Haploid or diploid DNA? A comparative analysis of methods for *de novo* assembly of hymenopteran genomes

Tal Yahav

Thesis submitted in partial fulfillment of the requirements for the master's degree

University of Haifa

Faculty of Natural Sciences

Department of Evolutionary and Environmental Biology

November, 2017

Haploid or diploid DNA? A comparative analysis of methods for *de novo* assembly of hymenopteran genomes

By: Mr. Tal Yahav

Supervisor: Dr. Eyal Privman

Thesis submitted in partial fulfillment of the requirements for the master's degree

University of Haifa

Faculty of Natural Sciences

Department of Evolutionary and Environmental Biology

November, 2017

Approved by: _____ Date: _____
(Supervisor)

Approved by: _____ Date: _____
(Chairperson of Master's studies committee)

Table of contents:

Abstract ..... **v**

List of tables ..... **vii**

List of figures ..... **viii**

1. Introduction **1**

1.1 Motivation for *de novo* sequencing and assembling whole genomes 1

1.2 Genomic and transcriptomic sequencing 2

1.3 Genome Assembly 5

1.4 Challenges and solutions in *de novo* assembling whole genomes 7

1.4.1 Challenges in *de novo* assembly due to genome structure 7

1.4.2 Experimental design using NGS 8

1.5 The *haplodiploid* life of Hymenoptera 9

1.6 The model organism 9

1.7 Transcriptome assembly i

1.8 Genome assembly quality assessment 10

1.9 Gene annotation 11

2. Research objective **13**

3. Materials and Methods **15**

3.1 Samples collection 15

3.2 Samples preparation and extraction 15

3.2.1 Extraction of male DNA and RNA 15

3.2.2 Extraction of worker DNA 17

3.2.3 Preparation and extraction of RNA pool 17

3.3 Sequencing 18

3.4 Pre-assembly quality control 18

3.5 DNA and RNA coverage reduction 19

3.6 Genome assembly 19

3.7 Assembly quality analyses 20

3.8 RNA mapping and transcriptome assembly 20

3.8.1 *De novo* transcriptome assembly 20
3.8.2 Transcriptome mapping to genome 20
3.9 Genome annotation 21
3.10 Extracting High Molecular Weight (HMW) DNA from *C. drusus* ants 22
3.10.1 HMW DNA role in the assembly 22
3.10.2 Extraction of worker DNA 22

4. Results 24
4.1 DNA and RNA sequencing 24
4.2 Haploidy confirmation 25
4.3 Quality analyses of *de novo* genome assemblies 25
4.3.1 Contiguity of the assemblies 25
4.3.2 Completeness of the assemblies 26
4.3.3 Contig and scaffold misassemblies 27
4.4 Quality analyses of *de novo* transcriptome assemblies 28
4.5 Transcriptome mapping 28
4.5.1 Completeness of the mapped transcripts 29
4.5.2 Polymorphism 34
4.5.3 Alternative splicing 37

5. Discussion 39
5.1 Genome assembly 39
5.2 Transcripts to genome mapping 40
5.2.1 Male and pool transcripts 40
5.2.2 Complexity of the pool sample 41

6. Conclusion 42

References 43

Appendices 47

Appendix 1: a satellite image of the research area in Betzet beach (33°4'40.88"N / 35°6'33.97"E; Google earth). Red dot marks nest BZT4B, from which the male and worker sample for the reference genome were taken. 47

Appendix 2: RNA pool composition with quantity and purity measurements using ND200. Total volume of pool sample for sequencing was 60µl. only four samples were diluted

(Larva_1, Larva_2, Larva_4, Larve_5). 260/280 ratio < 2 indicate protein residue contamination. 260/230 ratio < 2.2 indicates chemical contamination. 48

Appendix 3: All Prep Mini DNA and RNA modified protocol for *Cataglyphis* ants 49

Appendix 4: DNA and RNA ND2000 measurements results for the male candidate samples and Worker DNA measurements 51

Appendix 5: List of commands used in the various pipelines in the genome and transcriptome assembly as well as quality assessment..... 52

Appendix 6: Species homolog proteins used for MAKER annotation 53

Appendix 7: N50 of various *Camponotus* species for comparison to *Cataglyphis drusus de novo* transcriptome 53

Haploid or diploid DNA? A comparative analysis of methods for *de novo* assembly of hymenopteran genomes

Tal Yahav

Abstract

Whole genome *de novo* assembly is a crucial infrastructure for a wide range of genetic studies. Hymenoptera genome projects can make use of the haplodiploidy sex determination system, where females are diploid and males are haploid. Several hymenopteran genomes used DNA from a single haploid male sample that was assumed advantageous for genome assembly (e.g. *Acromyrmex echinatior* and *Solenopsis invicta*). For the purpose of gene annotation, full transcriptome sequencing is usually conducted using RNA from a pool of individuals. The present thesis is a comparative analysis of genome and transcriptome assembly, and annotation methods, using genetic sources of different ploidy: (1) for the DNA a haploid male or a diploid female (2) RNA from the same haploid male or a pool of individuals. Our approach is unique in that the core haploid genetic data, both DNA and RNA were derived from a single male. The diploid DNA sample was derived from a worker sister of the male. The working assumption is that the use of a haploid male as opposed to a diploid female, and the extraction of both RNA and DNA from the same male individual simplify the genome assembly and gene annotation thanks to the lack of heterozygosity. Pairing the source of the sequenced DNA and RNA is expected to provide more confidence in transcript-to-genome alignment, and ease the annotation of gene structure in terms of the exon/intron boundaries. This novel approach takes advantage a unique genomic characteristic of Hymenoptera, namely haplodiploidy.

Genome assemblies of the haploid and diploid samples were built by the assemblers SPAdes and SOAPdenovo2. Three quality assessment methods were used to compare the alternative assemblies: (1) calculating the N50 statistic for contigs and scaffolds; (2) evaluating the completeness of a conserved gene set in the assemblies; and (3) detecting misassemblies. For both assemblers, the haploid genome assemblies proved to be more contiguous, with both contig and scaffold N50 size at least threefold greater than their diploid counterparts. Completeness evaluation showed mixed results between the assemblers with the SPAdes haploid assembly having more complete genes, but a higher level of duplicates, and a greatly overestimated genome size.

Misassemblies detection showed mixed results with SOAPdenovo2 assemblies showing many more local misassemblies. *De novo* assembly of transcriptomes gave better N50 results as well as a more complete gene set for the pool transcriptome relative to the male. Lastly, when aligning the two transcriptomes against the male genome, the male transcriptome gave a more complete gene annotation. A possible explanation for this result can be the higher complexity of the pool sample due to polymorphism, alternative splicing and RNA editing events, which interfere with transcript assembly. In conclusion, the use of a haploid source material for *de novo* genome assembly provides a substantial advantage to the quality of the genome draft and the use of RNA from the same haploid individual for transcriptome to genome alignment produces a more complete gene annotation and predicted transcripts.

List of Tables

Table 1: Coverage reduction of DNA and RNA samples before assembly..... 19

Table 2: Illumina libraries constructed. Two genomic DNA libraries (300, 550bp) constructed from each of the source materials. For RNA sequencing, one library for each source was constructed. 24

Table 3: Microsatellites used for ploidy test of the male samples. 25

Table 4: N50 contig and scaffold sizes for the different genome assemblies 26

Table 5: BUSCO completeness results for the different genome assemblies against the Arthropoda (a) and Eukaryota (b) *odb9* datasets 26-27

Table 6: Misassemblies identification by QUAST for the different assemblies 27

Table 7: N50 results of two de novo transcriptome assemblies 28

Table 8: BUSCO completeness results for the transcriptome assemblies against the Eukaryota *odb9* dataset 28

Table 9: Completeness results done by BUSCO against Eukaryota *odb9* datasets 30

List of Figures

Figure 1: The decrease in cost per Mbp of DNA sequence between the years 2001-2015 4

Figure 2: an illustration of the comparisons for the transcript mapping to the haploid genome assembly 13

Figure 3: An example for a gene classified by BUSCO as complete in the male and fragmented in the pool. BUSCO gene 'EOG09370DXT'. The coverage data range is normalized to a range of 0-2500 reads per position. **..... 31**

Figure 4: An example for a gene classified by BUSCO as complete in the pool and fragmented in the male. BUSCO gene 'EOG093706PM'. The coverage data range is normalized to a range of 0-1000 reads per position**..... 32**

Figure 5: An example for a gene classified by BUSCO as complete in the male and missing in the pool. BUSCO gene 'EOG09370WZX'. The coverage data range is normalized to a range of 0-5000 reads per position**. 33**

Figure 6: An example for a gene classified by BUSCO as missing in the male and complete in the pool. BUSCO gene 'EOG09370MQ0'. The coverage data range is normalized to a range of 0-1000 reads per position. Black rectangles marks the BUSCO gene examined. Orange rectangle marks the 3' end exons overlapping the BUSCO gene. **..... 34**

Figure 7: An example of a SNP in the pool transcriptome. The male genome and transcriptome both have a G at this position, while the pool RNAseq reads, have either A or G. The black rectangles highlight multiple additional SNPs. **..... 35**

Figure 8: An example of alternative splicing in the pool transcriptome, but not the male. Curved arches (blue) below the center-line represent splice junctions on the negative strand of gene 'EOG093710JH'. Coverage is normalized to 0-2000 reads per position (a). Visualization of splice junctions using IGV Sashimi plots of male and pool transcripts of same gene (b). All the splice junctions are of on the negative strand of gene 'EOG093710JH'. Arcs represent splicing events. In orange circles are the number of reads splits across the splice junction. Height of bars between arcs represents exon coverage (reads per position**). 37**

1. Introduction

1.1 Motivation for *de novo* sequencing and assembling whole genomes

Whole genome *de novo* assembly is a crucial component in various types of genetic research. It is a starting point and the foundation for the development of genetic resources such as gene annotation, high resolution maps of polymorphism, genomic structural variation, etc., which can later be utilized for a wide range of applications and studies in fields including biomedicine, agriculture, biotechnology, molecular ecology, and evolutionary biology (Church et al. 2011).

Ideally, a fully sequenced genome, with long contiguous genomic segments anchored to full-length chromosomes should be produced. In order to produce such high-quality reference genomes, multiple advanced technologies are usually used in twain, a matter that requires a substantial funding and complex analysis. Naturally, greater funding and efforts are invested in high priority topics such as human healthcare, agriculture, industry etc. Therefore, genomes of model organisms, such as mouse (Waterston et al. 2002), fruit fly (Adams et al. 2000; Myers et al. 2000), as well as the human genome (Lander et al. 2001; Venter et al. 2001), are well studied and have well developed genomic infrastructures. Conversely, most non-model organisms are lacking such infrastructures. When starting a research project on a non-model organism a reference genome typically does not exist, requiring *de novo* sequencing and assembling the genome from scratch. Whole genome assembly provides a broad overview of an organism's genetic makeup and allows the researcher to conduct analyses and experiments on multiple levels from the structure of entire chromosomes to single point mutations in specific genes.

The requirements from the reference genome may differ by the type of study being conducted. The main difference is usually in the level of fragmentation of the assembled genome that can be tolerated. A fragmented assembly is less challenging and expensive, while a highly contiguous assembly requires substantial investment. On one end of the spectrum, for example, are Genome Wide Association Studies (GWAS) and Quantitative Trait Loci (QTL) mapping. GWAS and QTL mapping studies are commonly used in agriculture and biomedical research. In agriculture, QTL mapping can identify genetic variations in domesticated plant and animal populations that affect a certain trait of interest. Studies such as breeding assays to improve yield in farm animals (Soller et al. 2006; Lee et al. 2015) as well as the development of resistant plant strains, which can

withstand a larger array of parasites, diseases and extreme environmental conditions, are very common (Rispail et al. 2007). Among organisms with well-developed genomic infrastructures are the chicken (Hillier et al. 2004), cow (The Bovine Genome Sequencing and Analysis Consortium; Elsik et al. 2009), rice (Goff et al. 2002), and bread wheat (Brenchley et al. 2012). In these studies, genetic variants are examined in relation to the scoring of a certain quantitative phenotype or trait. Often such traits are polygenic, that is, they are affected by multiple polymorphic loci in multiple positions in the genome. Loci are genotyped in multiple individuals, and statistical analyses are applied to infer the relationship between genotypes and phenotypes in the population. For example, in biomedical research, GWAS identify associations between variations in certain genes and disease risk, thereby providing a broad overview of the genetic basis of complex diseases such as cancer, diabetes, and other diseases with a hereditary component (Chang et al. 2013; Sud et al. 2017). Mapping of the exact locus within a chromosome that functions as a QTL requires an accurate and contiguous reference genome. On the opposite end of the spectrum are, for example, studies that use differential gene expression analysis, or evolutionary studies that infer positive natural selection on specific genes (Warner et al. 2017). In studies such as these, each gene is analyzed independently from its neighbors, so chromosome-level completeness of the assembly is less important.

1.2 Genomic and transcriptomic sequencing

The last decade saw a rapid advancement and development of novel technologies for high throughput DNA sequencing. These Next Generation Sequencing (NGS) platforms became a common and powerful tool for an array of existing genomic, transcriptomic, and epigenomic applications, as well as multiple innovative methodologies and approaches.

The first generation of DNA sequencing technology, known as the chain termination method or Sanger sequencing, generated reads up to 800 base pairs (bp) long. This method allows the production of high fidelity sequence with very few sequencing errors. In Sanger sequencing the subject, double stranded DNA is first thermally desaturated to form a single stranded template DNA. Later DNA primers are attached to the template strand to initiate the sequencing reaction. A DNA polymerase adds deoxynucleosidetriphosphates (dNTPs) to the growing strand. The reaction mix

also contains modified di-deoxynucleotidetriphosphates (ddNTPs), lacking a 3'-OH group, which result in termination of the DNA strand elongation. These modified nucleotides are fluorescently labeled, each base with a different color. At the end of the sequencing process, the products are loaded into a capillary electrophoresis system for size separation. The fluorescent emission of the ddNTPs are read and recorded with an optical detector. The result is a chromatogram, in each colored peak represents one base position in the sequenced genomic fragment.

Among all NGS technologies, Illumina is by far the most popular short reads sequencing technology used in recent years. Illumina uses a method called "sequencing by synthesis". Each DNA fragment is amplified in a Polymerase Chain Reaction (PCR) while attached to the surface of a flow-cell. This allows spatial separation between millions to billions of distinct clusters of DNA fragments in the same flow-cell. During the amplification, fluorescently marked nucleotides are introduced in discrete cycles, which are read optically by a high-resolution camera in real time for each cluster that originated from a single DNA fragment. Illumina sequencers typically produce reads of up to 150bp, after which the base calling accuracy drops dramatically. Paired-end sequencing allows sequencing a 150bp read from each end of a DNA fragment of up to 800bp long. Prior to the actual sequencing process, libraries can be constructed according to various protocols. For whole transcriptome sequencing (RNAseq), RNA molecules are reverse transcribed and sequenced as complementary DNA (cDNA) using the same method used for genomic DNA. Sequencing of pairs of reads with longer insert sizes is highly useful for increased contiguity of a genome assembly. Thus, several long-insert Illumina sequencing protocols were developed (mate-pair, LJD, etc.), which can produce a pair of reads from the ends of fragments up to 20Kbp long. These pairs are used in the assembly process to scaffold short genomic fragments into longer fragments.

A different category of technologies allows the sequencing of longer sequence reads. The benefits of longer reads come to place in the bridging of gaps in the assembly and achieving better contiguity. One example is the Single Molecule Real Time (SMRT) sequencing technology developed by Pacific Biosciences (PacBio), which can visualize single DNA molecules as they are elongated by the polymerase. A nano photonic visualization chamber called a Zero Mode Waveguide (ZMW), containing attached DNA polymerase molecule at its bottom is the main engine of the sequencing process. During the polymerase synthesis action, light is emitted from the labeled

nucleotides and captured by a high detector. Each SMRT cell is populated by up to 1 million ZMWs, which can produce high throughput sequencing data. Thereby, PacBio can sequence much longer DNA reads than Illumina, with median read length greater than 20Kbp. The disadvantages of PacBio and other long read technologies are that they are far more prone to sequencing errors and are still much more expensive in terms of the cost per base pair.

This ongoing technological revolution over the last ten years resulted in a dramatic reduction in cost for sequencing projects. Comparing NGS to first generation sequencing technologies, we see a reduction of the cost per base pair by five orders of magnitude (Figure 1; https://www.genome.gov/27565109/the-cost-of-sequencing-a-human-genome/), thus, making these genomic tools increasingly available to a variety of researchers, labs and facilities.

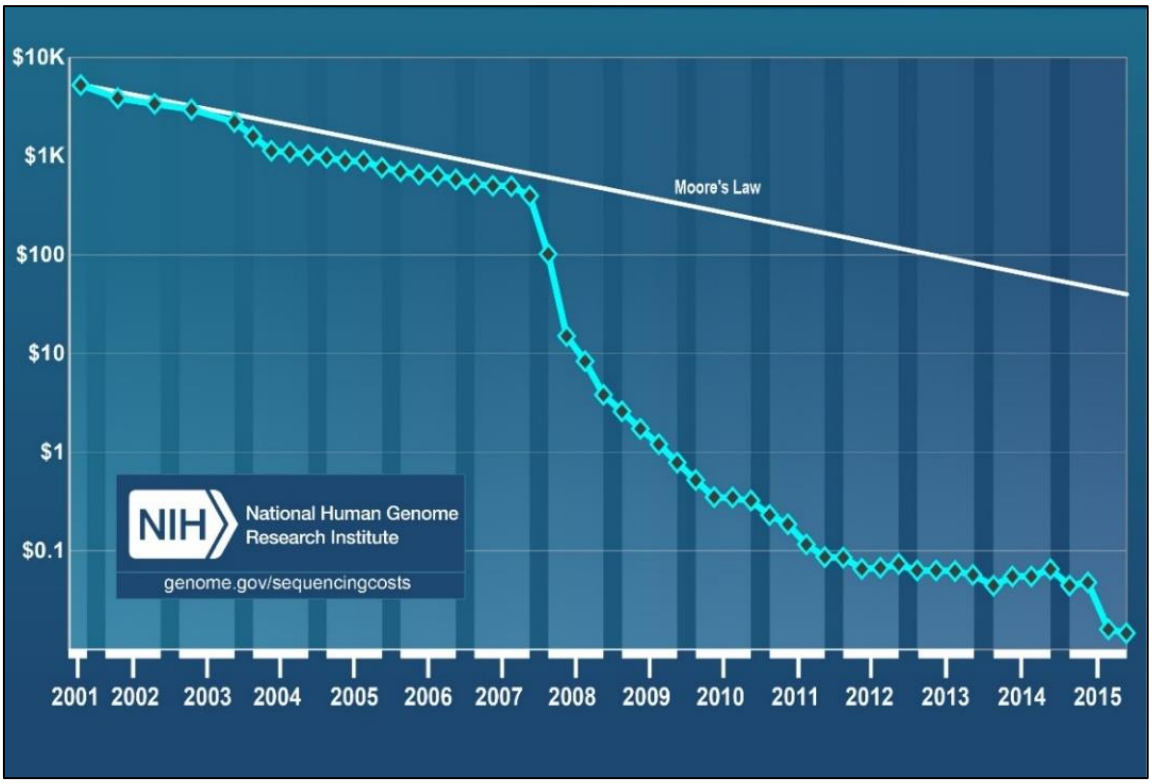

Figure 1: The decrease in cost per Mbp of DNA sequence between the years 2001-2015

1.3 Genome assembly

In spite of the advancements in sequencing technologies, they are still limited to short read lengths relative to the length of full chromosomes. Human chromosomes are between 50 and 250Mbp, and while some bacterial chromosomes are only a few Mbp long, no existing technology can sequence them in one read. Therefore, the challenge of genome assembly is to reconstruct as accurate and contiguous genome as possible from the data generated as short sequencing reads. In practice, assemblers use different statistical models to combine reads to form contiguous fragments called *contigs*. One common approach for genome assembly, used mainly in era of Sanger sequencing, is the Overlap Layout Consensus (OLC) graph method. In OLC, all the reads are scanned for overlapping one-size words of *k* nucleotides, referred as *k*-mers, followed by a construction of an overlay graph and eventually Multiple Sequence Alignment (MSA) according to the graph to create a consensus sequence. This method of all against all comparison for overlay discovery was feasible while the sequenced data contained relatively longer reads of several hundreds of base pairs. In addition, the number of reads assembled was significantly lower (Miller et al. 2010).

The rapid exponential growth in the amount of genomic data produced in short reads sequencing projects over the past decade required the development of new computational techniques. To this end, a multitude of computational approaches, algorithms, and software implementations were developed. Most short read assemblers use a *de-Bruijn* graph. Reads are broken down into overlapping *k*-mers. Unique *k*-mers are represented as nodes in the graph, and an edge connects every two *k*-mers that overlap each other with a shift of one base in some sequenced read. An Eulerian walk-through of the graph can provide a constructed genome. One major challenge lies in coping with repeated sequences in the genome. Sequence repeats create multiple walkthrough options. The assembler tries to decide which paths are the correct ones to form a consensus assembly, but often fails to extend contigs through repetitive sequences longer than the read length (Sims et al. 2014). After the construction of contigs, assemblers utilize data produced by long insert/reads protocols such as the Illumina mate-pair protocol or long PacBio reads to spatially associate and order different contigs to create larger fragments in a process called *scaffolding* (Simpson and Pop 2015).

In the present study, two commonly used assemblers were employed: SOAPdenovo2 (version r240; Luo et al. 2012) and SPAdes (version 3.9.1; Bankevich et al. 2012). Although both

SOAPdenovo2 and SPAdes rely on a *de-Bruijn* graph, they differ in their use of it to overcome challenges in the assembly process. SOAPdenovo2 takes a more standard approach by primarily using a single, user-defined *k*-mer for the contiging process (Luo et al. 2012). SPAdes was inspired by a theoretical approach termed Paired *de-Bruijn* Graphs (PDBG; Medvedev et al. 2011). In practice, SPAdes method, *k*-bimer adjustment, utilizes read-pairs to create a paired *de-Bruijn* graph. SPAdes assembles the genome using multiple *k*-mer sizes and eventually combines them into one consensus sequence. It was originally designed for prokaryotic genomes but was later developed to accommodate large eukaryotic genomes (Bankevich et al. 2012).

Complementary to sequencing methods are genetic and physical mapping methods, which can map the relative position of genomic markers along chromosomes. Mapping can assist in scaffolding *de novo* assemblies and obtain full-length chromosomes. Among popular physical mapping technologies are BAC fingerprinting and optical mapping. Bacterial Artificial Chromosome (BAC) cloning is a technology harnessing bacterial plasmid cloning for either sequencing or mapping genomic fragments. In BAC fingerprinting, BAC DNA is cleaved by restriction enzymes, the amplified fragments are sorted using gel electrophoresis for size separation, and the pattern of fragments is read as the BAC's fingerprint. Fingerprints are used to align and assemble BACs into a genome wide map (Howe and Wood 2015). Optical mapping is a mapping method that utilize the physical linearization of single DNA molecules and enables the reconstruction of the order and orientation of long genomic fragments up to several mega base pairs. The Irys technology developed by BioNano uses single-strand site-specific endonucleases that create nicks in DNA molecules, followed by the insertion of fluorescent-labeled dNTP's during correction of the nicks. The labeled DNA molecules are uploaded onto a nano-confinements chip, where they are stretched inside nano-tubes. These DNA molecules fluorescent signatures are then read using a fluorescent optical detector. The fluorescent patterns of each molecule are aligned to create a physical genome map.

1.4 Challenges and solutions in *de novo* assembling whole genomes

When dealing with a *de novo* assembly of large eukaryotic genomes some challenges rise from their architectural characteristics. The greatest two challenges are (1) repetitive sequences, including gene duplications, transposons, and microsatellites; and (2) polymorphism including single-nucleotide polymorphisms (SNPs), insertions and deletions, and large genome rearrangement polymorphisms. Although sequencing technologies have advanced greatly in the past decade, these issues still present a major hurdle for genome projects, resulting in highly fragmented assemblies.

1.4.1 Challenges in *de novo* assembly due to genome structure

All medium to large eukaryotic genomes contain large amounts of repetitive sequences that can vary from short sequence repeats of two or three nucleotides (microsatellites) to duplications of whole genes and even genomic regions of millions of base pairs. Repetitive DNA can be one long sequence containing many repeats in tandem, on the same chromosome, or many repeats dispersed in many different locations in the genome. These repetitive elements introduce a challenge for the assembler, as it cannot distinguish them if the read length is shorter than the repetitive sequence. During the assembly process, in the stage of the *de-Bruijn* graph walkthrough, the assembler deals with repetitive elements by creating alternative or circular paths ("bubbles"). These paths pose a dilemma for the assembler – how to continue the path through the graph? In some cases, these can be resolved based on the higher sequencing depth of the repetitive elements compared to the unique sequences. However, in many cases these problems are left unsolved, resulting in a highly fragmented assembly, consisting of non-repetitive fragments ending in unresolved repetitive sequences (Simpson and Pop 2015).

Genetic variations across the genome, such as SNPs, are another challenge for the assembler to tackle. The level polymorphism and heterozygosity vary considerably among different species and can be affected, among other factors, by the mutation rate and population size (Kimura 1983; Hartl et al. 1997). Therefore, highly polymorphic species such as the Amphioxus (Putnam et al. 2008) have significantly increased assembly complexity. The assembler attempts to recognize heterozygous sites in the genome, and collapse them so that only one allele is present in the resulting assembly. However, in many cases, the assembler will collapse sequences that are in fact slightly

different repeats of a repetitive sequence. Conversely, depending on the strictness of the assembler regarding variants, it can treat multiple variants as a duplicated version of the same region and add them to the assembled sequence (Steinberg et al. 2014).

Higher ploidy levels introduce an even greater challenge relative to the 'commonplace' diploid organisms. Genome projects of *polyploid* organisms such as the tetraploid African clawed frog (*Xenopus Laevis*) (Graf and Kobel 1991) or the *hexaploid* bread wheat (*Triticum aestivum*) (Brenchley et al. 2012) are considerably more difficult than diploid genomes. One of the challenges, when dealing with diploid genomes is *phasing* - the correct separation of the haplotypes in the genome. Higher ploidy levels increase the complexity of the genome by introducing more allelic-variants and typically more repetitive sequences. Conversely, assembling a genome from a haploid source is expected to facilitate higher contiguity in the genome assembly because of the lack of polymorphism (Steinberg et al. 2014).

The present study set out to evaluate the use of haploid male ants as opposed to diploid females, and the extraction of both RNA and DNA from the same male individual, to test their advantage in simplifying the assembly process thanks to the lack of polymorphism in the haploid sample. Moreover, pairing the source of the sequenced DNA and RNA is expected to provide greater confidence in transcript-to-genome alignment, and ease the annotation of gene structure in terms of the exon/intron boundaries.

1.4.2 Experimental design using NGS

The common design for an Illumina-based *de novo* sequencing project consists of a series of libraries with a gradation of insert sizes. Read length produced by the platform dictates the appropriate insert sizes. If we design an experiment with a read length of 125bp, we may choose paired end libraries with insert sizes of 300, 500 and 800bp. These may be combined with mate pair libraries of 2, 5, 10 and 20kb for maximum scaffolding. By using a mixture of insert lengths, the assembly process is optimized and a genome with a better contiguity can be assembled (Bonasio et al. 2010; Nygaard et al. 2011; Oxley et al. 2014). For both DNA and RNA sequencing the sequencing depth of the different libraries can be controlled by the use of *multiplexing*. Multiplexing enables the sequencing of multiple samples in a single lane, thus lowering the depth for each sample.

Knowledge of expected genome size can give some indication for the desired depth and assist in choosing the number of lanes needed. It can substantially reduce the cost of projects, which require sequencing a large number of samples such as genotyping and deferential expression studies.

1.5 The *haplodiploid* life of Hymenoptera

Unlike the more widespread chromosomal sex determination, haplodiploid sex determination is based on variation in ploidy. In this strategy, an unfertilized egg will develop into a haploid male while a fertilized egg will develop into a diploid female. The largest haplodiploid animal clade is the order Hymenoptera, with about 200,000 species. In hymenopteran social insects (ants, bees, and wasps), a fertilized egg can either develop into a worker or, if nutritionally directed, into a queen. The eggs are usually laid by the queen, who controls their fertilization by sperm from a specialized storage organ called *spermatheca*. In some species, and in certain conditions, unmated workers can also lay haploid eggs that will develop into males. A widespread molecular mechanism underlying haplodiploid sex determination is Complementary Sex Determination (CSD). In CSD, sex is determined by the zygosity of a particular locus. If an individual is heterozygous in the CSD locus it will developed into a female. A hemizygous (bearing only a single copy of the locus) or a homozygous genotype will develop into a male (van Wilgenburg et al. 2006). To date, over 20 hymenopteran genomes were fully sequenced and assembled (Elsik et al. 2016). A few, such as Nygaard et al. 2011 (*Acromyrmex echinatior*) and Wurm et al. 2011 (*Solenopsis invicta*), used haploid males as their main source for the assembly, alongside a pool of workers for mate pair libraries and RNAseq. Thus, haplodiploidy of Hymenoptera provides a unique solution to some of the main challenges in *de novo* genome assembly.

1.6 The target organism

The present study uses the ant *Cataglyphis drusus* from the Western Galilee, north of Mount Carmel. The Genus *Cataglyphis* consists of more than a 100 species that inhabit arid lands across Africa and the Mediterranean. *Cataglyphis* ants display variable and complex social structures. A colony consists of diploid workers in charge of everyday tasks while a queen functions as the female reproductive unit. The male reproductive unit consists of haploid drones, and queens often mate with multiple males. Furthermore, some species are monogyne while others are polygyne, having multiple reproductive queens in the same colony. The mating strategy and social structure (monogamy, polyandry, monogyne, polygyne, etc.) is species-specific and may be subject to environmental conditions and nest conditions (Amor et al. 2011; Leniaud et al. 2011). *C. drusus* is a monogynous (one queen per nest), polyandrous (queens are multiply mated) and a monodomous species (inhabits a single nest; Tali Reiner Brodetzki, unpublished data).

1.7 Transcriptome assembly

Transcriptomes are often used it several ways to annotate gene structures on a genome assembly, as well as to help assess the quality of the genome assembly. Two approached can be utilized. The first involves the alignment and construction of RNA sequencing (RNAseq) reads to the genome assembly. The RNAseq reads are aligned and mapped to an indexed genome and eventually assembled accordingly. The second approach is to *de novo* assemble the RNAseq reads into transcripts and then align them to the genome. Transcriptome assembly algorithms are based on the same principles of genome assembly. The most significant difference is in the need to allow multiple alternative splice isoforms for the same gene.

1.8 Genome assembly quality assessment

There are multiple methods of assessing the quality of a genome assembly, yet no single one of them is considered a standardized and complete benchmark for assembly quality comparison (Gurevich et al. 2013). The main methods used are calculating assembly statistics such as the N50

contig/scaffold size, assessing gene set completeness, and detecting misassemblies by comparison to the genome of a closely related species. N50 is a summary statistic similar to a median, which captures the contiguity of the assembly by the length the contig/scaffold that half of the total assembly length is found in contigs/scaffolds of that size or bigger (Simpson 2014). However, high N50 scores can result from misassemblies (Gurevich et al. 2013).

Misassemblies can occur during the assembly process in stages of contig construction or scaffolding, if the assembler erroneously links similar sequences from different regions of the genome. Tools have been developed to detect such possible errors. For example, QUAST is an assessment tool that identifies misassemblies using a related reference genome. This approach relies on the assumption of conservation of *synteny*, that is, that gene order is generally similar between closely related species. QUAST also puts functional elements (exons and other conserved sequence elements) into the context of the assessment for a more accurate evaluation of synteny (Gurevich et al. 2013).

Finally, in completeness evaluation, the genome assembly is searched for a pre-defined subset of conserved genes. This "inventory check" provides an indication of the completeness of the assembly. One such tool is the Core Eukaryotic Genes Mapping Approach (CEGMA), which uses a dataset of single-copy genes that are highly conserved across all eukaryotes (the euKaryotic Orthologous Groups (KOG)) database, alongside a gene finder-training algorithm (Parra et al. 2008). Another tool is Benchmarking Universal Single-Copy Orthologs (BUSCO). The BUSCO algorithm takes a similar approach as CEGMA but uses OrthoDB as its main database of conserved single-copy orthologues genes, and allows choice of gene sets conserved in different taxonomic levels, including Metazoa, Arthropoda, etc. (Simão et al. 2015).

1.9 Gene annotation

Annotating a genome involves identifying specific genes and their homology to known genes or other structural or functional elements from other (model) organisms. Moreover, gene structure such as exon-intron boundaries need to be identified, as well as alternative splicing that could potentially generated multiple isoforms. Various approaches and software tools were developed for genome annotation. One such tool is Cufflinks, the final stage of the transcriptome to genome alignment

pipeline implemented in the Tuxedo suite (Trapnell et al. 2012). Cufflinks constructs the exons and predicted transcripts of genes based on the alignment of RNAseq reads to the genome. MAKER implements a more comprehensive approach, which identifies genes and their exon-intron structure by aligning a newly assembled genome to protein sequences from other species and to transcriptomic sequencing of the same species (e.g. by the Tuxedo suite). Another approach implemented in MAKER is *ab initio* gene prediction – searching for regions in the enquired genome that contain the necessary elements of a protein coding gene, including promoter, translation initiation, splice junction motifs, stop codon, etc. (Cantarel et al. 2008).

2. Research objectives

In this study, we examine the advantage of haploidy for *de novo* assembling a genome and a transcriptome.

(1) Genome assembly analysis; The first goal of this study is to assemble a high quality draft of the *Cataglyphis drusus* genome and compare the effect of ploidy over the assembly process and the quality of the assembly. Thanks to the unique genomic characteristics in Hymenoptera, namely haplodiploidy, a comparison between a haploid male and a diploid female can be made. Genomes from two related sources, haploid (male) and diploid (worker sister from the same nest) are assembled using two popular assemblers: SOAPdenovo2 and SPAdes. These assemblers take somewhat different computational approaches based on a *de-Bruijn* graph. For quality assessment, alternative genome assemblies are compared using N50 statistics, misassemblies detection, and completeness.

(2) Transcriptome assembly analysis; the second goal is to compare the effect of different ploidy assemblies on transcript alignment to the genome. We intend to compare the quality of alignment to the genome of transcripts derived from different ploidy source material: the same haploid male or a pool of multiple samples (Figure 2). Third, we compare the quality of the transcriptome assembled *de novo* from the RNAseq data from the single haploid individual against the pool.

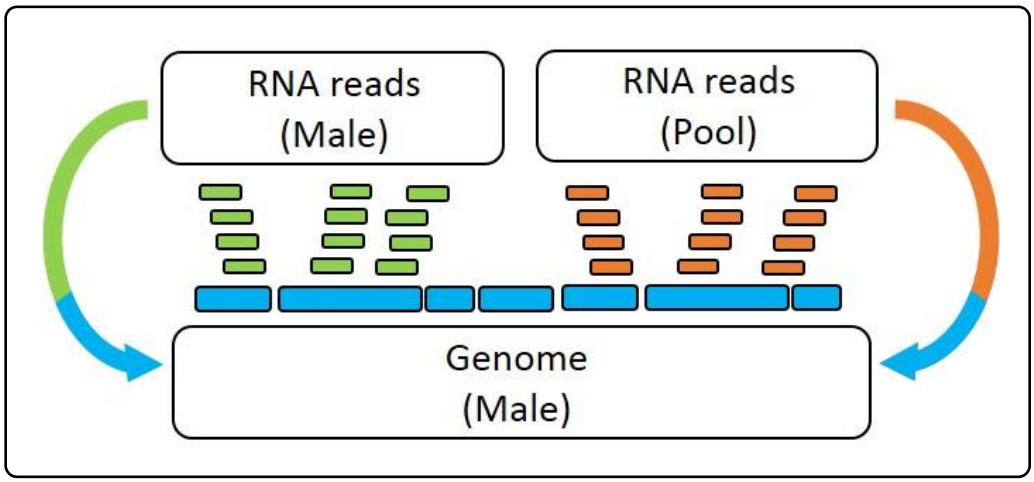

Figure 2: an illustration of the comparisons for the transcript mapping to the haploid genome assembly

(3) Future applications and research; The results of this study, in particular the genome and transcriptome assemblies, will form the basis of a high quality, high continuity annotated reference genome for the species *C. drusus*. *Cataglyphis* ants in general are a well-studied model for the study of nestmate recognition. This new genomic resource will facilitate future research to determine the genetic architecture of *cuticular hydrocarbons* (CHCs) synthesis in *C. drusus,* which serves as the basis for nestmate recognition in *Cataglyphis* ants. The genome is annotated in order to identify and predict the transcripts of all protein-coding genes. This would allow inspection of candidate genes in genomic regions identified by future QTL mapping and assist in future genomic research in the *Cataglyphis* genus.

The results both of the genome and transcriptome analysis may assist future studies in selection of source material and sequencing design for DNA and RNA sequencing in ants and Hymenoptera in general. This study will provide a first thorough evaluation of the contribution of the use of a haploid source of DNA and the advantage of using DNA and RNA from the same individual to the quality of a genome assembly and its gene annotation.

3. Materials and Methods

3.1 Samples collection

All samples (male, worker and pool samples) were collected during daytime using the same method of collection. In our research site near *Betzet* beach (Appendix 1), approximately 2 km southwest of *Rosh Hanikra*, Israel. Male and worker samples were collected from the same nest while samples for the RNA pool were collected from several additional nests in the site. The collection method first required the tracking down of workers. Bait (biscuit crumbs) was spread around their vicinity to lure them into carrying it to the nest entrance. Once a worker was detected carrying the bait, it was followed until the identification of its nest entrance. The nest was dug manually. Both male and worker sample were collected from the nest marked as 4B (33°4'40.88"N / 35°6'33.97"E). One haploid male sample was used for DNA and RNA sequencing, and one sister worker provided the diploid DNA sample. RNA pool was made from a collection of workers, males, gynes and different developmental stages from multiple nests (including larvae and pupae of various sizes; Appendix 2). After collection, the samples were brought to the lab, separated into individual 2ml LoBind (Eppendorf) tubes and put directly into liquid nitrogen while still alive.

3.2 Samples preparation and extraction

3.2.1 Extraction of male DNA and RNA

In this study, a main challenge was to extract both DNA and RNA from the same haploid individual (sample Male_BZT4B; male sample of nest BZT4B). Several preliminary experiments using different methods of extraction, eventually led to use of the All-Prep DNA & RNA mini extraction kit (QIAGEN). The manufacturer standard protocol was calibrated and modified for optimal use with our study organism, to produce a sufficient yield with adequate purity of DNA and RNA.

Initially three candidate males from three different nests (nests BZT4B, BZT4C and BZT7) were extracted. Males extraction was performed separately for each male, using the same method, beginning with the already snap frozen sample transferred to a sterile 2ml *LoBind* (Eppendorf) tube and inserted into liquid nitrogen with the aim of maximizing tissue disruption effect. The sample

was than disrupted manually with a micro pestle and underwent homogenization, lysis, and isolation of DNA and RNA separately (Appendix 3). DNA was suspended in 40µl of 2mM TRIS solution, 5µl of which was taken for the microsatellite analysis. RNA was suspended in 40µl of PCR RNase free water (Biological industries). After extraction, the sample was measured for nucleic acids concentration and purity (both of DNA and RNA) using the *NanoDrop* UV spectrophotometer (ND2000; Thermo-Fisher Scientific; results in Appendix 4).

As previously mentioned on section 1.5, under complementary sex determination, a diploid individual will develop into a male only if they are homozygous at the sex determination locus. Diploid males are rare, and yet it was important to verify the haploidy of the sample used for genome sequencing. Two common methods of determining ploidy are genotyping of microsatellites (which was used in this study) and flow cytometry. In the first, utilizing the difference in length between different microsatellite alleles, a male with heterozygous genotypes in one or more microsatellite loci will be identified as diploid, while a male with a single allele in each locus will be assumed haploid.

With flow cytometry, a laser beam is projected on a suspended solution with the cells in question for purpose of quantification the DNA content of each cell. Sheath fluid channel cells to flow in a single file. Cells are passing through the laser and light is scattered forward and sideways. Detectors detects the light and quantifies it electrically. A DNA-specific dye is used for the flow cytometer to quantify the DNA in the enquired cells. With this method, we can detect ploidy as well as quantify the size of the genome (Bohanec 2003; Doležel and Bartoš 2005).

The ploidy of the male samples was determined using four highly polymorphic microsatellite markers developed specifically for genotyping species of the *Cataglyphis* genus (for markers sequences see Table 3). These four markers were discovered in three different, closely related species and described previously for *C. niger* (Cn02 and Cn04), C. *hyspanica* **(**Ch08**)** (Darras et al. 2014) and *C. cursor* (Cc54) (Pearcy et al. 2004). PCR amplification of these loci was done as previously described (Timmermans et al. 2009). Four male samples were examined for haploidy. Three male samples were confirmed to have a single allele in all four microsatellites markers. Out of these, a single male (sample Male_BZT4B), which had the best overall extraction quality results (refers to quantity and purity of both DNA and RNA) and validated haploidy was chosen for sequencing.

3.2.2 Extraction of worker DNA

For the diploid DNA sample, two workers (sisters of the chosen male) were extracted using a modified protocol of the DNeasy Blood & Tissue kit (QIAGEN) and the same method described above. DNA was suspended in 30µl of 2mM TRIS solution. After extraction, the sample was measured for DNA concentration and purity using the ND2000. Out of two candidates, only one worker (sample Worker_BZT4B), which had the best overall extraction quality results was chosen for sequencing (Appendix 3).

3.2.3 Preparation and extraction of RNA pool

An RNA pool was constructed of several individuals from three different nests collected from the Betzet site. Various castes and developmental stages were taken as follows: one worker, one gyne (virgin queen), one drone (male), five larvae of three size groups and three pupae from two different size groups (larvae and pupae were chosen by size and their classification to different instar stages or castes was not checked). The diversity in the RNA pool composition provides an optimal representation of the potential gene expression in model organism, including genes and alternative isoforms, which are expressed only in certain life stages and/or castes. Sample extraction was performed separately and with equal effort, using the same method described above. A modified protocol of the RNeasy mini kit (QIAGEN) was used for the isolation and purification of RNA from the tissue. Each sample was suspended in 50µl of PCR RNase free water (Biological Industries). A measurement of RNA concentration and purity was done using the ND2000 on each of the samples. To achieve equal representation of each sample type (caste/developmental stage) in the final pool, RNA quantity was normalized by pooling different volumes corresponding to the RNA concentration of the sample. Overall, two duplicate pools (Pool-A and Pool-B) were constructed, containing 60µl of approximately 100ng/µl of RNA, of which Pool-B was chosen for sequencing (Appendix 2).

3.3 Sequencing

DNA and RNA sequencing was performed by *Eurofins Genomics GmbH* (Germany) using the HiSeq 2500 sequencing platform (Illumina) and the HiSeq SBS Kit v4 chemistry (Illumina). For genomic DNA sequencing, two paired-end libraries were constructed for each sample (haploid/diploid) with insert sizes of 300 and 550bp. Each pair of libraries was multiplexed in one lane, giving a total coverage for the 300bp library of 84/127X for the haploid/diploid samples respectively, and 93/108X for the 550bp library (based on the flow cytometry genome size estimate of 220Mbp). Following poly-A enrichment for mRNA sequencing, a 500bp paired-end, strand specific cDNA library was constructed for each sample. Each library was sequenced in a separate lane.

3.4 Pre-assembly quality control

Quality control of the raw sequence data was performed using FastQC (version 0.11.5; Andrews et al. 2010). The FastQC report contains number of reads (which was cross-referenced with the Eurofins project report), read quality distribution, adaptor contamination, over-represented sequences, etc. Mean Phred Q scores and %Q30 scores were provided as part of the Eurofins Genomic report. The Phred Q score is an estimate of the probability for a wrong base call. The percentage Q score of 30 (%Q30) represents the percentage of bases with a quality score of at least 30 (inferred base call accuracy of 99.9%). Post sequencing estimation of average insert size was done by Eurofins Genomics, using a TapeStation instrument (Agilent). The overall quality of all libraries is described in section 4.1 in the Results.

Trimmomatic (version 0.32; Bolger et al. 2014) was used to screen and remove or trim low quality reads and contamination of Illumina adaptors. Reads trimmed to less than 80bp were removed. A sliding window of four bases was applied and ends of reads were trimmed where the window-averaged quality score (Phred score) lower than 15. Elimination of Illumina adaptor contamination was made based on adaptor sequences taken from Illumina publications, Eurofins Genomics, and FastQC.

3.5 DNA and RNA coverage reduction

After trimming, a reduction of the number of reads (coverage) was made to equalize diploid and haploid samples, for a fair comparison of assembly quality. The reduction was on the higher read count coverage before assembly. For this purpose, the Linux command 'SED' was used until difference in read number was less than 1%. Reduction was made randomly according to the percentage needed for equal coverage (Table 1). At the end of the quality control pipeline, a second FastQC run was made to verify the quality of the remaining reads and to make sure that any Illumina adaptors were removed.

Table 1: Coverage reduction of DNA and RNA samples before assembly

Sample type	No. of reads in raw sequencing data [bp]	No. of reads after trimming [bp]	No. of reads after reduction [bp]	Percentage of reduction [%]
DNA				
Male (Haploid)	180,269,648	156,464,223	156,464,223	0%
Worker (Diploid)	225,419,503	206,886,942	157,694,240	23.7%
RNA				
Male	184,860,252	N/A	173,306,486	6%
Pool	173,787,824	N/A	173,787,824	0%

3.6 Genome assembly

Two popular assemblers, SOAPdenovo2 (version r240; Luo et al. 2012) and SPAdes (version 3.9.1; Bankevich et al. 2012), were used to assemble the genomes of both the haploid and diploid samples. The two assemblers were configured with default parameter settings regarding error correction cutoffs etc., as recommended by each assembler's manual. Optimal *k*-mer sizes were chosen based on preliminary assemblies and read length. With SOAPdenovo2, several assemblies were constructed using different *k*-mer sizes (35, 45, 63, 85, and 115). Among those, the 115 *k*-mer size was chosen being the one with the highest contig and scaffold N50. SPAdes was run using an array of *k*-mer sizes ranging from 13 to 123bp (base on a read size of 125bp), which was combined to form the final assembly by SPAdes.

3.7 Genome assembly quality analyses

To evaluate the overall quality of alternative genome assemblies, three methods were used: calculating the N50 size of contigs and scaffolds, evaluating the completeness of the assemblies, and detecting misassemblies in each assembled genome. Evaluation results were then compared between the different ploidy assemblies. The N50, for both contigs and scaffolds, was calculated using a customized PERL script (https://github.com/kakitone/finishingTool/blob/master/fasta-splitter.pl). For completeness assessment, BUSCO (version 1.22; Simão et al. 2015) was ran against the BUSCO Arthropoda dataset (http://busco.ezlab.org/). Misassemblies were detected using QUAST (version 4.4; Gurevich et al. 2013) with comparison to the genome assembly of *C.hyspanica* (Hugo Darras personal communication).

3.8 RNA mapping and transcriptome assembly

3.8.1 *De novo* transcriptome assembly

A *de novo* assembly of the transcriptomes of both male and pool was performed using the Trinity assembler (version 2.4.0; Grabherr et al. 2013). Prior to assembly, the raw RNAseq data was trimmed with Trimmomatic with a maximal read trim size of 60bp (version 0.36; Bolger et al. 2014; see Appendix 5 for parameters). Quality control of the data was made using FastQC before and after trimming. Reads were normalized using the Trinity build-in *In-silico read normalization utility* (https://github.com/trinityrnaseq/trinityrnaseq/wiki/Trinity-Insilico-Normalization). Completeness evaluation of the finished assemblies was done using BUSCO against the Eukaryota *odb9* dataset.

3.8.2 Transcriptome mapping to genome

The Tuxedo suite pipeline was used to map RNAseq reads of both the male and pool RNA to the haploid reference genome (Trapnell et al. 2012). First, an indexing of the reference genome was done with Bowtie2 (version 2.3.2; Langmead and Salzberg 2012). Moreover, Bowtie2 serves as an alignments engine for Tophat2 (version 2.1.1; Trapnell et al. 2009; Kim et al. 2013). Tophat2 aim is to overcome gaps (mainly introns) in the referenced genome, and mapping of the RNA reads

against the genome. It also identifies splice sites and align the reads accordingly. Transcript assembly was done using Cufflinks (version 2.2.1; Trapnell et al. 2010). Lastly, an analysis of the mapping quality for the male RNA and the pool RNA against the haploid genome was made. The goal of this analysis was to weigh the advantage of mapping RNA reads from the same haploid individual to its genomic sequence relative to the mapping of RNA from other individuals. Completeness of the resulting gene structures was evaluated by BUSCO against the Eukaryota *odb9* dataset. The use of the Eukaryota dataset, as opposed to larger gene datasets (i.e. Arthropoda, Hymenoptera, etc.), was to limit the analysis to universally conserved genes, which are expected to be expressed in all tissues and developmental stages, and thus should be well covered both in the male and the pool RNAseq data. In addition, Basic Local Alignment Search Tool (BLAST; *tbalastn* as part of the BUSCO pipeline) annotation results were evaluated by manually comparing the alignment of the RNA reads, to the haploid genome by Tophat2 and the resulting exon and transcript annotation by Cufflinks. Visualization was done using the Integrative Genomic Viewer (IGV) (version 2.3.97; Robinson et al. 2011).

3.9 Genome annotation

Transcripts annotation for all four genome assemblies was performed using the MAKER annotation pipeline (version 2.31.9; Cantarel et al. 2008). MAKER receives as input the genome assembly (.fasta file format) as well as Cufflinks transcripts (.gff file format) and sequences of homolog proteins of related species (.fasta file format; Appendix 6)

BUSCO was used to evaluate the completeness of the annotated gene set against the Eukaryota OrthoDB gene dataset (version 9.1; *odb9;* http://www.orthodb.org/). The same BUSCO evaluation was made on Cufflinks assembly results prior to the MAKER annotation. In addition to BUSCO, a BLAST (*tbalstn*) run was performed independently on all of the transcript assemblies against the same BUSCO gene dataset.

3.10 Extracting High Molecular Weight (HMW) DNA from *C. drusus* ants

3.10.1 HMW DNA role in the assembly

Besides the samples collected for the short-read sequencing by Illumina an additional DNA pool of *C. drusus* workers from a single nest was collected. This pool was intended to provide high molecular weight (HMW) DNA as the basis for long reads sequencing platforms such as PacBio, Nanopore or Illumina mate-pair protocol to produce long inserts. These long read/inserts data main application was to assist the construction of larger scaffolds and eventually improve the contiguity of the genome assembly. Moreover, as mentioned before, long read data can provide a solution for "holes" generated in long scaffolds during the assembly process. Both Mate-pair and PacBio sequencing required HMW DNA libraries with a mean size of 50-100kbp. Eventually the long insert/reads sequencing was not performed due to circumstances related to the sequencing provider (Eurofins Genomics GmbH).

3.10.2 Extraction of worker DNA

HMW DNA is sensitive and prone to breakage due to physical shredding force (such force is produced when using 'violent' disruption methods such as sonication, mechanical disruption with bead-beater, electric grinder and ever mixing by vortexing). Therefore, HMW DNA should be treated gently and every action of disruption and mixing should be performed manually. Nevertheless, the main challenge was getting both a sufficient size of HMW DNA fragments, as well as an appropriate concentration and purity level. For this purpose, a modified protocol based on the Blood and Cell Culture DNA Mini kit (QIAGEN) was developed. The kit includes the 20G columns, which were designed specifically for HMW DNA extraction. In order to achieve the amount of DNA material a pool of 20-30 live workers from the same nest (medium to large size) was collected in a 50 ml tube. 20ml of lysis buffer was then added to the tube and manual disruption of the samples was made using a rounded edge Teflon pestle. 40µl of pre-heated (37° C) RNase A was added to the mix alongside 2ml of Protease. The samples were then put into incubation for four hours. After the incubation, the samples were centrifuge for 20 minutes in high speed (14,680 rpm). The incubation resulted with a three-phased lysate, from which the middle clear layer was uploaded onto the Genomic tip columns. Prior to the upload of the samples, equalization of the column with

QF buffer was made. In each column, 1ml of lysate was uploaded, followed by four washes of 1ml of QC buffer. The extraction of the samples was done in three fractions; The first fraction was extracted with 200µl of QF buffer and discarded. The second and third fractions were each extracted with 200µl of 2mM TRIS buffer and were collected into individual sterile 2ml LoBind tubes. Before the precipitation stage each sample was measured for DNA concentration and purity using the NanoDrop ND2000. After elution, the samples were precipitated according to the standard protocol. Immediately after precipitation, the samples were dried for a five-minute cycle in a SpeedVac vacuum concentrator machine (Thermo-Fisher scientific) and eluted in 100-120 µl of 2mM Tris buffer. After the final elution, each sample was measured a second time for DNA concentration and purity, again using the ND2000.

4. Results

4.1 DNA and RNA sequencing

The main goal of this project was to compare the quality of whole genome assemblies and transcriptome-based gene annotation, considering the differences in the source material: haploid vs. diploid samples. For this, a whole genome assembly and transcriptome sequencing was done, using the HiSeq 2500 sequencing platform. Table 2 describes the sequenced genomic and transcriptomic libraries, their yield, and quality statistics. For each sample, two DNA libraries (300 and 550bp) were constructed. Average genome coverage was calculated with an estimated genome size of 220Mb based on flow cytometry measurements for *Cataglyphis hyspanica* (Hugo Darras personal communication). The quality of all the libraries, both the DNA and RNA was very high, with %Q30 scores above 88%. Estimated average insert size for the DNA libraries was very close to the desired size, at most ~7% different (for the 550bp haploid library). For the RNAseq libraries, the average insert size was much lower than the desired size, by about 50% (Table 2).

Table 2: Illumina libraries constructed. Two genomic DNA libraries (300, 550bp) constructed from each of the source materials. For RNA sequencing, one library for each source was constructed.

Sample type	Library type	Intended insert size [bp]	Estimated average insert size [bp]	No. of reads raw data [bp]	Depth [X]	%Q30[b]	Mean Q[a]
DNA							
Male (Haploid)	Paired-end 125b x 2	300	290	84,756,564	84	90.34	34.28
Male (Haploid)	Paired-end 125b x 2	550	510	95,513,084	93	88.63	33.90
Worker (Diploid)	Paired-end 125b x 2	300	340	120,165,717	127	93.15	35.21
Worker (Diploid)	Paired-end 125b x 2	550	560	105,253,786	108	90.44	34.56
RNA							
Male	Paired-end 125b x 2	500	210	184,860,252	N/A	93.07	35.13
Pool	Paired-end 125b x 2	500	240	173,787,824	N/A	93.50	35.22

(a) Mean Phred Q scores for each of the libraries, (b) %Q30 is the percentage of bases with a quality score of at least 30 (inferred base call accuracy of 99.9%) for each library (Eurofins genomics)

4.2 Haploidy confirmation

Haploidy was confirmed in the male sample by the lack of heterozygosity in four highly polymorphic microsatellites. This result strongly indicates that the sample was haploid, because a diploid sample has an expected probability of 2.76% or less to have homozygous genotypes in all four loci, based on the heterozygosity level of each of these loci in the Betzet population of *C. drusus* (Tali Reiner-Brodetzki, personal communication) (Table 3).

Table 3: Microsatellites used for ploidy test of the male samples.

Locus	Dye	Expected size	Homozygosity*	Primer sequence
Cn02	PET	115 bp	0.682	Forward 5'=>3' GAGGCCCCTGAAAAGAAGAT Reverse 5'=>3' TTCTATCTCTGCCGGCTTCT
Cn04	VIC	95 bp	0.251	Forward 5'=>3' GGAAACTCGTGCGAAAACTC Reverse 5'=>3' GAGCTCAGTGTGCATTCAACAT
Ch08	NED	135 bp	0.289	Forward 5'=>3' GCTGATAATCGCGTCTGGAT Reverse 5'=>3' CGACGTAAAGAGGAACGTGA
Cc54	FAM	210 bp	0.553	Forward 5'=>3' GAATTTGAATGGCTGATTGC Reverse 5'=>3' ATGGTCGTTTGGCATAAAGG

***** Proportion of homozygous out of a total of 708 samples in the *Betzet* site.

4.3 Quality analyses of *de novo* genome assemblies

Quality of the initial data was first evaluated using FastQC. Low quality reads and Illumina adaptor contaminations were either trimmed or removed using Trimmomatic. The haploid and diploid samples were assembled using two assemblers: SOAPdenovo2 with a *k*-mer size of 115bp, and SPAdes with *k*-mer size ranging between 13-123bp. Overall, four assemblies were assembled and compared.

4.3.1 Contiguity of the assemblies

N50 contig and scaffold sizes were at least threefold larger for the haploid relative to the diploid assemblies, both by SOAPdenovo2 and SPAdes (Table 4). The SPAdes haploid assembly had almost five fold larger contig N50 size than its SOAPdenovo2 counterpart did, while the same comparison on the diploid assemblies shows a factor of ten. Haploid scaffold N50 size was similar for both assemblers. Conversely, diploid scaffold N50 size was more than threefold higher in

SPAdes than in SOAPdenovo2. Moreover, the total size of the two haploid assemblies (219Mb for SPAdes; 296Mb for SOAPdenovo2) was much closer to the expected value of 220Mb based on the flow cytometry measurements in *C. hyspanica* (Hugo Darass, personal communication). The diploid SPAdes assembly was clearly inflated. The diploid assembly size by SOAPdenovo2 was not as high as the one by SPAdes, yet it was still substantially higher than the expected size.

Table 4: N50 contig and scaffold sizes for the different genome assemblies

	SPAdes		SOAPdenovo2	
	Male (Haploid)	Worker (Diploid)	Male (Haploid)	Worker (Diploid)
contigs	15,206	5,367	3,143	554
scaffolds	17,901	5,742	16,307	1,659
Total assembly size	296Mb	759Mb	219Mb	345Mb

4.3.2 Completeness of the assemblies

Completeness test by BUSCO showed mixed results (Table 5). While the haploid assembly performed by SOAPdenovo2 achieved better completeness than the diploid (63%), the SPAdes assembly completeness was better for the diploid assembly (84%). The percentage of fragmented genes was twice greater in the SPAdes haploid compared to the diploid assembly. With that being said, the percentage of duplicated genes in the SPAdes diploid assembly was ten times higher than its haploid counterpart. A similar trend can be seen in BUSCO results against the Eukaryota *odb9* dataset.

Table 5: BUSCO completeness results for the different genome assemblies against the Arthropoda (a) and Eukaryota (b) *odb9* datasets

(a) Arthropoda	SPAdes				SOAPdenovo2			
	Male (Haploid)		Worker (Diploid)		Male (Haploid)		Worker (Diploid)	
Complete	1832	68%	2256	84%	1706	63%	1276	47%
Single copy	1726	64.1%	1222	46%	1643	60.7%	1179	43.4%
Duplicated	106	3.9%	1034	38%	63	2.3%	97	3.6%
Fragmented	771	28%	376	14%	860	32%	1054	39%
Missing	72	2.6	43	1.6%	109	4%	345	12%
2675 total BUSCO genes								

(b) Eukaryota	SPAdes				SOAPdenovo2			
	Male (Haploid)		Worker (Diploid)		Male (Haploid)		Worker (Diploid)	
Complete	279	92%	296	97%	260	85%	223	73%
Single copy	244	81%	59	19%	248	81.1%	183	60%
Duplicated	35	11%	237	78%	12	3.9%	40	13%
Fragmented	19	6.2%	2	0.6%	32	10%	63	20%
Missing	5	1.6%	5	1.6%	11	3.6%	17	5.6%
303 total BUSCO genes								

4.3.3 Contig and scaffold misassemblies

Analysis of misassemblies using QAUST revealed that the total number of misassemblies in the SPAdes haploid assembly was higher than the diploid one (Table 6). An opposite trend was seen in the SOAPdenovo2 assembly, in which contig misassemblies in the diploid assembly were higher by a factor of almost 1.5 for the global misassemblies and more than 2.5 for the local ones. Relative to SPAdes, both haploid and diploid SOAPdenovo2 assemblies have a dramatically higher number of local misassemblies in contigs and scaffolds alike, with a factor of more than 30 in the diploid local contig misassemblies and a factor of more than six in haploid ones. There was an approximately equal number of global misassemblies in the haploid assemblies by SPAdes and the SOAPdenovo2, while the diploid SOAPdenovo2 assembly had more than twice misassemblies than the SPAdes assembly (Table 6).

Table 6: Misassemblies identification by QUAST for the different assemblies

	SPAdes				SOAPdenovo2			
	Haploid (M) [bp]		**Diploid (W) [bp]**		**Haploid (M) [bp]**		**Diploid (W) [bp]**	
	contigs	Scaff.	contigs	Scaff.	contigs	Scaff.	contigs	Scaff.
Global	420	438	231	272	482	504	698	546
Local	465	496	259	301	2964	766	8047	753

Misassemblies are classified as local if flanking sequences on both sides are gaped or overlapped by > 85bp and <1Kb. Global misassemblies are > 1Kb.

4.4 Quality analyses of *de novo* transcriptome assemblies

The pool's N50 contig size was larger by more than 30% of that of the male's. The total transcriptome size estimate for the pool was lower than the male by 25% (Table 7).

Table 7: N50 results of two de novo transcriptome assemblies

	Trinity	
	Male	Pool
N50	2115	2782
Number of genes	101559	56546
Total assembly size	214Mb	170Mb

According to BUSCO evaluation, the number of complete genes was higher in the pool transcriptome (Table 8). The number of fragmented genes was higher in the male than the pool transcriptome. Similarly, the number of missing BUSCO genes was higher for the male yet it was only 1.3% of the total number of genes checked.

Table 8: BUSCO completeness results for the transcriptome assemblies against the Eukaryota *odb9* dataset

	Trinity			
	Male		Pool	
Complete	263	86%	278	91%
Single copy	161	53%	166	55%
Duplicated	102	33%	112	36%
Fragmented	36	11%	23	7.5%
Missing	4	1.3%	2	0.6%
303 total BUSCO genes				

4.5 Transcriptome mapping

Genome annotation is usually done by mapping RNAseq reads to the reference genome assembly. DNA and RNA from the same haploid individual were used for the genome assembly and the annotation, which is expected to be advantageous because these RNA sequences should be easier to align relative to transcripts of other alleles. To evaluate the advantage of using the same individual as both DNA and RNA source, the completeness of gene annotations on the haploid genome

assembly was compared to annotations based on RNA from a different source, namely the pool RNAseq. Gene annotation was compared both after running the full MAKER annotation pipeline, which uses also evidence from alignment to proteins from other species, and also for the results of the RNAseq alignment alone before running MAKER (i.e., gene annotations and transcripts predicted by Cufflinks).

4.5.1 Completeness of the mapped transcripts

A fair comparison should be done for genes that are expressed in both males and other sample types (queens, workers, larvae, and pupae). Thus, completeness of the predicted transcriptomes was evaluated by BUSCO against the Eukaryota *odb9* database (Table 9), because these universally conserved, single copy genes are expected to be expressed in all tissues, developmental stages, sexes, and castes. Male RNA mapping to the Male genome (MvM) resulted in a higher count of complete genes before and after MAKER annotation than Pool RNA mapping to Male genome (PvM). Both of the mapped transcriptomes did not show any duplicated genes after MAKER annotation. Pre-MAKER BUSCO results found 15% and 19% duplicated genes for PvM and MvM, respectively. Fragmented gene count was higher for PvM both before and after MAKER annotation, by 30% and 41%, respectively. Missing gene counts in the pre-MAKER results were almost identical, with only one more gene missing in MvM. Post-MAKER missing gene counts were the same for both MvM and PvM. In both male and pool, MAKER succeeded in correcting some of the fragmented genes, and reduced their number from 34 to 17 and from 44 to 24 in the MvM and PvM respectively. A similar result was seen in the missing genes, which MAKER reduced in both samples to four genes. MAKER analysis also identified all genes as single copy in both samples.

Table 9: Completeness results done by BUSCO against Eukaryota *odb9* datasets

	Pre-MAKER					
	MvM[a]		PvM[b]		In both	
Complete	257	84%	248	81%	222	73%
Single copy	197	65%	200	66%	N/A	N/A
Duplicated	60	19%	48	15%	N/A	N/A
Fragmented	34	11%	44	14%	16	5%
Missing	12	3.9%	11	3.6%	5	2%
303 total BUSCO genes					**243**	**80%**
	Post-MAKER					
	MvM[a]		PvM[b]		Found in both	
Complete	282	93%	275	91%	275	90%
Single copy	282	93%	275	91%	N/A	N/A
Duplicated	0	0%	0	0%	N/A	N/A
Fragmented	17	5.6%	24	7.8%	16	5.2%
Missing	4	1.4%	4	1.2%	4	1.3%
303 total BUSCO genes					**296**	**96.8**

(a) MvM = male RNA mapped against the male genome assembly. (b) PvM=pool RNA mapped against the male genome. Pre-MAKER results are the transcripts produced by Cufflinks. "In both" refers to BUSCO genes, which are in the same state in both the MvM and PvM annotations.

Even though equal total amounts of RNAseq reads were used as input for mapping, most of the 303 Eukarya genes examined by BUSCO had higher coverage in the pool sample than the male. Figure 3 shows an example that BUSCO classified as complete in the male (transcript CUFF.9217) while fragmented in the pool. In the pool sample, the coverage decreases dramatically in the segment between the fragmented transcripts CUFF.9941 and CUFF.9925, leaving a gap of 110bp with no RNAseq reads mapped, compared to a gap of 22bp in the male transcriptome in the same position. This may have resulted in Cufflinks failing to recognize them as part of the same gene. Coverage reduction near the edges of gene fragments appear in most genes classified as fragmented by BUSCO.

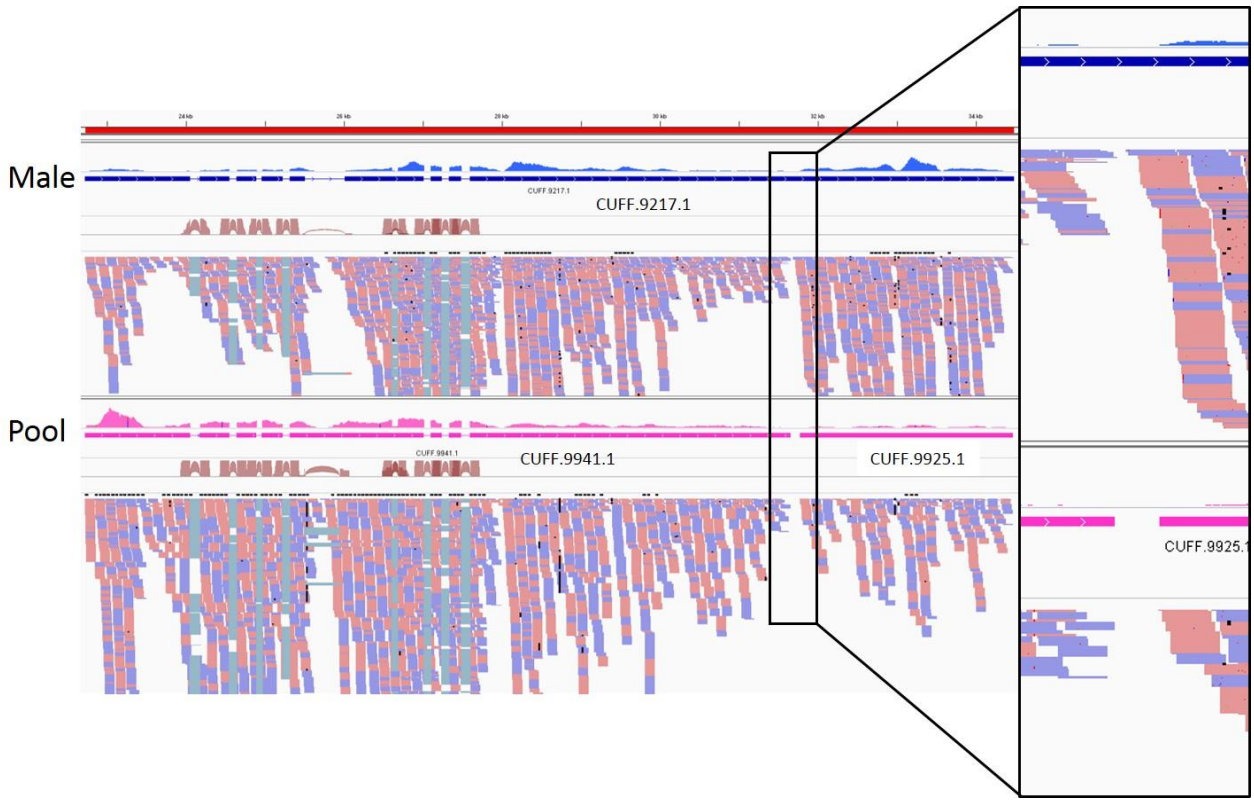

Figure 1: An example for a gene classified by BUSCO as complete in the male and fragmented in the pool. BUSCO gene 'EOG09370DXT'. The coverage data range is normalized to a range of 0-2500 reads per position.

Figure 4 shows a gene that was classified as fragmented in the male and complete in the pool sample. Transcript CUFF.11860.1 and CUFF.11860.2 in the male were not combined by Cufflinks as in the pool transcript CUFF.12758.1. In addition, the introns in segment two and three were not correctly defined.

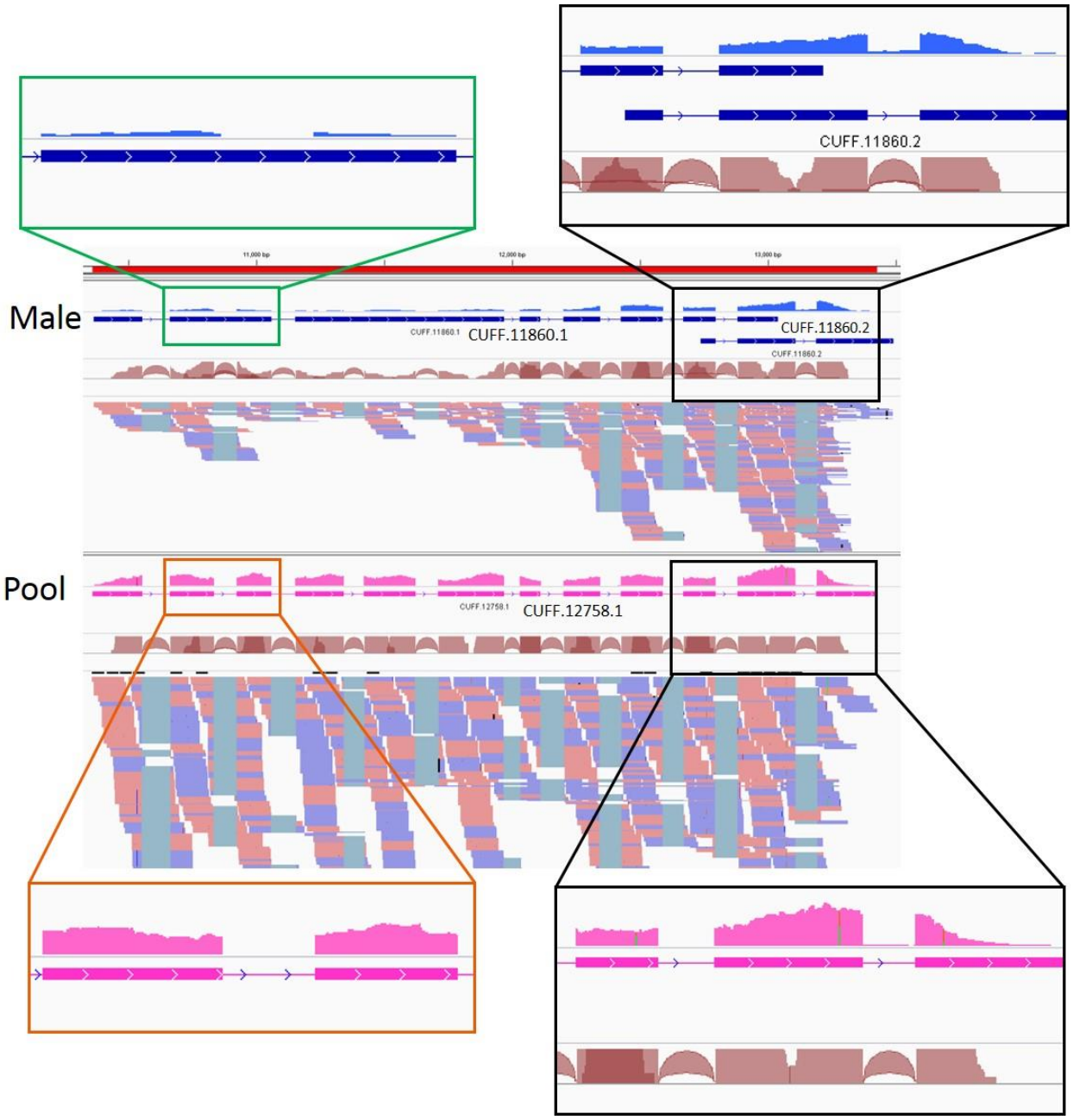

Figure 4: An example for a gene classified by BUSCO as complete in the pool and fragmented in the male. BUSCO gene 'EOG093706PM'. The coverage data range is normalized to a range of 0-1000 reads per position.

A third example was classified as complete in the male and missing in the pool (Figure 5). In this example, Cufflinks succeeded in constructing only the six last exons, in transcript CUFF.1340.1 in the pool sample, out of 13 in the male (CUFF.1095.1). A possible explanation is a drastic coverage drop in the pool from ~11K to less than a 1K reads per position.

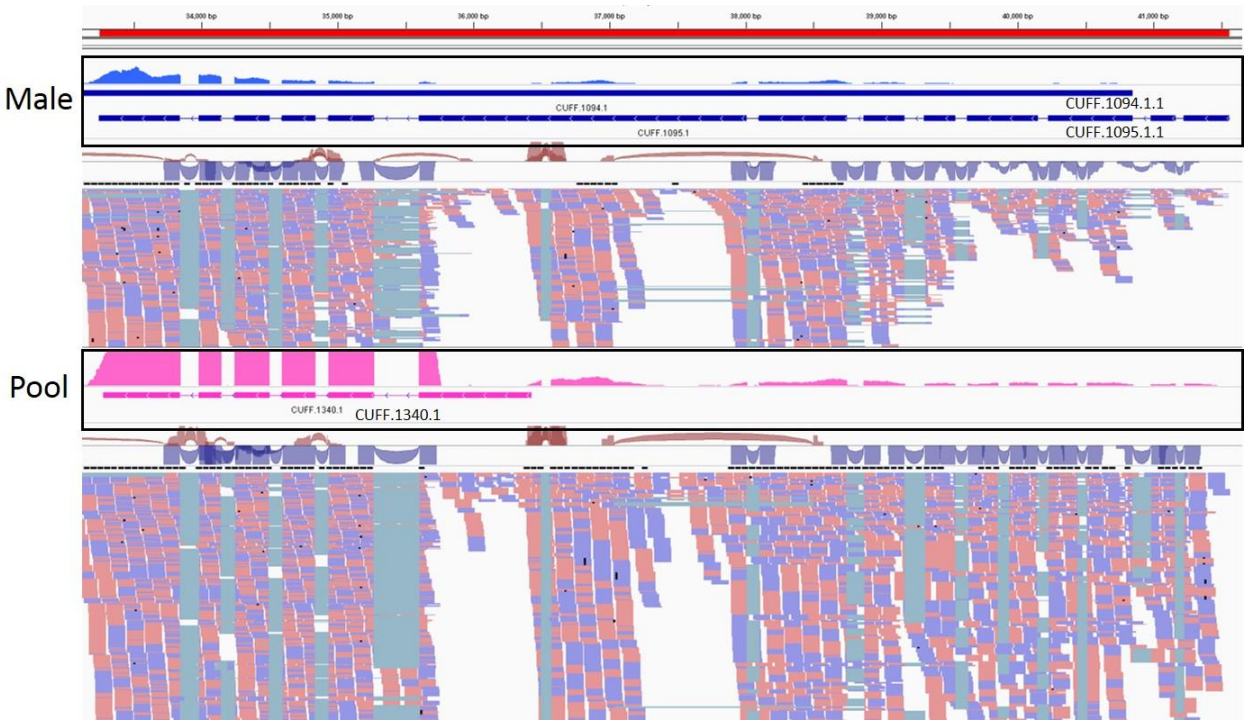

Figure 5: An example for a gene classified by BUSCO as complete in the male and missing in the pool. BUSCO gene 'EOG09370WZX'. The coverage data range is normalized to a range of 0-5000 reads per position.

Lastly, figure 6 shows a gene that was classified as missing in the male and complete in the pool. The coverage of the male sample dropped drastically (< 40 reads per position) and continued to be low along the gene, compared to the pool sample, which had higher coverage. Notice that the BUSCO gene examined is located on the negative strand (CUFF.43695.1) and its 3' end overlaps a different gene on the positive strand (CUFF.36947.1-2 on the male and CUFF.43694.1-2 on the pool; Figure 6).

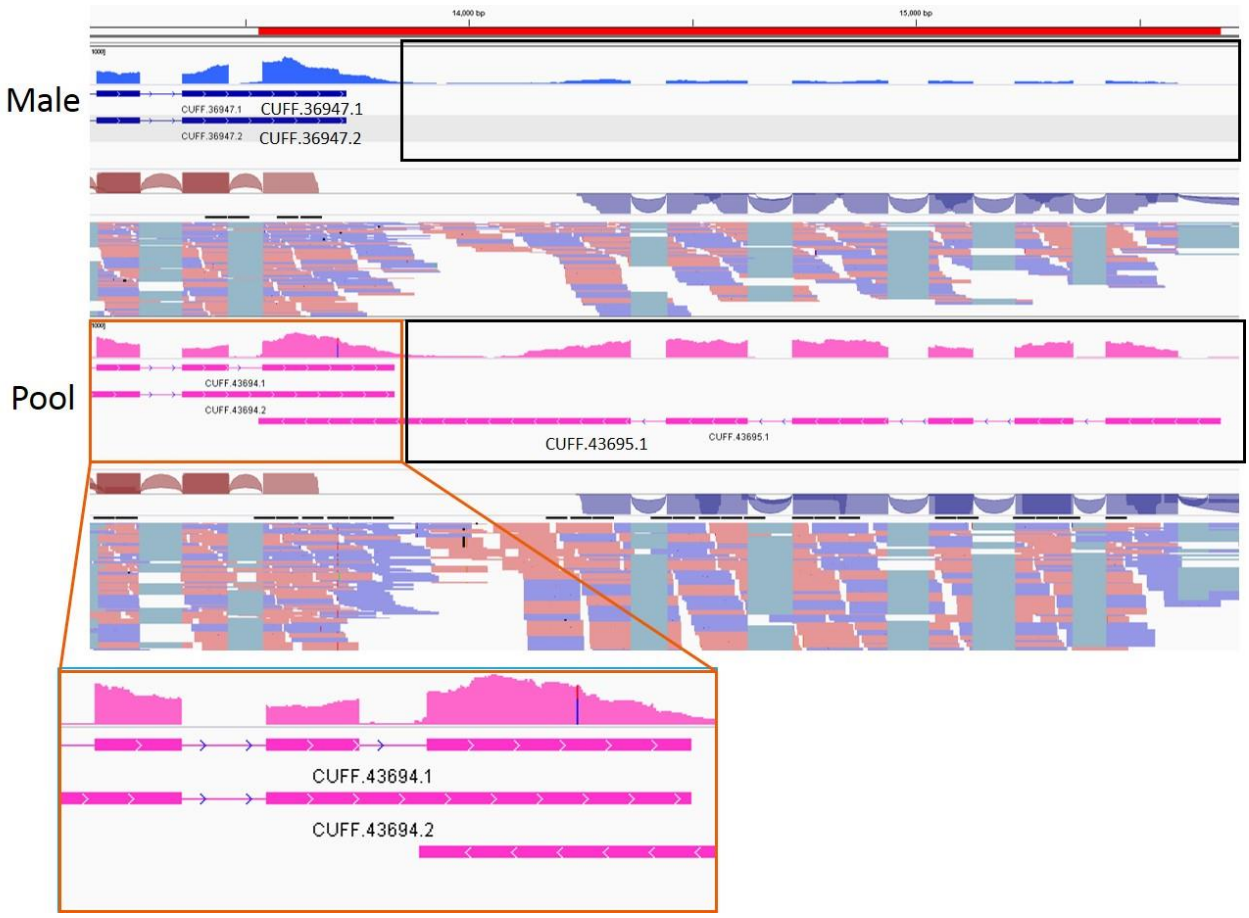

Figure 6: An example for a gene classified by BUSCO as missing in the male and complete in the pool. BUSCO gene 'EOG09370MQ0'. The coverage data range is normalized to a range of 0-1000 reads per position. Black rectangles marks the BUSCO gene examined. Orange rectangle marks the 3' end exons overlapping the BUSCO gene.

4.5.2 Polymorphism

As expected, many SNPs were observed in the pool sample and none in the male. Any base changes in the male sample could be associated with sequencing errors as they appear only in few of the reads. Figure 7 shows an example of several positions, which show SNPs in the pool sample only.

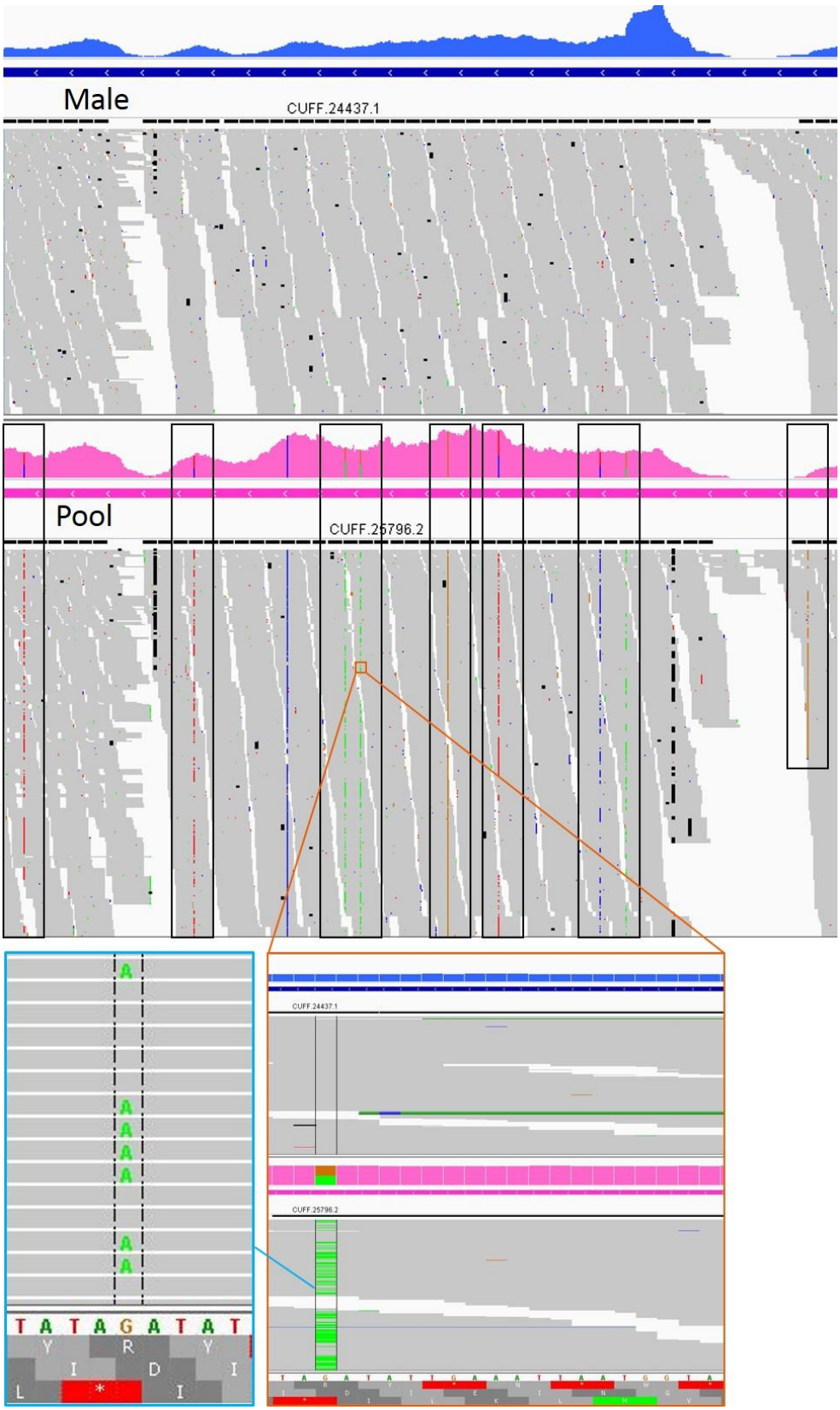

Figure 7: An example of a SNP in the pool transcriptome. The male genome and transcriptome both have a G at this position, while the pool RNAseq reads, have either A or G. The black rectangles highlight multiple additional SNPs.

4.5.3 Alternative splicing

Alternative splicing is a process in which genes can code for different transcripts and different protein isoforms by including or excluding exons in the mRNA. Complexity associated with alternative splice variants might be contributing to the fragmentation of transcripts assembled by Cufflinks. Figure 9 shows an alternatively spliced gene, and Figures 10a and 10b show a zoom-in on the alternative splice junctions. The number of splice variants, as seen by the number and location of the arcs, was higher in the pool sample. Read coverage in all splice junction in the pool was higher than the male sample. Among the genes examined only a few showed large differences in splicing, between male and pool.

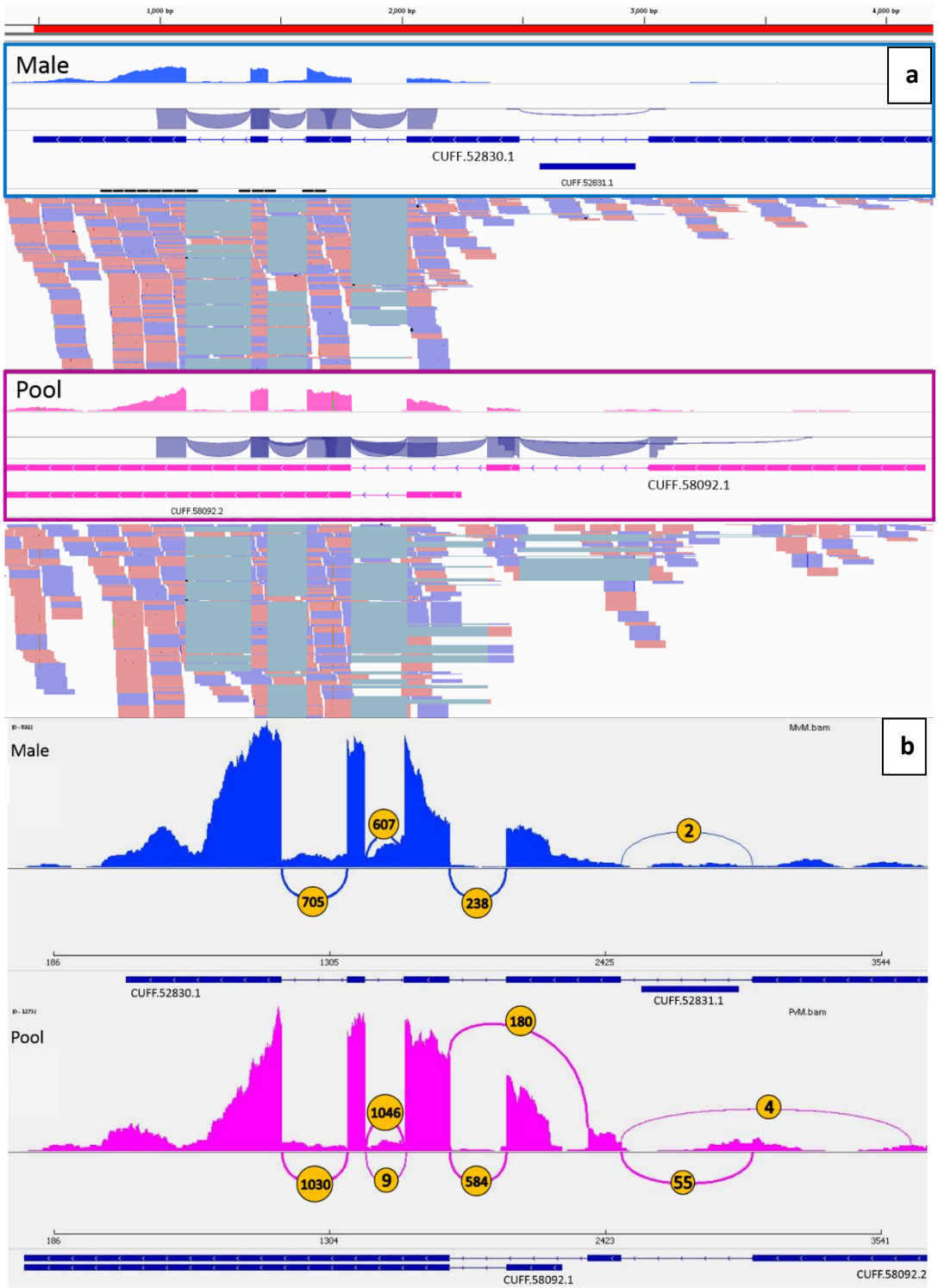

Figure 8: An example of alternative splicing in the pool transcriptome, but not the male. Curved arches (blue) below the center-line represent splice junctions on the negative strand of gene 'EOG093710JH'. Coverage is normalized to 0-2000 reads per position (a). Visualization of splice junctions using IGV Sashimi plots of male and pool transcripts of same gene (b). All the splice junctions are of on the negative strand of gene 'EOG093710JH'. Arcs represent splicing events. In orange circles are the number of reads splits across the splice junction. Height of bars between arcs represents exon coverage (reads per position).

5. Discussion

This thesis evaluated the utility of haploid samples as the source for both genomic and transcriptomic material in a *de novo* genome sequencing project. The results reveal which aspects of accuracy and completeness of the genomic draft benefit from the use of a haploid source.

5.1 Genome assembly

The greatest advantage of the haploid sample was as the source for genomic DNA for the genome assembly. Both SPAdes and SOAPdenovo2 achieved much greater contig and scaffold N50 sizes, by a three- to ten-fold factor, for the haploid relative to the diploid sample. This advantage does not seem to be accompanied by any noticeable cost such as more misassemblies. This result suggests that the lack of polymorphism in the haploid DNA sample facilitates assembly of longer contiguous genomic segments, whereas heterozygous sites in the diploid sample confuse the assembler by presenting multiple paths for contig extension.

Furthermore, the size of the diploid SPAdes assembly (759 Mb) is highly overestimated compared to the 220Mb flow cytometry estimate for the *C. hyspanica* genome. A likely main factor contributing to this bloating of the genome is polymorphic sequences (including SNPs, insertions/deletions, repetitive elements, rearrangements etc.) assembled by SPAdes separately for the two haplotypes of the diploid sample. This interpretation is supported by the high percentage (78%) of duplicated genes found by BUSCO (Table 5). The bloating of the genome and the number of duplicated genes are most likely due to the difficulty for the assembler in dealing with large eukaryotic, diploid genomes. SPAdes was originally designed for small, less repetitive bacterial genomes. It might not be able to cope as well as presumed with the complexity of diploid samples, which entails phasing or collapsing polymorphism. Apparently, the polymorphism in the *C. drusus* diploid sample was significantly more challenging for SPAdes relative to the haloid male sample. An extension of SPAdes, called dipSPAdes, that was designed for dealing with highly polymorphic diploid genomes, might have been a more capable solution (Safonova et al. 2014). The lack of longer insert size libraries in this study (i.e. mate-pair) or long read sequencing (i.e. PacBio) means the assembler had limited ability for scaffolding longer scaffolds and by that to increase the scaffold N50 size. Nevertheless, our study is a fair comparison of diploid and haploid DNA sources, and is

informative regarding the advantage of using a haploid DNA source, at least for the contiging stage of the assembly.

The large number of local misassemblies in both haploid and diploid SOAPdenovo2 assemblies, compared to SPAdes, can be associate with each assembler's application of the *de-Bruijn* graph approach. The unique PDBG approach of *k*-bimer adjustment, used by SPAdes, helps to overcome misassemblies. Analyzing pair reads reduces misassemlies created by chimeric read pairs (Bankevich et al. 2012). Apparently, the standard *de-Bruijn* approach used by SOAPdenovo2 tackles this issue less efficiently.

5.2 Transcripts to genome mapping

5.2.1 Male and pool transcripts

Overall, the male sample's gene annotation results, proved to be better than those of the pool. Both pre- and post-MAKER annotations had more genes classified as complete by BUSCO when using the male's RNA. Concomitantly, the number of fragmented genes is lower. Most of the genes classified as fragmented or missing by BUSCO are the result of split transcripts. This may be attributed to a drastic coverage drop, occurring at a certain point along the transcript, down to a critical level, which leads Cufflinks to split the gene to two fragments.

The improvement in the post-MAKER BUSCO results can be attributed to the use of BLAST against proteins databases for annotation based on homologous proteins from other species. For that reason, we decided to analyze the pre-MAKER transcript assembly by Cufflinks, as it allows evaluating gene annotation based by the RNAseq data alone. Conversely, there are genes that were classified as complete or fragmented pre-MAKER and post-MAKER were classified as missing. This should be checked in more in depth.

5.2.2 Complexity of the pool sample

As described in section 3.2.3, the RNA pool is composed of several castes and life stage from multiple nests (Appendix 2). This introduces additional complexity to the pool transcriptome relative to the simpler male transcriptome, in terms of both polymorphism and higher diversity of splice isoforms. Moreover, the multi-caste pool contains more splice isoforms because of caste-specific splicing. In several cases, BUSCO classifies transcripts as fragmented or missing, and while Cufflinks does not assemble the transcripts correctly or fully, the raw reads did align well to the genome (in the mapping by Tophat2). Cufflinks transcript reconstruction is guided by a model allowing alternative splicing. It constructs the most probable transcripts which are often left fragmented in order to explain all splice variants (Trapnell et al. 2012). Because of the complexity of the pool, it might be that Cufflinks did not associate the right isoforms as belonging to the same gene. This may explain some of the fragmented genes. For this reason, it may be advisable to avoid the use of a pooled sample. In order to lower the complexity of the pool each sample should be sequenced and assembled separately by Cufflinks. Later all assemblies can be combined using Cuffmerge (Trapnell et al. 2012).

Another factor which adds to the complexity of the pool RNA, and can affect Cufflink's decision process is the occurrences of RNA editing events. RNA editing is a post-transcriptional mechanism that adds complexity to the population of transcripts from a single genomics sequence. The most common RNA editing mechanisms is editing by *de-amination* of Adenosine to Inosine (A-to-I editing). Inosine is recognized by the ribosome as guanidine, and is also read as such by the polymerase (Laurencikiene et al. 2006; Li et al. 2014). Thus, in the RNAseq results, I is read as a G. When the complementary strand of the RNA is read, it appears as an edit from T to C. In ants, 8-23% of overall RNA editing sites are conserved and were suggested as a possible mechanism that contributed to the evolution of sociality. For example, RNA editing levels vary among castes and possibly underlie caste differences in morphology, physiology, and behavior (Li et al. 2014). Nevertheless, RNA editing cannot be distinguished from SNPs without using a designated analysis aimed to identify RNA editing. An initial impression out of the examined gene set is that there are putative RNA editing event apparent in the pool sample yet, none seen in the male sample.

In the *de novo* transcriptome assembly, the difference in insert sizes in the RNAseq libraries (larger for the pool) may have influenced the Trinity assemblies and caused the difference between the N50 of the male compare to the pool.

6. Conclusion

In conclusion, the ploidy of the source material used for *de novo* assembly of a genome can be an important factor in the design of a sequencing project. *de novo* assembling a genome using a haploid source (when possible as in the case of Hymenoptera) yields better results in terms of genome contiguity and correct representation of the full gene set without duplications. Furthermore, the use of RNA from the same individual for gene annotation will provide a more complete transcriptome. The approach of short read sequencing of DNA and RNA from the same individual, combined with other complementary long read sequencing and mapping methods, one can achieve optimal foundations for developing genomic infrastructures for new non-model organisms.

References

Amor F, Ortega P, Jowers MJ, Cerdá X, Billen J, Lenoir A, Boulay RR. 2011. The evolution of worker-queen polymorphism in Cataglyphis ants: Interplay between individual-and colony-level selections. Behav. Ecol. Sociobiol. 65:1473–1482.

Andrews S, others. 2010. FastQC: a quality control tool for high throughput sequence data.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, et al. 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J. Comput. Biol. 19:455–477.

Bohanec B. 2003. Ploidy determination using flow cytometry. Doubled Haploid Prod. Crop Plants A Man.:397–403.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120.

Bonasio R, Zhang G, Ye C, Mutti NS, Fang X, Qin N, Donahue G, Yang P, Li Q, Li C, et al. 2010. Genomic comparison of the ants Camponotus floridanus and Harpegnathos saltator. Science 329:1068–1071.

Brenchley R, Spannagl M, Pfeifer M, Barker GLA, Amore RD, Allen AM, Mckenzie N, Kramer M, Kerhornou A, Bolser D, et al. 2012. Analysis of the bread wheat genome using whole-genome shotgun sequencing. Nature 491:705–710.

Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, Holt C, Alvarado AS, Yandell M. 2008. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 18:188–196.

Chang CQ, Yesupriya A, Rowell JL, Pimentel CB, Clyne M, Gwinn M, Khoury MJ, Wulf A, Schully SD. 2013. A systematic review of cancer GWAS and candidate gene meta-analyses reveals limited overlap but similar effect sizes. Eur. J. Hum. Genet. 22:402–408.

Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, Chen HC, Agarwala R, McLaren WM, Ritchie GRS, et al. 2011. Modernizing reference genome assemblies. PLoS Biol. 9:1–5.

Darras H, Leniaud L, Aron S. 2014. Large-scale distribution of hybridogenetic lineages in a Spanish desert ant. Proc. R. Soc. B 281:20132396.

Doležel J, Bartoš J. 2005. Plant DNA flow cytometry and estimation of nuclear genome size. Ann. Bot. 95:99–110.

Elsik CG, Tayal A, Diesh CM, Unni DR, Emery ML, Nguyen HN, Hagen DE. 2016. Hymenoptera Genome Database: Integrating genome annotations in HymenopteraMine. Nucleic Acids Res. 44:D793–D800.

Goff SA, Ricke D, Lan T-H, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, et al. 2002. A Draft Sequence of the Rice Genome (Oryza sativa L. ssp. japonica). Science (80-.). 296:92–100.

Grabherr MG., Brian J. Haas, Moran Yassour Joshua Z. Levin, Dawn A. Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, Zehua Chen, Evan Mauceli, Nir Hacohen, Andreas Gnirke, Nicholas Rhind, Federica di Palma, Bruce W. N, Friedman and AR. 2013. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nat. Biotechnol. 29:644–652.

Graf J-D, Kobel HR. 1991. Chapter 2 Genetics of Xenopus laevis. In: Kay BK, Peng HBBT-M in CB, editors. Xenopus laevis: Practical Uses in Cell and Molecular Biology. Vol. 36. Academic Press. p. 19–34.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: Quality assessment tool for genome assemblies. Bioinformatics 29:1072–1075.

Hartl DL, Clark AG, Clark AG. 1997. Principles of population genetics. Sinauer associates Sunderland

Hillier LW, Miller W, Birney E, Warren W, Hardison RC, Ponting CP, Bork P, Burt DW, Groenen MAM, Delany ME, et al. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432:695–716.

Howe K, Wood JM. 2015. Using optical mapping data for the improvement of vertebrate genome assemblies. Gigascience 4:10.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2 : accurate alignment of transcriptomes in the presence of insertions , deletions and gene fusions. :1–13.

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge University Press

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9:357–359.

Laurencikiene J, Källman AM, Fong N, Bentley DL, Ohman M. 2006. RNA editing and alternative splicing: the importance of co-transcriptional coordination. EMBO Rep. 7:303–307.

Lee YS, Jeong H, Taye M, Kim HJ, Ka S, Ryu YC, Cho S. 2015. Genome-wide association study (GWAS) and its application for improving the genomic estimated breeding values (GEBV) of the berkshire pork quality traits. Asian-Australasian J. Anim. Sci. 28:1551–1557.

Leniaud L, Heftez A, Grumiau L, Aron S. 2011. Multiple mating and supercoloniality in Cataglyphis desert ants. Biol. J. Linn. Soc. 104:866–876.

Li Q, Wang Z, Lian J, Schiøtt M, Jin L, Zhang P, Zhang Y, Nygaard S, Peng Z, Zhou Y, et al. 2014. Caste-specific RNA editomes in the leaf-cutting ant Acromyrmex echinatior. Nat. Commun. 5:4943.

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 1:18.

Mark D. Adams, Susan E. Celniker, Robert A. Holt, Cheryl A. Evans JDG, Peter G. Amanatides, Steven E. Scherer, Peter W. Li, Roger A. Hoskins, Richard F. Galle, Reed A. George, Suzanna E. Lewis, Stephen Richards, Michael Ashburner, Scott N. Henderson GGS, Jennifer R. Wortman, Mark D. Yandell,, et al. 2000. The genome sequence of Drosophila melanogaster. Science (80-.). 287:2185—2195.

Medvedev P, Pham S, Chaisson M, Tesler G, Pevzner P. 2011. Paired de Bruijn graphs: A novel approach for incorporating mate pair information into genome assemblers. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 6577 LNBI:238–251.

Miller JR, Koren S, Sutton G. 2010. Genomics Assembly algorithms for next-generation sequencing data. Genomics 95:315–327.

Myers EW, Sutton GG, Delcher a L, Dew IM, Fasulo DP, Flanigan MJ, Kravitz S a, Mobarry CM, Reinert KH, Remington K a, et al. 2000. A whole-genome assembly of Drosophila. Science 287:2196–2204.

Nygaard S, Zhang G, Schiøtt M, Li C, Wurm Y, Hu H, Zhou J, Ji L, Qiu F, Rasmussen M, et al. 2011. The genome of the leaf-cutting ant Acromyrmex echinatior suggests key adaptations to advanced social life and fungus farming. Genome Res. 21:1339–1348.

Oxley PR, Ji L, Fetter-Pruneda I, McKenzie SK, Li C, Hu H, Zhang G, Kronauer DJC. 2014. The genome of the Clonal raider ant Cerapachys Biroi. Curr. Biol. 24:451–458.

Parra G, Bradnam K, Ning Z, Keane T, Korf I. 2008. Assessing the gene space in draft genomes. 37:289–297.

Pearcy M, Clémencet J, Chameron S, Aron S, Doums C. 2004. Characterization of nuclear DNA microsatellite markers in the ant Cataglyphis cursor. Mol. Ecol. Notes 4:642–644.

Putnam NH, Butts T, Ferrier DEK, Furlong RF, Hellsten U, Kawashima T, Robinson-Rechavi M, Shoguchi E, Terry A, Yu J-K, et al. 2008. The amphioxus genome and the evolution of the chordate karyotype. Nature 453:1064–1071.

Rispail N, Dita MA, González-Verdejo C, Pérez-De-Luque A, Castillejo MA, Prats E, Román B, Jorrín J, Rubiales D. 2007. Plant resistance to parasitic plants: Molecular approaches to an old foe: Research review. New Phytol. 173:703–712.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. Nat. Biotechnol. 29:24.

Safonova Y, Bankevich A, Pevzner PA. 2014. dipSPAdes: Assembler for Highly Polymorphic Diploid Genomes. In: Sharan R, editor. Research in Computational Molecular Biology: 18th Annual International Conference, RECOMB 2014, Pittsburgh, PA, USA, April 2-5, 2014, Proceedings. Cham: Springer International Publishing. p. 265–279.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. 2015. BUSCO: user guide. Bioinformatics 31:3210–3212.

Simpson JT. 2014. Exploring genome characteristics and sequence quality without a reference. Bioinformatics 30:1228–1235.

Simpson JT, Pop M. 2015. The Theory and Practice of Genome Sequence Assembly. Annu. Rev. Genomics Hum. Genet. 16:153–172.

Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. 2014. Sequencing depth and coverage: key considerations in genomic analyses. Nat. Rev. Genet. 15:121–132.

Soller M, Weigend S, Romanov MN, Dekkers JCM, Lamont SJ. 2006. Strategies to assess structural variation in the chicken genome and its associations with biodiversity and biological performance. Poult. Sci. 85:2061–2078.

Steinberg KM, Schneider VK, Graves-lindsay TA, Schneider VA, Robert S, Agarwala R, Huddleston J. 2014. Single haplotype assembly of the human genome from a hydatidiform mole Single haplotype assembly of the human genome from a hydatidiform mole. :2066–2076.

Sud A, Kinnersley B, Houlston RS. 2017. Genome-wide association studies of cancer : current insights and future perspectives. Nat. Publ. Gr. 17:692–704.

The Bovine Genome Sequencing and Analysis Consortium, Elsik, C. G.; Tellam, R. L.; Worley KC. 2009. The Genome Sequence of Taurine. Science (80-.). 324:522–529.

Timmermans I, Grumiau L, Hefetz a., Aron S. 2009. Mating system and population structure in the desert ant Cataglyphis livida. Insectes Soc. 57:39–46.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: Discovering splice junctions with RNA-Seq. Bioinformatics 25:1105–1111.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. 7:562–578.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotech 28:511–515.

Venter J, Adams M, Myers E, Li P, Mural R, Sutton G, Smith H, Yandell M, Evans C, Holt R, et al. 2001. The Sequence of the Human Genome. Science (80-.). 291:1304.

Warner MR, Mikheyev AS, Linksvayer TA. 2017. Genomic Signature of Kin Selection in an Ant with Obligately Sterile Workers. Mol. Biol. Evol. 34:1780–1787.

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562.

van Wilgenburg E, Driessen G, Beukeboom L. 2006. Single locus complementary sex determination in Hymenoptera: an “unintelligent” design? Front. Zool. 3:1–15.

Wurm Y, Wang J, Riba-Grognuz O, Corona M, Nygaard S, Hunt BG, Ingram KK, Falquet L, Nipitwattanaphon M, Gotzek D, et al. 2011. The genome of the fire ant Solenopsis invicta. Proc. Natl. Acad. Sci. U. S. A. 108:5679–5684.

Appendices

Appendix 1: a satellite image of the research area in *Betzet* beach (33°4'40.88"N / 35°6'33.97"E; Google earth). Red dot marks nest BZT4B, from which the male and worker sample for the reference genome were taken.

Appendix 2: RNA pool composition with quantity and purity measurements using ND200. Total volume of pool sample for sequencing was 60µl. only four samples were diluted (Larva_1, Larva_2, Larva_4, Larve_5). 260/280 ratio < 2 indicate protein residue contamination. 260/230 ratio < 2.2 indicates chemical contamination.

Sample type	Nucleic Acid concentration [ng/µl]	260/280 ratio	260/230 ratio	Total amount of RNA in sample [ng/50µl]	concentration After x10 dilusion [ng/µl]	volume taken after dilusion [µl]
Larva_1 (small size)	1052.3	2.13	2.24	52615	105.23	5
Larva_2 (small size)	1790	2.13	2.09	89500	179	3
Larva_3 (medium size)	661.6	2.08	2.01	33080	N/A	1
Larva_4 (medium size)	1243.2	2.11	1.77	62160	124.32	5
Larva_5 (large size)	1356.3	2.12	1.97	67815	135.63	5
Pupa_1 (small size)	83.8	2.08	0.92	4190	N/A	7
Pupa_2 (small size)	102.7	2.11	0.94	5135	N/A	5
Pupa_3 (large size)	215.1	2.11	1.15	10755	N/A	5
Gyne	57.2	2.06	0.89	2860	N/A	20
Male	146.5	2.2	1.85	7325	N/A	10
Worker	105.7	2.02	1.1	5285	N/A	10
Total RNA_POOL-A sample	106.6	2.03	1.08			60
Total RNA_POOL-B sample	102	2.01	1.11			60

Appendix 3: All Prep Mini DNA and RNA modified protocol for *Cataglyphis* ants

Amount of starting material Volume of Buffer RLT
20–30 mg 600 µl (an average *cataglyphis* worker is ~20 mg)

Tissue disruption:

1. Disrupt the tissue and homogenize the lysate in RLT.
*** add 10 µl of β-ME to 1ml of RLT buffer before use.

Disruption and homogenization using the Tissue Lysser >>> 3 metal beads + shredded glass powder + glass beads (SIGMA) inside 1.5-2 ml tube. Snap frozen tissue is put inside tube and in liquid nitrogen before disruption (can be repeated).

Tissue Lysser >>> 30 Hz for cycles of 20 sec (or 50 sec) **dry cycles** + **wet cycles** with addition of 100ul of RLT+ β-ME.

After Tissue Lysser add 700 µl of RLT and incubate in 65°C for 30-60 min.

2. Centrifuge >>> lysate for 3 min, max speed.
Carefully remove the supernatant by pipetting, and transfer it to the AllPrep DNA spin column placed in a 2 ml collection tube.

Centrifuge >>> 30 sec, max speed (can use ~ 10,000 g)

3. Place the AllPrep DNA spin column in a new 2 ml Eppendorf tube and store at 4°C for later DNA purification. **Do not freeze the column**.

***Use the flow-through for RNA purification.

Total RNA purification:

4. Add 1 volume (~700 µl) of 100% frozen ethanol (instead of 70%) to the flow through, and mix well by pipetting. **Do not centrifuge.** Incubate in -20°C for 10-15min.

5. Transfer up to 700 µl of the sample, including any precipitate that may have formed, to an RNeasy spin column placed in a 2 ml Collection tube.

Centrifuge >>> 30 sec, max speed (Discard the flow through).

*** If the sample volume exceeds 700 µl, centrifuge successive aliquots in the same RNeasy spin column. Discard the flow-through after each centrifugation.

DNase Treatment (optional):

E1. Add 350 µl Buffer RW1 to the RNeasy spin column, and centrifuge for 15 s at ≥8000 x *g* (≥10,000 rpm) to wash the spin column membrane. Discard the flow-through.*

Reuse the collection tube in step E4.

E2. Add 10 µl DNase I stock solution (see above) to 70 µl Buffer RDD. Mix by gently inverting the tube, and centrifuge briefly to collect residual liquid from the sides of the tube.
Buffer RDD is supplied with the RNase-Free DNase Set.
Note: DNase I is especially sensitive to physical denaturation. Mixing should only be carried out by gently inverting the tube. Do not vortex.

E3. Add the DNase I incubation mix (80 µl) directly to the RNeasy spin column membrane, and incubate at room temperature (20–30°C) for 15 min.

***Note:** Be sure to add the DNase I incubation mix directly to the RNeasy spin column membrane. DNase digestion will be incomplete if part of the mix sticks to the walls or the O-ring of the spin column.

E4. Add 350 µl Buffer RW1 to the RNeasy spin column, and centrifuge for 15 s at ≥8000 x *g* (≥10,000 rpm). Discard the flow-through.*
Continue with step 9 of the protocol on page 26 (i.e., the first wash with Buffer RPE).
Reuse the collection tube in step 9.

Optional:
6. Add 700 µl Buffer RW1 to the RNeasy spin column to wash the spin column membrane.

Centrifuge >>> 30 sec, max speed (Discard the flow-through).

7. Add 450 µl RPE to the RNeasy spin column (can do 2-3 washes)
Reuse the collection tube in step 10.

8. Add 450 µl RPE to the RNeasy spin column.

Centrifuge >>> 2 min, max speed (Discard the flow-through).

9. Place the RNeasy spin column in a new 2 ml collection tube (discard the old collection tube with the flow through).

Centrifuge >>> 1 min, max speed.

10. Place the RNeasy spin column in a new 1.5-2 ml collection tube Add 40 µl RNase-free water directly to the spin column membrane. Incubate for 10-15 min in RT.

Centrifuge >>> 1 min, max speed.

Genomic DNA purification:

11. Add 500 µl Buffer AW1 to the AllPrep DNA spin column from step 5.

Centrifuge >>> 30 sec, max speed. (Discard the flow-through).

Note: Buffer AW1 is supplied as a concentrate. Ensure that ethanol is added to Buffer AW1 before use.

12. Add 500 µl Buffer AW2 to the column.
Centrifuge >>> 2 min, max speed. (Discard the flow-through).

13. Place the AllPrep DNA spin column in a new 1.5 ml collection tube
Add 50 µl Buffer EB (preferably use 2mM tris heated to 55°C) directly to the spin column membrane and close the lid. Incubate 10-15 min at RT.

Centrifuge >>> 1min, max speed.

Appendix 4: DNA and RNA ND2000 measurements results for the male candidate samples and Worker DNA measurements

Sample type	Nucleic Acid concentration [ng/µl]	260/280 ratio	260/230 ratio	Total amount of DNA/RNA in sample [ng/40µl]
DNA				
Male_BZT4B	121.824	2.229	0.649	4872.96
Male_BZT4C	45.652	2.227	0.735	1826.08
Male_BZT7	44.861	2.186	0.899	1794.44
Worker_BZT4B	278.8	2.12	1.9	8364
RNA				
BZT4B_M1	660.113	2.274	1.512	26404.52
BZT4C_M2	1038.168	2.26	2.247	41526.72
BZT7_M2	1065.382	2.277	1.518	42615.28

Appendix 5: List of commands used in the various pipelines in the genome and transcriptome assembly as well as quality assessment.

```
'java -jar trimmomatic-0.36.jar PE -phred33 ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:80'

'java -jar trimmomatic-0.36.jar PE -phred33 ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:60'

'spades.py --pe1-1 < 300bp R1.fastq > --pe1-2 < 300bp R2.fastq > --pe2-1 < 550bp R1.fastq > --pe2-2 < 550bp R2.fastq > -k 13,23,33,43,53,63,73,83,93,103,113,123 -t 20 -o <output folder>'

'SOAPdenovo-127mer all -s cdru_W4B.config -K 115 -R -p 20 -o cdru_W4B 1>cdru_W4B.log 2>cdru_W4B.err'

'python3 BUSCO_v1.22.py -o dru_M4B_4_3_2017_SOAP -in /data/home/ASSEMBLIES/SOAPdenovo2/cdru_M4B_28_2_2017/cdru_M4B.scafSeq  -l /data/home/SOFTWARES/BUSCO_v1.22/arthropoda -m genome -c 20'

'python quast.py -e -t 20 -o /data/home/ASSEMBLIES/QUAST/cdru_W4B_spades_4_3_2017 /data/home/ASSEMBLIES/SPades/W4B_K13-123_28_2_2017_B/scaffolds.fasta /data/home /ASSEMBLIES/SPades/W4B_K13-123_28_2_2017_B/contigs.fasta -R /data/home /ANTS_RAW_DATA/ANTS_ASSEMBLIES/Chis1_v1.0.sorted.fa'

'bowtie2-build --threads 20 -f /data/home/ASSEMBLIES/SPades/W4B_K13-123_28_2_2017_B/scaffolds.fasta cdru_W4B'

'gffread transcripts.gff3 -g /data/home/C_DRUSUS_TAL/ASSEMBLIES/DNA/SPades/M4B_K13-123_28_2_2017_B/scaffolds.fasta -y transcripts_2.fasta -M'

'sed -e 1~16d;2~16d;3~16d;4~16d'

'tophat -p 20 M4B_K13-123_28_2_2017_B_scaffolds /data/home/C_DRUSUS_TAL/7_RNA_M4B_PE125_IS500/RAW_DATA/RNA_M4B_Pst_R1.fastq /data/home/C_DRUSUS_TAL/7_RNA_M4B_PE125_IS500/RAW_DATA/RNA_M4B_Pst_R2.fastq'

'cufflinks --no-update-check -p 20 /data/home/ASSEMBLIES/RNA/bowtie2_spades_12_8_2017/M4B_Vs_M4B_asm/tophat_out/accepted_hits.bam'
```

Appendix 6: Species homolog proteins used for MAKER annotation

Species name	Databae
Acromyrmex echinatior	http://www.antgenomes.org
Atta cephalotes	http://www.antgenomes.org
Harpegnathos saltator	http://www.antgenomes.org
Lasius niger	http://www.antgenomes.org
Linepithema humile	http://www.antgenomes.org
Pogonomyrmex barbatus	http://www.antgenomes.org
Drosophila melanogaster	https://www.ncbi.nlm.nih.gov/
Apis mellifera	https://www.ncbi.nlm.nih.gov/
Polistes dominula	https://www.ncbi.nlm.nih.gov/

Appendix 7: N50 of various *Camponotus* species for comparison to *Cataglyphis* drusus *de novo* transcriptome.

specie	N50	Number of genes
Camponotus aethiops	1542	35185
Camponotus japonicus	2271	43035
Camponotus ligniperdus	1853	34839
Camponotus castaneus	2831	51328

הפלואידי או דיפלואידי? ניתוח השוואתי של שיטות הרכבה לגנום של חרקים מסדרת הדבוראים

טל יהב

תקציר

הרכבה מחדש או אסמבלי מחדש (*de novo* assembly) של גנום שלם מהווה תשתית חיונית למגוון רחב של מחקרים גנטיים. מסדרת הדבוראים כבר רוצפו מספר גנומים ובמינים מסוימים (*Acromyrmex echinatior* ו- *Solenopsis invicta*) אף השתמשו לריצוף בדגימת דנ"א של זכר הפלואידי, מתוך הנחה כי הפלואידיות הגנום הזכרי מקנה יתרון להרכבת הגנום. עבור ריצוף רנ"א השתמשו במקבץ (pool) של פרטים מקאסטות ושלבי חיים שונים.

מחקר זה הוא ניתוח השוואתי של אסמבלי גנום וטרנסריפטום, בנוסף לאנוטציה של הגנום (annotation), עבור הנמלה *'נווטת שחורה'* (*Cataglyphis drusus*), באמצעות שימוש בחומר גנטי ממקורות בעלי פלואידיות שונה: (1) עבור ריצוף דנ"א נעשה שימוש בזכר הפלואידי ובנקבה דיפלואידית; (2) עבור ריצוף רנ"א נעשה שימוש באותו זכר הפלואידי ובמקבץ דגימות של פרטים מקאסטות ושלבי חיים שונים, מקנים שונים. הגישה שלנו ייחודית בכך שנעשה שימוש בזכר יחיד עבור ריצוף דנ"א ורנ"א. דגימת הדנ"א הדיפלואידית נלקחה מפועלת אחות של הזכר. הנחת העבודה העיקרית היא שהשימוש בדגימת זכר הפלואידי לעומת נקבה דיפלואידית, והפקת דנ"א ורנ"א מאותו פרט הפלואידי, יפשטו את תהליך האסמבלי של הגנום וכן את תהליך האנוטציה, זאת הודות לחוסר בהטרוזיגויות בדגימה הזכרית. שימוש באותו מקור לדנ"א ולרנ"א אמור לספק תוצאה טובה יותר בהעמדת הטרנסקריפטים כנגד הגנום, ולהקל על שיערוך מבנה הגנים במונחים של גבולות אקסון/אינטרון. גישה חדשה זו מנצלת מאפיין גנומי ייחודי של סדרת הדבוראים, קרי, האפלודיפלואידיות (haplodiploidy).

בהערכת האיכות הכוללת של כל אסמבלי ממקור שונה, נעשה שימוש בשלוש שיטות הערכה: (1) חישוב רציפות הגנום (N50) של קונטיגים (contigs) וסקאפולדים (scaffolds); (2) הערכת השלמות (completeness) של האסמבלי; (3) זיהוי טעויות הרכבה באסמבלי (misassemblies). האסמבלי ההפלואידי נמצא רציף יותר, עם גודל N50 לפחות פי שלושה ממקבילו הדיפלואידי. הערכת השלמות הראתה תוצאות מעורבות בעוד שבאסמבלי של SPAdes היו יותר גנים שלמים, היתה גם רמה גבוהה יותר של דופליקציות, יחד עם ניפוח משמעות של גודל הגנום. גילוי טעויות אסמבלי הראה תוצאות מעורבות, ונמצאו מספר גדול בהרבה של טעויות אסמבלי מקומיות ע"י SOAPdenovo2. אסמבלי הטרנסקריפטום של המקבץ נתן תוצאות N50 טובות יותר, כמו כן שלמות טובה יותר. לבסוף, בהשוואה בין שני אסמבלי טרנסקריפטום של הדגימה הזכרית והמקבץ, בזכר נמצאה שלמות טובה יותר. תוצאה זו יכולה להיות

מוסברת ע״י המורכבות של דגימת המקבץ לרבות פולימורפיזם, שחבור חלופי (alternative splicing) ועריכת רנ"א שהקשו על הרכבת הטרנסקריפטום.

לסיכום, הפלואידיות של חומר המקור המשמש לאסמבלי של גנום הוא בעל תרומה משמעותית לאיכות האסמבלי. יתרה מכך, השימוש של דנ"א ורנ"א מאותו מקור הפלואידי מביא לאנוטציה טובה יותר של הגנום.

הפלואידי או דיפלואידי?
ניתוח השוואתי של שיטות הרכבה
לגנום של חרקים מסדרת הדבוראים

מאת: טל יהב

מנחה: דר' איל פריבמן

עבודת גמר מחקרית (תזה) המוגשת כמילוי חלק מהדרישות לקבלת התואר "מוסמך האוניברסיטה"

אוניברסיטת חיפה

הפקולטה למדעי הטבע

החוג לביולוגיה אבולוציונית וסביבתית

נובמבר, 2017

הפלואידי או דיפלואידי?
ניתוח השוואתי של שיטות הרכבה
לגנום של חרקים מסדרת הדבוראים

טל יהב

עבודת גמר מחקרית (תזה) המוגשת כמילוי חלק מהדרישות לקבלת התואר "מוסמך האוניברסיטה"

אוניברסיטת חיפה

הפקולטה למדעי הטבע

החוג לביולוגיה אבולוציונית וסביבתית

נובמבר, 2017