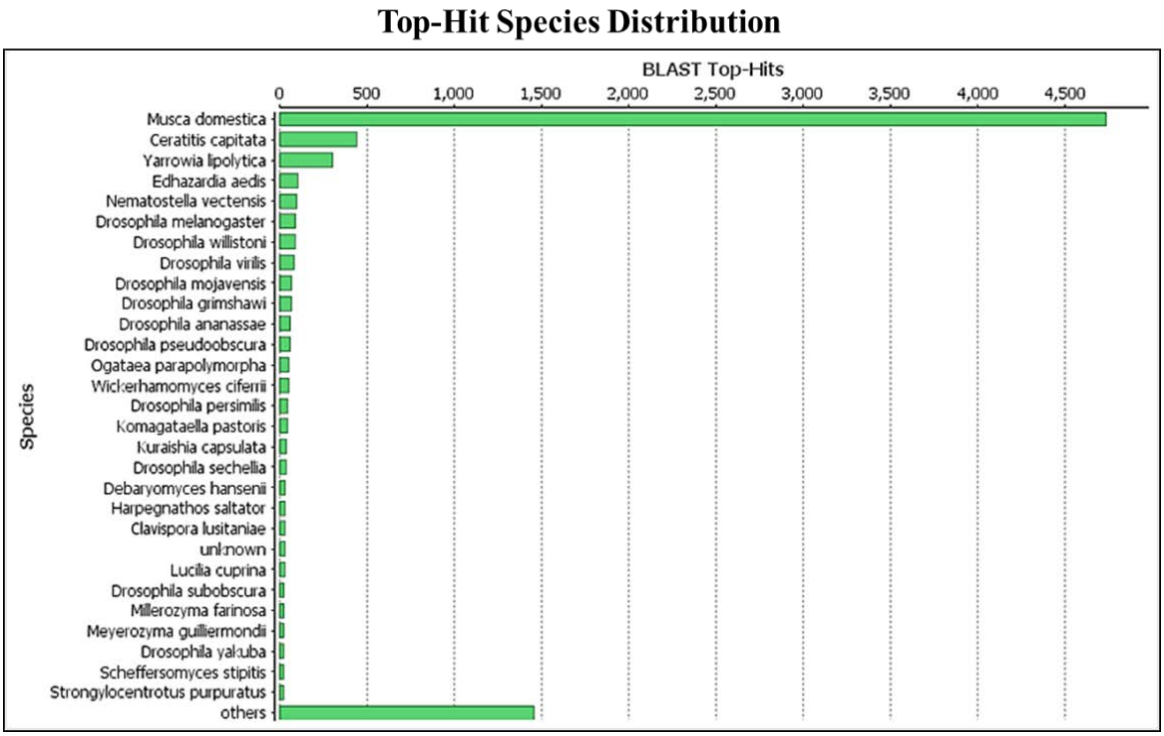


Figure 3.6 Top-hit species distribution of the BlastP results. The list shows the different species distribution of the top blast hits. The top hit was dominated by *Musca domestica*, the common house fly.

There are three categories of GO, namely biological processes, molecular function and cellular components, which are all attributes of genes and gene products [120]. Biological processes refer to a biological objective to which the gene or gene product contributes accomplished by ordered assemblies of molecular function (e.g. mitosis). Molecular function is defined as the biochemical activity or tasks performed by individual gene products at the molecular level (e.g. DNA helicase). Cellular components refer to the place in the cell where a gene product is active such as subcellular structures and locations (e.g. nucleus) [120-122].

Filtering can be applied to the annotated blast results in Blast2GO (a tool for functional annotation and analysis) by setting a user-specified annotation weight score (sequence abundance) in order to obtain a representative summary of the annotation [121] which can be visualized in pie charts or bar graphs. For example, if the sequence filter value is set to 5, it will report the annotations with >5 sequence alignments [121].

The distribution of the mapped and annotated sequences is shown in Figure 3.7. Of the 9,611 blasted protein sequences, 7,472 (~78%) were annotated with GO terms. Approximately 1,300 of the genes (without blast hits) did not match the BlastP NR protein database.

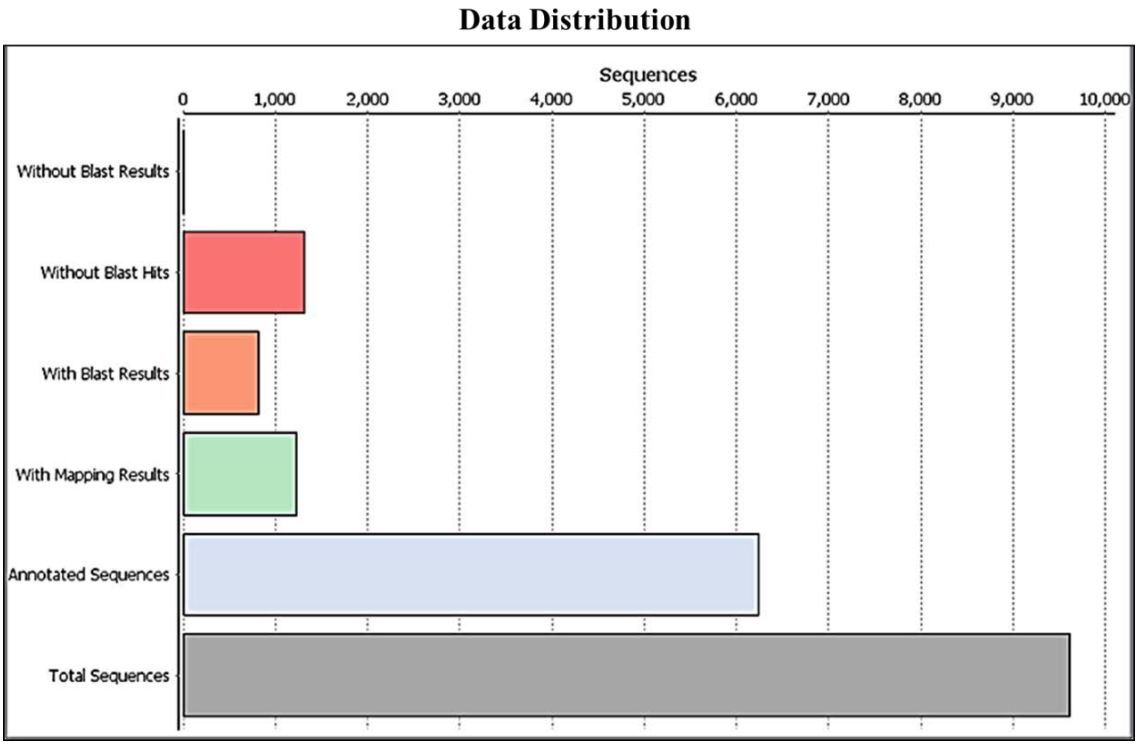


Figure 3.7 Data distribution of the mapped and annotated sequences

The assigned GO terms to the BlastP results were summarized into the three gene attributes (molecular, cellular and biological). The score filters were set to 5 for each of the three attributes.

A total of 2599 GO terms were categorized under the molecular function category. ATP and zinc ion binding molecular functions dominated with a total of 622 and 291 GO terms respectively (Figure 3.8).

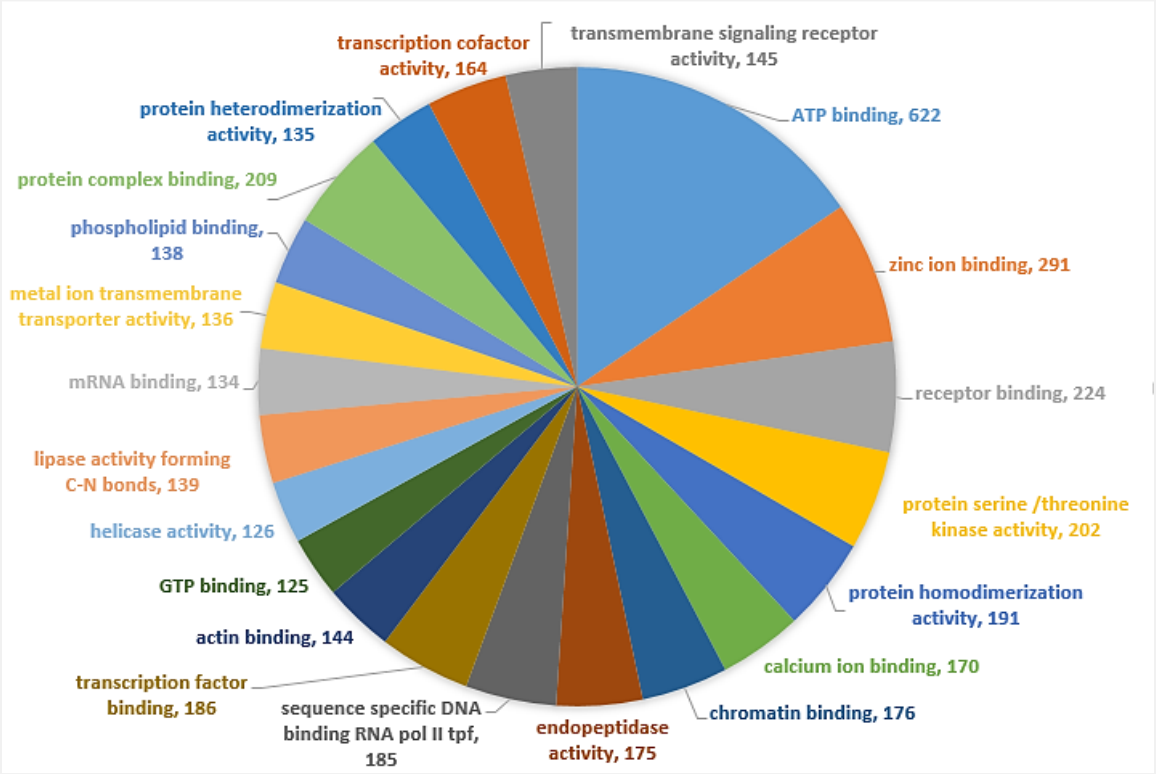


Figure 3.8 Molecular function sequence distribution of the Gene Ontology (GO). The molecular function of the identified proteins in the BlastP results, selected with a sequence cutoff score of 5. The numbers in parenthesis represents the number of protein sequences included in each term.

A total of 1204 GO terms were categorized under the cellular components category (Figure 3.9). The identified proteins were predicted to cover some of the main organelles in the cell, of which the nucleolus had the most GO terms assigned (390 GO terms).

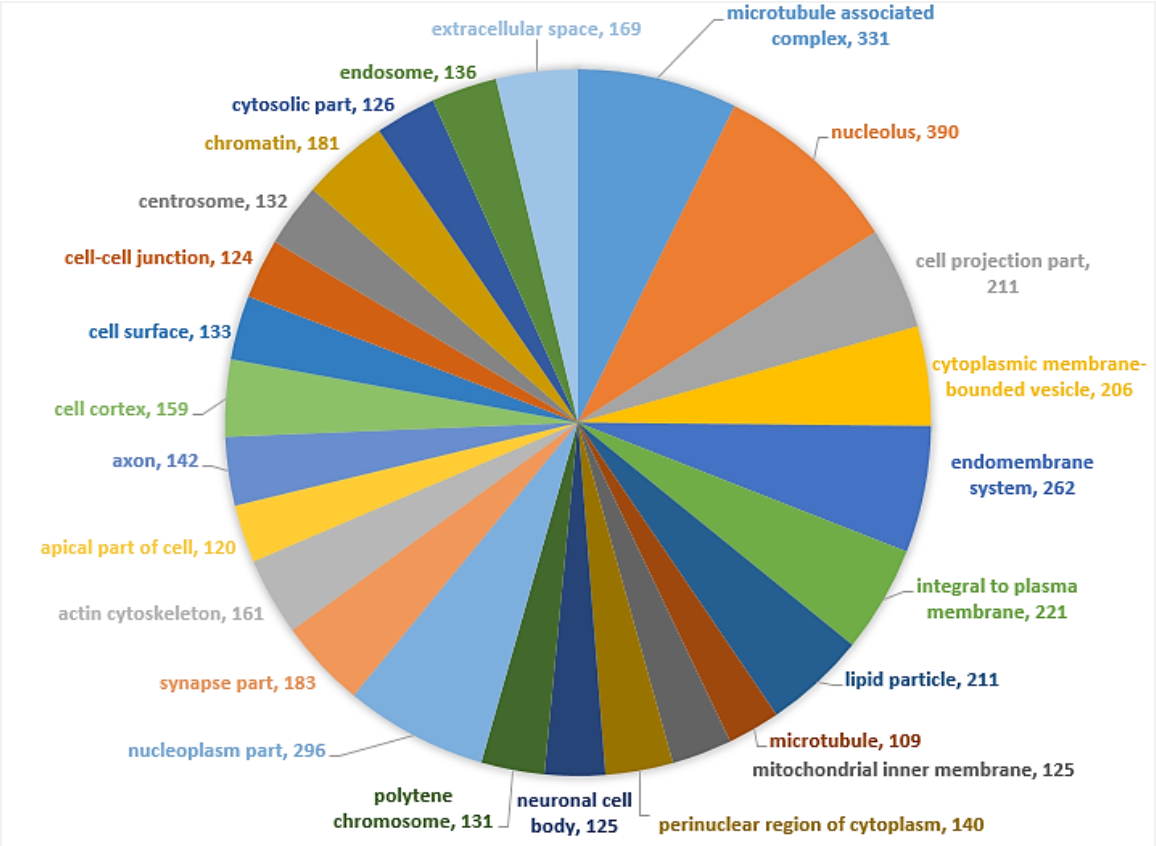


Figure 3.9 Cellular component sequence distribution of the gene ontology (GO). Annotation of the identified proteins from the BlastP results, selected with a cutoff score of 5. The numbers in parenthesis represents the protein sequences included in each term.

Out of all the three gene attributes, biological processes had the most number of GO terms assigned to it, a total of 9339. Of the predicted genes, most were mainly assigned to organ development (1,564) and anatomical structure morphogenesis (1,698) (Figure 3.10).

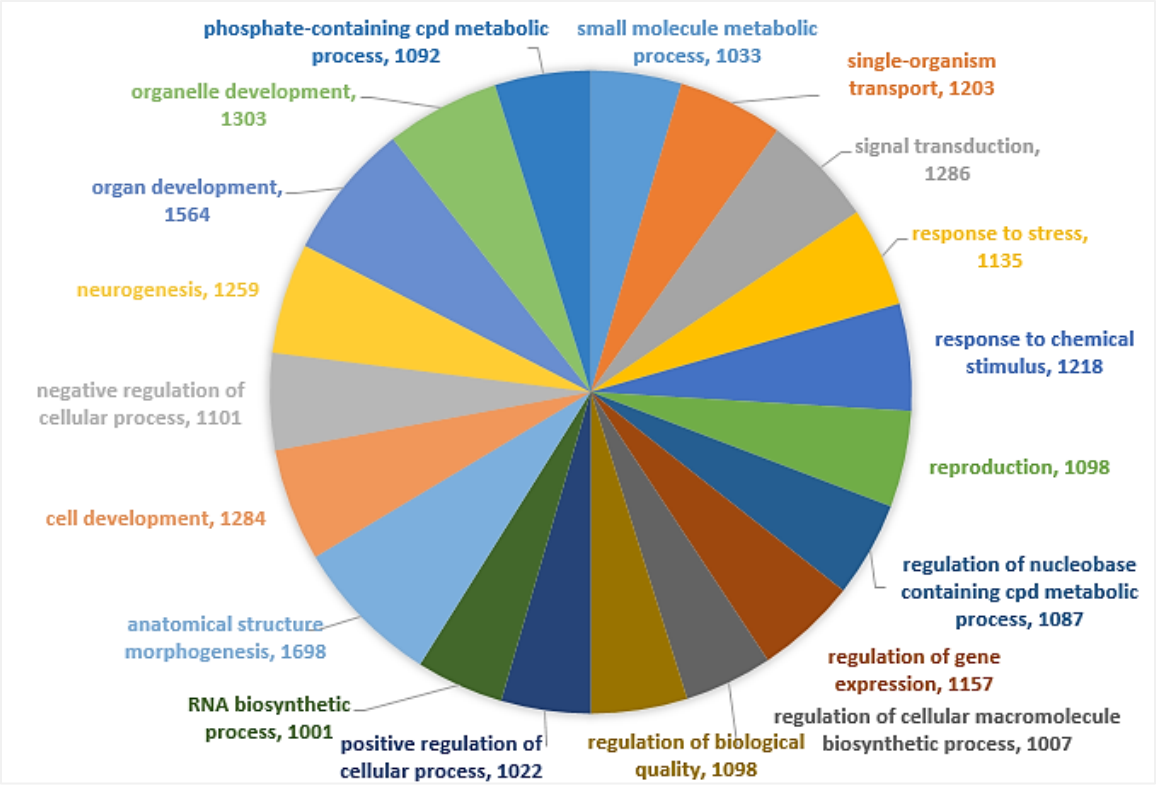


Figure 3.10 Biological process sequence distribution of the gene ontology (GO). Annotation of the identified proteins from the BlastP results, selected with a cutoff score of 5. The numbers in parenthesis represents the protein sequences included in each term.

A comparison of all three gene attributes (Figure 3.11**)** reveal that the GO annotated genes mainly represented biological processes, as seen by the number of sequences represented in each of the terms in Figure 3.10. These findings reveal that the main GO classifications obtained from the results are responsible for fundamental biological regulation and metabolism.

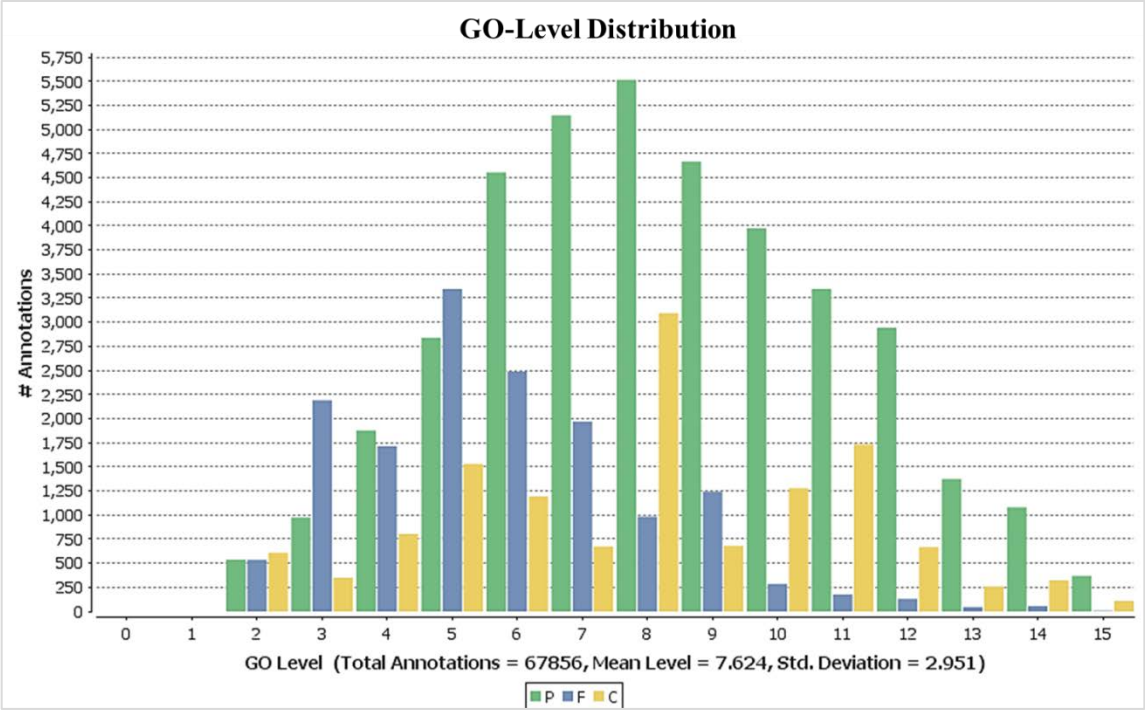


Figure 3.11 GO-level distribution. P = biological processes (green bars); F = molecular function (blue bars); C = cellular component (orange bars). The three categories are sub divided into levels, from level 2 to level 15 that determine coverage and specificity [122]. These levels correspond to the depth of the classifications with the higher levels being more general and the lower ones being more specific. The GO-level distribution of the results shows that in all of the 15 GO-levels, many of the GO annotated genes represented biological processes, indicated by the green bars.

CHAPTER 4. CONCLUSION AND FUTURE STUDIES

This study has succeeded to provide the first draft genome of *P. regina* using a combination of high quality paired-end reads sequenced by the Illumina HiSeq2000 sequencing platform supplemented with longer reads from the Roche 454 sequencing platform. Assessment of the draft genome reveals that it is approximately 524 Mbp in size and ~92% of the core eukaryotic genes have been detected suggesting it is a relatively complete genome. Although it is probable not all genes in *P. regina*'s genome were detected, the 11,037 genes predicted will serve as a starting point for further annotation of the genome.

Further studies on the draft genome will facilitate the extraction of genetic markers such as transposable elements (TEs), microsatellites and SNPs. These markers will enhance the study of genetic variation between blow fly populations and species. In addition, the detection of TEs in the *P. regina* draft genome will boost the study of gene regulation and evolution in blow flies. The completion of the mitochondrial genome assembly will be an additional source of information for phylogenetic and evolutionary studies of blow fly species.

Completion and refinement of the male and female draft genomes by annotation and genome comparison studies will be of great importance as it will enhance the study of sex determining genes, which will enable the determination of mechanisms that govern sexual determination in blow fly species. Since studies have shown that blow flies exhibit variable developmental rates due to various factors such as temperature, available food sources and geographical location; the detection and analysis of developmental genes will be of great value, especially to developmental and forensic studies. The study of these genes will enable the detection of markers that would characterize and categorize different developmental stages especially in the larval and pupae phases.

Much remains to be done with the draft genome; however, the presence of this first draft will be of great importance in blow fly and other studies. It will act as a source of reliable genomic data for comparative genomic studies between blow flies and other insect species. Further analysis of the predicted genes will enhance the detection of specific genetic markers (SNPs, TEs and microsatellites) that play a role in sexual determination, environmental adaptation and species identification. The draft genome, incorporated with restriction site associated DNA sequences will enable the identification of polymorphic SNPs to be used in mapping and population genetic analysis.

REFERENCES

REFERENCES

1. Rosenberg, D., H.V. Danks, and D. Lehmkuhl, *Importance of insects in environmental impact assessment.* Environmental Management, 1986. **10**(6): p. 773-783.
2. Byrd, J.H. and J.L. Castner, *Insects of Forensic Importance*. 2 ed. Forensic Entomology: The Utility of Arthropods in Legal Investigations, ed. J.H. Byrd, Castner J. L. 2010, Boca Raton, FL: CRC Press.
3. Scudder, G.G.E., *The Importance of Insects*, in *Insect Biodiversity: Science and Society*, R.G. Foottit and P.H. Adler, Editors. 2009, Wiley-Blackswell: Oxford, UK.
4. Dufour, D.L., *Insects as Food - a Case-Study from the Northwest Amazon.* American Anthropologist, 1987. **89**(2): p. 383-397.
5. Schoenly, K., R.A. Beaver, and T.A. Heumier, *On the Trophic Relations of Insects - a Food-Web Approach.* American Naturalist, 1991. **137**(5): p. 597-638.
6. Thompson, S.N., *Nutrition and culture of entomophagous insects.* Annu Rev Entomol, 1999. **44**: p. 561-92.
7. Hipolito, J., et al., *Pollination biology and genetic variability of a giant perfumed flower (Aristolochia gigantea Mart. and Zucc., Aristolochiaceae) visited mainly by small Diptera.* Botany-Botanique, 2012. **90**(9): p. 815-829.
8. Bommarco, R., L. Marini, and B.E. Vaissiere, *Insect pollination enhances seed yield, quality, and market value in oilseed rape.* Oecologia, 2012. **169**(4): p. 1025-32.
9. Schadler, M., et al., *Palatability, decomposition and insect herbivory: patterns in a successional old-field plant community.* Oikos, 2003. **103**(1): p. 121-132.
10. Shao, Z. and F. Vollrath, *Surprising strength of silkworm silk.* Nature, 2002. **418**(6899): p. 741.
11. Rinderer, T., A. Collins, and K. Tucker, *Honey production and underlying nectar harvesting activities of Africanized and European honeybees.* J. Apic. Res, 1985. **24**: p. 161-167.
12. Simpson, P. and S. Marcellini, *The origin and evolution of stereotyped patterns of macrochaetes on the nota of cyclorraphous Diptera.* Heredity (Edinb), 2006. **97**(3): p. 148-56.
13. Francesconi, F. and O. Lupi, *Myiasis.* Clin Microbiol Rev, 2012. **25**(1): p. 79-105.
14. McGuire, T.R. and J. Hirsch, *Behavior-genetic analysis of Phormia regina: conditioning, reliable individual differences, and selection.* Proc Natl Acad Sci U S A, 1977. **74**(11): p. 5193-7.
15. Volney, W.J.A. and R.A. Fleming, *Climate change and impacts of boreal forest insects.* Agriculture Ecosystems & Environment, 2000. **82**(1-3): p. 283-294.

16. East, I.J. and C.H. Eisemann, *Vaccination against Lucilia cuprina: the causative agent of sheep blowfly strike.* Immunol Cell Biol, 1993. **71 (Pt 5)**: p. 453-62.
17. Campobasso, C.P., G. Di Vella, and F. Introna, *Factors affecting decomposition and Diptera colonization.* Forensic Sci Int, 2001. **120**(1-2): p. 18-27.
18. Greenberg, B., *Flies as forensic indicators.* J Med Entomol, 1991. **28**(5): p. 565-77.
19. Norris, K.R., *The Bionomics of Blow Flies.* Annu. Rev. Entomol., 1965. **10**: p. 47-68.
20. Marineau, M.L., et al., *Maggot debridement therapy in the treatment of complex diabetic wounds.* Hawaii Med J, 2011. **70**(6): p. 121-4.
21. Sherman, R.A., *Maggot therapy takes us back to the future of wound care: new and improved maggot therapy for the 21st century.* J Diabetes Sci Technol, 2009. **3**(2): p. 336-44.
22. Singh, B. and J.D. Wells, *Chrysomyinae (Diptera:Calliphoridae) is monophyletic: a molecular systematic analysis.* Systematic Entomology, 2011: p. 1365-3113.
23. Brundage, A., S. Bros, and J.Y. Honda, *Seasonal and habitat abundance and distribution of some forensically important blow flies (Diptera: Calliphoridae) in Central California.* Forensic Sci Int, 2011. **212**(1-3): p. 115-20.
24. Byrd, J.H. and J.C. Allen, *The development of the black blow fly, Phormia regina (Meigen).* Forensic Sci Int, 2001. **120**(1-2): p. 79-88.
25. Jordaens, K., et al., *DNA barcoding and the differentiation between North American and West European Phormia regina (Diptera, Calliphoridae, Chrysomyinae).* Zookeys, 2013(365): p. 149-74.
26. Boehme, P., J. Amendt, and R. Zehner, *The use of COI barcodes for molecular identification of forensically important fly species in Germany.* Parasitol Res, 2012. **110**(6): p. 2325-32.
27. Cowan, F.A., *A Study of Fertility in the Blow Fly, Phormia regina Meigen.* Ohio Journal of Science, 1932. **32**(4): p. 389-392.
28. Nisimura, T., et al., *Experiential Effects of Appetitive and Nonappetitive Odors on Feeding Behavior in the Blow Fly, Phormia regina: A Putative Role for Tyramine in Appetite Regulation.* The Journal of Neuroscience, 2005. **25**(33): p. 7507-7516.
29. Walker, F.F., Haub, J. G., *Digestion in Blowfly Larvae, Phormia Regina Meigen, Used in the Treatment of Ostermyelitis.* The Ohio Journal of Science, 1933. **33**(2): p. 101-109.
30. Ali-Khan, F.E. and Z. Ali-Khan, *A case of traumatic dermal myiasis in Quebec caused by Phormia regina (Meigen) (Diptera: Calliphoridae).* Can J Zool, 1975. **53**(10): p. 1472-6.
31. Harrison, H.H. and D.J. Joslyn, *Gene expression patterns in the black blowfly (Phormia regina) as revealed by two-dimensional electrophoresis of proteins. I. Developmental stage-specific and sex-specific differences.* Biochemical Genetics, 1992. **29**.
32. Cockburn, A.F., T. Jensen, and J.A. Seawright, *Detection of Cryptic Species.* 1998. **29**(19): p. 31-36.
33. Catts, E.P. and M.L. Goff, *Forensic entomology in criminal investigations.* Annu Rev Entomol, 1992. **37**: p. 253-72.

34. Catts, E.P., *Problems in Estimating the Postmortem Interval in Death Investigations.* J. Agric. Entomol, 1992. **9**(4): p. 245-255.
35. Tarone, A.M., et al., *Population and temperature effects on Lucilia sericata (Diptera: Calliphoridae) body size and minimum development time.* J Med Entomol, 2011. **48**(5): p. 1062-8.
36. Hebert, P.D., et al., *Biological identifications through DNA barcodes.* Proc Biol Sci, 2003. **270**(1512): p. 313-21.
37. Wells, J.D. and F.A. Sperling, *DNA-based identification of forensically important Chrysomyinae (Diptera: Calliphoridae).* Forensic Sci Int, 2001. **120**(1-2): p. 110-5.
38. Wallman, J.E., T. Leys, and K. Hogendoorn, *Molecular systematics of Australian carrion-breeding blowflies (Diptera : Calliphoridae) based on mitochondrial DNA.* Invertebrate Systematics, 2005. **19**(1): p. 1-15.
39. Stevens, J. and R. Wall, *Genetic relationships between blowflies (Calliphoridae) of forensic importance.* Forensic Sci Int, 2001. **120**(1-2): p. 116-23.
40. McDonagh, L.M. and J.R. Stevens, *The molecular systematics of blowflies and screwworm flies (Diptera: Calliphoridae) using 28S rRNA, COX1 and EF-1alpha: insights into the evolution of dipteran parasitism.* Parasitology, 2011. **138**(13): p. 1760-77.
41. Nelson, L.A., J.F. Wallman, and M. Dowton, *Using COI barcodes to identify forensically and medically important blowflies.* Med Vet Entomol, 2007. **21**(1): p. 44-52.
42. Sonet, G., et al., *Why is the molecular identification of the forensically important blowfly species Lucilia caesar and L. illustris (family Calliphoridae) so problematic?* Forensic Sci Int, 2012. **223**(1-3): p. 153-9.
43. Picard, C.J. and J.D. Wells, *Survey of the genetic diversity of Phormia regina (Diptera: Calliphoridae) using amplified fragment length polymorphisms.* J Med Entomol, 2009. **46**(3): p. 664-70.
44. Stevens, J. and R. Wall, *Species, sub-species and hybrid populations of the blowflies Lucilia cuprina and Lucilia sericata (Diptera:Calliphoridae).* Proc Biol Sci, 1996. **263**(1375): p. 1335-41.
45. Florin, A.B. and N. Gyllenstrand, *Isolation and characterization of polymorphic microsatellite markers in the blowflies Lucilia illustris and Lucilia sericata.* Molecular Ecology Notes, 2002. **2**(2): p. 113-116.
46. Stevens, J. and R. Wall, *Genetic variation in populations of the blowflies Lucilia cuprina and Lucilia sericata (Diptera: Calliphoridae). Random amplified polymorphic DNA analysis and mitochondrial DNA sequences.* Biochemical Systematics and Ecology, 1997. **25**(2): p. 81-&.
47. Vos, P., et al., *AFLP: a new technique for DNA fingerprinting.* Nucleic Acids Res, 1995. **23**(21): p. 4407-14.
48. Picard, C.J., M.H. Villet, and J.D. Wells, *Amplified fragment length polymorphism confirms reciprocal monophyly in Chrysomya putoria and Chrysomya chloropyga: a correction of reported shared mtDNA haplotypes.* Med Vet Entomol, 2012. **26**(1): p. 116-9.

49. Kelkar, Y.D., et al., *What is a microsatellite: a computational and experimental definition based upon repeat mutational behavior at A/T and GT/AC repeats.* Genome Biol Evol, 2010. **2**: p. 620-35.
50. Kelkar, Y.D., et al., *The genome-wide determinants of human and chimpanzee microsatellite evolution.* Genome Res, 2008. **18**(1): p. 30-8.
51. Downing, T., et al., *Genome-wide SNP and microsatellite variation illuminate population-level epidemiology in the Leishmania donovani species complex.* Infect Genet Evol, 2012. **12**(1): p. 149-59.
52. Guichoux, E., et al., *Current trends in microsatellite genotyping.* Mol Ecol Resour, 2011. **11**(4): p. 591-611.
53. Nabity, P.D., L.G. Higley, and T.M. Heng-Moss, *Effects of temperature on development of Phormia regina (Diptera: Calliphoridae) and use of developmental data in determining time intervals in forensic entomology.* J Med Entomol, 2006. **43**(6): p. 1276-86.
54. Byrne, A.L., et al., *Forensic implications of biochemical differences among geographic populations of the black blow fly, Phormia regina (Meigen).* J Forensic Sci, 1995. **40**(3): p. 372-7.
55. Mardis, E.R., *Next-generation DNA sequencing methods.* Annu Rev Genomics Hum Genet, 2008. **9**: p. 387-402.
56. Pop, M. and S.L. Salzberg, *Bioinformatics challenges of new sequencing technology.* Trends Genet, 2008. **24**(3): p. 142-9.
57. Lorizzo, M., et al., *De novo assembly of the carrot mitochondrial genome using next generation sequencing of whole genomic DNA provides first evidence of DNA transfer into an angiosperm plastid genome.* BMC Plant Biology, 2012. **12**(61).
58. Li, R., et al., *The sequence and de novo assembly of the giant panda genome.* Nature, 2010. **463**(7279): p. 311-7.
59. Adams, M.D., et al., *The genome sequence of Drosophila melanogaster.* Science, 2000. **287**(5461): p. 2185-95.
60. Pegadaraju, V., et al., *De novo sequencing of sunflower genome for SNP discovery using RAD (Restriction site Associated DNA) approach.* BMC Genomics, 2013. **14**: p. 556.
61. Powers, J.G., et al., *Efficient and accurate whole genome assembly and methylome profiling of E. coli.* BMC Genomics, 2013. **14**: p. 675.
62. Maxam, A.M. and W. Gilbert, *A new method for sequencing DNA.* Proc Natl Acad Sci U S A, 1977. **74**(2): p. 560-4.
63. Sanger, F., S. Nicklen, and A.R. Coulson, *DNA sequencing with chain-terminating inhibitors.* Proc Natl Acad Sci U S A, 1977. **74**(12): p. 5463-7.
64. Earl, D., et al., *Assemblathon 1: a competitive assessment of de novo short read assembly methods.* Genome Res, 2011. **21**(12): p. 2224-41.
65. Liu, L., et al., *Comparison of next-generation sequencing systems.* J Biomed Biotechnol, 2012. **2012**: p. 251364.
66. Baker, M., *De novo genome assembly: what every biologist should know.* Nature America, 2012. **9**(4): p. 333-337.

67. Henson, J., G. Tischler, and Z. Ning, *Next-generation sequencing and large genome assemblies.* Pharmacogenomics, 2012. **13**(8): p. 901-15.
68. Margulies, M., et al., *Genome sequencing in microfabricated high-density picolitre reactors.* Nature, 2005. **437**(7057): p. 376-80.
69. Ronaghi, M., et al., *Real-time DNA sequencing using detection of pyrophosphate release.* Anal Biochem, 1996. **242**(1): p. 84-9.
70. Venter, J.C., et al., *The sequence of the human genome.* Science, 2001. **291**(5507): p. 1304-51.
71. Ewing, B. and P. Green, *Base-calling of automated sequencer traces using phred. II. Error probabilities.* Genome Res, 1998. **8**(3): p. 186-94.
72. Ewing, B., et al., *Base-calling of automated sequencer traces using phred. I. Accuracy assessment.* Genome Res, 1998. **8**(3): p. 175-85.
73. Churchill, G.A. and M.S. Waterman, *The accuracy of DNA sequences: estimating sequence quality.* Genomics, 1992. **14**(1): p. 89-98.
74. Weber, J.L. and E.W. Myers, *Human whole-genome shotgun sequencing.* Genome Res, 1997. **7**(5): p. 401-9.
75. Cock, P.J., et al., *The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants.* Nucleic Acids Res, 2010. **38**(6): p. 1767-71.
76. Miller, J.R., S. Koren, and G. Sutton, *Assembly algorithms for next-generation sequencing data.* Genomics, 2010. **95**(6): p. 315-27.
77. Treangen, T.J. and S.L. Salzberg, *Repetitive DNA and next-generation sequencing: computational challenges and solutions.* Nat Rev Genet, 2012. **13**(1): p. 36-46.
78. Phillippy, A.M., M.C. Schatz, and M. Pop, *Genome assembly forensics: finding the elusive mis-assembly.* Genome Biol, 2008. **9**(3): p. R55.
79. Narzisi, G. and B. Mishra, *Comparing de novo genome assembly: the long and short of it.* PLoS One, 2011. **6**(4): p. e19175.
80. Paszkiewicz, K. and D.J. Studholme, *De novo assembly of short sequence reads.* Brief Bioinform, 2010. **11**(5): p. 457-72.
81. Yandell, M. and D. Ence, *A beginner's guide to eukaryotic genome annotation.* Nat Rev Genet, 2012. **13**(5): p. 329-42.
82. Bosco, G., et al., *Analysis of Drosophila species genome size and satellite DNA content reveals significant differences among strains as well as between species.* Genetics, 2007. **ada177**(3): p. 1277-90.
83. Brent, M.R. and R. Guigo, *Recent advances in gene structure prediction.* Curr Opin Struct Biol, 2004. **14**(3): p. 264-72.
84. Parra, G., et al., *Assessing the gene space in draft genomes.* Nucleic Acids Res, 2009. **37**(1): p. 289-97.
85. Xie, Y., et al., *SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads.* Bioinformatics, 2014.
86. Zerbino, D.R., *Using the Velvet de novo assembler for short-read sequencing technologies.* Curr Protoc Bioinformatics, 2010. **Chapter 11**: p. Unit 11 5.
87. Butler, J., et al., *ALLPATHS: De novo assembly of whole-genome shotgun microreads.* Genome Research, 2008. **18**(5): p. 810-820.

88. Zimin, A.V., et al., *The MaSuRCA genome assembler.* Bioinformatics, 2013. **29**(21): p. 2669-2677.
89. Bradnam, K.R., et al., *Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species.* Gigascience, 2013. **2**(1): p. 10.
90. Li, Z., et al., *Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph.* Brief Funct Genomics, 2012. **11**(1): p. 25-37.
91. Simpson, J.T., et al., *ABySS: a parallel assembler for short read sequence data.* Genome Res, 2009. **19**(6): p. 1117-23.
92. Myers, E.W., et al., *A whole-genome assembly of Drosophila.* Science, 2000. **287**(5461): p. 2196-204.
93. Melsted, P. and J.K. Pritchard, *Efficient counting of k-mers in DNA sequences using a bloom filter.* BMC Bioinformatics, 2011. **12**: p. 333.
94. Luo, R., et al., *SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler.* Gigascience, 2012. **1**(1): p. 18.
95. Zerbino, D.R. and E. Birney, *Velvet: algorithms for de novo short read assembly using de Bruijn graphs.* Genome Res, 2008. **18**(5): p. 821-9.
96. Seabury, C.M., et al., *A multi-platform draft de novo genome assembly and comparative analysis for the Scarlet Macaw (Ara macao).* PLoS One, 2013. **8**(5): p. e62415.
97. Zimin, A.V., et al., *A whole-genome assembly of the domestic cow, Bos taurus.* Genome Biol, 2009. **10**(4): p. R42.
98. Salzberg, S.L., et al., *GAGE: A critical evaluation of genome assemblies and assembly algorithms.* Genome Res, 2012. **22**(3): p. 557-67.
99. Yang, X., C.P. Chockalingam, and S. Aluru, *A survey of error-correction methods for next-generation sequencing.* Oxford Journals, 2012.
100. Alkan, C., S. Sajjadian, and E.E. Eichler, *Limitations of next-generation genome sequence assembly.* Nat Methods, 2011. **8**(1): p. 61-5.
101. Wang, B., et al., *Whole genome sequencing of the black grouse (Tetrao tetrix): reference guided assembly suggests faster-Z and MHC evolution.* BMC Genomics, 2014. **15**: p. 180.
102. Nederbragt, A.J., et al., *Identification and Quantification of Genomic Repeats and Sample Contamination in Assemblies of 454 Pyrosequencing Reads.* Sequencing, 2009. **2010**.
103. Parra, G., K. Bradnam, and I. Korf, *CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes.* Bioinformatics, 2007. **23**(9): p. 1061-7.
104. Finn, R.D., J. Clements, and S.R. Eddy, *HMMER web server: interactive sequence similarity searching.* Nucleic Acids Res, 2011. **39**(Web Server issue): p. W29-37.
105. Stanke, M., et al., *AUGUSTUS: a web server for gene finding in eukaryotes.* Nucleic Acids Res, 2004. **32**(Web Server issue): p. W309-12.
106. Camacho, C., et al., *BLAST+: architecture and applications.* BMC Bioinformatics, 2009. **10**: p. 421.
107. Cock, P.J., et al., *Galaxy tools and workflows for sequence analysis with applications in molecular plant pathology.* PeerJ, 2013. **1**: p. e167.

108. Brockman, W., et al., *Quality scores and SNP detection in sequencing-by-synthesis systems.* Genome Res, 2008. **18**(5): p. 763-70.
109. Picard, C.J., J.S. Johnston, and A.M. Tarone, *Genome sizes of forensically relevant Diptera.* J Med Entomol, 2012. **49**(1): p. 192-7.
110. Kumar, S., et al., *Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots.* Front Genet, 2013. **4**: p. 237.
111. Lessinger, A.C., et al., *The mitochondrial genome of the primary screwworm fly Cochliomyia hominivorax (Diptera: Calliphoridae).* Insect Mol Biol, 2000. **9**(5): p. 521-9.
112. Gnerre, S., et al., *Assisted assembly: how to improve a de novo genome assembly by using related species.* Genome Biol, 2009. **10**(8): p. R88.
113. Cameron, S.L., *Insect Mitochondrial Genomics: Implications for Evolution and Phylogeny.* Annual Review of Entomology, Vol 59, 2014, 2014. **59**: p. 95-117.
114. Howland, D.E. and G.M. Hewitt, *Phylogeny of the Coleoptera Based on Mitochondrial Cytochrome-Oxidase-I Sequence Data.* Insect Molecular Biology, 1995. **4**(3): p. 203-215.
115. Kozol, A.J., J.F.A. Traniello, and S.M. Williams, *Genetic-Variation in the Endangered Burying Beetle Nicrophorus-Americanus (Coleoptera, Silphidae).* Annals of the Entomological Society of America, 1994. **87**(6): p. 928-935.
116. Sikes, D.S., R.B. Madge, and S.T. Trumbo, *Revision of Nicrophorus in part: new species and inferred phylogeny of the nepalensis-group based on evidence from morphology and mitochondrial DNA (Coleoptera : Silphidae : Nicrophorinae).* Invertebrate Systematics, 2006. **20**(3): p. 305-365.
117. Mohamed, S. and J.F. Wen, *Molecular identification of forensically relevant Diptera inferred from short mitochondrial genetic marker.* Libyan Journal of Medicine, 2013. **8**.
118. Tweedie, S., et al., *FlyBase: enhancing Drosophila Gene Ontology annotations.* Nucleic Acids Res, 2009. **37**(Database issue): p. D555-9.
119. Holt, R.A., et al., *The genome sequence of the malaria mosquito Anopheles gambiae.* Science, 2002. **298**(5591): p. 129-49.
120. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.
121. Conesa, A., et al., *Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.* Bioinformatics, 2005. **21**(18): p. 3674-6.
122. Dennis, G., Jr., et al., *DAVID: Database for Annotation, Visualization, and Integrated Discovery.* Genome Biol, 2003. **4**(5): p. P3.

APPENDICES

Appendix A Mitochondrial Nucleotide Sequences

>Contig 1, Length = 15,801 bp
ATTCAAGCTCCTAATCATGAATTAGCTGAAATAGAAATTAAAGTTCCTATTATCAAAATTCCGAAAAATAT
AATTTTTGATGAATTATTAAAAATTAAAAAGAAAAGGATTAGAACCTTTATAAATGGGGTATGAGCCCAGT
AGCTTAATTAGCTTATCTTTTTATTGTATATTTTGGTGTATGATGCACAAAAGTTTTTGATACTTTTAGAA
ATAGTTTAATTCTATTAAATATAATAAATTATAATGAATGTAATATTATTACATGATTTACYCTATCAAGG
TAATCCTTTTTATCAGGCAATTCATTGATTACTAATTTAATTTATTTAATAAAATATAATTTATATTGTTT
ATTATTTCATGTTTTTTAATTCATTGTATAAAGTTTTATTTTGGCTTGTAAATTTATTATTAGTTTAATTT
ATATGTAAATTTTTGTGTGAATTTTTATTTATTTTAATAGTAAATAAATTAATTATATTCGCAGTAATTAA
TATTATTAATTAGGGAAACCTAGAAATAGTAATACTAAAGAGTATTGGCTAAATTTGTGCCAGCAGCCGCG
GTTATACGAATAATGCAAATAAATTTTTTTAGTTAAAGTTAAATTGTTTATTTATAAAATAAAAATAAATT
TATTAGGTGAAATTTTAAATTTATAAATTTTTTATTTAAATTAATTGAAGCTTGAAAATTTTAATAATAAA
CTAGGATTAGATACCCTATTATTTAAAATGTAAATTTAAAAACTAAGGTAGTAGTAGTTATGTTCTTGAAA
CTTAAAAAATTTGGCGGTATTTTAGTCTGTTCAGAGGAACCTGTTCTGTAATCGATAATCCACGATGGACC
TTACTTAAGTTTGTAATCAGTTTATATACCGTCGTTATTAGAATATTTTATAAGAATGTTAATTTTCAGAA
TTTTATAAAAGAAAATATATCAGATCAAGGTGTAGCTAATATTTAAGTAGAAATGGGTTACAATAAAATTA
TTTATACGAATATAAATTTGAAATGTTTATTGAAGGTGGATTTGATAGTAAAATTATAAAGATTAATAATT
TGATTTTAGCTCTAAAATATGCACACATCGCCCGTCACTCTTACTACTAAGGTAAGATAAGTCGTAACATA
GTAGATGTACTGGAAAGTGTACCTAGAATGCAATTTAAAGCTTATTTAGTAAAGCATTTCATTTACATTGA
AAAGATTTTTGTGCAAATCAATATAAATTGAATAGATTTTATTTATTAATTATTTTTGTTTTTAAATTAAA
TTATTAAAAATAATTATAAAATGAATTGTTTTAGTATTTAATAAAGAAAAAATAATATTAATTATAGTGTA
ATAGTATTGTGAAAGAAAATTGAAATAATTTGAAAAATTTTTATTATAAAAGAAAATTTAATTTATTGTAC
CTTGTGTATCAGGGTTTATCAAATAAAAATTATTTATTATAATTTTCTCGATTTTAAAAGAGTTAATATAT
TATCAAAGTTAATGTGGCAAAATTATTTTTAATAATATATTAGAAATGAAATGTTATTCGTTTTTAAAGGT
ATCTAGTTCTTTAAGAAATAAATTTAATTTAGAAATTTTATATTATTTAATTAAATAAATTAATTAAATAA
ATATTTAATTTTAATATTTTATGGGATAAGCTGTAAAATAAATTTTTAAAAATAAATAAATAAATTAATAA
ATATAAGCTTAGAATTAGCTATTATTAATAAAAGTGTTATAATTTATTTTATAATAATTATTATTTATTAA
TATTTTAAATTTTTATTAAAGTATTTATTTTAATTAAAAAAATAAATAAATAATGATAAAATTAGTATATT
TAATTGTATAAATAAATATAAATGAAAAGTTTTTATAAAGAACTCGGCAAAAATAATGTTCGCCTGTTTAA
CAAAAACATGTCTTTTTGAATTTTTTTTAAAGTCTAGCCTGCCCACTGAATTTTTTTAAATGGCCGCAGTA
TACTAACTGTGCAAAGGTAGCATAATCATTAGTCTTTTAATTGAAGGCTGGTATGAATGGTTGGACGAGAT
ATTAACTGTTTCATAAAAATTTATAATAGAATTTTATTTTTTAGTCAAAAAGCTAAAATATATTTAAAAGA
CGAGAAGACCCTATAAATCTTTATATTTAAATTATTATAATTTTATAGATTTATTTTGTTATAATAGTTGA
TAATATTTTATTGGGGTGATATTAAAATTTAATAAACTTTTAATTGTTTAAATCATTAATTTATGAATAAT
TGATCCGTTATTAGCGATTAAAAAAATAAGTTACTTTAGGGATAACAGCGTAATTTTTTTGGAGAGTTCAT
ATCGATAAAAAAGATTGCGACCTCGATGTTGGATTAAGATATAATTTTAGGTGTAGCCGCTTAAATTTTAA
GTCTGTTCGACTTTTAAATTCTTACATGATCTGAGTTCAAACCGGCGTAAGCCAGGTTGGTTTCTATCTTT
AAAAAATTAAAATATTTCAGTACGAAAGGACCTAATATTTGATAAATTATTATTTTTATATTGAATATAAT
TAATATAATACTATTTTGGCAGATTAGTGCAATAAATTTAGAATTTATGTATATGATTTTTATCATAAATA
GTACTTGTTTTTTATAGAAAATTTGTTATTTATTATTGGTAGATTATTACTAATTATTTTTGTATTAGTAA
GAGTAGCTTTTTTAACTCTTTTAGAACGGAAGGTATTAGGTTAYATTCAAATTCGGAAGGGTCCTAATAAG
GTAGGAATTGCAGGGATTCCTCAACCTTTTTGTGATGCAATTAAGTTATTTACTAAGGAACAAACTTATCC
TTTATTGTCTAATTATATCTCTTATTATTTTTCACCAATTTTTTCTTTATTTTTATCTTTATTAGTATGAA
TATGTATACCAATATTTGTAAAATTATTTTCATTTAATTTAGGTTTATTATTTTTTTTATGTTGTACTAGT
TTAGGGGTTTATACAGTAATAATTGCTGGTTGATCTTCTAATTCAAATTATGCTTTATTAGGAGGATTACG
AGCTGTTGCTCAAACAATTTCTTATGAAGTAAGTTTAGCTTTAGTTTTATTAAGTTTTATTTTTTTAATYG
GAGGGTATAATATATTAATATTTTATAAGTATCAAATGTTTATTTGATTTTTATTTATTATGTTTCCTATA
GCGTTAGTGTGATTTAGTATTTCTTTAGCTGAAACTAATCGTACACCATTTGATTTTGCTGAAGGAGAATC
AGAATTAGTTTCTGGATTTAATGTAGAATATAGAAGAGGAGGATTTGCTTTAATTTTTTTAGCTGAATATG
CAAGAATTTTATTTATAAGAATATTATTTTGTGTTATATTTTTGGGAAGTGATGTATTTTCTTTTTTTTTT
TATATTAAGTTAACTTTTGTTTCATTTATATTTATTTGAGTTCGAGGGACTTTACCTCGTTTTCGGTATGA
TAAATTAATATATTTAGCTTGAAAAAGTTTTTTACCTTTTTCATTAAATTTTTTATTATTTTTTGTTGGAT

TTAAAATTTTTTTAATTTATTTAATTTAATGTATTATTTTTAGTAAAGTTAATAGAAAATTTGATTTCTAT
CTTATGTTTTCAAAACATATGCTTATTTCAAGCTCATTAACTAATTTAATAAATTATCTCATCATTTAATA
ATTAGTGGGTTAAKTATAAAATAAGAAAAGTATACTACAGTTAAAATTTGACCTACTAATACGTAAGGTTC
TTCAACTGGTCGAGCTCCAATTCATGTTAGTAGAATTACTGTAACTACTATTACTCAGAATAAAATTTGAT
TAATAGGATAGAATTGAATTCCTCGAAATTTACTTAAATGGTAAAATGGTAAAATTGCTAAAATTGCAATA
GATAAAACTAATGCAATTACTCCTCCTAGCTTATTAGGAATAGATCGYAGAATTGCATAAGCGAATAGGAA
GTATCATTCTGGTTNNNNNNNNNNNNNNNNNNGGATTAATTAGAACTAGTAGAATTAGAATTATTGTTATT
ACAATGAACCCTACAATATCCTTGTAAGTAAAATAAGGATGAAATGGAATTTTATCAATATTTGAATTTAA
TCCTATTGGGTTATTTGATCCTGTTTCATGAAGGAATAAAATATGAATTAAAGTAGCGGCTAATACAATAA
AAGGTAGAATAAAGTGGAAAGTAAAGAATCGTGTTAATGTTGCATTATCWACAGCAAATCCTCCTCATACT
CATTGTACTAAGTCGATCCCTAAATATGGAATAGCAGATAATAAATTAGTAATAACTGTTGCTCCTCAGAA
AGATATTTGTCCTCAAGGTAGTACATATCCTATAAAGGCTGTTCCTATTACTAAAAATAGGATAATTACTC
CTACTAATCAAGTTGGAGTAAATAAATATGAACCATAATAGATTCCTCGTCCTACATGTAAATAAATACAA
ATAAAAAAGAATGATGCACCATTAGCATGTATAGTTCGTAATAATCATCCATAATTTACATCTCGACAAAT
ATGATTTACTCTATTAAAGGCTAAATTAATATCTGCAGTGTAATGTATAGCTAAAAATAATCCAGTTAAGA
TTTGAATCATTAAACATAAAAATAATAATGATCCAAAGTTTCATCAAGCTGAAATATTAATAGGGGCAGGT
AAATCTACTAAAGCACTATTAGCAATTCTAAAAATTGGGTGTTTAATTCGTAAAGGTTTGTTCATTAGTTG
AATATTGGTCGAAGAGGTCCCTTAAATAGCTTAGTAATTTTTACAACAGCAATTAATGTAATTAATAAATA
ATTTATTAATAAAATTGTTAATAAATTAGTTGGATAATTATATAATTTGTTAAGTGATAGGGAATTTTCTG
TTAAATATGAGTTTATATCATAAATTGATTTAACTTCTAAATTTTGATATTGAAGTAGTAAGTTTTTATCT
ATGAAATATAAAGATAAAATTATAAAYACAAAAATTGATAAGGAGATGAATATGAGTTTGATTGAAAATGA
AAATATTTCATTTGATGCAAGAGAAGTTACATAAATGAATAATACTAATATCCCTCCTAAAAATACTAGGA
ATAAAATATAAGAAAATCAGAATCTTTTAGTTATTAGACCTGAAGTTAAACAAACTAAAGTTGTTTGAATA
AGTAATGTTAATCCTATAGCTAGAGGGTGTTTCATATTTATGAAAACAAAATTAAAAATAATTAAAAGTGA
TATTAAAATTCATTGTATAATATCAAGAGGTAGTTTAAATAAAATATTAATTTTGGGGATTAATGATAAAG
AAATTTCTTTTCTCTTGAAGTTTTAAAAGAAATAATCTTATTTTTGATTTACAAGACCAATGTTTTTATTA
AACTATTAAAACTAATGATAGCAATTTTAAATTGAAGCTTACCAAGTATTTTATTTATTATAGGAGTATTT
ACTTTTGTTTCTAATCGTAAGCATTTATTATCAATATTATTGAGATTAGAATATATTGTATTGAGTTTATT
CCTTTTATTATTTATTTATTTAAATATGTTAAATTATGAAAATTTTTTTAGTATAATATTTTTAACTTTTA
GAGTTTGTGAAGGTGCACTTGGACTTTCAATTTTAGTTTCAATAATTCGTACACATGGAAATGATTATTTT
CAAAGATTTAATGTTTTACAATGTTAAAAATTATTATTAGAATTTTATTTTTATTTCCATTGTGTTTAATA
CATAATACTTATTGAATGGTTCAAAGTTTTTTATTTTTATTAAGTTTTATTTTTATTTTAATAAATATATA
TAGAAATTATTTTATATCAATTTCTTATTTATTTGGATGTGATATAATTTCTTATGGATTAATTTTATTGA
GCTTATGAATTGTTTCTTTAATATTAATGGCTAGTGAGTCAGTTTATAAGTATAGAAATTATACAAATTTA
TTTTTATTAAATATTGTTTTATTATTAGTTTTATTAGTTCTTACTTTTAGAAGAATGAGATTATTTATATT
TTATTTATTTTTTGAAAGTAGTTTAATTCCTACTTTATTTTTAATTTTAGGTTGAGGGTATCAACCAGAAC
GATTACAGGCTGGAGTATATTTATTATTTTATACTTTGTTAGTGTCTTTGCCWATATTAATTGGTATTTTT
TATTTATATAAGGTTACGGGAACTTTGAATTTTTATTTATTAAATAATTATATATTTAATTATGAATTTCT
TTATTTTTCTTTAGTGATAGCTTTTTTAGTAAAAATACCTATATTTTTAGTTCATTTATGATTACCTAAGG
CTCATGTAGAAGCTCCTGTTTCAGGTTCAATAATTTTAGCTGGAATTATATTAAAATTAGGGGGTTATGGA
TTATTACGAGTATTTCCTTTTTTACAGATTATAGGATTAAAGTTTAATTTTATTTGAATTAGAATTAGATT
AGTAGGAGGAGTACTAGTTAGTTTAATTTGTTTACGACAAACAGATTTAAAGGCATTAATTGCTTATTCTT
CAGTTGCTCATATAGGAATTGTTTTAGCTGGGTTAATAACTTTAACTTATATAGGAATTTGTGGTTCTTAT
ACTTTAATAATTGCTCATGGTTTATGTTCTTCAGGACTTTTTTGTTTAGCTAATATTTCTTATGAACGAAT
GGGTAGTCGTAGATTATTAATTAATAAGGGTATATTAAATTTTATACCTTCAATGGCATTATGATGRTTTT
TATTAAGATCAGCTAATATAGCTGCTCCTCCTACTTTAAATTTATTAGGAGAAATTTCTTTAATTAATAGT
ATTGTTAGTTGATCTTGAGTTTCAATATTAATGTTATCTTTATTATCTTTTTTCAGAGCTGCTTATACGTT
ATATTTATACGCTTATAGTCAACATGGAAAGATTTTTTCTGGTGCTTATTCATTTAGAGGAGGTTCAGTTC
GTGAATTTTTACTTTTATTTTTACATTGATTTCCTTTAAATTTGTTAATTTTAAAGGGAGATATATGTATA
TTGTGATTATATTTAAATAGTTTAATAAAAAATATTGATTTGTGGTGTCAATGATAAGAAAATTTCTTTTT
AAATCGTGAAATATTTATCAATTTGTACAATTAGTTTTGTAAGTTTATTTTTTTTTAGATTATTATCTTTT
TTAATAGGGATAATTTTTATTATAAATGATTATAGAATTTTTATTGAATGAGAAGTGGTTTCTATAAATTC
TTTAAGAATTGTTATAACTTTATTATTAGATTGAATAAGATTAACTTTTATATCTTTTGTTTTAATAATCT
CTTCTTTGGTTATTTTTTATAGAAAGGAGTATATAGAAAGTGATTATAAAATTAATCGATTTATTATATTA
GTTTTAATATTTGTTATATCAATAATGTTATTAATTATTAGTCCTAATTTGATTAGAATTTTATTAGGATG

AGATGGATTAGGACTTGTATCTTATTGTTTAGTTATTTATTTTCAAAACGTCAAGTCTTATAATGCTGGRA
TATTAACTGCTTTATCAAATCGGATTGGAGATGTTGCGTTATTATTAGCTATTGCTTGAATGTTAAATTAT
GGTAGTTGAAATTATATTTTTTATTTGGAAGTAATAAAAAGTGATTTTGAAATAATAATTGTAGGAAGATT
AGTTATATTAGCAGCTATAACTAAAAGTGCTCAAATTCCTTTTTCTTCTTGATTACCAGCTGCTATAGCAG
CTCCTACGCCTGTTTCTGCTTTAGTTCATTCTTCAACTTTGGTAACGGCAGGAGTATATTTATTAATTCGA
TTTAATATTTTATTAAATAGATCATGAATGGGTAATTTATTATTATTATTATCTGGGTTAACAATATTTAT
AGCTGGGTTAGGAGCAAATTATGAATTTGATTTAAAGAAGATTATTGCTYTATCTACTTTAAGTCAGTTAG
GTTTAATAATAAGAATTTTATCTATAGGTTATTATAAATTAGCTTTTTTTCATTTATTAACTCATGCTTTA
TTTAAGGCTTTACTTTTTATATGTGCTGGAGCTATTATTCATAATATAAATAATTGTCAAGATATTCGTTT
AATAGGAAGATTAAGTTTAATAATACCACTTACATCTTCTTGTTTTAATGTTGCTAATTTAGCTTTATGTG
GTATACCTTTTTTAGCTGGGTTTTATTCTAAGGATTTAATTTTAGAAACAGTATCTTTGTCTTATATTAAC
ATGTTTTCTTTTTTTTTTATATTTTTTTTCTACAGGTTTAACTGTTTGTTATTCTTTTCGGTTGGTATATTA
TACAATAACTGGAGATTCAAATTTTTCATCTTTGAACATGCTTAATGATGAAGGTTGAGTGATATTAAAAA
GTATAYTAGGTTTATTGATTTTAAGAATTTTTGGGGGAAGAATGTTAAGATGGTTGATTTTTCCTACACCA
ATAGTAGTTGTTCTTCCTTTTTATTTAAAGTTGTTAACTTTATTCGTTTGTATTGTAGGGGGATTAATAGG
TTACATGATTTCTCATGTTTCTTTATTTTTTTATAATAAGGCTTTAAATAATTATCATTCTTCTTATTTTT
TAGGTTCTATATGATTTATACCTTATATTTCTACTTATGGAATTATTAATTATTCATTAGTTGTAGGTAAT
ATAGTGGTAAAGTCTTTTGATCAAGGTTGATCAGAATATTTTGGAGGTCAACAATTATATTTAAATTTAGT
AAAAAATTCTCAATTAAATCAAATATTACAAAATAATAATTTAAAAATTTATTTATTAAGTTTTATTTTTT
GAATTATAATTTTATATATGTATATAATTTGTTTTTATTCAAATAGCTTATATTTAGAGTATGACACTGAA
GATGTTAGGGAGATTAATTTAATCTTTGAATATAATGAATTATTTTTAGTAATTTATAAAAATAAAAAATT
TTAACTTTACAATGAAAATGTAATGTTTTATTTAAACTATATAAACTAGAAGTATAAATGGAATTTAACCA
TTAAAAGAAAAAAGTTAGCAGCTTTTACTTGAACATCATACTTCTTTAATTGGAGTTTTGATCTCAATTCT
ATCATTAACAGTGATAAGCCTCTTTTTGGCTTCAATTAAGTAGAATAAGGGTAGAAATTACCTAAGATTAG
GTCGAAACTAATTGCAATATATCGCTTCATATTCAAGGTAGATTGAAAAATCAATACTTTTTGAATGCAAA
TCAAATGTTATACTTAACTACAACCCTAAAAATTAATTTGATCAATTTAGTATACCTTGATTTCATTCGTG
GTAAAGTCCTAATAATAAAATTAAAATAAAAATAATTGATGTAATTGTTCATAYTATTAAATTTGAAAATT
TAATAATTAAGATAATAGGAAGAATTAAAGCAATTTCTACATCAAAAATTAAAAAAATAATTGTAATTAGA
AAGAATCGTAAGGAGAAAGGTAATCGAGAAGAAGATTTTGGGTCGAATCCACATTCAAATGGAGATGCTTT
TTCTCGATCAACAAGTGTTTTTTTCGAAAGAATTGAAGCCAATAATATAACTACGATTGAGATTAATAAAA
TAATTGAACTAATTGTTAAAATTGATAAAATTATCTATACTATTAATAATAGACCATATGATTGGAAGTCA
AATATACTTTTTATACTATATAGATAATTAATTATCCTCCTCATCAATAAATTGAAACATAAAGAAATAAT
CAAACAATATCAACAAAGTGTCAGTATCAAGCAGCGGCTTCAAATCCGAAGTGATGATTTTTAGAAAAATG
ATTATTTAAATGTCGGATTAAACAGATTAATAAGAAAGTTGTTCCAATTAAAACATGAATTCCATGGAATC
CTGTTGCTATAAAAAATGTTGATCCATAAACAGAGTCAGCAATTGTAAATGGGGCTTCAATGTATTCATAA
GCTTGTAAAATTGTAAAATAAACTCCTAAAATAACTGTAAAAAATAATCCTTGAGTAGTTTGGGAGTGATT
TCCTTCCATTAAACTGTGGTGAGCTCAAGTTACAGTAATTCCTGAAGTTAATAAAATAACTGTATTTAAAA
GAGGAATTTGGAAGGGATTAAAAGGAGTAATTCCTATAGGAGGTCATATAGCTCCTAATTCAATTGATGGT
GAAAGACTACTGTGAAAAAAAGCTCAGAAGAATGAAACAAAGAATAAAACTTCTGATAAAATAAATAAAAT
TATTCCTCATCGTAATCCTGTAGTTACTGCATCAGTATGAAGTCCTTGGAATGTACCTTCTCGTGAAACAT
CTCGTCATCATTGATAAACAGTTAAAATAGTGATAATATTTCCTAAAAAGAATAATGATGTGTCATATTGA
TGAAATCATTTTACTATACCAGCAACAGTTGTTATAGCTCCGATTGAAGCTGTTAATGGTCATGGACTGTA
GTCTACTAAATGGAATGGGTGATTTGAGTGAGTTGACATTAGTTTACTTCACTAGAATATAAAGTTCTTAA
AACAGCAAATACGTATGATTGAATTATAGCAACAGCTGATTCTAGAACTAATAAAGCAATTTGTGTAATAA
TTAATAAACTTAATAAAATTGTTGATATAGAAGGTCCAGTATTTCCTAAAAGAGTTAATAGTAAATGTCCT
GCAATTATATTAGCTGTTAGTCGAACTGCTAAAGTTCCAGGTCGAATTACATTACTAATAGTTTCAATGCA
TACTATAAAAGGTATTAAAACTGCTGGGGTTCCTTGAGGAACTAAGTGGGCAAATATATGTTGAGTGTTAT
TAATTCATCCAAATAATATAAAACAAAGTCATAGGGGAAGAGCTAATGTTAAAGTTAGAGTTAAGTGACTT
GTTCTTGTAAAAATATAAGGGAATAATCCTATAAAATTATTAAATAAAATTATAGAAAATAAAGAAACAAA
AATAAATGTAGAACCGTTTGCTCCYATAGGTCCTAAAAGAGTTTTGAATTCCTTATGAAGTGTTAATAAAA
TATTATTTCAGAAAATATGATATCGGGAAGGTATTAATCAATATATTGATGGAATTATTAAAATTCCTAAG
AATGTTCTTAATCAATTTAATGATAAATTAAAAATACTTGAAGAAGGATCAAATACTGAAAATAAATTTGT
TATCATTTTCAATTTAATGAATTTGTAGATTGTGTTTTGTTTACTAAATCAGATTTAGGTATAGAAGGAAT
AAATGAATAATAATTTATTATATTAAAAATTACGAATGCAATTGAAAAGATAATAAATAAACTTAATCAAC
CAATAGGTGCTATTTGAGGAATTAAAAAATATCAAAAAMCTGATAATTTTAGTTTGACAAACTAATGTTAT

AAATTTAACTAATTTTTTCATTAGAAGTAAGTGCTAATTTACTATAAAATGGTTTAAGAGACCAGTACTTG
CTTTCAGTCATCTAATGAAGAGTTTACATTATTAGAAATTCATTTGATAAAGTAATTTACTGGGATTCTTT
CGATTACAATTGGTATAAAACTATGATTAGCTCCACAAATTTCTGAACATTGTCCATAAAATAAACCTGGT
CGATTAATTAAGAAATTAGTTTGATTTAGTCGTCCAGGTGTACCATCAACCTTAACTCCTAAAGCTGGAAT
AGTTCATGAATGAATTACATCGGCTGCTGTCACTAAAATTCGAATTTGTGAATTTATTGGTAAAACTACTC
GATTATCAACGTCTAATAAACGAAAACTATCAATTGATAATTCATTTGTAGGAATTATATAAGAATCAAAC
TCAATGTTTGCAAAATCAGAATATTCATAACTTCAATATCATTGATGTCCAATTGCCTTTAAAGTAATTGA
AGGTTCATTAATTTCATCTAGTAAGTAAAGAAGTCGTAAAGAAGGAAAAGCAATAAATAACAAAATAATTG
CAGGTAAAATTGTTCAAATAATTTCAATAGTTTGTCCATGGAGTAGATATCGATTTACATATTTATTAAAG
AATAATATAAATATTAAATAACCTACTAGAACAGTAATTATTACTAAAATTAAAAGTGCATGGTCATGGAA
GAAAATCAATTGTTCTATTAAAGGAGAAGAACTATCTTGTAAACCTAAATTTGCTCATGTTGACATTTATT
TTCTAATAAAAGTAAAATACTTTATATATGGAGCTTAAATCCATTGCACTAATCTGCCATATTAGAAATTA
GTTAATAAAGGTAATTCACTATAGCTGTGTTCAGCTGGTGGAGTATTTTGTAATCATTCAATAGATGAATT
TAATTGTACAGGGAATAAAACTTGACGTTGAGATACTAAACTTTCTCAAATAATAAAAAAGAAAAATAAAA
TTCCTAATAATGAGATTGTTGAACCAATTGTAGAGATTACGTTTCAAGCCGTGTAAGCATCTGGGTAATCT
GAGTATCGTCGAGGCATTCCAGCTAATCCTAAGAAATGTTGAGGGAAGAATGTTAAATTTACCCCAATAAA
TATAATAGCAAATTGACTTTTTAATAACTTATTATTTAATGTTAATCCAGTAAATAAAGGGAATCAGTGGA
CAAATCCAGCTATAATAGCAAATACAGCTCCTATTGATAATACATAGTGGAAATGAGCTACTACATAATAT
GTATCATGAAGAATAATATCGATTGATGAATTAGCTAAAACAACACCAGTTAATCCTCCTACAGTAAATAA
AAATACAAATCCTAAAGCTCATAGAGTTGCTGGAGAGTAATTTAATTGAGTTCCATAAAGAGTTGCTAGTC
AACTGAAAATTTTAATTCCAGTTGGTACAGCAATAATTATAGTTGCTGAAGTAAAGTAAGCTCGTGTATCA
ACGTCTATTCCAACAGTAAATATATGATGAGCTCATACAATAAATCCTAATAGACCAATAGCTAATATAGC
ATAAATTATTCCTAATGATCCAAAAGTTTCCTTCTTTCCTGATTCTTGACTAATAATATGAGAAATTATTC
CAAATCCAGGTAGAATTAAAATATAAACTTCAGGGTGACCAAAGAATCAGAATAAGTGTTGATATAAAATA
GGATCTCCTCCTCCTGCTGGGTCAAAGAATGAAGTGTTTAAATTTCGATCAGTTAATAATATAGTAATAGC
ACCGGCTAATACAGGTAAAGATAATAAAAGTAATAGAGCAGTAATAACTACAGATCAAACAAATAAAGGTA
TTCGATCAAATGTAATTCCAGTTGATCGTATATTAATTACAGTTGTAATGAAATTTACAGCTCCTAAAATT
GAAGAAATTCCTGCTAAGTGAAGAGAGAAAATAGCTAGATCAACAGATGCTCCTCCATGAGCAATATTAGA
TGATAAAGGTGGGTAAACAGTTCATCCTGTTCCAGCCCCATTTTCTACTATACTACTAACTAACAATAGAG
TTAATGCAGGAGGTAAAAGTCAAAAACTTATATTGTTTATTCGTGGGAAAGCTATATCAGGAGCCCCTAAT
ATTAAAGGAACTAATCAATTTCCAAATCCTCCAATTATAATTGGTATAACTATAAAGAAAATTATAATAAA
AGCATGAGCTGTTACAATTACGTTATAAATTTGGTCATCTCCAATTAGAGCTCCAGGGTGCCCTAGTTCAG
CTCGAATTAGAATTCTTAATGAAGTTCCAATTATTCCAGATCAAGCTCCGAAAATAAAATATAAAGTACCA
ATATCTTTATGATTAGTAGAAAATAACCATTGTCGCGATTAAATGGCTGAAGTTTAGGCAATAGACTGTAA
ATCTATTTATGAGAATTATTCTCTTTAATCAATAGGCTTTATAGTCAATAATGACATTAGACTGCAATTCT
AAAGGAGTAATAAAATTACTAAGGCTTAAAAGATTATTCTTATATTTATAGCTTTGAAGGCTATTAGTTTA
TTTAACTTAAAGCCTTAGAATAAAAAGTATAAAGAAGAAATTAGGAATAACCCAAATGAAGAGAAAAATGA
ACAAATTATAAAAATTTTTATATAATTTGATTTATATACAGAAGAAGTTAATCAATTATTTTCATAATAAT
TTAATATAAATGCTCTGTAACATAATCGTATATAAAAATATAATGTGATTAGAGTCATTAAAACTATAAAT
GTTAATAAAAAAAAATTGACTATTTATAGTTAATGATTGAATAACAATTCATTTTGGGAAAAATCCTAAGAA
AGGAGGCAATCCTCCTAACGATAGTAAATTAAAAAATAAAAAAAATTTTATAACCTTTGAGTGGAAAAATA
AAGAAAATAATTGGTTAATATGTGAAGTTTTAAATATATTAAATATAAAAATTATACAAAAAGTTAAAAAA
GTATAAAATATAAAATAAGTTATTCACAATAAATTACTATTATATATAGCTGCTAATATTCAACCTAAATG
GTTAATTGAAGAATACGCTATTAATTTACGTAATGAAGTTTGATTTAATCCTCCTAATGCACCAATTAATC
TAGAAAGAATAATTCTTGTAATAATTATTGGTTTAAAAATAATATAAGAAATTAATATTAAAGGAGCAATT
TTTTGTCAAGTTATTAAAATTAATGCATTTAGTCATGATAGTCCTTCTATTACATTAGGAAATCAAAAATG
AAAAGGAGCTGACCCTCTTTTTAATAAAAGAGAGGAGAAAATTATTATTTCTATAAAGTAATTTGAATTTC
TTTTACTTGAATTTAGTAAAAATAAAATTACTGCAAATAAAAATACTGAAGACGCTAATGCTTGAGTTAAG
AAATACTTTAATGAGGCTTCTGTTGATATTAATTTATTATCTCTTATTAGGGGGATAAAAGCTAACAAATT
AATTTCTAAACCTATTCAAGCTCCTAATCATGAATTAGCTGAAATAGAAATTAAAGTTCCTATTATCAAAA
TTCCGAAAAATATAATTTTTGATGAATTATTAAAAATTAAAAAGAAAAGGATTAGAACCTTTATAAATGGG
GTATGAGCCCAGTAGCTTAATTAGCTTATCTTTTTATTGTATATTTTGGTGTATGATGCACAAAAGTTTTT
GATACTTTTAGAAATAGTTTAATTCTATTAAATATAATAAATTATAATGAATGTAATATTATTACATGATT
TACCCTATCAAGGTAATCCTTTTTATCAGGCAATTCATTAAAAGTAAGTTTCTCATGTTTTTGTTTTTTTT
TTTTTWTTYTTTATWYTTATTTTTTTGTTAATTAATAAATTATAGAATATAGTTTATTAATTAATTAGATA

ATTATAATTAGTATTARTTTATTAGATTAAATAAATTATAAATATAAATAATTAAATTTTGTATATATATA
TATATATATATATWTATATATATWTATWYATATATTTATAATATATTAAAATTTGTATACAATGTTATATA
ATTAATAATTAATAAATTATTTATATAAAACTATAAATATAATTCACATACATTATTAAATAATTTTTTTT
AATAAATCATATTTCTACATAATAATTATATAAATCATTATTAATAATGCTTTAATAACATGATATTCTTA
ATTGGATTTAAATTTTTTTTTTTATCTAAATTAATAGATATATATTAATTAATTAAATATTTATATATTAA
TAGATATCTATTAATCTTATTTGGTATATAGACTAAAAATAAATTTTTGCATCGCTCAATATAAAATTGGA
GAGGTATATATAAATGAATTATATATTATAATTTTTARAWAWAAAWAAAAAAAAAAAAAAAAAAAATWWW
AAAAATTHATTTWTTTTTAATAAATTTTTGCTAATTTTT
>Contig 2, Length = 628 bp
ATCGTTCATATTATTAAATTGGAGAATTTAATAATTAAAATAATAGGTAGAATTAAAGCAATTTCTACATC
GAAAATTGGGAAAATAATTGTGATTAAAAAGAATCGTAAAGAAAAAGGTAATCGGGATGACGATTTTGGAT
CAAATCCGCATTCAAATGGAGATGTTTTTTTTCRATCAACRAGTGTTTTTTTTTTTTTTTAADAATTGAAG
CTAAAARTATTACTACAATTGAAATTAATCAAATAATTGAACTAATTGTTANNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNAAAACATGAATTCCGTGGAATCCTGTCGCTATAAAAAATGTCGATCCGTATACAGAGTCAGCGAT
TGTAAATGGAGCTTCAACGTTTTCATAAAAAATCATCCTTGAGCAGTTCGGGAATGATTT

Appendix B Eukaryotic Orthologous Groups (KOG's) Proteins IDs.

List of 442 Eukaryotic Orthologous Groups (KOGs) ID's detected by CEGMA in the Draft Genome								
KOG0002.7	KOG0327.3	KOG0563.1	KOG0952.21	KOG1358.9	KOG1646.4	KOG2014.14	KOG2732.3	KOG3174.2
KOG0003.6	KOG0328.2	KOG0567.1	KOG0959.85	KOG1367.4	KOG1647.3	KOG2017.1	KOG2738.6	KOG3180.4
KOG0018.5	KOG0329.8	KOG0602.18	KOG0960.10	KOG1370.2	KOG1654.3	KOG2035.4	KOG2749.1	KOG3185.2
KOG0019.2	KOG0330.1	KOG0622.5	KOG0964.10	KOG1373.5	KOG1662.2	KOG2036.5	KOG2754.4	KOG3188.7
KOG0025.1	KOG0331.1	KOG0625.4	KOG0969.12	KOG1374.4	KOG1664.2	KOG2044.1	KOG2757.5	KOG3189.2
KOG0047.1	KOG0344.7	KOG0631.6	KOG0985.6	KOG1390.2	KOG1668.2	KOG2047.16	KOG2767.1	KOG3204.7
KOG0062.18	KOG0346.5	KOG0650.16	KOG0989.9	KOG1393.2	KOG1678.1	KOG2067.3	KOG2772.4	KOG3205.1
KOG0073.6	KOG0357.9	KOG0659.3	KOG0991.2	KOG1394.3	KOG1688.4	KOG2104.3	KOG2775.2	KOG3218.4
KOG0077.4	KOG0358.5	KOG0675.6	KOG0996.14	KOG1415.4	KOG1691.7	KOG2189.21	KOG2781.5	KOG3222.17
KOG0084.5	KOG0359.2	KOG0679.2	KOG1036.2	KOG1430.4	KOG1692.5	KOG2270.9	KOG2783.4	KOG3229.8
KOG0092.6	KOG0361.2	KOG0683.6	KOG1047.17	KOG1433.2	KOG1712.5	KOG2276.14	KOG2784.3	KOG3232.3
KOG0094.4	KOG0362.11	KOG0687.10	KOG1058.13	KOG1439.2	KOG1722.1	KOG2292.4	KOG2785.3	KOG3237.1
KOG0100.5	KOG0363.9	KOG0688.7	KOG1062.18	KOG1448.6	KOG1723.3	KOG2303.3	KOG2803.13	KOG3239.1
KOG0102.4	KOG0364.22	KOG0727.6	KOG1068.1	KOG1458.1	KOG1727.6	KOG2309.14	KOG2807.6	KOG3271.5
KOG0103.2	KOG0365.15	KOG0728.6	KOG1077.9	KOG1463.6	KOG1728.5	KOG2311.4	KOG2825.6	KOG3273.5
KOG0122.4	KOG0366.2	KOG0729.3	KOG1078.7	KOG1466.1	KOG1733.10	KOG2321.3	KOG2833.7	KOG3275.5
KOG0142.8	KOG0367.2	KOG0734.6	KOG1088.10	KOG1468.2	KOG1742.4	KOG2387.13	KOG2851.6	KOG3283.3
KOG0173.12	KOG0371.5	KOG0741.9	KOG1098.4	KOG1487.5	KOG1746.10	KOG2415.6	KOG2854.2	KOG3284.1
KOG0174.3	KOG0372.1	KOG0756.6	KOG1099.7	KOG1491.5	KOG1750.1	KOG2451.3	KOG2855.15	KOG3285.1
KOG0175.2	KOG0373.5	KOG0758.5	KOG1112.6	KOG1494.4	KOG1753.6	KOG2467.5	KOG2874.5	KOG3291.11
KOG0176.3	KOG0376.12	KOG0767.1	KOG1123.9	KOG1498.1	KOG1754.3	KOG2472.2	KOG2877.12	KOG3295.1
KOG0177.9	KOG0394.6	KOG0780.3	KOG1131.6	KOG1506.4	KOG1755.3	KOG2481.5	KOG2906.2	KOG3297.6
KOG0179.4	KOG0397.5	KOG0784.1	KOG1137.9	KOG1523.1	KOG1758.1	KOG2509.11	KOG2908.8	KOG3301.5
KOG0180.5	KOG0400.1	KOG0785.9	KOG1145.8	KOG1526.7	KOG1760.9	KOG2519.5	KOG2909.4	KOG3311.1
KOG0181.7	KOG0402.4	KOG0787.11	KOG1148.8	KOG1531.3	KOG1762.3	KOG2529.5	KOG2916.7	KOG3313.4
KOG0182.3	KOG0407.5	KOG0815.5	KOG1149.13	KOG1532.4	KOG1769.5	KOG2531.1	KOG2930.4	KOG3318.1
KOG0183.7	KOG0418.9	KOG0817.7	KOG1158.6	KOG1533.4	KOG1770.4	KOG2535.6	KOG2948.9	KOG3320.9
KOG0184.11	KOG0419.16	KOG0820.7	KOG1159.11	KOG1534.2	KOG1772.9	KOG2537.2	KOG2952.4	KOG3330.3
KOG0185.10	KOG0420.17	KOG0829.5	KOG1185.2	KOG1535.6	KOG1774.1	KOG2555.9	KOG2957.6	KOG3343.1
KOG0188.2	KOG0424.4	KOG0852.2	KOG1211.12	KOG1539.23	KOG1779.3	KOG2572.3	KOG2967.2	KOG3361.2
KOG0190.16	KOG0434.3	KOG0853.7	KOG1235.9	KOG1540.8	KOG1780.2	KOG2574.2	KOG2971.1	KOG3372.7
KOG0209.28	KOG0441.4	KOG0857.12	KOG1241.15	KOG1541.9	KOG1781.12	KOG2575.2	KOG2981.2	KOG3387.9
KOG0211.3	KOG0450.14	KOG0861.3	KOG1255.13	KOG1549.2	KOG1782.6	KOG2613.1	KOG3013.8	KOG3400.8
KOG0225.4	KOG0460.6	KOG0862.6	KOG1268.5	KOG1555.5	KOG1784.1	KOG2617.3	KOG3022.3	KOG3404.6
KOG0233.1	KOG0462.5	KOG0871.6	KOG1272.9	KOG1556.4	KOG1800.8	KOG2623.24	KOG3031.6	KOG3405.3
KOG0258.4	KOG0466.8	KOG0876.2	KOG1291.5	KOG1562.10	KOG1816.4	KOG2636.5	KOG3049.3	KOG3411.2

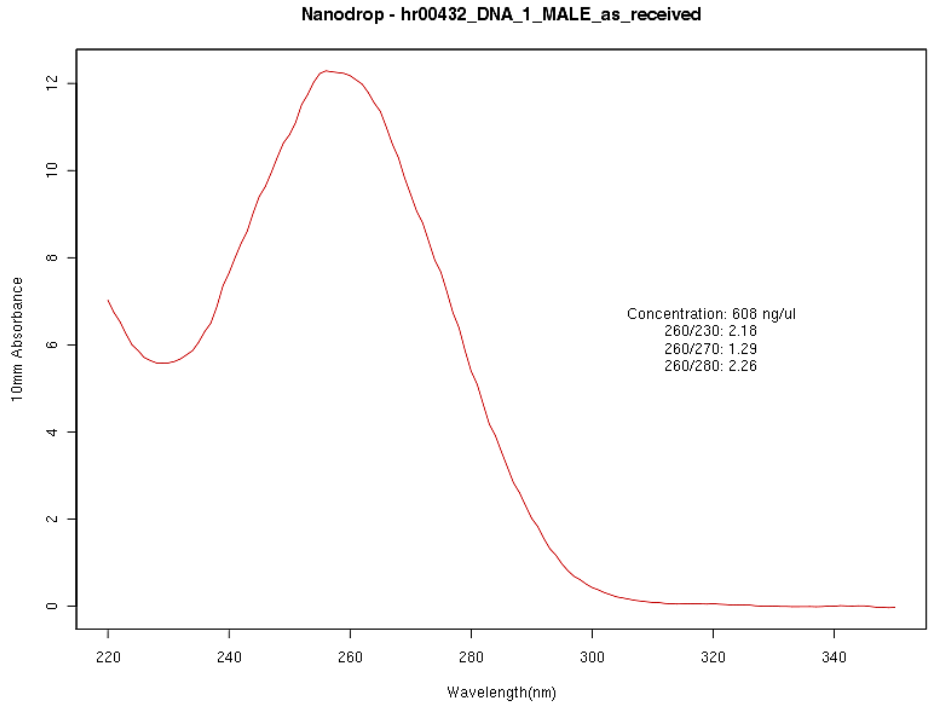
KOG0264.2	KOG0468.3	KOG0878.6	KOG1299.6	KOG1566.10	KOG1872.12	KOG2638.2	KOG3052.1	KOG3418.8
KOG0271.16	KOG0469.6	KOG0880.8	KOG1300.3	KOG1567.5	KOG1885.7	KOG2653.14	KOG3064.3	KOG3428.10
KOG0276.8	KOG0477.7	KOG0888.1	KOG1301.3	KOG1568.4	KOG1889.12	KOG2670.2	KOG3079.6	KOG3430.6
KOG0279.11	KOG0481.3	KOG0894.5	KOG1322.4	KOG1596.1	KOG1915.39	KOG2680.3	KOG3090.2	KOG3432.4
KOG0285.5	KOG0495.1	KOG0898.1	KOG1335.3	KOG1597.5	KOG1936.6	KOG2700.9	KOG3106.2	KOG3436.2
KOG0289.18	KOG0523.9	KOG0922.2	KOG1342.34	KOG1626.2	KOG1942.1	KOG2703.1	KOG3147.9	KOG3442.7
KOG0291.27	KOG0524.4	KOG0927.8	KOG1349.1	KOG1636.1	KOG1979.4	KOG2707.3	KOG3149.3	KOG3448.9
KOG0292.3	KOG0530.9	KOG0934.3	KOG1350.3	KOG1637.9	KOG1980.20	KOG2711.7	KOG3157.4	KOG3449.3
KOG0302.1	KOG0534.1	KOG0935.5	KOG1351.5	KOG1641.2	KOG1986.7	KOG2719.3	KOG3163.6	KOG3453.14
KOG0313.9	KOG0544.3	KOG0937.4	KOG1353.2	KOG1643.1	KOG1992.18	KOG2726.4	KOG3164.2	KOG3457.1
KOG0318.20	KOG0556.5	KOG0938.6	KOG1355.1	KOG1644.10	KOG2004.8	KOG2728.3	KOG3167.4	KOG3459.7
KOG3463.7	KOG3475.1	KOG3480.1	KOG3489.6	KOG3497.2	KOG3499.4	KOG3503.6	KOG3855.8	KOG3974.3
KOG3464.4	KOG3479.4	KOG3482.2	KOG3493.14	KOG3498.1	KOG3502.2	KOG3506.2	KOG3954.1	KOG4392.1
KOG4655.6								

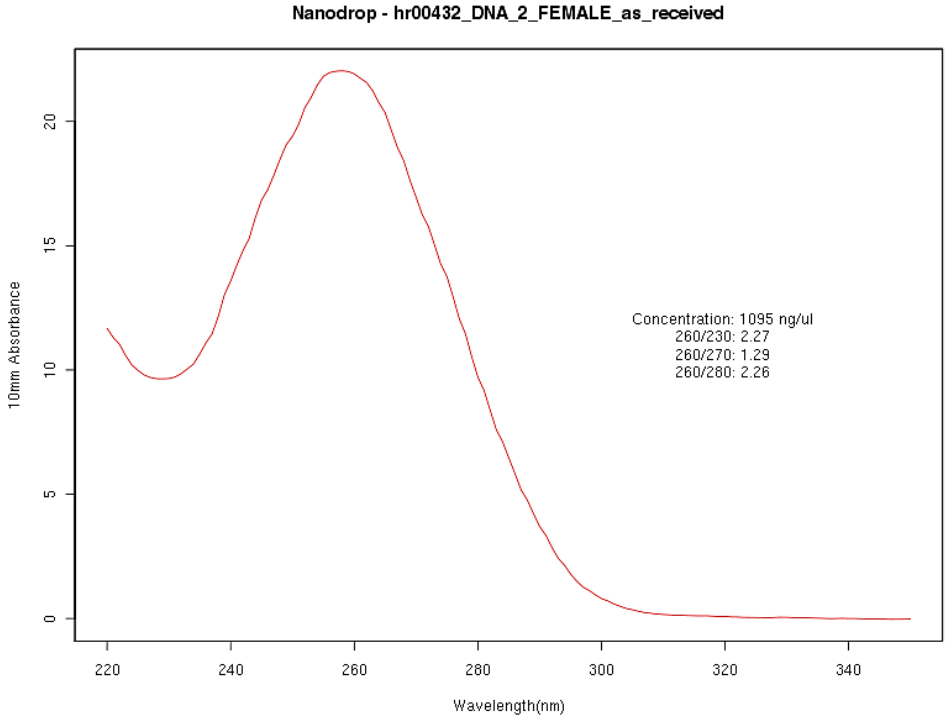
Appendix C Purdue University Genomics Core Facility - Protocol followed in Genomic DNA Library Preparation

The overall protocol used was based on the Illumina TruSeq DNA sample preparation guide (Catalog #PE-940-2001, Part # 15005180 Rev. A, November 2010), which can easily be found on the Illumina website (www.illumina.com). It was mostly followed stepwise; however the amplicons were not gel purified and only 6 cycles of amplification were done. The following is some of the details of the preparation steps from the Purdue Univ. genomics core facility.

Request Data	**Libraries**
Analysis Type: Denovo Assembly Genome **Library Type:** Paired End **Number of Lanes:** 1 **Read Length #1**: 100 **Read Length #2:** 100 **Request Name:** Phormia_regina **Sequence Engine:** hiseq2000 **Source Genomic:** DNA **Source Test:** 2 **Species:** Phormia regina **Version:** v3	**Sample # 1** **Library Name** - 1- MALE **Accession#** - 002843 **Sample Type** - DNA **Amount** - 1/2 lane **Species** - Phormia regina **Control** - No **Sample # 2** **Library Name** - 2- FEMALE **Accession#** - 002844 **Sample Type** - DNA **Amount** - 1/2 lane **Species** - Phormia regina **Control** - No

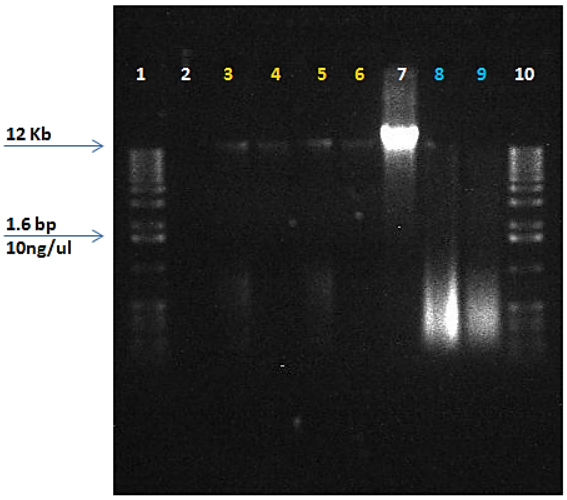
NANODROP


Nanodrop - hr00432_DNA_1_MALE_as_received
10mm Absorbance
Wavelength(nm)
Concentration: 608 ng/ul
260/230: 2.18
260/270: 1.29
260/280: 2.26


Nanodrop - hr00432_DNA_2_FEMALE_as_received
10mm Absorbance
Wavelength(nm)
Concentration: 1095 ng/ul
260/230: 2.27
260/270: 1.29
260/280: 2.26

Sample ID	ng/ul	260/280	260/230	ul TV
1 hr00432_1-MALE gDNA	608.83	2.26	2.18	80
2 hr00432_2-FEMALE gDNA	1105.78	2.26	2.278	74

GEL

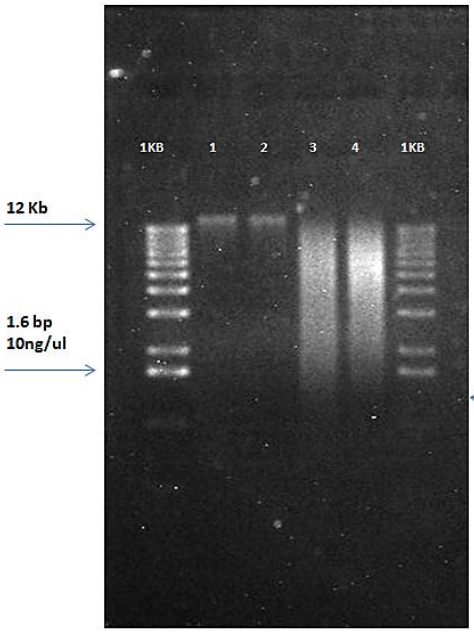


0.8% LMP agarose gel
40min
100V

Cresol Red 6x	2ul
Water	2ul
DNA sample	1ul
1Kb 10x	2ul

Lanes:
1-1Kb
2-empty
3-4296 before Riboshredder and Zymo clean up
4-4296 after Riboshredder and Zymo clean up
5-4295 before Riboshredder and Zymo clean up
6-4295 after Riboshredder and Zymo clean up
7-hr 435 Soybean DNA 3ul out of 37ul
8-hr 00432 FEMALE
9-hr00432 MALE
10-1Kb

12.20.12



0.8% LMP agarose gel
64 min
100V

Cresol Red 6x	2ul
Water	9ul
DNA sample	1ul
1Kb 10x	2ul

Lanes:
1-hr431 F8-4296
2-hr431 F8-4295
3-hr 00432 FEMALE
4-hr00432 MALE

All DNA samples treated with Riboshredder, then purified and Zymo clean up Total volume eluted- 10ul

Tubes are in -80C box hr432

AGILENT

Library Construction

1-14-2013

Library sheared using 400bp shearing protocol on Covaris
Duty Cycle 10%
Intensity 4
Cycles/burst 200
Time (sec) 55

Amounts sheared listed below.

PI	HR	Acc	Sample Name	ng/ul	Total V	Total ng	Shear ul	uL EB
cpicard	432	2843	1	608	80.0	48640.0	10.0	120.0
		2844	2	1105.8	74.0	81829.2	10.0	120.0

After shearing samples transferred to 1.5ml low binds tubes

0.8% Ampure XP on all samples.
130ul sample + 104ul of beads
15min RT
15 min RT on magnet, discard supernatant
200ul 80% EtOH wash, discard supernatant
200ul 80% EtOH wash, discard supernatant
Dry at RT for 15 mins
Re-suspend in 30ul RSB- Remove and save 28 ul.

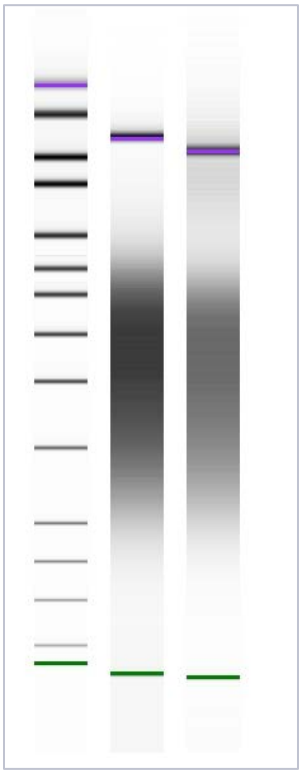
Samples measured on Nanodrop.

1-15-2013

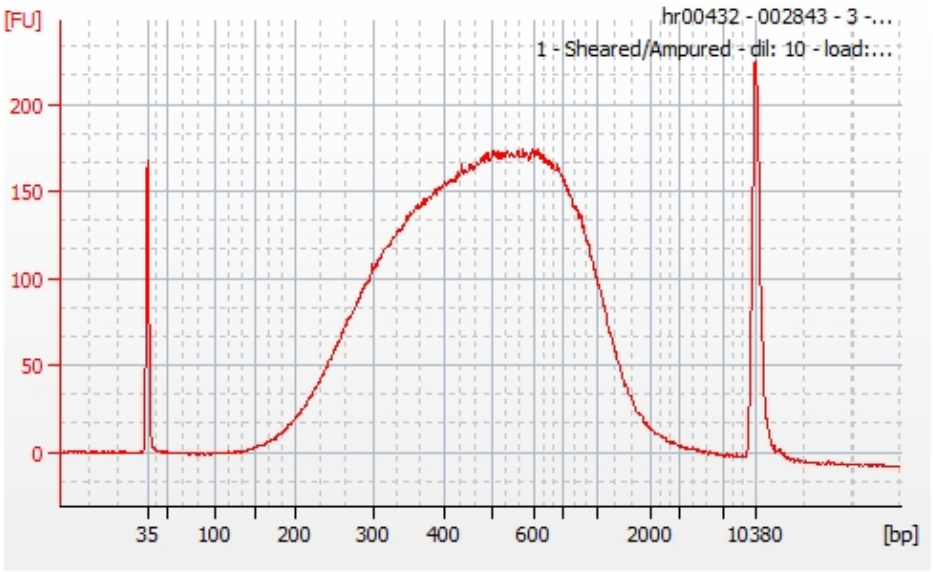
High sensitivity chip ran to check shearing products.

Chip ID 287 or 288.

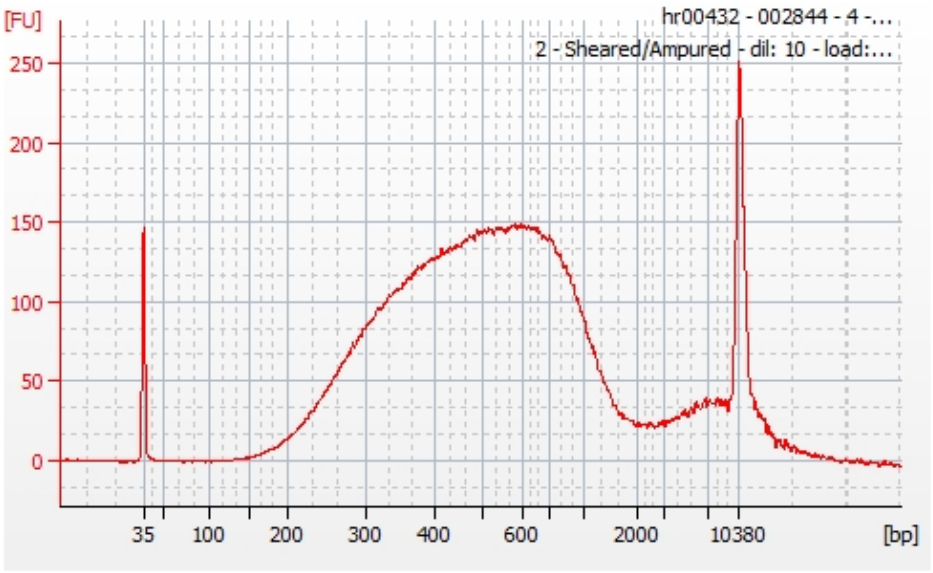
DNA High Sensitivity Chip run by Allison Sorg on 2013-01-15 at 10:39:57
Faux Gel Images



Electropherograms



Lane	Accession	Version	Library Name	Dilution	ul Load	Total Volume	RIN	Concentration
3	002843	1	1	10	1	28		



Lane	Accession	Version	Library Name	Dilution	ul Load	Total Volume	RIN	Concentration
4	002844	1	2	10	1	28		

Concentrations determined after shearing (want 500ng/ul of input DNA for procedure):

PI	HR	Acc	Sample Name	ng/ul	Amount need	EB added
cpicard	432	2843	1	26.0	19.2	10.8
		2844	2	16.0	31.3	3.0

** For sample 2844 only 27ul available after the shearing ampure so all of sheared product was used. **

End repair of sheared DNA.

Added 5ul of RSB to each sample because we are not using End Repair Control.
Added 20ul End Repair Mix.
Thermal cycler, 30^0C for 30 mins.

0.8% Ampure XP on all samples.

~50 sample + 40ul of beads
15min RT
15 min RT on magnet, discard supernatant
200ul 80% EtOH wash, discard supernatant
200ul 80% EtOH wash, discard supernatant
Dry at RT for 15 mins
Re-suspend in 9ul RSB- Remove and save 7.5 ul.

A-Tailing

Add 1.25ul RSB to each sample because no A-Tailing Control.
Add 6.25ul A-Tailing Mix to each sample.
Thermal cycler: 37^0C for 30 mins.
70^0C for 5 mins.

Ligated Adapters

Added 1.25ul of RSB because no Ligation Control.
Added 1.25ul of Ligation Mix.
Added 1.25ul of Adapter. Adapter assignments listed below.

PI	HR	Acc	Sample Name	Barcode
cpicard	432	2843	1	AD013
		2844	2	AD014

Thermal cycler: 30^0C for 10 mins.

Added 2.5ul Stop Ligation Mix and mixed well.

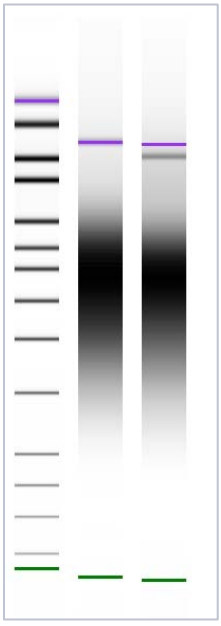
Double Ampure XP

Full Ampure- 21.25ul sample + 21.25ul Ampure XP beads.
Mix 10X.
RT for 15 mins.
RT 5 mins on magnet.
Remove and discard supernatant.
200ul 80% EtOH wash, discard supernatant.
200ul 80% EtOH wash, discard supernatant.
Dry at RT for 15 mins.
Re-suspend in 26.5 ul RSB- Remove and save 25 ul.
0.8% Ampure- 25ul sample + 20ul Ampure XP beads.
Mix 10X.
RT for 15 mins.
RT 5 mins on magnet.
Remove and discard supernatant.
200ul 80% EtOH wash, discard supernatant.
200ul 80% EtOH wash, discard supernatant.
Dry at RT for 15 mins.
Re-suspend in 22.5 ul RSB- Remove and save 20 ul.

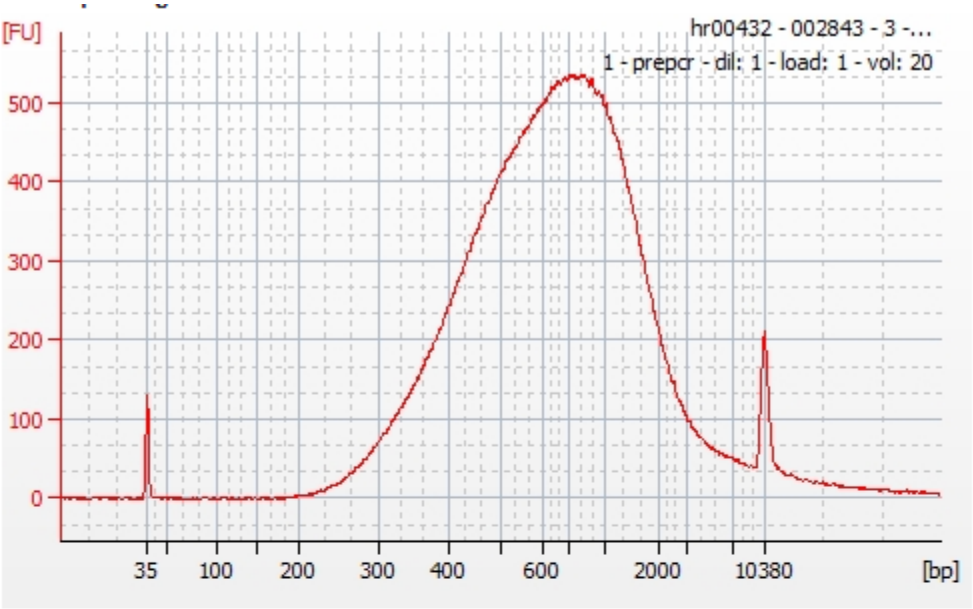
Ran High Sensitivity Chip.

Chip ID 291

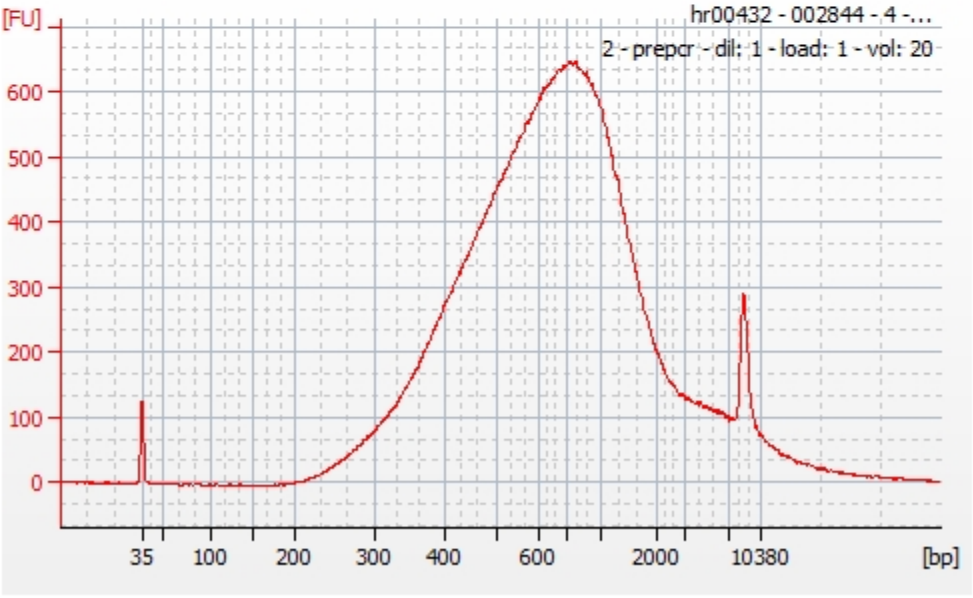
DNA High Sensitivity Chip run by Allison Sorg on 2013-01-15 at 18:01:26
Faux Gel Images



Electropherograms



Lane	Accession	Version	Library Name	Dilution	ul Load	Total Volume	RIN	Concentration
3	002843	1	1	1	1	20		



Lane	Accession	Version	Library Name	Dilution	ul Load	Total Volume	RIN	Concentration
4	002844	1	2	1	1	20		

Samples frozen at -20 until 1-17-2013.

1-17-2013

PCR Reaction

Added 2.5ul PCR Primer Cocktail
Added 12.5ul PCR Master Mix.
Mix 10X.
Thermal cycler: 6 cycles
98^0C for 10 secs
60^0C for 30 secs
72^0C for 30 secs
72^0C for 5 mins
Hold at 10^0C

0.8% Ampure XP

0.8% Ampure- 25ul sample + 20ul Ampure beads.
Mix 10X.
RT for 15 mins.
RT 5 mins on magnet.
Remove and discard supernatant.
200ul 80% EtOH wash, discard supernatant.
200ul 80% EtOH wash, discard supernatant.
Dry at RT for 15 mins.
Re-suspend in 32 ul RSB- Remove and save 30 ul.

1-18-2013

High sensitivity chip to check library construction/PCR prior to QPCR.

Appendix D Permission from Elsevier Publishing Company

ELSEVIER LICENSE
TERMS AND CONDITIONS

Jun 03, 2014

This is a License Agreement between Anne Andere ("You") and Elsevier ("Elsevier") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Elsevier, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

Supplier	Elsevier Limited The Boulevard,Langford Lane Kidlington,Oxford,OX5 1GB,UK
Registered Company Number	1982084
Customer name	Anne Andere
Customer address	723 W Michigan St, SL 306 INDIANAPOLIS, IN 46202
License number	3401460460720
License date	Jun 03, 2014
Licensed content publisher	Elsevier
Licensed content publication	Trends in Genetics
Licensed content title	Bioinformatics challenges of new sequencing technology
Licensed content author	Mihai Pop,Steven L. Salzberg
Licensed content date	March 2008
Licensed content volume number	24
Licensed content issue number	3
Number of pages	8
Start Page	142
End Page	149
Type of Use	reuse in a thesis/dissertation
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
Format	both print and electronic
Are you the author of this Elsevier article?	No
Will you be translating?	No
Title of your thesis/dissertation	De novo Genome Assembly of the Blow Fly Phormia regina (Diptera: Calliphoridae)
Expected completion date	Aug 2014
Estimated size (number of pages)	105
Elsevier VAT number	GB 494 6272 12
Permissions price	0.00 USD
VAT/Local Sales Tax	0.00 USD / 0.00 GBP
Total	0.00 USD
Terms and Conditions	