NASONIA VITRIPENNIS: AN EMERGING MODEL ORGANISM FOR CHRONOBIOLOGY

Thesis submitted for the degree of

Doctor of Philosophy

At the University of Leicester

2016

Nathaniel Jude Davies
Department of Genetics
University of Leicester
Abstract

The circadian clock co-ordinates a wide range of biological processes in animals and as such has implications for a range of research topics, including human disease. Recent research has shown that *Drosophila*, the most widely used insect clock model, may be somewhat esoteric, suggesting that other insects may make more suitable models. The work presented in this thesis advances the jewel wasp *Nasonia vitripennis* as a model for circadian research. Firstly, the response of the *Nasonia* circadian clock to temperature was characterised, revealing the presence of a temperature compensation mechanism more strict in constant darkness than in constant light. This work was followed up by the profiling of the circadian transcriptome in *Nasonia* in both of these constant conditions. This work revealed fundamental differences in the dynamics of circadian transcription between *Drosophila* and *Nasonia*, as well as identifying temporal separation of biological function in the wasp, most notably of anabolic and catabolic processes. Secondly, conserved upstream non-coding sequences were identified and analysed, shedding light on the mechanisms of transcriptional and translational control in *Nasonia*. This analysis revealed conservation of the regulatory elements which control regulatory genes, indicating the presence of conserved regulatory cascades. This work also reports the identification and analysis of conserved regulatory mechanisms in RNA, including conserved secondary structures, carrying implications outside of chronobiology. Thirdly and finally, a gene-focused database was created for *Nasonia* to facilitate such research. Genome annotation data was processed and combined with functional data from published RNA-seq and microarray datasets, along with the original analysis of a genome-wide DNA methylation dataset, and the implementation of several tools.
Acknowledgements

Firstly, I would like to thank my supervisor Dr Eran Tauber for his invaluable help and advice throughout my PhD. The (many) lessons taught to me by Eran will doubtlessly stay with me throughout my career, and I will continue to build on them. One day I may even be able to entirely rid my writing of unnecessary semicolons; but then again perhaps not.

Dr Sascha Ott deserves my thanks for his supervision during my MIBTP mini-project, and for many interesting and useful conversations both during and after the project.

I’d like to thank my family and friends. If you don’t see your name specifically elsewhere, this is where I’ve thanked you.

Special thanks go to: Mirko, for interesting chats, help, and advice; Merav, for making my time at Leicester that bit more pleasant; Najwa; for helping with the collection of the temperature compensation samples; Laura, without whom my PhD (and LiveGene!) would have been a far more solitary experience; ‘The Lunch Group’, for giving me something to look forward to at 12:30; my mum, without whom none of this would have been possible; my late grandma, who I wish could see me graduate; Georgia, for so many things; and Ben, my work wife.

Super special thanks go to Raju for his wisdom, courage, and unwavering support.
As part of this PhD, work in this thesis has been published in peer-reviewed journals. Only analyses and writing which were a product of my own efforts have been included in this thesis. Contributions to each publication are outlined below. Publications related to this PhD are listed below.

Chapter 3

http://dx.doi.org/10.1101/048736

Contributions: NJD: Performed experiments, performed analysis, guided project, wrote & revised manuscript. ET: Supervision, guided project, revised manuscript.

Chapter 4


Contributions: NJD: Performed analysis, wrote software, guided project, wrote & revised manuscript. PK: Wrote software, revised manuscript. ET: Supervision, guided project, revised manuscript. SO: Supervision, wrote software, guided project, revised manuscript.
Chapter 5


Contributions: NJD: Wrote software, performed analysis, guided project, wrote & revised manuscript. ET: Supervision, guided project, revised manuscript.


Contributions: MP: Performed experiments, performed analysis, wrote & revised manuscript. AB: Performed experiments, performed analysis, wrote & revised manuscript. NJD: Performed RRBS analysis, wrote RRBS section of manuscript, revised manuscript. DMS: Supervision, guided project, revised manuscript. ET: Supervision, guided project, revised manuscript.

This thesis is my original work and has not been submitted, in whole or in part, for a degree at this or any other university. I confirm that all of the work recorded in my thesis is original unless otherwise acknowledged in the text or by references.
# Table of contents

Abstract ii  
Acknowledgements iii  
Preface iv  
Table of contents vi  
List of figures viii  
List of tables xi  
Common abbreviations xii  

## 1 General introduction 1-29

1.1 Circadian rhythms 1  
1.2 The genetic basis of animal circadian rhythms 3  
1.3 Clock network and outputs 7  
1.4 Circadian research in other insects 10  
1.5 *Nasonia*: a solitary parasitoid wasp 24  
1.6 Making up ground on *Drosophila* 27  

## 2 General methods 30-37  

## 3 Circadian behaviour and transcription 38-70  

3.1 Introduction 38  
3.2 Methods 40  
3.3 Results 47  
3.4 Discussion 65
<table>
<thead>
<tr>
<th>4</th>
<th>Conserved regulatory elements</th>
<th>71-107</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Introduction</td>
<td>71</td>
</tr>
<tr>
<td>4.2</td>
<td>Methods</td>
<td>75</td>
</tr>
<tr>
<td>4.3</td>
<td>Results</td>
<td>87</td>
</tr>
<tr>
<td>4.4</td>
<td>Discussion</td>
<td>104</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>5</th>
<th>Genomic database development</th>
<th>108-134</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>Introduction</td>
<td>108</td>
</tr>
<tr>
<td>5.2</td>
<td>Methods</td>
<td>113</td>
</tr>
<tr>
<td>5.3</td>
<td>Results</td>
<td>124</td>
</tr>
<tr>
<td>5.4</td>
<td>Discussion</td>
<td>132</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>6</th>
<th>General discussion</th>
<th>135-146</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1</td>
<td>The diversity of insect molecular clocks</td>
<td>137</td>
</tr>
<tr>
<td>6.2</td>
<td>The <em>Nasonia</em> clock in context</td>
<td>139</td>
</tr>
<tr>
<td>6.3</td>
<td>Beyond the clock: wider implications</td>
<td>141</td>
</tr>
<tr>
<td>6.4</td>
<td>Future work</td>
<td>142</td>
</tr>
<tr>
<td>6.5</td>
<td>Concluding remarks</td>
<td>146</td>
</tr>
</tbody>
</table>

| 7 | Appendices | 147-153 |
| 8 | Bibliography | 154-204 |
List of figures

Chapter 1

Figure 1.1. *Drosophila* circadian clock. 5
Figure 1.2. Mammalian circadian clock 6
Figure 1.3. Clock output pathways. 10
Figure 1.4. Phylogenetics of major insect orders. 11
Figure 1.5. *Nasonia* life cycle. 25

Chapter 2

Figure 2.1. Diagram of DAM Trikinetics setup. 31
Figure 2.2. Survival curves for various foods. 32
Figure 2.3. Hierarchical clustering method. 35
Figure 2.4. Reciprocal best blast hits. 36
Figure 2.5. Ortholog identification methods. 37

Chapter 3

Figure 3.1. Wasp activity in experimental conditions. 42
Figure 3.2. Effect of temperature on circadian period. 47
Figure 3.3. Expression clusters in DD. 50
Figure 3.4. Expression clusters in LL. 51
Figure 3.5. Rhythmic expression clusters. 54
Figure 3.6. Dynamics of circadian transcription. 55
Figure 3.7. Phase distributions in JTK cycle. 56
Figure 3.8. Cycling amplitudes in JTK cycle. 57
Figure 3.9. Transcriptional profiles of clock genes. 59
Figure 3.10. Functions of rhythmic genes. 61
Figure 3.11. Evolution and expression of opsin genes. 62
Figure 3.12. Differences between DD and LL. 64

Chapter 4
Figure 4.1. Alignment strategy. 75
Figure 4.2. CNE bundling strategy. 79
Figure 4.3. Phylogeny of comparator species. 88
Figure 4.4. CNE sequence properties. 90
Figure 4.5. CNEs and nucleosome occupancy. 91
Figure 4.6. TFBS and GC content. 91
Figure 4.7. ATG underrepresentation in CNEs. 92
Figure 4.8. Conservation of position. 93
Figure 4.9. Conserved RNA structure in ferritin. 95
Figure 4.10. Conserved hairpin in paramyosin. 96
Figure 4.11. Conserved hairpin in not1. 97
Figure 4.12. Conserved CNEs in the osiris cluster. 99
Figure 4.13. Phylogenetic distribution of ancient CNEs. 100
Figure 4.14. Evolution of RPLP1 CNE. 102
Figure 4.15. Evolution of RPLP2 CNE. 103

Chapter 5
Figure 5.1. PFAM overlap procedure. 115
Figure 5.2. RRBS protocol. 119
Figure 5.3. Overview of gene information page. 125
Figure 5.4. Analysis with WaspAtlas. 126
Figure 5.5. RNA-seq tissue clustering. 129
Figure 5.6. DNA methylation in *Nasonia*. 131

Chapter 6

Figure 6.1. Annotated clock output pathways. 136
Figure 6.2. Insect clock gene prevalence by order. 138
List of tables

Chapter 3
Table 3.1. Calculation of Q10 values. 48
Table 3.2. FDR thresholds for rhythmic transcription. 52
Table 3.3. JTK cycle comparison. 53
Table 3.4. Putative clock gene q-values. 58

Chapter 4
Table 4.1. Numbers of RBHs among insect species 76
Table 4.2. BLASTX filtering of CNE sets. 80

Chapter 5
Table 5.1. *Nasonia* genome annotations. 109
Table 5.2. RNA-seq cross-correlation. 128
## Common abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>Amino Acid</td>
</tr>
<tr>
<td>BLAST</td>
<td>Basic Local Alignment Search Tool</td>
</tr>
<tr>
<td>CCG</td>
<td>Clock-Controlled Gene</td>
</tr>
<tr>
<td>CCS</td>
<td>Combined Conservation Score</td>
</tr>
<tr>
<td>CNE</td>
<td>Conserved Non-coding Element</td>
</tr>
<tr>
<td>CT</td>
<td>Circadian Time</td>
</tr>
<tr>
<td>DD</td>
<td>Constant Darkness</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic Acid</td>
</tr>
<tr>
<td>FDR</td>
<td>False-discovery rate</td>
</tr>
<tr>
<td>FPKM</td>
<td>Fragments Per Kilobase Mapped</td>
</tr>
<tr>
<td>GO</td>
<td>Gene Ontology</td>
</tr>
<tr>
<td>LD</td>
<td>Light/Dark cycle</td>
</tr>
<tr>
<td>LL</td>
<td>Constant Light</td>
</tr>
<tr>
<td>mRNA</td>
<td>Messenger RNA</td>
</tr>
<tr>
<td>RBH</td>
<td>Reciprocal Best BLAST Hit</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic Acid</td>
</tr>
<tr>
<td>RNA-seq</td>
<td>RNA sequencing</td>
</tr>
<tr>
<td>SCN</td>
<td>Suprachiasmatic Nucleus</td>
</tr>
<tr>
<td>RRBS</td>
<td>Reduced-Representation Bisulphite Sequencing</td>
</tr>
<tr>
<td>TFBS</td>
<td>Transcription Factor Binding Site</td>
</tr>
<tr>
<td>ZT</td>
<td>Zeitgeber Time</td>
</tr>
</tbody>
</table>
1 General introduction

1.1 Circadian rhythms

As a consequence of the Earth’s rotation about its axis relative to the sun, the environment undergoes rhythmic 24-hour changes. To co-ordinate biological processes and behaviours with the environmental cycle, organisms from all domains of life have evolved endogenous pacemakers which produce internal rhythms of roughly 24 hours (Johnson, Elliott & Foster 2003), allowing aspects of biology to be performed with evolutionarily advantageous timing (Vaze, Sharma 2013). Due to their approximate 24-hour rhythmicity, these endogenous pacemakers are termed circadian (Latin: circa; around, diēm; day) clocks. Rhythms produced by circadian clocks persist in constant conditions without rhythmic environmental stimulus (Saunders, Steel & Lewis 2002), but typically with periodicities of slightly more or less than 24 hours. To keep synchrony with environmental rhythms, circadian clocks therefore make use of predictable environmental cues termed zeitgebers (German: time-givers).

The first recorded study of a diurnally-regulated phenotype is contained within the fourth century BC writings of Androsthenes, in which he observed the daily openings and nightly closings of the leaves of the tamarind tree (Bretzl 1903). Although several similar observations had been made throughout the course of history, it was not until 1729 that the endogenous nature of these rhythms was established. To distinguish between endogenously generated rhythms and those resulting simply as a response to daily fluctuations in light, de Mairan (de Mairan 1729) transferred plants to constant darkness (DD) and monitored daily leaf movement. The continuation of rhythmic leaf movement in constant darkness
demonstrated for the first time the existence of an internal timekeeping device in the natural world.

Since the demonstration of the first circadian rhythm, almost every imaginable aspect of behaviour and physiology has been shown to be under the influence of the circadian clock. In animals, clock control has been demonstrated of processes such as feeding (Xu, Zheng & Sehgal 2008), mating (Sakai, Ishida 2001), locomotor activity (Allada, Chung 2010), and sleep (Cirelli, Bushey 2008, Huang et al. 2011), producing dramatic alterations of the face of the ecosystem in a time-dependent manner. In addition to these more obvious circadian phenotypes, the clock has also been implicated in, for example, learning and memory in organisms as diverse as mammals (Valentinuzzi, Menna-Barreto & Xavier 2004) and fruit flies (Lyons, Roman 2008), immunity (Scheiermann, Kunisaki & Frenette 2013), olfactory response (Krishnan, Dryer & Hardin 1999), and metabolism (Huang et al. 2011).

Circadian research is not confined to the animal kingdom; significant bodies of work exist on the circadian rhythms of, for example, plants (McClung 2013, McClung 2006), Fungi (Hevia, Canessa & Larrondo 2016), and Cyanobacteria (Cohen, Golden 2015). Unsurprisingly given their importance for fundamental biological processes in almost every organism, the study of circadian rhythms has implications for a wide range of human diseases. As well as intuitively clock centred ailments such as jet-lag (Waterhouse 1999), notable examples of diseases linked to the circadian clock include cancer (Sahar & Sassone-Corsi, 2009), Alzheimer’s disease (Musiek, Xiong & Holtzman 2015), cardiovascular disease (Takeda, Maemura 2011), obesity (Maury, Ramsey & Bass 2010), diabetes (Maury, Ramsey & Bass 2010), and depression (Quera Salva et al. 2011).
It was from chronobiology that the first link between genetics and behaviour was established, through the discovery of the gene *period* in 1971 (Konopka, Benzer 1971). The genetic underpinnings of the circadian clock remain today our best understood example of how genes link to behaviour. This, along with the ubiquity of circadian rhythms across life and the clear links between circadian rhythms and human disease, constitutes the motivation for the study of circadian rhythms in animals.

### 1.2 The genetic basis of animal circadian rhythms

The genetic mechanisms underlying the animal circadian clock were first elucidated through studies of model animals, primarily the fruit fly *Drosophila*. The first clock gene to be identified, *period (per)*, was discovered through mapping the genetic basis of *Drosophila* mutants with aberrant locomotor and eclosion rhythms (Konopka, Benzer 1971, Reddy et al. 1984, Zehring et al. 1984, Bargiello, Jackson & Young 1984). The discovery of *period* was followed by the discovery of its heterodimeric partner *timeless (tim)* (Sehgal et al. 1994). In *Drosophila*, these two genes are key components of a transcriptional/translational feedback loop (TTFL), a conserved feature of almost all known circadian clocks (Brown, Kowalska & Dallmann 2012), which I describe below.

#### 1.2.1 The Drosophila clock

In *Drosophila*, *period* and *timeless* are joined by a roster of other genes working together to produce robust internal rhythms (Figure 1.1). The transcription of *period (per)* and *timeless (tim)* is promoted by CLK and CYC heterodimers which bind to E-boxes in the promoters of these genes (Darlington et al. 1998). During the night, PER and TIM accumulate in the cytoplasm (Saez, Young 1996) and form
heterodimers (Gekakis et al. 1995). The formation of PER-TIM heterodimers affords PER protection from phosphorylation by DBT (Price et al. 1998, Kloss et al. 1998) which would otherwise lead to its degradation (Grima et al. 2002) and thus prevent cytoplasmic PER accumulation. Gradual PER and TIM accumulation in the cytoplasm leads to their mutually-aided migration into the nucleus, where they suppress the transcription of their own genes by binding to CLK-CYC heterodimers (through a PER-CLK interaction (Ashmore et al. 2003, Curtin, Huang & Rosbash 1995, Lee, Bae & Edery 1999, Rothenfluh, Young & Saez 2000, Yu et al. 2006) and impeding the ability of CLK to bind DNA (Yu et al. 2006, Lee, Bae & Edery 1999) CLK-CYC promoted genes are thus transcriptionally repressed in the presence of nuclear PER. During the daytime, the product of the gene cryptochrome, CRY, is activated by light (Ceriani et al. 1999, Rosato et al. 2001). Once activated, CRY binds to TIM and promotes its degradation (Busza et al. 2004). PER, now without its heterodimeric partner, is vulnerable to degradation, mediated through phosphorylation by DBT and subsequent ubiquitination by SLIMB (Grima et al. 2002, Ko, Jiang & Edery 2002a). TIM is phosphorylated by SHAGGY (SGG) (Martinek et al. 2001). This phosphorylation is counterbalanced PP2A and PP1, which dephosphorylate PER and TIM respectively (Sathyanarayanan et al. 2004, Fang, Sathyanarayanan & Sehgal 2007). When PER levels are low enough, CLK-CYC mediated transcription of per and tim resumes, and the cycle begins anew. For more in depth reviews of the Drosophila molecular clock, see (Rosato, Tauber & Kyriacou 2006, Peschel, Helfrich-Forster 2011)

This negative feedback loop is interlocked with at least one other loop (Glossop, Lyons & Hardin 1999) which centres on Clk. The functioning of this loop depends on the fact that CLK-CYC heterodimers also promote the transcription of
two genes, *vri* (Blau, Young 1999, McDonald, Rosbash 2001) and *pdp1ε* (McDonald, Rosbash 2001). The protein products of these genes in turn bind V/P boxes in the promoter region of *Clk*, VRI inhibiting *Clk* transcription and PDP1ε promoting it (Cyran et al. 2003). Although one could conceivably envisage a far simpler mechanism for generating circadian rhythms, this apparently convoluted system is likely to be necessary to enhance the stability and precision of the rhythms produced, as reviewed by (Albrecht 2010), based on the behaviour of synthetic clocks. The presence of partially redundant genes is also likely to be useful for this purpose, for example the gene *clockwork orange*, which also serves to inhibit CLK-CYC mediated transcription of genes associated with E-boxes (Kadener et al. 2007).

**Figure 1.1.** Schematic diagram of the circadian clock. CLK and CYC heterodimers promote the transcription of E-box genes, including *per* and *tim*. PER and TIM repress their own transcription. Light activates CRY which binds to and promotes the degradation of TIM. DBT and SGG phosphorylate PER and TIM leading to their degradation by SLIMB, and thus releasing transcriptional repression. A second
feedback loop is centred on CLK. The transcription of *pdp1ε* and *vri* is promoted by CLK and CYC. PDP1ε promotes the transcription of *Clk*, whereas VRI represses it.

### 1.2.2 The mammalian clock

The discoveries made in *Drosophila* have been and are being used to investigate the mechanisms of the circadian clock in mammals, which has been shown to involve many of the same key genes identified in *Drosophila*, although these often fulfil different roles (Yu, Hardin 2006).

**Figure 1.2.** Schematic diagram of the main components of the mammalian circadian clock. Genes largely fulfil roles as in *Drosophila* (Figure 1.1), but with paralogy (e.g. both CRY 1 and 2 as opposed to simply CRY1), and a different interlocking loop (with rev-erb and ror genes as opposed to pdp and vri).

The mammalian clock features many of the same key genes as the *Drosophila* clock in similar roles, with a high degree of paralogy, redundancy, and functional divergence. The mammalian orthologue of CYC, BMAL1, heterodimerises with CLK (Gekakis et al. 1998) and promotes the transcription of PER (Gekakis et al. 1998) and CRY (Kume et al. 1999). Similarly to PER and TIM in *Drosophila*, mammalian orthologues of PER and CRY accumulate in the
cytoplasm, heterodimerise, and re-enter the nucleus to suppress their own transcription by interfering with the actions of the CLK-BMAL1 heterodimer (Kume et al. 1999, Jin et al. 1999). Mammalian homologues of DBT, CKIδ and CKIɛ, perform a similar action in mammals, as they are involved in the degradation and nuclear localisation of PER (Camacho et al. 2001, Keesler et al. 2000, Vielhaber et al. 2000). This feedback loop in mammals is interlocked with another feedback loop acting on BMAL1 with positive (RORα, RORβ, and RORγ) and negative (REV-ERBα and REV-ERBβ) regulatory elements promoted by BMAL1 and feeding back into its own transcription levels, as well as the transcription of CRY1 (Guillaumond et al. 2005, Akashi, Takumi 2005, Triqueneaux et al. 2004, Preitner et al. 2002, Matsumoto et al. 2007). More detailed reviews of the mammalian circadian clock can be found here (Partch, Green & Takahashi 2014), and detailed comparisons of the mammalian and Drosophila circadian clocks can be found here (Yu, Hardin 2006).

1.3 Clock network and outputs
To understand how these molecular rhythms ultimately result in circadian behavioural and physiological rhythms, it is important to understand how the central clock network works and how its outputs control biological rhythms. The Drosophila central clock consists of around 150 neurons (70-80 per hemisphere) which generate rhythms using the molecular clockwork described in the section above (Hermann-Luibl, Helfrich-Förster 2015). These neurons are located in the lateral and ventral brain, and are subdivided into groups based on their anatomical location (Hermann-Luibl, Helfrich-Förster 2015). In Drosophila these neurons can now be examined and manipulated individually (Guo et al. 2014, Cao et al. 2013) to
investigate their functions. These pacemaker neurons differ in their functions and in their neurochemistry. For example, it has been demonstrated that different groups of neurons can be desynchronised by certain light conditions, resulting in the separation of different bouts of activity (Rieger et al. 2006). This result, along with other similar studies (Yoshii et al. 2004) demonstrates that the clock integrates information from multiple oscillators controlling different aspects of behaviour. The synchrony between clock neurons is controlled, in part, by the neuropeptide PDF. PDF is the most powerful neuropeptide in the Drosophila circadian clock, working as a putative output factor (Hermann-Luibl, Helfrich-Förster 2015, Shafer, Yao 2014), and capable of speeding up or slowing down the clock of individual cells to maintain synchrony (Yoshii et al. 2009). For detailed reviews of the Drosophila clock network see (Shafer, Yao 2014, Hermann-Luibl, Helfrich-Förster 2015, Helfrich-Forster, Nitabach & Holmes 2011).

The mammalian clock network is more complex than that of Drosophila, and thus more difficult to research. The circadian oscillator in mammals is located in the suprachiasmatic nucleus (SCN), which contains around 20,000 cells (Mohawk, Takahashi 2011). The majority of cells from the SCN are able to produce rhythms in vitro, even in solitude (Welsh et al. 1995). Both electrical and humoural signals from other rhythmic neurons in the SCN work together to produce a synchronised and robust rhythm which is resilient against environmental and genetic perturbations (Liu et al. 2007, Welsh, Takahashi & Kay 2010). Rhythmic behavioural and physiological outputs are accomplished through a variety of means, including both rhythmic neuronal firing and humoural output (Welsh, Takahashi & Kay 2010, Gachon et al. 2004).
Ultimately, the clock aims to regulate biological processes in a circadian manner. Very few output pathways have thus far been fully described, but progress is being made. Recently, for example, a pathway linking the circadian clock to rest/activity rhythms has been identified in *Drosophila*, demonstrating output through a neural circuit and ultimately acting through a clock output molecule, DH44 (Cavanaugh et al. 2014). The outputs from the molecular clock are translated into rhythmic neural firing and humoral outputs through complex regulatory networks of clock-controlled genes (Bozek et al. 2009).

The regulation in these networks involves gene transcription programs (Hughes et al. 2009), post-transcriptional regulation (Kojima, Shingle & Green 2011), and post-translational modifications (Kramer, Merrow 2013). In order to further understand the methods through which different aspects of biology are rhythmically controlled, efforts have been made to identify rhythmically expressed genes in both *Drosophila* and mammals. Microarray (Hughes et al. 2012, Keegan et al. 2007) and RNA-seq (Akhtar et al. 2002, Hughes et al. 2009) studies have revealed widespread circadian gene regulation in mammals, opening up new avenues for research into clock output pathways. A simplified illustration of the general principle of how signals from the central pacemaker are translated into behavioural rhythms is shown in Figure 1.3.
Figure 1.3. Simplified illustration of how signals from the molecular clock are translated into behavioural and physiological rhythms. Signals from the molecular clock induce rhythmic expression in sets of genes through a complex gene regulatory network. This rhythmic gene expression is converted into extracellular signals through neural firing or circadian output molecules (e.g. hormones, neuropeptides). These signals may then affect gene transcription in other cells, perhaps inducing rhythmic gene transcription through the gene regulatory network present in the signal-receiving cell.

1.5 Circadian research in other insects

As circadian research into non-drosophilid insects has advanced, it has been shown that there are several different ways of constructing circadian clocks in insects (Yuan et al. 2007). It is also clear that Drosophila can not by itself showcase the full diversity of insect behavioural genetics. In fact, comparisons of these insects with mammalian clocks suggests that some of these alternative clock constructions may
be able to better model mammalian clocks than *Drosophila*, and that *Drosophila* may have a somewhat esoteric clock construction. Below I discuss the major insect orders used in circadian research. The relationship between these orders is shown in Figure 1.4.

![Figure 1.4. The phylogenetic relationship between the five main insect orders used in circadian research. Branch lengths from (Hedges, Dudley & Kumar 2006).]

1.5.1 **Diptera**

Diptera (the true flies) is the order to which *Drosophila* belongs, and contains many other species of interest for circadian research. As well as containing many species important in the food chain, Diptera also contains disease vectors such as mosquitos and midges, economically damaging pests such as the tsetse fly (Alsan 2015), and predators of pests such as the tiger fly. Therefore, as well as the general aims of
understanding the evolution of insect clocks, an important goal of circadian research in Diptera is to understand how the circadian clock interacts with phenotypes such as blood feeding, host preference, and insecticide resistance.

A motivating example of the direct influence of Dipteran circadian clocks on human health is the case where *Anopheles funestus* mosquitos were shown to rapidly evolve a change in the timing of their biting behaviour in order to avoid the timing of net deployment (Moiroux et al. 2012). Understanding the genetic basis of this behaviour may thus have direct consequences for malaria control. A large number of dipteran behaviours have been examined in a circadian context; behaviour and biochemical assays on various dipteran species have revealed circadian rhythms in emergence (Wool, Kugler 1969, Kureck 1979), locomotor activity (Meireles-Filho et al. 2006), blood-feeding (Das, Dimopoulos 2008), host seeking and landing behaviour (Fall et al. 2015b, Fall et al. 2015a), and insecticide resistance (Yang et al. 2010). Notably, as well as being rhythmic in constant darkness (DD), mosquitos are also rhythmic in constant light (LL) (Jones 1976), a departure from the *Drosophila* model, which behaves arrhythmically when moved into constant light conditions (Aschoff 1979).

Behavioural work has been supplemented with molecular work. The presence of all canonical clock genes has been confirmed in several mosquito species (Meireles-Filho, Kyriacou 2013, Tormey et al. 2015), along with *cry2*, the mammalian-like cryptochrome which is absent from *Drosophila*. The presence of *cry2* within even Dipteran lineages adds further weight to the view that *Drosophila* has a highly diverged and perhaps esoteric clock mechanism.

The primary output of the clock is rhythmic transcription. Therefore in order to understand the outputs of the circadian clock, a crucial part of circadian research
is identifying those genes which are rhythmically transcribed (Clock-controlled genes; CCGs). Circadian microarray studies have been performed in \textit{Aedes aegypti} (Ptitsyn et al. 2011) and \textit{Anopheles gambiae} (Rund, Gentile & Duffield 2013), revealing clock control of genes involved in roles such as immunity, detoxification/pesticide resistance, and development.

Though there is a wealth of ecological observation of dipteran species and many genomes have been sequenced, the molecular workings of the clock have not yet been fully investigated. The role of \textit{cry2} in the clock has not been tested within Diptera, though work with \textit{cry2} in other insects strongly suggests a likely transcriptionally repressive role for this protein (Yuan et al. 2007). Though the majority of the clock genes in \textit{Drosophila} are present in other Diptera, a paucity of research into how the various clock components interact may hide variations in clock mechanism hidden by this superficial similarity.

### 1.5.2 Lepidoptera

Circadian research in Lepidoptera has focussed primarily on a few models. The most studied Lepidopterans in terms of the circadian clocks are the monarch butterfly \textit{Danaus plexippus}, the giant silkmoth \textit{Antheraea pernyi}, and Manduca sexta (the Carolina sphinx moth). Although the bulk of lepidopteran circadian research has been focussed on these three models, other well-studied lepidopterans such as the commercial silkmoth \textit{Bombyx mori} have also been the subject of circadian research.

Whilst the underpinnings of the \textit{Drosophila} clock were being elucidated, work was also underway to identify whether the same mechanisms were responsible for circadian rhythms in other insects. An early circadian phenotype intensely
studied in Lepidoptera was gated egg hatching in *A. pernyi*. PER was shown to have a role in this behaviour (Sauman et al. 1996, Sandrelli et al. 2007), and through tissue ablation it was discovered that the brain was responsible for the hatching signal (Sauman, Reppert 1998). In *A. pernyi*, PER was found to localise to eight neurons, with no nuclear expression detected at any timepoint (Sauman, Reppert 1996). Further to this, experiments in *B. mori* identified co-expression of PER, CRY, CYC, and DBT in groups of neurons (Sehadova et al. 2004), marking out putative pacemaker cells in the brain.

The unexpected absence of PER expression in the nucleus of putative pacemaker cells in *B. mori* and *A. pernyi* was further confirmed in *D. plexippus* (Sauman et al. 2005). So, although all of the clock proteins have been shown to be present in lepidopteran species, and these proteins have been shown to constitute a working clock in vitro (i.e. PER was shown to be able to repress CLK/CYC mediated transcription) (Chang et al. 2003), there are clearly fundamental differences between the mechanism of the *Drosophila* clock and the lepidopteran clock. If PER does not enter the nucleus to constitute the negative arm of the clock, how is transcription repressed?

The answer to this comes with the identification of *cry2*. All lepidopteran species thus far analysed have been found to possess both an orthologue of the *Drosophila*-like *cry1*, and the mammalian-like *cry2*. *cry2* has been shown to be a powerful repressor of CLK-CYC mediated transcription (Yuan et al. 2007, Zhu et al. 2005), thus constituting the negative limb of the clock.

A behaviour that has motivated the intense study of the *D. plexippus* clock is their spectacular annual migration across North America. As the monarch butterfly primarily uses the sun to navigate, time compensation of these cues is vital to ensure
a consistent migratory direction (Reppert 2007). cry2-mediated links from the clock to the sun compass have been investigated (Zhu et al. 2008). Consistent with the identification of PER expression in the antennae of *M. sexta* (Schuckel, Siwicki & Stengl 2007), independent peripheral clocks have been identified in the antennae of *D. plexippus*. These antennal clocks independently feed into the navigation mechanism, as elegantly demonstrated by (Guerra et al. 2012). Painting one antenna with an opaque paint (and thus de-synchronising the two antennae) results in disruption of navigatory abilities. Ablation of the painted antenna restores normal navigation, showing that the signal from the single antenna is sufficient for normal functionality in the absence of confounding signal.

Although the circadian transcription of core clock genes has been investigated, for example *B. mori* cyc, (Markova et al. 2003), *per* and *tim* (Iwai et al. 2006)), as have some rhythmic clock outputs, for example *B. mori* rhythmic Hydroxyindole-O-methyltransferase activity (Itoh, Nomura & Sumi 1997) and rhythmic melatonin synthesis (Itoh et al. 1995), relatively little work has thus far been done to understand clock outputs in these species. Overall, lepidopteran species offer a well-studied clock model with opportunities for further research to deepen our understanding of what has been proposed to represent the ‘ancestral’ clock state of insects (Reppert 2007).

### 1.5.3 Coleoptera

Coleoptera (beetles) is the most diverse order on Earth. Consistent with this, coleopteran circadian clocks are of interest for a wide range of reasons. Coleoptera contains a great many pest species, such as the boll weevil, which attacks cotton crops (Greenberg et al. 2006), or the rhinoceros beetle, an invasive palm pest
Beetles also prey on a range of pests themselves, making them useful in biological pest control. Interestingly for circadian research, some cave-dwelling beetles which have adapted to their environment have lost their colour vision (Friedrich et al. 2011), and some are in the process of adaptation to a troglobiont lifestyle (Pasquali, Sbordoni 2014), making interesting targets to study the adaptation of circadian rhythms in low light environments. Beetles have even been sent to space to study the effects of low gravity conditions on circadian rhythms, which have been known to cause desynchrony in humans (Hoban-Higgins et al. 2003).

Circadian-controlled behaviours which have been studied in the Coleoptera include locomotion (Evans 1976), oviposition and emergence (Greenberg et al. 2006), cuticle deposition (Ikeno et al. 2011), oxygen consumption (Chiba, Cutkomp & Halberg 1973), and even the prey-attracting bioluminescence of coleopteran larvae (Merritt, Aotani 2008). Interestingly, circadian rhythms have been demonstrated in a species currently undergoing the transition to a cave-dwelling lifestyle, revealing two distinct groups - a rhythmic group of animals which show normal circadian behaviours, and an arrhythmic group of animals which exhibit disrupted clocks, including ultradian rhythms (Pasquali, Sbordoni 2014).

Beetles have also been used to investigate the links between the circadian clock and photoperiodism, an important question in biological rhythms research (Pegoraro et al. 2014). Photoperiodism refers to the internal clock which measures changes over the year (i.e seasonal changes), as opposed to circadian rhythms which measure time over the space of a day. RNAi studies on this topic have helped us to understand how the clock genes work together in the molecular clockwork of the Coleoptera; as in Drosophila, CLK and CYC promote transcription on the positive
limb of the clock (Ikeno, Numata & Goto 2011), whereas CRY3 and PER have been shown to repress transcription (Ikeno et al. 2011, Ikeno, Numata & Goto 2011). Coleoptera do not possess cry1, but unlike the Hymenoptera, do possess timeless (Yuan et al. 2007). The light input pathway is therefore an open question, constituting a clock model not seen elsewhere in the insect kingdom. Some work has been conducted on this topic, and there is evidence for the presence of non-visual photosensitive organs which may make use of opsins, potentially as an input to the clock (Fleissner, Fleissner & Frisch 1993).

So far, studies in coleopteran circadian rhythms have been useful, allowing researchers to better understand when light traps may be effective against invasive pest species (Ehsine, Belkadhi & Chaieb 2014), or even to better understand the walking behaviour of flightless pest control species (Nakayama et al. 2010). Thus far, a limited amount of work has been conducted into temporal or spatial expression of circadian genes, nor of the neural structures involved in the coleopteran clock. Peripheral clocks also remain relatively unstudied in beetles. The diversity of coleopteran biology, along with the development of tools such as CRISPR transgenics in model beetles such as Tribolium castaneum marks the coleopterans as fruitful targets for future circadian research.

1.5.4 Orthoptera

The Orthoptera (crickets, locusts, et al.) offer advantages for research for several reasons. One of these reasons is that locusts are perhaps the most well-known pest species - swarms of locusts cause a great deal of economic damage (Steedman 1990). Also, the large size of many orthopterans (e.g. crickets) make the brains of these insects very amenable to surgical manipulation. Tools such as RNAi are also
available to enable investigation of circadian mechanisms (Mito et al. 2007). Finally, orthopterans demonstrate interesting phenotypes which provide opportunities to study novel clock mechanisms, such as dramatic plastic dimorphic phenotypes in *Schistocerca gregaria* (Nishide, Tanaka & Saeki 2015), strong photoperiodic responses (Tomioka 2014), circatidal rhythms in mangrove crickets (Satoh, Yoshioka & Numata 2008), and migratory behaviour in locusts (Helfrich-Forster, Nitabach & Holmes 2011). The primary model organism used for circadian research in Orthoptera is the field cricket *Gryllus bimaculatus*.

Examples of phenotypes under circadian control in Orthoptera include locomotion (Sokolove, Loher 1975, Tomioka, Okada & Chiba 1990), egg hatching (Itoh, Sumi 2000), stridulation (Sokolove 1975), photoreceptor membrane turnover (Williams 1982), insecticide resistance (Onyeocha, Fuzeau-Braesch 1991), and melatonin synthesis (Itoh, Sumi 1998). Tidal rhythms have also been shown to be present in some orthopteran species, such as the mangrove cricket *Apteronomobius asahinai* (Satoh, Yoshioka & Numata 2008). This species was used to investigate the potential link between the circadian clock. RNAi knockdown of *Clk* in *A. asahina* disrupted circadian rhythms, but failed to disrupt the circatidal mechanism, demonstrating a separation in mechanism between the two clocks (Satoh, Yoshioka & Numata 2008).

*Schistocerca gregaria* is a locust which has two distinct morphs: locusts raised in solitude (solitary locusts) walk with a different gait, have different diets, and are morphologically different to those raised alongside other locusts (gregarious locusts) (Gaten et al. 2012). The switch from solitary to gregarious occurs within hours of transfer to a social environment, though the morphological traits associated with being gregarious take generations to fully appear (Pener 1991). Phenotypic
plasticity in the circadian clock has been demonstrated in *S. gregaria*; locomotor activity and egg hatching peaks differ between the two morphs (Nishide, Tanaka & Sacki 2015), as does the rhythm of interneuron response to looming stimuli (Gaten et al. 2012). Clock plasticity has also been demonstrated in juvenile hormone (JH) rhythms: flight-capable morphs of *Gryllus firmus* exhibit high amplitude JH rhythms in constant darkness which are almost entirely undetectable in flightless morphs of the same species (Zhao, Zera 2004).

The ease of surgical manipulation in crickets has been exploited in circadian research. The circadian clock in the cricket has been shown to reside in the optic lobes (Abe, Ushirogawa & Tomioka 1997). Ablation of both optic lobes results in the abolition of circadian locomotor rhythms (Sokolove, Loher 1975). One lobe is sufficient for circadian rhythms in locomotor activity, demonstrating functional redundancy. Circadian oscillations in the optic lobe have been demonstrated in vitro (Tomioka, Chiba 1992). The neurons responsible for the coupling between the two optic lobe pacemakers are more photo-responsive at night than during the day. These have been experimentally manipulated; injections of pdf into these neurons results in greater photo-responsiveness (the ‘night-state’) (Saifullah, Tomioka 2003) whereas injections of serotonin result in a weaker photo-responsiveness, the ‘day-state’ (Saifullah, Tomioka 2002). Bilateral optic nerve severance (i.e. severing the connection between the optic lobe and the compound eye) results in a lengthened circadian period (Abe, Ushirogawa & Tomioka 1997), suggesting a role for the compound eye in regulation of the circadian period.

The clock genes have also been investigated in Orthoptera. The Orthoptera contain a copy of *cry2* (mammalian-like *cry*), but not *cry1*. In *G. bimaculatus*, *period*, *timeless*, and *cycle* have all been shown to exhibit rhythmic expression, but
not *Clock* (Uryu, Tomioka 2014). In-situ hybridisation experiments in various cricket species have localised the expression and co-expression of clock genes in putative pacemaker neurons (Shao, Hiragaki & Takeda 2008, Shao et al. 2006, Lupien et al. 2003). RNAi knockdown has been used to investigate the mechanism of the circadian clock in *G. bimaculatus*. RNAi knockdowns of *per* and *Clk* both abolish circadian rhythms (Uryu, Karpova & Tomioka 2013). Knockdown of *cyc* lengthened the free-running period, whereas knockdown of *itim* shortened it (Danbara et al. 2010). For a review of cricket circadian clocks, see (Tomioka 2014).

From the studies so far conducted, our understanding of the orthopteran circadian clock is generally good, but questions still remain. We have clues to the molecular workings of the orthopteran circadian clock, but do not yet know the precise interactions between genes. For example we do not yet know which genes constitute the negative limb of the clock. This question is particularly pertinent given the fact that all in-situ hybridisation experiments thus far performed have shown only cytoplasmic localisation of the core clock genes and no nuclear translocation has yet been seen (Shao et al. 2006). As the clock has been shown to be able to oscillate without timeless or cycle (Danbara et al. 2010, Uryu, Karpova & Tomioka 2013), it is clear that we are looking at a different mechanism to *Drosophila*, despite the involvement of the expected candidate genes. Working out the precise mechanism of the clock in Orthoptera will be paramount to a full understanding of the diversity of insect clocks.

### 1.5.5 Hymenoptera

Hymenoptera is the third largest insect order and is comprised of the bees, ants, wasps, and sawflies. As well as being ecologically important in, for example,
pollination, the Hymenoptera are of particular interest due to the repeated independent evolution of eusociality (Hughes et al. 2008, Andersson 1984), a rare trait in the rest of the animal kingdom. These facts together underscore the potential importance of circadian research in the Hymenoptera. Consistent with this, there is a growing body of Hymenopteran circadian research.

A large number of behaviours (eusocial and otherwise) have been shown to be under circadian control in the Hymenoptera. Beginning with locomotor rhythms, circadian control of diurnal foraging behaviour has been demonstrated in, for example, *Bombus terrestris* (bumblebee) and its more northerly cousin *Bombus pascuorum* (Stelzer, Chittka 2010), and the honey bee *Apis mellifera* (Bloch, Sullivan & Robinson 2002). In *B. terrestris*, foraging begins at dawn and is concluded at dusk. This rhythmic behaviour continues at high altitudes in natural constant light conditions, apparently being entrained by variations in light intensity and temperature, rather than free-running in response to constant light. Constant light (and intensity) conditions in the laboratory produce a smaller free-running period, and constant darkness conditions produce a longer free-running period (Stelzer, Chittka 2010). The circadian clock is also important for sun compass navigation in *Apis mellifera* (Bloch, Sullivan & Robinson 2002).

Ants also exhibit rhythmic foraging behaviour, as demonstrated by the ant *Dinoponera quadriceps*, which primarily forages diurnally, with seasonal changes in daily foraging peaks (Medeiros et al. 2014). Differential activity is also seen in the weaver ant (*Oecophylla longinoda*), which is more active during the day than at night (Vayssières et al. 2011). Locomotor circadian rhythms have been studied in paper wasps (Giannoni-Guzman et al. 2014), and the parasitoid wasp *Nasonia vitripennis* exhibits sexually dimorphic circadian control of locomotor rhythms with
a longer free running period in DD than in LL (Bertossa et al. 2013). Eclosion from
the host is also under circadian control in, for example, the *N. vitripennis* (Bertossa
et al. 2010) and *Trichogramma embryophagum* (Reznik et al. 2008).

In some Hymenoptera, circadian rhythms are intricately intertwined with
eusociality. Circadian rhythms in honey bees are highly plastic, varying through, for
example, ontogeny (Spangler 1972), and the task of the individual within the colony
(Bloch, Robinson 2001). A well-studied example of this is that foragers showing
normal diurnal patterns of foraging and rest will behave arrhythmically if induced
into nursing behaviour within the hive (Bloch, Robinson 2001). In this case, the
adaptive significance of clock plasticity is clear; while foraging is best performed
during the day (and timing flower visits to match peaks in nectar secretion (Moore
2001)), honey bee larvae benefit from the 24-hour care provided by an arrhythmic
nurse (Bloch, Sullivan & Robinson 2002). There are many similar cases of
circadian plasticity in honey bees. For more examples, see (Bloch et al. 2013,
Moore 2001).

1.5.6 Time for a new insect model?

From research in other insects, it is clear that there exist major differences between
the clocks of various insects, and that some insects may better model mammalian
circadian phenomena than others. A major difference between the various clock
models in insects concerns the light input pathway. One light input pathway to the
clock in *Drosophila* is mediated through *cryptochrome* (*cry1*) which is activated in
response to light (Ceriani et al. 1999, Rosato et al. 2001), binds to and promotes the
degradation of *tim* (Busza et al. 2004), ultimately resulting in the degradation of *per*
(Ko, Jiang & Edery 2002b, Grima et al. 2002). Mammalian-like *cryptochrome*
(cry2) is not light-sensitive (Yuan et al. 2007), instead forming part of the core transcriptional feedback loop suppressing its own transcription (and that of per) by interfering with the actions of the CLK-BMAL1 heterodimer (Kume et al. 1999, Jin et al. 1999). Mammals also lack an ortholog for timeless, possessing only a homolog (an ortholog of the Drosophila gene timeout) (Benna et al. 2000), a gene whose potential role in the clock is less clear and less crucial than that of timeless (Gustafson, Partch 2015, Benna et al. 2010).

The Lepidoptera harbour both types of cryptochrome (Drosophila-like cry1 and mammal-like cry2) (Tomioka, Matsumoto 2010) as well as homologs of both timeless and timeout (Tomioka, Matsumoto 2015). The two cryptochromes have been shown to act in a similar way to their drosophilid and mammalian counterparts; cry1 functions as a light receptor and cry2 functions as a transcriptional repressor (Zhu et al. 2008). The Coleoptera also possess cry2 and timeless, but lack cry1 (Tomioka, Matsumoto 2015). The light input pathway to the clock in the case of the Coleoptera is therefore hypothesised to ultimately act through timeless.

Of the major insect orders, the Hymenoptera arguably exhibits the most mammalian-like core clock architecture, possessing cry2 and timeout but neither cry1 nor timeless (Tomioka, Matsumoto 2015, Yuan et al. 2007). In addition to these molecular similarities, there is evidence that the transcriptional profiles of these genes match more closely the mammalian model than the Drosophila model (Rubin et al. 2006). Light-entrained circadian rhythms have been demonstrated in the Hymenoptera, but the question of light detection in the Hymenopteran clock is currently open.
Given the similarities between the hymenopteran and mammalian molecular clocks, if one were to pick a circadian model then a hymenopteran would seem a potentially fruitful choice. However, one of the chief advantages of Drosophila is the ease with which one can keep stocks. The eusocial nature of many Hymenoptera makes this more difficult, and the complex interplay between eusociality and clocks may add an extra level of complexity in initial investigations which would be likely to present difficulties. These considerations suggest that a non-social hymenopteran such as the Jewel wasp Nasonia would be an ideal choice for a new clock model system.

1.6 Nasonia: a solitary parasitoid wasp

Parasitoids are an extremely diverse group of insects, making up roughly 10% of all described insects (Foottit, Adler 2009). The largest group of insect parasitoids, and thus an important cornerstone of animal biodiversity, is the parasitoid wasp. The parasitoid wasp life cycle (Figure 1.5) involves acquisition of nutrients from a host organism. Adult wasps sting the host organism, then lay their eggs onto or into the host’s body. When the eggs hatch, the larvae consume the host and typically develop inside the host, eclosing as adults and beginning the cycle anew. The great diversity of parasitoid wasps is matched by a great diversity of hosts, many of which are pest species. Parasitoid wasps are thus widely used commercially in environmentally-friendly biological pest control.
Figure 1.5. The *Nasonia* life cycle. (A) *Nasonia* are strongly sexually dimorphic. Males are smaller than females and are incapable of flight. (B) The female wasp stings the pupa of the host fly, and (C) lays eggs. These eggs hatch into larvae, (D), which eat the host from the inside. The larvae mature into pupae (E), develop into adults, and eclose (F) from the host pupa. The whole cycle takes around two weeks at 25°C.

*Nasonia vitripennis* is one such parasitoid wasp, and is a generalist able to parasitize a wide variety of flies including the blowfly (Werren 1984), a common
pest. *Nasonia* is an important model for parasitism, and has been used extensively in studies on host-parasitoid relationships, e.g. (Cornell, Pimentel 1978, Danneels, Rivers & de Graaf 2010, Rivers, Ruggiero & Hayes 2002, Desjardins et al. 2010). *Nasonia* is also becoming an important model organism in many other areas of study as an alternative to *Drosophila* and other models, as it offers several advantages over existing model systems. Firstly, as a hymenopteran, *Nasonia* has a haplodiploid sex determination system (Werren, Loehlin 2009), making the wasp ideal for genetic screens as males lack heterozygosity. As a solitary wasp *Nasonia* are simple to rear, requiring only the pupae of their fly host, although in vitro methods of cultivation are now available (Brucker, Bordenstein 2012). *Nasonia* possess a fully functional DNA methylation system (Park et al. 2011, Wang et al. 2013, Beeler et al. 2014), robust circadian (Bertossa et al. 2010, Bertossa et al. 2013, Bertossa et al. 2014) and photoperiodic (Saunders 1965) responses, a systemic RNAi response (Lynch, Desplan 2006, Werren, Loehlin & Giebel 2009), and a fully sequenced genome (Werren et al. 2010). A further advantage as a model organism is that the closely related species of *Nasonia* (whose genomes have also been sequenced) can be crossed once cured of *Wolbachia* infection (Werren et al. 2010).

1.6.1 *Nasonia* circadian research

Circadian behavioural responses have been characterised in *Nasonia*. *Nasonia* exhibits rhythmic locomotor activity and emergence from the host pupa, but arrhythmic eclosion from its own pupal stage (Bertossa et al. 2010), presumably due to adaptive neutrality (i.e. no evolutionary advantage to a more precise timing of pupal eclosion). Examination of several *Nasonia* species shows that different
species differ significantly in their circadian phenotypes, as do males and females. Males of *Nasonia vitripennis* show a more robust circadian response, and are generally rhythmic in LL, whereas females of *N. vitripennis* are generally arrhythmic in LL. Other species, such as *Nasonia giraulti*, show the opposite trend (Bertossa et al. 2010). The period in LL is always shorter than the period in DD, regardless of sex or species. The link between the circadian clock and photoperiodism has been investigated in *Nasonia*; interruptions of the circadian rhythm with chilling (Saunders 1967) and light pulses (Saunders 1970) have been shown to have time-dependent effects on the diapause response, and *Period* RNAi has been shown to abolish the photoperiodic response (Mukai, Goto 2016). Circadian transcription has been demonstrated for a handful of genes, *per* and *cry2* (Mukai, Goto 2016, Bertossa et al. 2014), and *pdp* (Mukai, Goto 2016) have been shown to be transcribed rhythmically, and *cyc* and *cwo* have analysed without any significant cycling found (Mukai, Goto 2016).

1.7 Making up ground on *Drosophila*

As the principal circadian model in insects, *Drosophila* has been used to great effect to model circadian phenomena in humans (Rosato, Tauber & Kyriacou 2006). The majority of literature reviews on the topic draw comparisons between mammalian and insect clocks, using *Drosophila* as a reference point for these comparisons. A principal reason for using *Drosophila* as a clock model in the first place was the greater availability of genetic tools and a deeper general understanding of *Drosophila* than others insects. The same is true today and can be seen through the great availability and depth of bioinformatics resources available for *Drosophila*. 
As mentioned previously, the circadian transcription of the entire *Drosophila* transcriptome has been characterised, as have extensive swathes of regulatory elements. New methods such as RNA-seq and the availability of genomic data for cross-species comparisons make these analyses much quicker to perform than they were in the first instance. Also, the large number of bioinformatic resources available for *Drosophila* greatly facilitate these analyses.

FlyBase, for example, is a bioinformatics database containing functional gene data from twelve sequenced *Drosophila* species (Attrill et al. 2016). The database covers gene sequences and annotation, mutant alleles, phenotypic data including correlations with disease, orthology data, and expression data. Tissue specific expression data is also available for *Drosophila* genes on the database FlyAtlas (Robinson et al. 2013), including in-situ hybridisation data. A whole host of other databases are also available, including gene and protein interaction data (Murali et al. 2011) and information on conserved and experimentally verified regulatory elements (Gallo et al. 2011). These databases greatly facilitate bioinformatic investigation, and thus increase the utility of investigations carried out through the generation of large datasets.

A great many genetic tools are now available for insects other than *Drosophila*, and with the ever decreasing cost of genome sequencing, more and more insect genomes are being added to our collective repertoire. Therefore, as well as being able to conduct more detailed studies of individual insects using newly available tools such as CRISPR transgenics and RNAi, we are also able to generate large datasets for practically any insect species. The utility of these large datasets does however depend on the presence and quality of genome assemblies and annotations, and availability of databases and species specific bioinformatics tools.
Due to the increase in number of insect genomes, large scale bioinformatics comparisons between insect species are now possible, from which evolutionary conclusions may be drawn.

In order to further *Nasonia* circadian research, I took a three pronged approach. Firstly, I investigated both behavioural aspects and the transcriptomic basis of circadian rhythms in *Nasonia*, gaining insights into the aspects of physiology under circadian control in the wasp (Chapter 3). Secondly, I examined conserved regulatory elements in *Nasonia*, gaining insights into transcriptional and post-transcriptional regulation of genes (Chapter 4). Finally, I developed a gene-based bioinformatics resource for *Nasonia*, integrating multiple genome-wide datasets, including the analysis of DNA methylation in the wasp (Chapter 5).
2 General methods

2.1 Wasp maintenance

*Nasonia vitripennis* AsymCX wasps were maintained in plastic *Drosophila* vials at 25°C. Wasps were maintained in 12:12 light:dark light conditions. Upon eclosion of each new generation, ~20 wasps of mixed sex were transferred to *Drosophila* vials containing fresh pupal hosts.

2.2 The DAM Trikinetics system

Behavioural monitoring was performed using the DAM Trikinetics (Trikinetics inc) system. Wasps were inserted into glass tubes, one end of which contains food and is sealed with a rubber cap, and the other end of which was closed with cotton wool to allow airflow. An infra-red beam intersects the middle of the tube, which is interrupted when the wasp passes through. These interruption events were summed in 30 minute bins to provide a profile of the wasp’s activity over time. This set-up is illustrated in Figure 2.1.
Monitoring the behaviour of individual wasps using the DAM Trikinetics system is not possible alongside a pupal host, and changing hosts mid-experiment is likely to disturb circadian behaviour. Given these difficulties, an alternative food source was developed. Sucrose-based foods have been shown to extend the life of *Nasonia* (Davies, 1975) and other parasitoid wasps (Williams, et al, 2015), and sucrose-based foods are recommended for keeping *Nasonia* stocks alive in the absence of pupal hosts (Ward’s, 2002). Based on this information, combined with the results of

**Figure 2.1.** DAM Trikinetics set-up. Individual wasps are inserted into tubes containing food, and activity is monitored using an infra-red beam which passes through the centre of the tube.

### 2.3 Optimisation of food for behavioural monitoring

Monitoring the behaviour of individual wasps using the DAM Trikinetics system is not possible alongside a pupal host, and changing hosts mid-experiment is likely to disturb circadian behaviour. Given these difficulties, an alternative food source was developed. Sucrose-based foods have been shown to extend the life of *Nasonia* (Davies, 1975) and other parasitoid wasps (Williams, et al, 2015), and sucrose-based foods are recommended for keeping *Nasonia* stocks alive in the absence of pupal hosts (Ward’s, 2002). Based on this information, combined with the results of
some small pilot experiments, a range of sucrose/water based foods were prepared. Pilot experiments showed that concentrations of 20% and 30% sucrose appeared to improve survival. These concentrations of sucrose were therefore included along with a 0% sucrose control. 1.5% bioagar was added to the food so that it formed a solid gel food suitable for use in DAM glass tubes. 1% nipagen (20% in ethanol) was added as an anti-fungal agent. Behaviour of 32 male wasps per condition were monitored over one month (Figure 2.2) in constant light, after 4 days of 12:12 light-dark entrainment, using the DAM Trikinetics system.

The time of last movement was noted for each wasp and used as the time of death. The survival of wasps given 20% food is better than the survival of wasps given only a water/agar food preparation. Wasps given a 30% sucrose solution survived for significantly longer than both other groups, with next to no wasps dying in the first 20 days of analysis, more than adequate for circadian experiments.

![Figure 2.2](image)

**Figure 2.2.** Survival curves for *Nasonia vitripennis* wasps on three different sucrose-based foods. 64 individual wasps in each condition were monitored and times of death recorded.
2.4 Behavioural monitoring

Wasps used for behavioural monitoring were isolated upon eclosion and transferred to DAM Trikinetics tubes with a 30% sucrose food. For the temperature compensation experiments, wasps were maintained in LED light boxes with white light. Light intensity for the temperature compensation experiments was 2-17 lum/ft². Wasps were entrained for four days in 12:12 light-dark cycles before being subjected to 8 days of constant conditions, either constant light (LL) or constant darkness (DD), then returned to 12:12 light-dark for four days. The temperatures tested were 16°C, 19°C, 21°C, 24°C, and 27°C, though temperatures in constant conditions varied slightly depending on whether lights were on or off. This is reflected in the analysis of the data - when modelling the response of circadian period to temperature, the exact recorded temperature within each light box was used rather than the target temperature set on the incubator.

2.5 Hypergeometric tests

For the analysis in chapters 3 and 5, GO term overrepresentation was performed within WaspAtlas (Davies and Tauber, 2015) using the Nvit_2.1 NCBI annotation dataset. Given the resources available at the time, GO term overrepresentation analysis in chapter 4 was performed on both the CNEs and pseudo CNEs using BiNGO (Maere, et al, 2005), a plugin for Cytoscape (Shannon, et al, 2003). Annotation obtained from the Gene Ontology Consortium (Ashburner, et al, 2000) was used (data-version: 2013-08-23) and tested against a background annotation set, formed by combining annotation derived from the D. melanogaster Ensembl

All additional hypergeometric tests were performed within R using the ‘phyper’ function. Mfuzz expression clusters with rhythmic components were identified by collapsing the fuzzy clusters into hard clusters using the ‘cluster’ property of the Mfuzz object and performing hypergeometric tests to identify clusters with enrichment for rhythmic transcripts. Thirty tests were performed for each condition (i.e. one for each cluster), and were corrected per-experiment (i.e. separately for DD and LL) using the Benjamini-Hochberg method in R (R Development Core Team, 2008).

### 2.6 Hierarchical clustering of hypergeometric p-values

For a given analysis, GO terms were selected for hierarchical clustering if they were significantly (q < 0.01) overrepresented in at least one condition. This analysis was conducted in order to display groups of GO terms which exhibit similar regulation across groups of conditions. For each selected GO term, the absolute difference between p-values (within each condition) was summed over all conditions (Figure 2.3). Hierarchical clustering of hypergeometric p-values was performed in R using the hclust function. This was used as the distance metric for clustering. GO terms were sorted based on the results of the clustering and presented using the ggplot R package.
Figure 2.3. Schematic diagram of hierarchical clustering of p-values, showing four go terms and their associated p-values for four conditions. All pairwise comparisons are shown, similar p-value profiles are shown in green, dissimilar comparisons in red. The resulting hierarchical clustering dendrogram is shown below, with ordered nodes.

2.7 Reciprocal best BLAST hits (RBH)

RBHs are calculated by a two-way comparison; for example if the best BLAST hit for a gene A in *Nasonia* is gene X in *Drosophila*, then it is called an RBH if and only if the best BLAST hit for gene X in *Drosophila* is gene A in *Nasonia* (Figure 2.4). This method is probably the most common working definition of orthology (Ward and Moreno-Hagelsieb, 2014). The number of *N. vitripennis* orthologs calculated for each genome is shown in Appendix 2.1.
Figure 2.4: An illustration of a reciprocal best blast hit strategy for ortholog identification, with a single ortholog identified (in white). Best hits from organism 1 to organism 2 are shown in red, from organism 2 to organism 1 are shown in blue.

The RBH method was judged more appropriate, for the purposes of this research, than more comprehensive ortholog search methods such as orthoMCL (Li, et al, 2003) through a performance comparison of *Apis mellifera* and *Nasonia vitripennis* ortholog detection. While orthoMCL detects orthologs for more genes than the RBH method, RBH detects orthologs for most of these genes (>70%, Figure 2.5). Due to the higher sensitivity of orthoMCL to events such as gene duplication, the pairwise comparisons necessitated by an orthoMCL search (217,485) compared to those (best BLAST hits only) necessitated by an RBH search (7,435) would have been computationally prohibitive for the analysis presented in chapters 4 and 5.
Figure 2.5. Comparison of ortholog identification methods.
3 Circadian behaviour and transcription

3.1 Introduction

For a biological rhythm to be classified as a circadian rhythm, it must fulfil three criteria (Dunlap, Loros & DeCoursey 2004). Firstly, the rhythm must be endogenous, and have a period of roughly 24 hours. Secondly, the rhythm must be entrainable by external zeitgebers, such as light and/or temperature. Thirdly and finally, the rhythm must be temperature compensated, i.e. the 24 hour rhythm must be approximately maintained at a range of temperatures. As biochemical reactions generally speed up at higher temperatures and slow down at lower temperatures, a temperature compensation mechanism is necessary to ensure that the clock can maintain its periodicity with variations in temperature.

The temperature independence of the *Drosophila* circadian period was demonstrated early on in the field (Pittendrigh 1954). The presence of temperature compensated rhythms has been demonstrated in other hymenopterans, for example the Japanese honeybee *Apis cerana japonica* shows temperature compensation of its circadian period within the range shown by other animals (Fuchikawa, Shimizu 2007), but less strictly than *Drosophila*. Interestingly, the period in *Apis mellifera* is approximately 24h at 35°C, but drops to 22.7h at 25°C (Giannoni-Guzman et al. 2014), perhaps indicating relaxed selection for a temperature compensated clock on an animal which maintains a constant 35°C in the hive. The molecular basis by which temperature compensation of the clock is achieved is currently unknown and remains an important question in circadian research.

---

1 The research presented in this chapter is under review (Davies & Tauber, 2016, Proc. R. Soc. B)
As well as working to understand the fundamental properties of circadian rhythms in *Nasonia*, there is also the motivation to understand its molecular workings and its outputs. A primary output of the clock is circadian regulation of transcription, a trait which has been demonstrated in mammals (Hughes et al. 2009), insects (McDonald, Rosbash 2001a), plants (Schaffer et al. 2001), and even bacteria (Woelfle, Johnson 2006). Whole-transcriptome circadian studies in *Drosophila* and mammals have shown that while the majority of clock-controlled genes cycle at relatively low amplitudes, a small group of genes is found to cycle with a high amplitude (Hughes et al. 2009, Hughes et al. 2012, Li et al. 2015). This small group of high amplitude cyclers includes most of the core clock genes. It would therefore be reasonable to expect that a transcriptome-wide study of *Nasonia* would provide a set of candidate genes likely to be involved in the core clock mechanism. Aside from this aim of identifying core clock genes, analysing transcriptional oscillations in clock-controlled genes (CCGs) is a key step in understanding how the daily rhythms produced by the clock are ultimately linked to behavioural phenotypes.

This work presented in this chapter aims to achieve two goals. Firstly, to characterise the temperature compensation phenotype (or lack of) in the *Nasonia* clock by profiling circadian behaviour at a range of temperatures in both sexes and in both constant darkness and constant light. Secondly, I aimed to characterise circadian gene expression in the *Nasonia* head in both of these constant conditions using RNA-seq. This is the first circadian RNA-seq study performed in an insect other than *Drosophila*, and the first study to profile the circadian transcriptome oscillating under constant light.
3.2 Methods

3.2.1 Temperature compensation analysis

The circadian data was analysed using the BeFly excel package (Allebrandt et al. 2013). The bin size for activity monitoring was set to 30 minutes. Dead wasps were removed from the data before analysis. The first two days of constant conditions were removed from the data to get a better estimate of the period once the wasps had acclimatised. The Cosinor method (Cornelissen 2014) was used to fit and test a range (8 hr to 30 hr) of periods to the data. A Monte Carlo approach was used to estimate significance using 100 permutations. If the F-value of the cosinor fit was in the top 5% of simulations, the wasp was classified as rhythmic and the circadian period for each wasp was the period of best fit.

For the temperature compensation analysis, the Q10 was calculated for each condition and sex. The Q10 metric represents the rate by which a reaction increases with each 10°C increase in temperature. A Q10 of below 1 therefore represents a decrease in the rate of a reaction, above one represents an increase, and a Q10 of close to 1 represents a reaction that is controlled with respect to temperature. The Q10 is calculated using the reaction rates at two different temperatures using the formula:

\[
Q10 = \left( \frac{R_2}{R_1} \right)^{\frac{10}{T_2-T_1}}
\]

Where \(T_1\) and \(T_2\) represent the lower and upper temperatures at which the reaction rate was measured, and where \(R_1\) and \(R_2\) represent the reaction rates measured at the lower and upper temperatures.
3.2.2 RNA-seq sample collection

To obtain male wasps for collection, groups of eight females were isolated at the yellow pupal stage and transferred onto fresh hosts upon eclosion. The resulting male progeny were collected upon eclosion and moved onto 25 x 95 mm plastic vials with a 30% sucrose agar medium, in groups of 20.

During entrainment (four full days in an LD 12:12 cycle) and collection, wasps were kept in four LED light boxes in the same incubator at 19°C. Starting at CT1 (one hour after the wasps were released into free-run conditions), wasps were collected every four hours over a period of 48 hours and snap-frozen in liquid nitrogen and immediately transferred to storage at -80°C. Wasps were collected sequentially from different light boxes every four hours to minimise disturbance of wasps, and so that wasps were collected from each light box once every 16 hours, thereby minimising the effect of variations within light boxes. Light intensity recordings showed inter-light-box variation of 6-21 lum/ft² during the entrainment period of DD and 0-13 lum/ft² during the entrainment and constant period of LL. Normal circadian behaviour was verified in conditions matching the experimental conditions using groups of 32 male wasps (Figure 3.1).
**Figure 3.1.** Median activity of 32 isolated wasps in each experimental condition, DD (top) and LL (bottom). Lights on/off are shown in orange, as recorded by activity monitors. Colours of lines respond to the four lightboxes used in the experiment.

### 3.2.3 RNA extraction, sequencing, and read mapping

RNA was extracted from pooled groups of 50 heads for each sample, using Trizol RNA extraction protocol, and followed by clean-up using the RNAeasy spin column kit (Qiagen). Samples were polyA selected and sequenced at Glasgow Polyomics (University of Glasgow, United Kingdom) on the Illumina NextSeq500 platform, resulting in approximately 20 million 75bp paired-end reads per sample.
This number of reads per sample was chosen as appropriate for this experiment as this was the maximal number of reads available under cost limitations, rendering each library robust to potential variations in library quality.

Read mapping was achieved with Tophat2 (v2.1.0) (Trapnell et al. 2012) against the Nasonia Nvit_2.1 NCBI annotation. As the purpose of this study was not to identify novel splice variants or improve on existing annotation, novel junction detection was disabled for accurate quantification of known transcripts. Mean mapping efficiency was above 90% for both conditions. Read quantification was performing using the DEseq normalisation method (Anders, Huber 2010). This method normalises between-sample expression values by calculating a scaling factor for each sample, based on differences in the geometric means. The assumption of this method is therefore that the majority of genes are not differentially expressed between samples. All 24 samples from both conditions were grouped together to allow comparison between as well as within experiments.

3.2.4 Expression profile clustering

Isoform expression profiles were first filtered to include only those isoforms with no missing values at any time-point in either condition. Expression values were standardised using the ‘Standardise’ function in Mfuzz (Kumar, E Futschik 2007). The ‘cselection’ function in Mfuzz was used to select an appropriate c-value for the c-means clustering (default parameters; m=1.25), resulting in a value 30 clusters per condition based on the number of empty clusters at each iteration. Based on this analysis, thirty fuzzy clusters were generated for each condition using the fuzzification parameter m=1.25. By definition of fuzzy clustering, all genes belong to all clusters with varying membership values.
3.2.5 Identifying rhythmic expression

To detect rhythmic isoforms at a period of 24 hours, RAIN (Thaben, Westermark 2014) was used on all filtered isoforms (i.e. those with no missing values at any timepoint) in either condition. As a non-parametric method, RAIN only facilitates detection of rhythmic isoforms with periods which are a multiple of the sample resolution (in this case 4 hr). The p-values produced by RAIN were corrected to q-values using the Benjamini-Hochberg method (Benjamini, Hochberg 1995). This method was repeated using expression values for genes rather than transcripts for the clock gene analysis (i.e. the summed expression values for all known transcripts of a particular gene). The rhythmicity analysis was repeated using JTK cycle, also searching for a 24 hour periodicity. JTK cycle is a non-parametric algorithm designed to detect rhythmic transcripts in large datasets, based on the ‘Jonckheere-Terpstra’ (JT) test which detects monotonic orderings of data over independent groups (Hughes, Hogenesch & Kornacker 2010). While JTK cycle and RAIN are both non-parametric, RAIN enables the detection of waveforms with asymmetrical rises and falls.

3.2.6 Analysis of circadian dynamics of rhythmic transcripts

Maximum fold changes in expression were calculated by normalising per-condition expression values by the median value and calculating the ratio from the lowest expression over 48 hours to the highest. Reliably quantified transcripts were defined as those transcripts where the absolute FPKM value is 5 or above at all timepoints. The threshold for this was set at a similar level to other analyses (Hughes et al. 2012).
To analyse the period of rhythmic transcripts, sine waves with a variety of periods were fitted (20 to 28 hrs in steps of 0.2 hrs) to all transcripts identified as rhythmic (q < 0.1) in both conditions. This FDR threshold is in line with, or more strict, than thresholds chosen in other similar studies (Hughes et al. 2012, Huang et al. 2013, Keegan et al. 2007). Those transcripts (85 in total) which showed a significant (q < 0.1) fit to the model in both conditions were analysed in terms of their best fitting period.

3.2.7 Phylogenetic analysis of opsin genes

Opsin genes were searched for using NCBI BLASTP using six species; *Apis mellifera*, *Bombyx mori*, *Drosophila melanogaster*, *Mus musculus*, *Nasonia vitripennis*, and *Homo sapiens*, using the *Nasonia LOP1* protein sequence as a query. BLAST results were inspected and 7e-19 was chosen as an appropriate cut-off to include all opsin sequences based on manual inspection of BLAST results. Sequences scoring below 7e-19 in this test were found to be unrelated gPCR proteins.

3.2.8 Comparison with Drosophila microarray studies

For comparison to microarray studies, rhythmic genes in *Drosophila melanogaster* were obtained from a meta-study of circadian microarray data (Keegan et al. 2007). The 214 obtained FlyBase identifiers were converted to the latest identifiers using the validation tool on FlyBase, resulting in 218 unique identifiers (this increase in identifiers can be attributed to previous identifiers referring to multiple genes in the current annotation). Orthologs for these *Drosophila* genes were obtained through WaspAtlas, retrieving orthologs for 135 genes which mapped to 173 unique
Nasonia genes due to gene duplications, etc. This set of 173 genes was compared with the number of genes with rhythmic transcripts that would be expected by chance using a hypergeometric test.

3.2.9 Treatment of alternative transcripts

All analyses (except where otherwise indicated) were performed at the transcript level, using cuffnorm quantifications based on the Nvit_2.1 NCBI annotation of alternative transcripts. This was done to capture those genes which may, in theory, have alternative transcripts which are differentially regulated in a circadian manner (for example a gene for which one transcript is more highly expressed during the subjective day, and another more highly expressed in the subjective night). All alternative transcripts were included in both the clustering and RAIN analyses.
3.3 Results

3.3.1 Temperature compensation

Taking advantage of the opportunity in *Nasonia* to study circadian rhythms in both LL and DD, behaviour was monitored in both males and females, in both constant conditions, at various temperatures (16°c, 19°c, 21°c, 24°c, and 27°c). 32 wasps were monitored in each condition, resulting in behavioural profiles for a total of 640 individual wasps. Period in response to temperature of wasps identified as rhythmic was analysed (Figure 3.2), splitting by both sex and constant condition (LL or DD).

**Figure 3.2.** Effect of temperature (x-axis) on free-running circadian period (y-axis) for both female wasps (left) and male wasps (right) in both constant darkness (blue) and constant light (red). The fit of the linear model is shown in black for each sex and condition. Individual data points represent individual wasps, in many cases overlapping with one another.
A three-way ANOVA with all interactions considered was performed on rhythmic wasps, analysing period as a function of sex, condition, and temperature. Temperature was included as a quadratic term to model the curved response in period seen in the data. Significant effects of temperature (p < 4e-6) and condition (p < 2e-16) were observed, as well as a significant interaction between temperature and condition (p < 0.05). No significant effect of sex (p > 0.05) was found on the circadian period, nor any interactions between sex and any other independent variable (p > 0.05).

The q10 for each condition and sex was calculated using the difference between the lowest and highest temperatures examined, i.e. between 16°C and 27°C. The q10 was calculated as described in 3.2.1 and compared with results from similar studies. The results of this are shown in Table 3.1, showing that Nasonia have tight temperature compensation.

**Table 3.1.** Temperature compensation Q10 values for both sexes in both conditions. T1 and T2 represent the temperatures between which the Q10 was calculated, and R1 and R2 represent the mean period among all rhythmic wasps at the respective temperatures.

<table>
<thead>
<tr>
<th>Condition</th>
<th>T1</th>
<th>T2</th>
<th>R1</th>
<th>R2</th>
<th>Q10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Males, LL</td>
<td>17.18</td>
<td>29.79</td>
<td>21.63</td>
<td>19.96</td>
<td>0.94</td>
</tr>
<tr>
<td>Females, LL</td>
<td>17.18</td>
<td>29.79</td>
<td>22.77</td>
<td>21.21</td>
<td>0.95</td>
</tr>
<tr>
<td>Males, DD</td>
<td>16.45</td>
<td>28.3</td>
<td>24.3</td>
<td>24.26</td>
<td>1</td>
</tr>
<tr>
<td>Females, DD</td>
<td>16.45</td>
<td>28.3</td>
<td>24.75</td>
<td>23.56</td>
<td>0.96</td>
</tr>
</tbody>
</table>
RNA sequencing sample collection

*N. vitripennis* females show considerably less robust rhythms in constant light (Bertossa et al. 2013). Based on this data, I entrained newly-eclosed male *N. vitripennis* wasps in two experiments for four days and collected wasps every four hours in constant conditions beginning at CT1. LL and DD were profiled independently, i.e. these two experiments were not run in parallel. For each experiment, groups of 50 male wasp heads were collected at each time point, RNA was extracted and sequenced, and one 75bp paired-end sample was obtained for each time point in each condition. Reads were mapped to the *N. vitripennis* genome and transcripts quantified. This ultimately resulted in 24 libraries - 12 libraries for LL and 12 libraries for DD, representing 12 timepoints (separated by 4 hours) over a total of 48 hours. An average of 20 million reads per sample was obtained, with an average mapping efficiency of above 90% for all samples.

### 3.3.2 Identifying rhythmic transcription

An unbiased clustering analysis was first performed to ascertain the kinds of expression patterns present in each experiment, and in order to select an appropriate algorithm for analysis of the data. To this end, Mfuzz (Kumar, E Futschik 2007) was used to carry soft c-means clustering, a method which is less sensitive to biological noise than traditional clustering (Futschik, Carlisle 2005). After filtering (see Methods 3.2.4), thirty clusters were generated for each condition (Figures 3.3 and 3.4), revealing a variety of potentially rhythmic and non-rhythmic expression trends. Potential asymmetric wave forms were detected in LL (i.e. wave forms with a steeper rise than fall, or vice versa, e.g. Figure 3.4, clusters 22 and 26). The
obtained expression patterns in LL, in general, were similar to the obtained expression clusters for DD.

With fuzzy clustering, each transcript belongs to all clusters, but with varying membership values for each cluster. Stronger membership values represent transcripts with longitudinal expression closer to the core (average) expression of that cluster. When considering the best fit for each transcript, the median size of each cluster was 692 for DD, and 674 for LL.

**Figure 3.3.** Normalised expression of thirty clusters produced by a fuzzy c-means clustering analysis in DD. Deeper red colours represent transcripts with strong membership to each individual cluster, and blue colours represent transcripts with...
weaker membership values. The x-axis represents time, and the y axis represents expression.

**Figure 3.4.** Normalised expression of thirty clusters produced by a fuzzy c-means clustering analysis in LL. Colours and labelling are as in Figure 3.3.
To identify rhythmic transcripts, the RAIN algorithm was used (Thaben, Westermark 2014). RAIN is a non-parametric test which treats the rising and falling portions of waves separately, allowing for the detection of asymmetric wave forms. The number of rhythmic transcripts at various false discovery thresholds is shown in Table 3.2.

Table 3.2. Number of rhythmic transcripts detected by RAIN at various thresholds

<table>
<thead>
<tr>
<th>FDR</th>
<th>RAIN rhythmic isoforms</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DD</td>
<td>LL</td>
<td>Overlap</td>
</tr>
<tr>
<td>0.05</td>
<td>334</td>
<td>59</td>
<td>4</td>
</tr>
<tr>
<td>0.1</td>
<td>1057</td>
<td>929</td>
<td>150</td>
</tr>
<tr>
<td>0.2</td>
<td>2696</td>
<td>2552</td>
<td>695</td>
</tr>
<tr>
<td>0.3</td>
<td>3942</td>
<td>4086</td>
<td>1409</td>
</tr>
</tbody>
</table>

At a false discovery rate (FDR) threshold of 0.1 (adjusted p-value, i.e. q-value) I identified 1,057 rhythmic transcripts in DD and 929 in LL. 77% of RAIN-identified rhythmic transcripts in DD and 52% of RAIN-identified rhythmic transcripts LL were detected by RAIN at an FDR threshold of 0.6, showing both that RAIN is more sensitive in the case of this data and that both algorithms identify a similar set of transcripts. (Table 3.3).
Table 3.3. Rhythmic transcripts identified at various FDR thresholds by JTK_CYCLE, along with the number of RAIN-rhythmic transcripts identified by JTK at each threshold.

<table>
<thead>
<tr>
<th>FDR</th>
<th>JTK</th>
<th>DD RAIN rhythmic transcripts</th>
<th>JTK</th>
<th>LL RAIN rhythmic transcripts</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>1</td>
<td>1 / 1057</td>
<td>0</td>
<td>0 / 929</td>
</tr>
<tr>
<td>0.1</td>
<td>6</td>
<td>6 / 1057</td>
<td>0</td>
<td>0 / 929</td>
</tr>
<tr>
<td>0.2</td>
<td>9</td>
<td>9 / 1057</td>
<td>0</td>
<td>0 / 929</td>
</tr>
<tr>
<td>0.3</td>
<td>161</td>
<td>154 / 1057</td>
<td>0</td>
<td>0 / 929</td>
</tr>
<tr>
<td>0.4</td>
<td>461</td>
<td>394 / 1057</td>
<td>29</td>
<td>29 / 929</td>
</tr>
<tr>
<td>0.5</td>
<td>1078</td>
<td>701 / 1057</td>
<td>343</td>
<td>266 / 929</td>
</tr>
<tr>
<td>0.6</td>
<td>1561</td>
<td>812 / 1057</td>
<td>917</td>
<td>484 / 929</td>
</tr>
<tr>
<td>0.7</td>
<td>2110</td>
<td>898 / 1057</td>
<td>1933</td>
<td>683 / 929</td>
</tr>
<tr>
<td>0.8</td>
<td>2755</td>
<td>960 / 1057</td>
<td>2699</td>
<td>779 / 929</td>
</tr>
<tr>
<td>0.9</td>
<td>3539</td>
<td>1003 / 1057</td>
<td>3527</td>
<td>858 / 929</td>
</tr>
</tbody>
</table>

To test whether the rhythmic transcripts identified in RAIN were overrepresented in any particular Mfuzz cluster, hypergeometric tests were employed. Clusters found to have an overrepresentation of rhythmic transcripts (q < 0.05, Appendices 3.1 and 3.2) are shown in Figure 3.5. These clusters represent
Figure 3.5. Normalised expression of clusters with significant ($q < 0.01$) overrepresentations of rhythmic genes. Each transcript profile in each cluster is coloured by that gene's membership of the cluster, as in Figures 3.3 to 3.4. Normalised and standardised expression is shown on the y-axis of each plot. Clusters are ordered according to the number of transcripts present in each cluster.

Rhythmic transcripts ($q < 0.1$) were sorted by phase, peak shape, and significance, and plotted (Figure 3.6A). Examining the phase distribution (Figure 3.6B), it is apparent that the majority of transcripts show peak expression early in the subjective morning or in the subjective night, with fewer transcripts peaking at intermediate times. This trend is also apparent in the phases reported by JTK cycle (Figure 3.7), and so is not algorithm specific. This disparity in phase is greater in
the transcripts which show rhythmic expression in both DD and LL; less than 12% of transcripts in DD and less than 5% in LL show peak expression at intermediate times (Figure 3.6B). The majority of these transcripts (~87%) exhibit a similar (+/-4 hrs) phase in LL to their phase in DD.

Figure 3.6. Circadian transcriptional rhythms. (A) Heatmap of median-normalised expression of rhythmic (q < 0.1) transcripts in both constant darkness and constant light over time. High expression of each transcript is shown in yellow, low expression in blue. (B) Histograms and heatmap of phases of rhythmic transcripts (q < 0.1 in both conditions), showing bimodal phase distribution through the histograms and showing overlap between the two conditions through the central heatmap.
Figure 3.7. Phase histograms of phase distributions produced by JTK cycle, showing a bimodal distribution of phases in DD (blue, left) and LL (red, right). Each observation in these histograms represents the phase of the waveform that JTK fit for a single transcript.

Similarly to *Drosophila* (Hughes et al. 2012) and mammals (Hughes et al. 2009), the majority of transcripts show only small cyclic changes in expression amplitude over the day; over 80% of reliably quantified (see Methods 3.2.6) transcripts in both conditions have amplitudes (peak expression divided by trough expression) of 2 or less. In both DD and LL, transcripts with exceptionally high amplitudes (> 4) are transcripts with unusually low or high measurements at isolated time-points with no obvious specific shared function. This trend is also visible in the JTK cycle rhythmic transcripts; JTK reports that all reliably quantified cycling transcripts cycle with a median-normalised amplitude of less than 0.6, and more than 90% of transcripts cycle with a median normalised amplitude of less than
0.2 (Figure 3.8). This is in contrast with results in *Drosophila* and mammals, where some core clock genes exhibit very high amplitude oscillations (Hughes et al. 2009, Hughes et al. 2012, Li et al. 2015).

![Histograms showing median-normalised amplitudes](chart)

**Figure 3.8.** Median-normalised amplitudes as reported by JTK cycle, showing low cycling amplitudes for all transcripts in DD (blue, left) and LL (red, right), mirroring results shown in Figure 3.6.

### 3.3.3 Canonical clock genes, comparison with *Drosophila*, and regulatory elements

The canonical clock genes were examined for rhythmicity both at the transcript level and via an additional RAIN analysis at the gene level. The q-values for the canonical clock genes are shown in Table 3.4, and their expression levels across LL and DD are shown in Figure 3.9. Evidence for rhythmicity was found in some clock genes, which included *pdp1e* (q ~ 0.1, LL and DD), *shaggy* (q < 0.1, DD), and *Clock* (q ~ 0.1, LL).
Table 3.4. Clock genes along with the lowest RAIN q-value among all transcripts analysed.

<table>
<thead>
<tr>
<th>Gene name</th>
<th>Lowest RAIN q-value (transcript)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DD</td>
</tr>
<tr>
<td>period</td>
<td>0.34</td>
</tr>
<tr>
<td>cryptochrome</td>
<td>0.94</td>
</tr>
<tr>
<td>Clock</td>
<td>0.44</td>
</tr>
<tr>
<td>cycle</td>
<td>0.49</td>
</tr>
<tr>
<td>timeout</td>
<td>0.97</td>
</tr>
<tr>
<td>dbt</td>
<td>0.23</td>
</tr>
<tr>
<td>pdp1ε</td>
<td>0.11</td>
</tr>
<tr>
<td>vrille</td>
<td>0.88</td>
</tr>
<tr>
<td>shaggy</td>
<td>0.08</td>
</tr>
<tr>
<td>clockwork orange</td>
<td>0.25</td>
</tr>
</tbody>
</table>
Figure 3.9. Transcriptional profiles of clock genes in DD (blue) and LL (red) over time (x-axis), showing expression on the y-axis.

The transcripts identified as cycling in Nasonia heads were compared with the transcripts identified as cycling in Drosophila heads. For these purposes, I used a list of genes identified in a meta-analysis study of Drosophila circadian microarray data as being rhythmically expressed in either LD or DD (Keegan et al. 2007). Of 173 genes identified as rhythmic in Drosophila, 33 genes (Appendix 3.3) were found to also be rhythmic in Nasonia (either in LL or DD, q < 0.1), no more than would be expected by chance (p = 0.11, hypergeometric test).
Additionally, the rhythmic genes were compared with a study examining the *Nasonia* photoperiodic response. Four genes were found to be rhythmic in either DD or LL and also differentially methylated in long and short photoperiods (Pegoraro et al. 2016). These genes are the UPF0430 protein CG31712, Inositol hexakisphosphate and diphosphoinositol-pentakisphosphate kinase, Carboxy-terminal domain RNA polymerase II polypeptide A small phosphatase 1, and the endoplasmin gene.

In order to identify potential shared regulatory transcription factor motifs responsible for the rhythmic transcription of groups of genes, the 2 kbp region upstream of each gene in each cluster (using the cluster of best fit) was extracted, and analysed using the de-novo sequence motif elicitation software MEME (Bailey et al. 2009), using default settings. None of the 30 clusters analysed for either condition returned a significant motif.

### 3.3.4 Functions of rhythmic genes

To capture the general functions that rhythmic genes may fulfil in *Nasonia*, a broader set of rhythmic genes (FDR < 0.2 in RAIN) was used for GO term overrepresentation (Davies, Tauber 2015, Ashburner et al. 2000), revealing 94 GO terms overrepresented for genes rhythmic in DD (including ‘response to light stimulus’, ‘proteasome complex’, and ‘generation of neurons’, and 123 terms for genes rhythmic in LL (including ‘locomotion’, ’proteasome complex’, and ‘response to external stimulus’. Hierarchical clustering of hypergeometric p-values (Figure 3.10) does not reveal any clear distinction between DD and LL in terms of overrepresented functions, with differences between the two conditions primarily consisting of terms with few associated genes (i.e. liable to disruption by noise).
From this analysis, it appears that genes cycling in DD and LL fulfil similar functions. Significantly overrepresented GO terms shared between both conditions include those related to neurons, signal transmission, and responses to stimuli. Notably, all four *Nasonia* opsins were found to exhibit similar transcriptional profiles in LL and DD, with low expression in the morning and high expression in the evening (Figure 3.11).

**Figure 3.10.** Hierarchical clustering of hypergeometric p-values for terms significantly overrepresented in at least one condition, showing terms represented in DD-only cycling genes (q < 0.2, left column), LL-only cycling genes (q < 0.2, middle column), and genes cycling in both conditions (q < 0.2, right column) - each column therefore represents a distinct set of genes and thus may result in three distinct functional profiles. Dark blocks represent those GO-terms highly
overrepresented in the given condition, light blocks represent those GO-terms not found to be highly overrepresented.

![Expression patterns of all four Nasonia opsin genes over 48 hours](image)

**Figure 3.11:** Expression patterns of all four *Nasonia* opsin genes over 48 hours are shown here, showing expression in DD (blue) and higher expression in LL (red).

It has previously been demonstrated that the timing of different (or indeed opposing) biological processes can be controlled through the circadian regulation of groups of genes (Sancar et al. 2015, Zhang et al. 2014). Unsupervised clustering methods have previously been established as a useful method for functional characterisation of circadian genes (Nguyen et al. 2014). To establish whether temporal separation of functions occurs in *Nasonia*, I therefore returned to the expression clustering analysis, using all 30 clusters identified in each condition. Examples of clusters with enriched functions include clusters DD7 and LL20 which are significantly enriched for catalytic activity GO terms (and therefore may
compromise homologous clusters), especially genes involved in the proteasome, and clusters DD24 and LL6 which are both involved in circadian and neural processes. Non-rhythmic clusters such as DD10 and LL24 were found to be overrepresented for cytosolic ribosomal genes and cellular anabolism genes. Other clusters (DD1 and DD2) did not turn up any overrepresented GO terms and are thus likely comprised of genes with a wide range of functions.

3.3.5 Transcriptional differences between constant darkness and constant light

To examine whether differences in circadian period seen in locomotor activity between DD and LL could also be detected in transcriptional rhythms, I fitted parametric models with a range of periods to transcripts rhythmic in both conditions (q < 0.1). For those transcripts with statistically significant fits to the model in both conditions (q < 0.1, see Methods 3.2.6), I took the period with the best fit and compared these periods between conditions. Overall, transcripts in LL showed a significantly (p < 3.9e-09, Wilcoxon rank sum test) shorter (median 24) period than those in DD (median 25.4), mirroring the behavioural differences in period. Despite the independent nature of the two experiments and the lack of biological replicates, I also tested for differential expression between DD and LL to give preliminary indications of potential transcriptional differences between these two conditions. DESeq2 was used to examine differential expression analysing 26,595 transcripts in total, using a threshold of an FDR-adjusted p-value of 0.05, yielding 3,246 transcripts expressed higher in DD than LL and 3,193 transcripts expressed higher in LL than DD (Figure 3.12). Transcripts more highly expressed in DD were significantly enriched (q < 0.01) for genes involved in various forms of catalytic activity, including the vast majority of proteasome genes (>85%). Transcripts more
highly expressed in LL were enriched terms including cell-cell signalling terms, and
terms related to stimulus responses, including light. These data together show that
metabolism and defence response are overrepresented in genes associated with
transcripts more highly expressed in DD, and that terms involved in detection of
light and cell-cell signalling are overrepresented in genes more highly expressed in
LL.

Figure 3.12. (A) FPKM (log2) expression of transcripts in DD (x axis) and LL (y
axis), showing transcripts classified (> 1.5 median fold change) as differentially
expressed up in DD (blue) and up in LL (red). (B) Overrepresentation of selected
GO terms up in DD, showing overrepresentation of genes up-regulated in DD
relative to the genome annotation.
3.4 Discussion

In the first section of this chapter, I have shown that the *Nasonia* circadian clock is temperature compensated; the circadian period remains relatively constant over the temperatures tested. The temperature compensation response in *Nasonia* is better represented by a curved model than a linear one. Period initially lengthens with increased temperature, but begins to shorten at around 22°C and by 28°C is significantly shorter. The initial lengthening in period is superficially counter-intuitive: if the speed of biochemical reactions increases at higher temperatures, why would the inherent period of the clock increase rather than decrease? This phenomenon could perhaps be explained by the compensation mechanism acting to lengthen the period, with initial overcompensation and gradual decline in effectiveness after 22°C.

The complexity of the interaction between temperature and circadian period is not something generally examined in studies on the topic. The majority of studies simply test the circadian period at two temperatures (typically 10°C apart) and calculate the Q10 value. Here, I tested a range of temperatures, revealing a curved response to temperature. The Q10 value was nevertheless calculated between the lowest and highest temperature measured, revealing that *Nasonia* has tight temperature compensation, well within the standard temperature compensation range (0.8-1.2) reported by Brady (Brady 1979).

A significant interaction was shown between condition and temperature, showing that the temperature compensation mechanism differs in action between DD and LL. Performing regression for each condition suggests that temperature is controlled more tightly in DD than in LL (DD: p > 0.05, LL: p < 0.0005). This observation is backed up by the Q10 values, which are closer to 1 for the DD
samples than in the LL samples. Also shown in this data is the observation that the period in LL is always shorter than the period in DD. The Japanese honey bee also shows a shorter period in LL than in DD, and also exhibits differences in temperature compensation between these two conditions (Fuchikawa, Shimizu 2007), suggesting that this is a conserved feature of the temperature compensation mechanism in Hymenoptera. The circadian period across temperature in the honeybee appears to be more tightly controlled in constant light than in constant darkness, leading (Fuchikawa, Shimizu 2007) to remark that this implies that the honey bee temperature compensation mechanism works more strictly in the light than in the dark. Further characterisation of the differences in the circadian clock between DD and LL would help the investigation of this issue.

This chapter also provides the first insights into global transcriptional oscillation in *Nasonia*. With RNA-seq, I profiled the circadian transcription of >26,000 transcripts in *Nasonia* in either DD or LL. At a relatively stringent FDR (q < 0.1), 1,057 cycling transcripts were identified in DD and 929 cycling transcripts in LL. These transcripts correspond to a cycling fraction of 6.7% and 5.9% of all transcripts analysed in DD and LL respectively. These figures are comparable to cycling fractions reported in various organisms and tissues, generally between 2% and 10% of the transcriptome (Michael, McClung 2003). This is the first study to examine the circadian transcriptome in *Nasonia*, previous studies (e.g. (Bertossa et al. 2014, Mukai, Goto 2016) have focussed on a few genes likely to be involved in circadian function, but concord with the data presented here in terms of peak expression and cycling amplitude.

In both conditions, cycling transcripts were found to cycle at low amplitudes (mostly < 2 fold) and with a limited, bimodal, range of phases. This is in contrast to
microarray/RNA-seq studies in *Drosophila*, where transcripts were found to cycle with a broader range of phases (Rodriguez et al. 2013) and studies in both mammals and *Drosophila*, which have identified a group of high-amplitude (> 4-fold) cycling genes among the low-amplitude majority (Akhtar et al. 2002). High amplitude cyclers typically include clock genes (Akhtar et al. 2002, Hughes et al. 2012). It is possible that the low cycling amplitudes of CCGs in *Nasonia* is due to cycling in a relatively small proportion of cells in the *Nasonia* head, however this does not explain why no differences are observed between core clock genes and other CCGs (i.e. why there are no high-amplitude cyclers even on a relative level). In *Drosophila,* much of this amplitude difference is applied post-transcriptionally, as demonstrated by nascent-seq (Rodriguez et al. 2013). The low cycling amplitudes in *Nasonia* could therefore imply differences in post-transcriptional regulation of CCGs. The low oscillations of the *Nasonia* head transcriptome render the expression profiles of the canonical clock genes difficult to resolve, as demonstrated through comparison of plant microarrays by (Covington et al. 2008). This issue may also contribute to the discordance between the various circadian microarray studies in *Drosophila* (Keegan et al. 2007).

It is possible that the *Nasonia* clock is able to function without high amplitude cycling, raising the question of its utility in other organisms and suggesting that the *Nasonia* circadian niche renders this function unnecessary. Alternatively, there exists the possibility that cycling in *Nasonia* occurs only in relatively few neurons, suggesting the existence of a simpler clock network. Another possibility is that high amplitude mRNAs do exist in *Nasonia,* but that there is desynchrony between cells which express these transcripts, resulting in a more constitutive readout at the level of the head.
An emerging property of the circadian transcriptome in *Nasonia* is the temporal separation of function by phase (Fig 3.5). Notably, genes involved in catalytic activity were strongly overrepresented in morning-peaking transcripts in both experiments (Results 3.3.4). This is in line with other studies which show catalytic activity confined to the morning in Fungi (Sancar et al. 2015) in agreement with a general observation that an important (or even primary) function of circadian clocks (Hurley, Loros & Dunlap 2015) is to temporally separate catabolism and anabolism. Although no overrepresentation of anabolic genes was detected within the cyclic transcripts, expression clusters DD10 and LL24 (Figures 3.2 and 3.3) did show strong overrepresentation for genes involved in cytosolic ribosomal genes ($q < 3.7.56$) and cellular anabolism ($q < 2e-06$) (3.3.4). These clusters exhibit an antagonistic expression pattern to the expression clusters containing the catabolic genes, suggesting that catabolism and anabolism are indeed separated by the circadian clock in *Nasonia*.

Though this analysis consisted of two independent experiments, the comparison of expression between LL and DD reveals that a majority of genes (>85%) involved in the proteasome and a broader set of genes involved in catabolism were more highly expressed in DD than LL. As turnover rates of clock proteins have shown to be coupled with changes in the circadian period (Syed, Saez & Young 2011, He, Liu 2005) pending confirmation, up-regulation of the proteasome may provide an explanation for differences in period observed between DD and LL. Given that temperature compensation is fundamentally a response to the effect of an increased metabolic rate on the clock, this differential expression may provide a potential testable hypothesis for the mechanism of difference in the temperature compensation phenotype between DD and LL. This conclusion would,
however, require further experimental validation (e.g. through the manipulation of proteasome expression and subsequent measurement of the circadian period).

A cursory analysis was performed for conserved sequence motifs upstream of genes involved in the Mfuzz expression clusters. No particular cluster was enriched for any sequence motif. Though this analysis does not rule out the presence of shared transcription factor binding sites, it does perhaps point to the absence of a single transcription factor being responsible for the rhythmic transcription of these clusters, revealing that this job is likely fulfilled by several regulatory mechanisms (an observation explored further in Chapter 4). A more detailed analysis could be performed using known transcription factor binding sites in other species, perhaps limiting this search to a smaller upstream sequence length or to genes only in the cores of each expression cluster.

The observation that the opsin genes are expressed more highly in constant light than in constant darkness is in line with other research from fungi to animals (Yan et al. 2014, Kihara et al. 2009). The opsin expression pattern in Nasonia is characterised by a rise in the subjective day, followed by a decline in the subjective night. In the light-dark cycles produced by the sun, the coupling of circadian control and light response would likely produce robust rhythms with higher amplitudes than seen in constant darkness and constant light. This prediction is in line with Drosophila, which sees dampened cycling amplitude in constant conditions compared to LD (Hughes et al. 2012). Given that the light input pathway in Nasonia is thus far unknown, the demonstration by this finding that the opsins in the Nasonia head respond to light is important, highlighting a likely light-input pathway and providing a start to research in this area.
The genes which cycle in *Drosophila* largely differ from those cycling in *Nasonia*. These differences could result from the lack of an authoritative set of cycling genes in the *Nasonia* head, which will be resolved through further RNA-seq/microarray studies. Alternatively, the difference in CCG identity between *Drosophila* and *Nasonia* could represent a true biological difference, and therefore would demonstrate plasticity in terms of which genes can be rhythmically transcribed in order to regulate a particular function in a circadian manner.

Examples of functions shared by CCGs in the *Drosophila* (based on the literature) and *Nasonia* heads (as per GO overrepresentation tests in 3.3.4) are: various aspects of metabolism (Rodriguez et al. 2013, Ueda et al. 2002, Ceriani et al. 2002, Claridge-Chang et al. 2001), phototransduction (Ueda et al. 2002, Rodriguez et al. 2013), synaptic/nervous functions (McDonald, Rosbash 2001b, Ceriani et al. 2002, Claridge-Chang et al. 2001), oxidoreductase activity (Claridge-Chang et al. 2001), mating behaviour (Rodriguez et al. 2013), and immunity (McDonald, Rosbash 2001b, Ceriani et al. 2002). The set of functions here shown to be under circadian control may open the door to the identification of further circadian phenotypes in *Nasonia*.

Generally, the circadian dynamics (phase, amplitude, etc) and functions of cycling genes were found to be similar between DD and LL. Differences were observed in the circadian period of cycling transcripts which reflected differences in the behavioural circadian period (longer period in DD than in LL), providing evidence that these transcriptional oscillations underlie behavioural phenotypes.
4 Conserved regulatory elements

4.1 Introduction

The orchestration of gene expression is accomplished through a wide variety of regulatory mechanisms. One of the most well-characterized of these mechanisms is regulation of transcription through the binding of transcription factors to regulatory DNA sequence (Weake, Workman 2010). The DNA sequences bound by transcription factors are generally short, with the average binding site length being approximately 10bp in eukaryotes (Stewart, Hannenhalli & Plotkin 2012). Other types of regulatory sequences are less well characterized; for example those sequences which become functional when transcribed into RNA. The most well-known example of this kind of regulatory element is the IRE (iron response element), a hairpin loop found in the mRNA of many genes involved in iron metabolism which helps to maintain iron homeostasis (Piccinelli, Samuelsson 2007).

Detecting regulatory elements experimentally is time consuming (Haeussler, Joly 2011), and identifying appropriate experimental targets may be difficult. Using computational methods for prediction of regulatory elements also presents issues; for example, prediction of transcription factor binding sites is usually accomplished by scanning a sequence of interest for matches to position-specific scoring matrices (PSSMs). These PSSMs (Stormo et al. 1982) describe the types of, generally short (Stewart, Hannenhalli & Plotkin 2012) sequence motifs bound by these proteins. As such, the probability of finding a chance match in a sequence of any

---

1 The study presented in this chapter has been published (Davies et al. 2015 BMC Evolutionary Biology 515:227).
considerable length is high, and the majority of predicted transcription factor binding sites are therefore likely to have no functional role; a concept dubbed the ‘futility theorem’ (Wasserman, Sandelin 2004). Many other regulatory elements are also characterized by short sequence motifs, and so identification of these elements through straightforward sequence scanning methods is subject to the same problem.

Phylogenetic footprinting is a method that can greatly reduce the search space when looking for functional regulatory elements (Visel, Bristow & Pennacchio 2007). It is based on the principle that functionally important sequence elements are more likely to be conserved over time than less (or non-) functional elements, leaving behind a ‘footprint’ of functionality. This approach can be highly successful at identifying functional regulatory elements, the specificity of detection increasing with the divergence times of the species used; for example >40% of conserved non-coding elements (CNEs) detected through a human-fugu (454.6 Myr divergence) (Hedges, Dudley & Kumar 2006) comparison showed enhancer activity when tested (Visel, Bristow & Pennacchio 2007), as opposed to only 5% of human-rodent CNEs tested (Nobrega et al. 2004). CNE detection tends to drop sharply with increased evolutionary time, for example sensitive BLAST analysis shows a clear alignment signal between similar loci of two Drosophila species (~60 Myr), but an almost complete lack of alignment between two more diverged dipteran species (~75 Myr) (Kazemian et al. 2014).

The most deeply conserved CNEs detected to date originated before the divergence of deuterostomes and protostomes, only four examples of which have been reported. The first two sequences of this kind to be discovered were found conserved between a variety of deuterostomes and a cnidarian, *Nematostella vectensis* (Royo et al. 2011), dating back over 670 million years (Peterson et al.
2008). The other two conserved sequences that predate this split were, unlike the other two sequences, found to be present in species belonging to both Deuterostomia and Protostomia (Clarke et al. 2012) and date back at least 600 million years (Peterson et al. 2008).

Here, I took advantage of the recent releases of various insect genomes to identify novel regulatory elements conserved across large (180-700 myr) evolutionary distances. The majority of phylogenetic footprinting studies in insects use the model organism *Drosophila melanogaster* as a central comparison species, aimed at finding regulatory elements conserved within the fast-evolving (Wyder et al. 2007) order Diptera. *Nasonia vitripennis*, a member of the more slowly evolving order Hymenoptera (Wyder et al. 2007), was used as a central comparison species to identify conserved regulatory elements both in the emerging model wasp and in other animal species. The aim of this chapter was to characterize a small subset of deeply conserved sequences in the upstream region of genes, thus potentially capturing both novel transcriptional and translational regulatory elements.

By using a sensitive alignment algorithm (see 4.2.2) (Krusche, Tiskin 2009) and ensuring that the analysis was conducted with a low false discovery rate, I identified a set of conserved sequences. Among the sequences identified are both known regulatory elements and a variety of novel regulatory elements in or near genes with core regulatory or developmental roles, some of which could potentially represent novel classes of RNA regulatory elements. The set of CNEs was used to examine the nature of conserved regulatory elements and their evolution. This chapter also reports the discovery of the two most deeply and ubiquitously conserved regulatory elements yet identified in the animal kingdom which date back
to the radiations of basal animal phyla and are likely over 670 (Peterson et al. 2008) and 700 million years old (Hedges, Dudley & Kumar 2006) respectively.

The research described in this chapter aims to characterise regions of DNA which regulate genes, exploring the ways in which regulatory mechanisms may act. This is an important concept in the circadian clock (See Introduction 1.3), as the transcriptional oscillations produced by the clock are ultimately aimed at the rhythmic regulation of gene output (See Chapter 3).
4.2 Methods

4.2.1 Data used for detection of conservation

Genomes of *Nasonia vitripennis*, *Apis mellifera*, *Atta cephalotes*, *Solenopsis invicta*, *Drosophila melanogaster*, *Megaselia scalaris*, *Aedes aegypti*, *Bombyx mori*, *Danaus plexippus*, *Heliconius melpomene*, *Dendroctonus ponderosae*, *Tribolium castaneum*, and *Acyrthosiphon pisum* were obtained from the core databases in Ensembl metazoa release 21 (Flicek et al. 2014). To search for conserved non-coding sequences, the 2 kb sequence upstream of each gene’s translation start site in *N. vitripennis* was extracted and compared with the sequence upstream of the orthologous gene in each comparator organism (Figure 4.1).

![Figure 4.1. A schematic diagram of the alignment strategy used in this study.](image)

The choice of 2 kb as an appropriate sequence length to analyse was twofold; firstly, based on the OGS v1.2 gene annotation, 2 kb is enough to cover over 95% of 5′ UTR sequence, secondly, it is a computationally tractable amount of sequence given time and computational resource constraints. If the 5′ end of a 2 kb sequence
overlapped with a nearby gene, the sequence was truncated. If the sequence was entirely contained within another gene, then it was removed from the analysis entirely. *Nasonia* sequences were compared against sequences from each other species.

Orthologs of *N. vitripennis* genes were computed using a pairwise reciprocal best BLAST hit (RBH) approach (Table 4.1) (Tatusov, Koonin & Lipman 1997) based on the protein sequences (protein sequences from Ensembl metazoa release 21) of all transcripts in each genome. Under this approach, only the best hit among many:many, many:1, and 1:many orthologs is included. Although this may result in false negative calls for genes with many orthologs and paralogs, the aim of this study was to characterise a few strongly-conserved elements rather than to provide a broad characterization of every conserved element in the *Nasonia* genome. RBH therefore provides a high-confidence set of orthologs suitable for the purposes of this study.

**Table 4.1.** Number of reciprocal best blast hit (RBH) genes assigned per species analysed.

<table>
<thead>
<tr>
<th>Species name (vs <em>Nasonia</em>)</th>
<th>Total annotated genes</th>
<th># genes with RBH</th>
<th>% annotated genes with RBH</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Acyrthosiphon pisum</em></td>
<td>36939</td>
<td>6623</td>
<td>17.93%</td>
</tr>
<tr>
<td><em>Aedes aegypti</em></td>
<td>17391</td>
<td>6738</td>
<td>38.74%</td>
</tr>
<tr>
<td><em>Apis mellifera</em></td>
<td>10675</td>
<td>7434</td>
<td>69.64%</td>
</tr>
<tr>
<td><em>Atta cephalotes</em></td>
<td>18534</td>
<td>8008</td>
<td>43.21%</td>
</tr>
<tr>
<td><em>Bombyx mori</em></td>
<td>14623</td>
<td>6472</td>
<td>44.26%</td>
</tr>
<tr>
<td><em>Danaus plexippus</em></td>
<td>16260</td>
<td>6842</td>
<td>42.08%</td>
</tr>
<tr>
<td><em>Dendroctonus ponderosae</em></td>
<td>13407</td>
<td>6426</td>
<td>47.93%</td>
</tr>
<tr>
<td><em>Drosophila melanogaster</em></td>
<td>15682</td>
<td>6760</td>
<td>43.11%</td>
</tr>
<tr>
<td><em>Heliconius melpomene</em></td>
<td>16203</td>
<td>6325</td>
<td>39.04%</td>
</tr>
<tr>
<td><em>Megaselia scalaris</em></td>
<td>11784</td>
<td>4262</td>
<td>36.17%</td>
</tr>
<tr>
<td><em>Solenopsis invicta</em></td>
<td>17377</td>
<td>7249</td>
<td>41.72%</td>
</tr>
<tr>
<td><em>Tribolium castaneum</em></td>
<td>16541</td>
<td>7091</td>
<td>42.87%</td>
</tr>
</tbody>
</table>
4.2.2 Aligning orthologous upstream sequences

The seaweeds algorithm provides optimal alignment scores for all pairs of possible windows across the two 2 kbp sequences, i.e. about 4 million short optimal alignments. The window length was chosen to be 50bp. The alignment score was set to 1 for a match, 0 for a mismatch, and -0.5 for any gap. The rationale for this scoring matrix, and for using alignments in general, was to perform a search without a priori knowledge of the regions in question or the types of regulatory elements likely to be found. The choice of 50 bp is a variation on a previous study (Baxter et al. 2012) which allows for greater sensitivity while maintaining specificity. An advantage of the window-based seaweeds algorithm (Krusche, Tiskin 2009) over other algorithms such as Smith-Waterman (Smith, Waterman 1981) is the avoidance of the “shadow effect” (Arslan, Egecioglu & Pevzner 2001). The shadow effect describes the phenomenon where longer, but biologically less significant alignments may be computed while different, shorter alignments are ignored. Instead all windows are considered equally and results can be easily compared and tested for statistical significance as all sequences are of equal length.

4.2.3 Scoring conservation

To calculate a conservation score for a given pairwise comparison, sequence conservation is taken into account, punitively applying information about annotated repetitive elements to produce intermediate scores. These intermediate scores are then scaled from 0 to 1 using an lower \((L)\) and upper \((U)\) threshold. Scores below \(L\) are assigned a conservation score of 0, scores above \(U\) are assigned a conservation score of 1, and scores in between are defined on a sigmoid curve to reflect an initially exponential increase in confidence as scores increase above the lower
threshold, which levels off as scores approach the upper threshold reflecting saturating confidence. The results of all the pairwise alignments were bundled together to form one inclusive dataset. Essentially, this step involves identifying individual small CNEs which actually map to the same sequence, and combining them into a single CNE (illustrated in Figure 4.2).

Figure 4.2. A schematic diagram of the bundling strategy. Three pairwise CNEs found conserved between *N. vitripennis* and two other species are bundled into a single CNE.

During the bundling process, significantly overlapping alignments, i.e. individual CNEs which score above the lower bound but which are not disjoint from one another, were merged together into longer regions and significant regions in two or
more species which mapped to the same subsequence in *N. vitripennis* were identified and merged into a single CNE. Each potential merged CNE is then assigned a *combined conservation score* (CCS); the combination of the conservation scores from each pairwise comparison. The CCS is computed using the following formula:

\[
1 - \prod_i (1 - P_i)
\]

Where \( P \) is the maximum conservation score for a potential CNE in a species, and \( i \) indexes the species.

### 4.2.4 Calculating conservation score thresholds

To set the appropriate parameters for detecting conservation, randomly paired upstream sequences (*pseudo-orthologs*) were aligned to get an idea of the sort of scores which could be expected by chance. Using real non-orthologous genomic sequence as a control is preferable to using randomized or ‘shuffled’ sequence as it maintains the complex sequence makeup of true genomic sequences, whereas in shuffled sequences these motifs will be depleted, providing a less stringent control. Candidate CNEs were first identified by running the pseudo ortholog sequences and the true ortholog sequences with lower \( L \) and \( U \) parameters \((L: 80 \ U: 94)\), and setting a CCS threshold at the level where no conservation was detected in the pseudo ortholog set. The lower and upper bound parameters were initially set based on examination of pairwise comparison histograms; the lower threshold was set at the point where scores were unlikely to be meaningful (i.e. where the control set shows a similar number of CNEs), and the upper threshold at a level where no
sequences were reported in the control, as performed by Baxter et al. (Baxter et al. 2012). These candidate CNEs were filtered for similarity to known coding sequences, and after increasing the parameters to the level where no conservation was detected in the pseudo-ortholog control \((L: 87 \ U: 100)\), these candidate CNEs were scored for conservation producing a final filtered set of 322 CNEs. At this point of filtering, the lower and upper bound are simply used to define a continuum of confidence scores for CNEs already known to be significant.

4.2.5 Filtering for coding sequences and pseudogenes

To ensure that the CNEs were unlikely to contain unannotated coding sequences or pseudogenes, BLASTX 2.2.27+ (Altschul et al. 1990) was utilised to filter out such sequences. To set an appropriate filtering threshold, the nucleotide sequences in the set to be filtered were first randomly permuted. These sequences were then scored against the nr protein database (Pruitt, Tatusova & Maglott 2007) using BLASTX, and the minimum (most significant) E-value score was noted as the most significant hit likely to be produced by random sequences with identical nucleotide composition to the set to be filtered. Once this threshold was set, the true sequences were scored against nr using BLASTX, and any sequence scoring below this threshold was discarded. The thresholds and number of sequences filtered by this method can be seen in Table 4.2.

Table 4.2. BLASTX filtering of CNE sets.

<table>
<thead>
<tr>
<th>Set</th>
<th>No. CNEs (pre-filtering)</th>
<th>Significance (E-value) threshold</th>
<th>No. CNEs (post-filtering)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNE candidates</td>
<td>383</td>
<td>0.096</td>
<td>334</td>
</tr>
<tr>
<td>Pseudo-CNEs</td>
<td>382</td>
<td>0.049</td>
<td>359</td>
</tr>
</tbody>
</table>
4.2.6 Creating pseudo CNEs

To elucidate properties of the CNEs through comparison to a control, a set of *pseudo CNEs* were generated. The set of 359 pseudo CNEs obtained were created by aligning pseudo orthologs with relatively low threshold parameters \((L: 80 \ U: 94)\), applying a CCS cutoff of 0.528 to retrieve a similar number of sequences to the true CNEs, and performing BLASTX filtering. Pseudo CNEs constitute a good control as they are similar to true CNEs in every way but for the fact are not assumed to represent true orthologous sequences. Given the method of orthology chosen, it is possible that many:many orthologs may be present within these pseudo-CNEs, but if these do occur then, due to the scoring method, this will only result in an increase of false negative results as opposed to false positives, thus resulting in an even more conservative analysis. By comparing the CNEs with these pseudo CNEs, we can therefore identify properties likely to be characteristic of the truly conserved non-coding sequences.

4.2.7 Analysis of CNE properties

All comparisons were performed on the *N. vitripennis* versions of each CNE. Statistical comparisons of the distributions of GC content, CpG observed/expected, CNE length, and CNE position were performed using the Wilcoxon rank sum test in R (R Development Core Team 2010). P-values of 2.2e-16 are at the floating point precision limit (\(p \sim 0\) at machine precision). For each CNE, the distance from translation start site was calculated as the distance from the 3’ end of the CNE from the 5’-most annotated translation start site of each gene. Unless indicated otherwise, all comparisons are between the *N. vitripennis* true CNEs and the *N. vitripennis* pseudo CNEs.
4.2.8 Nucleosome occupancy prediction and GC content analysis

The nucleosome occupancy prediction software (nucleosome_prediction.pl) described by Kaplan et al., (Kaplan et al. 2009) was used to predict nucleosome occupancy within each CNE and in the flanking region. The sequence of each CNE along with flanking sequence were extracted in order to obtain a 10 kb sequence centered on each CNE. Sequences containing Ns were removed as per the software requirements. This step removed a significant proportion of CNEs - 142 of 322 CNEs (44%). Control sets were produced by extracting, for each CNE, a region with the same properties (length, and distance from translation start site) upstream of a randomly selected gene. 10 random sets were created, and the results from these controls were averaged and the standard deviations calculated. For the GC content comparison, sequences were extracted in the same way and GC content was measured in a sliding window (50bp window size, 10bp step size) along each sequence using a custom Perl script.

4.2.9 Identifying transcribed CNE portions

To get an estimate of how many sequences were conserved in transcribed regions as opposed to non-transcribed regions, each CNE was split into all possible 20-mers and bowtie2 v2.0.5 (Langmead, Salzberg 2012) was used to map each sequence to the N. vitripennis evidential gene transcriptome dataset (Munoz-Torres et al. 2011, Werren et al. 2010) supplemented with RNA-seq data (data available on http://www.waspatlas.com), as well as the Ensembl version of the official gene set OGS v1.2 augmented with the same data. A 20-mer was counted as transcribed if it mapped to either transcriptome, and the percentage of mapped to unmapped reads was calculated to give an overlap percentage for each CNE.
4.2.10 Identifying conserved secondary structures

To test for the presence of conserved secondary structures, the SCI (structure conservation index) was calculated for each CNE alignment and used to compute the probability of a conserved secondary structure. The SCI is defined as the minimum free energy of the consensus folding structure divided by the mean minimum free energy (MFE) of each sequence in the alignment folded independently (Gruber et al. 2008). Sequences in alignments with high structural conservation will show similar energies whether allowed to fold independently or forced into the consensus structure. A high SCI (close to 1) therefore indicates a well-conserved structure, and an SCI of more than 1 may indicate the presence of compensatory mutations. SCI calculation was implemented in Perl using RNAfold (Hofacker et al. 1994) and RNAalifold (Bernhart et al. 2008), and code was used from (Washietl, Hofacker 2004) to implement a shuffling procedure as a control. Alignments are shuffled on a column-by-column basis, keeping the overall conservation pattern intact. By generating sets of shuffled alignments in this way, we can thus calculate the probability that the conservation of RNA secondary structure in the true alignment is significantly strong. For each CNE, 1000 control sequences were generated, the Z score distribution was calculated, and this was used to generate an empirical p-value. 29 CNEs showed conserved structure at p<0.05, falling to 11 at p<0.01.

4.2.11 Motif overrepresentation

To test for overrepresentation of transcription factor binding site motifs, a set of 1038 position weight matrices was used, including matrices from JASPAR (Sandelin et al. 2004) and PLACE (Higo et al. 1999) and the procedure described
by (Baxter et al. 2012) was followed. Redundancy was reduced by performing hierarchical clustering based on the Hellinger distances between matrices, yielding a set of 735 representative matrices at a threshold of 1.5. The matrix with the median entropy was selected to be the representative of each cluster. Motifs were tested for overrepresentation using a binomial test taking into account the strength of the matches. 100 control sets were produced using the same method as in the nucleosome occupancy prediction/GC content analysis section, with the exception that if Ns were found in the true CNEs then Ns were inserted into each control CNE in the same positions. A matrix was counted as over-represented if the binomial p-value in the true CNE set was lower than the p-value in all control sets, and underrepresented if it was higher than the p-value in all control sets. Of the 735 representative matrices, 88 were found to be under-represented compared to the controls, and 35 were over-represented. GC content of a matrix was measured proportionally with the weight of each position; a matrix with several highly weighted G or C nucleotides will thus have higher GC content than a matrix with G or C nucleotides that have low weight.

4.2.12 Not1 CNE characterisation

To test for presence or absence of the Not1 CNE, the consensus sequence from the original CNE alignment was scored against the 100 bp upstream sequences of Not1 homologs in all organisms available in Ensembl metazoa using BLAST (Altschul et al. 1990). The E-value distribution was plotted, and sequences with E-values lying outside of the main distribution (E < 0.001) were aligned, and the 30 bp hairpin loop sequences extracted. The stabilities of the hairpins were predicted using RNAfold (Hofacker et al. 1994) and sequence logo diagrams prepared using the seqLogo
(Bembom 2007) package in R. All genes and organisms used in this analysis can be found in Appendix 4.1.

### 4.2.13 Motif elicitation and RPLP1/RPLP2 analysis

MEME (Bailey et al. 2009) was used to elicit motifs from the *RPLP1* and *RPLP2* CNEs. After initial exploratory analysis, a search was performed for the 3 best motifs in the sequences of both CNEs, looking for motifs from 6 to 13 bp in the *RPLP1* CNE, and from 6 to 15 bp in the *RPLP2* CNE. For the *RPLP1* CNE, 500 bp sequences upstream of all *RPLP1* homologs analyzed were used, and 2 kb sequences for the *RPLP2* homologs analyzed. Due to total sequence length restrictions, the *RPLP2* motif elicitation analysis was first done with a seed elicitation followed by analysis of the remaining sequences. 3 motifs were extracted for the *RPLP1* CNE, and 2 for the *RPLP2* CNE, one of which was manually split into two upon further inspection. Sequences were inspected for presence or absence of the CNE motif components, and sequences containing the CNEs were aligned based on the motif positions. Distances of the motifs from translation start sites were calculated and plotted. Sequence logo diagrams were prepared using the SeqLogo (Bembom 2007) package in R.

The motif elicitation analysis was repeated using a phylogeny-aware method, Phylogibbs (Siddharthan, Siggia & van Nimwegen 2005). The motif-based alignments produced by MAST (Bailey et al. 2009) were reduced to 17 sequences due to memory constraints, and used as inputs for the algorithm. Phylogenetic trees were constructed using divergence estimates from timetree (Hedges, Dudley & Kumar 2006), and blanket proximity values were assigned to branches based on the approximate of 0.85 proximity given in the Phylogibbs documentation for the
mouse-rat divergence time of 22.6 Myr. The analysis for the *RPLP1* CNE revealed three significant motifs overlapping the motifs produced by MEME. Similar results were also obtained for the *RPLP2* CNE.
4.3 Results

4.3.1 Identification of deeply conserved non-coding elements

In order to identify conserved regulatory elements, a comparative analysis of 13 highly diverged insect genomes (Figure 4.3A) was performed on a locus-by-locus basis, scanning the 2 kb non-coding region upstream of the translation start site of each gene and comparing it with that of N. vitripennis. N. vitripennis was thus used as the central species compared to all other species in a series of pairwise comparisons. The “seaweed algorithm” (Krusche, Tiskin 2009) was used to perform alignments, performing over 3.8 million optimal alignments of short sub-sequences per pair of 2 kb sequences upstream of orthologous genes, optimal as described by Krusche and Tiskin (Krusche, Tiskin 2009), and 3.8 million due to all possible sequence combinations of 50 bp along two 2 kb sequences. Significantly aligned, overlapping sub-sequences were merged and regions in other species that mapped to the same sub-sequence in N. vitripennis were identified to yield one inclusive dataset (see 4.2.3).

As a control, pairs of randomly matched upstream non-coding sequences were aligned. The number of ‘conserved’ sequences detected in the control set at various alignment score thresholds can therefore be used to estimate the false discovery rate. The algorithm parameters were adjusted such that no conservation at all was detected in the control, and then these parameters were used to align the putatively orthologous sequences. Sequences were pre-filtered for repetitive regions, and post-filtered for similarities to known coding sequences. At this very strict level of false discovery, 322 CNEs were detected within or near 276 genes (Figure 4.3B). Each of these genes was given a combined conservation score (CCS)
in the interval 0-1, where anything above zero is considered statistically significant and one represents particularly strong conservation (see 4.2.3).

Figure 4.3. Phylogenetic relationships between species and genomic CNE distribution. (A) Phylogenetic tree showing the relationships and approximate divergence times between the insects used in the analysis. The number of conserved CNEs at different branching points is plotted on the figure. Phylogenetic relationships and divergence times from: inter-order (Hedges, Dudley & Kumar 2006); Hymenoptera (Schmieder, Colinet & Poirie 2012); Lepidoptera (You et al. 2013); Diptera (Wiegmann et al. 2011); Coleoptera (Hedges, Dudley & Kumar 2006). (B) *Nasonia vitripennis* genome diagram showing the locations of the sequences analyzed (red, outer circle, 3.66 % of genome) in contrast with the sequences identified as conserved (black, inner circle, 0.0064 % of genome). Lower CNE density on chromosome U reflects lower gene density on these unplaced scaffolds.

Since the most closely related species in the analysis (the three Hymenopteran species) diverged from *Nasonia* approximately 180 million years ago (Schmieder, Colinet & Poirie 2012), all of the 322 CNEs have been conserved for at least this long. The CNEs found in Hymenoptera tend to be found in more than just two species; ~58% of the hymenoptera-specific CNEs are conserved in three species or more (*N. vitripennis* and two others). Focusing on the 276 genes
with an associated CNE, the analysis was expanded to a wider range of animals. A handful of CNEs were found to be conserved at greater evolutionary distances; 20 CNEs in or near 18 genes were found to have been conserved for at least 350 million years (i.e. the common ancestor of Holometabola) (Wiegmann et al. 2009). Of these, one CNE dates back to the common ancestor of Mandibulata (myriapods, crustaceans, and hexapods), and 2 CNEs date further back to the radiations of basal animal phyla (Cnidaria, Placozoa, Ctenophora, and Porifera). These 20 anciently conserved CNEs exhibit a high degree of overlap, with only two being specific to *N. vitripennis* and one other species.

### 4.3.2 Relative position of CNEs is conserved along with sequence

In order to investigate the properties of the CNEs, a series of analyses were performed comparing the CNEs with a control set of sequences. To obtain these control sequences, the parameters of the algorithm were adjusted to allow for the capture of false discoveries, and the alignments were run on randomly matched (pseudo-orthologous) pairs of sequences. By setting an appropriate threshold, a similar (but not identical, due to the nature of the score distribution) number of sequences to the CNEs from the control set were extracted and post-filtered, representing the highest scoring non-orthologous sequence alignments. We term these sequences ‘*pseudo-CNEs*’, as they are sequences that have high alignment scores, albeit below our conservation threshold, but lack orthology. As high-scoring non-orthologous sequences, these pseudo-CNEs can be used as a comparison to elucidate important sequence properties about the true, orthologous CNEs, as opposed to comparisons with sequences with randomized properties.
The GC content of the CNEs contrasts starkly with the GC content of the pseudo CNEs; the CNEs have a mean GC content of 51%, compared to 27% in the pseudo CNEs (Figure 4.4A). This pattern of GC content is strongly associated with a peak of predicted nucleosome occupancy in the center of the CNE (Figure 4.5), a markedly different population of over/under-represented transcription factor binding sites (Figure 4.6), and an underrepresentation of ATG trinucleotides (Figure 4.7), although whether these are a cause or effect of the observed GC content disparity is unclear. The expected number of CpG dinucleotides based on the GC content (CpG O/E) in the CNEs does not significantly differ from the control (Figure 4.4B), suggesting that there is no suppression or special use of these methylation-related dinucleotides in the conserved regions. The length of the CNEs is strongly biased towards shorter (~90 bp) sequences, a trend which is not generally apparent in the pseudo-CNEs (Figure 4.4C), perhaps indicative of the mode of mechanism of these conserved sequences.

**Figure 4.4.** Comparative analysis of CNE sequence features. Analysis of CNE (blue) properties in comparison with pseudo-CNE controls (red). Distributions were compared using the Wilcoxon rank-sum test. (A) GC content. (B) CpG O/E. (C) CNE length distribution. (D) CNE position.
Figure 4.5. Association of CNEs with predicted nucleosome occupancy.

Figure 4.6. Relationship of transcription factor binding site motifs with GC content.
When we consider the distance of the sequences from the translation start site of their associated gene, we see enrichment for adjacency among both the CNEs and the pseudo-CNEs, decreasing with distance (Figure 4.4D). Although the relative positions of the CNEs and the pseudo-CNEs are similar, the conservation of these distances across species is not. Comparing the translation start site distance in *N. vitripennis* with the translation start site distance in the comparator species reveals that the distance of each pseudo-CNE in *N. vitripennis* is completely uncorrelated.
with its distance in the comparator organism (Figure 4.8A), whereas in the set of CNEs there is a significant correlation between distances (Figure 4.8B). This result shows that the position of CNEs is important as well as the conserved sequence itself.

**Figure 4.8.** Conservation of relative CNE positions. Scatter plot showing the conservation of CNE positions, comparing the CNE position in *Nasonia* (x-axis) with its position in each comparator organism (y-axis). Conservation of position relative to translation start site is insignificant in the control (A, \( p = 0.47 \), n=359) but significant among CNEs (B, \( p < 2.2\text{e}-16 \), n=322).

### 4.3.3 CNEs are tightly associated with developmental and regulatory genes

The 322 CNEs identified here are associated with a specific class of genes. Overrepresentation of gene ontology (GO) terms against the genomic background using the annotation information available (see Methods 2.5) was performed using each *N. vitripennis* gene associated with a conserved region. 319 terms were significantly overrepresented with a q-value below 0.01. The most overrepresented term in the set was ‘regulation of gene expression’ (q < 2.7e-31) which was
associated with over a third (36.7%) of the genes tested. In addition, many significant terms such as ‘nucleic acid binding transcription factor activity’ (q < 3.7e-28, 21.9% of genes tested) and ‘developmental process’ (q < 8.8e-30, 51.5% of genes tested) were returned, suggesting that genes associated with upstream conserved regions often themselves have regulatory and/or developmental roles. A set of 28 terms were overrepresented for the set of pseudo-CNEs, albeit with lower significance compared to the CNE set. This suggests that the long highly AT-rich sequences that are picked up in this control have a weak, but detectable association with gene expression and specific processes – an observation not further explored here.

The 20 most deeply conserved sequences (>= 350 Myr) also appear to be associated with a specific class of genes. 14 of these CNEs were found to lie completely within transcribed regions (see 4.2.9), and all 20 were found to overlap transcribed regions by at least a third of the length of the CNE. This enrichment is significant (p < 6.5e-04, hypergeometric test) when compared to the full set of 322, of which only ~70% overlap transcribed regions by this amount. Remarkably for such a small set of genes, a GO term overrepresentation test turned up 39 significant terms. The 17 genes associated with these 20 CNEs are enriched for genes active in processes such as post-transcriptional regulation of gene expression (q < 6.1e-4), regulation of translation (q < 4.8e-3), and translational elongation (q < 1.2e-2). This list of overrepresented GO terms is, unlike that obtained from the full set of 322 CNEs, devoid of terms relating to transcriptional regulation, matching the shift towards putative translational regulatory CNEs.
4.3.4 5’ UTR CNEs contain conserved secondary structures

Among the CNEs identified were previously-studied regulatory elements, as well as many unidentified novel putative regulatory elements. As the majority of CNEs overlap 5’ UTRs, the likelihood of there being a conserved RNA secondary structure in each CNE was calculated. This analysis revealed several conserved secondary structures, including an example of the well-characterized iron response element (IRE) in the 5’ UTR of the Ferritin gene (Figure 4.9), a conserved hairpin loop bound by iron response proteins (IRPs) to help maintain iron homeostasis.

Figure 4.9. Ferritin 5’ UTR iron response element (IRE) alignment diagram. A) Sequence alignment, along with sequence conservation bar-plot. Mismatching (from the consensus) nucleotides are coloured. B) Secondary structure plot. Ribonucleotides are coloured by the level of sequence conservation found in the alignment.
Novel conserved RNA structures were also identified, including a conserved, strong (-52.60 kcal/mol) hairpin loop found in the 5' UTR of the *Paramyosin* gene (Figure 4.10) identified in all four Hymenoptera species, and a hairpin loop with perfect stem complementarity but variable apical sequence conserved in the 5' UTR of the *Not1* gene (Figure 4.11).

**Figure 4.10.** Conserved CNE is a novel hairpin in the 5' UTR of the *Paramyosin* gene. A) Sequence alignment, along with sequence conservation bar-plot. Mismatching (from the consensus) nucleotides are coloured. B) Secondary structure plot. Ribonucleotides are coloured by the level of sequence conservation found in the alignment.
**Figure 4.11.** A highly conserved hairpin loop in the 5' UTR of Not1. **a** Upstream CNE-containing Not1 sequences. Three footprints of conservation are clearly visible; the first two (from left) constitute the stem sequence of the hairpin and are shown as motif 1. The third footprint (shown as motif 2) is the conserved sequence adjacent to the translation start site and contains an ATG upstream of the translation start site. Hairpin loop stabilities are shown in red and outliers (disrupted loops) are marked with blue asterisks. *Acyrthosiphon pisum* represents a putative Not1 paralog. (A) *N. vitripennis* Not1 CNE predicted RNA folded structure colored by sequence conservation, showing highly variable apical sequence and conserved stem.

These three hairpins differ in their fundamental characteristics. IREs are characterized by a highly conserved apical sequence (CAGUGY; clearly demonstrated in the three hymenopteran species) with a more variable stem.
sequence (Piccinelli, Samuelsson 2007). In contrast, the 4-nucleotide apical sequence (IUPAC code of HVHN) of Not1 appears to be highly variable, whereas the stem sequence is almost perfectly conserved. More sequences are necessary to be able to reliably characterize the Paramyosin hairpin, although there does appear to be at least one variable nucleotide in the hairpin apex. The positions of the hairpins also appear to be of functional importance; all three hairpins are conserved in their position relative to the translation start site, particularly the Not1 hairpin (Figure 4.11A).

The Not1 hairpin loop has a stem sequence of 12bp, and the CNE containing it is found directly adjacent to the translation start site. The CNE contains two conserved stem sequences with near-perfect complementarity, a weakly conserved apical sequence, and a highly conserved, upstream ATG-containing motif directly adjacent to the translation start site. In N. vitripennis, this CNE is present in the 5′ UTR of all four known transcripts. As the position of this CNE is so strongly conserved, scanned the first 100bp of every orthologous transcript in all Ensembl Metazoa species was scanned for presence of either the conserved hairpin or for the conserved sequence adjacent to the translation start site. The results of this search (see 4.2.12) indicated that in all cases where the hairpin loop is present the conserved sequence adjacent to the translation start site is present too, but not vice-versa (i.e. the sequence adjacent to the translation start site may exist on its own). The presence of the sequence in the Antarctic krill Euphasia superba (Hunt and Rosato, unpublished data) and in a centipede (Strigamia maritima) shows that this CNE was an early arthropod adaptation.
4.3.5 An uncharacterized gene cluster contains several CNEs

Conserved putative regulatory sequences were identified in six individual genes of the insect-specific *Osiris* gene cluster (Figure 4.12). The analysis indicates that these regions are Hymenoptera-specific, and are generally conserved in position relative to their associated gene. Since the conserved regions are associated with a specific class of genes with core functions, the fact that conserved promoter regions were identified near to six genes in the same cluster is perhaps indicative of an important developmental or regulatory role for this as-yet uncharacterized gene cluster.

---

**Figure 4.12.** Uncharacterized *Osiris* gene cluster contains several CNEs. Conserved upstream sequence location relative to the coding sequences and combined conservation scores of six genes in the *Osiris* cluster.
4.3.6 Ribosomal stalk gene CNEs date back to early animals

Two conserved sequences were identified in the 5′ UTRs of the two ribosomal stalk heterodimer genes, *RPLP1* and *RPLP2*. Given that parts of these sequences were found to be perfectly conserved over several nucleotides, I looked for the presence of the same sequences in more distant phyla. A motif elicitation analysis (see 4.2.13) revealed three separate sequence motifs in the *RPLP1* CNE, and three in the *RPLP2* CNE. These motifs are present in many different phyla (Figure 4.13), including both deuterostomes and protostomes, making these the third and fourth known examples of bilaterian conserved regulatory elements (Bicores) (Clarke et al. 2012).

![Figure 4.13. Distribution of most ancient CNEs across phyla. Cladogram showing presence (green), absence (red), or mixed presence-absence (orange) of the deeply conserved RPLP1 and RPLP2 CNEs. Numbers in brackets show number of species analyzed per group for the RPLP1/RPLP2 CNEs respectively. Groups are outlined.](image-url)
by color: blue (Protostomia), orange (Deuterostomia), green (Cnidaria, Ctenophora, and Placozoa), pink (Porifera)

These two conserved sequences were both early innovations in Animalia. The RPLP1 CNE was found in the genomes of the placozoan Trichoplax adhaerens and the cnidarian Nematostella vectensis (starlet sea anemone). N. vectensis also contains the RPLP2 CNE, as do the ctenophore Mnemiopsis leidyi (warty comb jelly) and the poriferan Amphimedon queenslandica (a demosponge). Both CNEs were present in the majority of species that I analyzed (RPLP1: 33/38 species analyzed, RPLP2: 23/38). Previously, the most ancient CNEs identified were found conserved between Deuterostomia and Cnidaria (Royo et al. 2011), thus dating back over 670 million years (Peterson et al. 2008). The CNE on RPLP2 that reported here appears to have originated even earlier, being found in the Porifera. This CNE is thus likely over 700 million years old (Hedges, Dudley & Kumar 2006). The CNE on RPLP1 may also be older than 670 million years, depending on how the deep splits in the phylogeny of animals are eventually resolved (Dohrmann, Worheide 2013).

The conserved regions paint an interesting evolutionary story. Firstly, in the RPLP1 CNE (Figure 4.14), there are three distinct conserved sequences. The first conserved sequence appears to have two distinct forms; one found in protostomes (motif 1a) and another in deuterostomes (motif 1b), which appears to be the ancestral form as it is found in Cnidaria and Placozoa. The second and third motifs are found conserved across both deuterostomes and protostomes, and are variably spaced; for example all mammalian species analyzed share a similar insertion
between these two motifs. The relative position of the CNE is found conserved across all phyla (Figure 4.14B), remaining within 150bp of the translation start site.

**Figure 4.14.** Evolution of RPLP1 CNE over at least 670 million years. (A) Alignment of RPLP1 CNE in all organisms where detected. Sequence logo diagrams of each conserved motif are shown below the alignment. Motif 1a is protostome-specific whereas motif 1b appears to be the ancestral and deuterostome form. Motifs 2 and 3 are variably spaced and are present in all phyla examined. Species name colour scheme matches that of Fig. 4.13. (B) Diagram showing the position and spacing of each motif in each organism in relation to the translation start site. Genus/species abbreviations are defined in Appendix 4.2.

The *RPLP2* CNE (Figure 4.15) also appears to be described best as three distinct motifs. Motif 1 is exceptionally well conserved, with no variation at all
across 10 bp. Motif 2 comprises a conserved region, generally followed by a short stretch of adenine nucleotides. Motif 3 is short and does not appear to be present in either *D. melanogaster* or *Mnemiopsis leidyi*. These observations make clear that these CNEs are functionally complex, being comprised of several discrete elements punctuated by less evolutionarily constrained sequence. This is in contrast to other kinds of conservation such as ultraconserved regions, where long stretches of nucleotides (>200bp) are found perfectly conserved between human, rat, and mouse, (Bejerano et al. 2004) which can be in some cases deleted without a clear critical loss of function (Ahituv et al. 2007). As a whole, the complexity, shared associated gene function, and age of these CNEs marks them as interesting targets for future study.

**Figure 4.15.** Evolution of the *RPLP2* CNE over at least 700 million years. a Alignment of *RPLP2* CNE in all organisms where detected. The three distinct sequence motifs are shown aligned below the main alignment. Species name colour scheme matches that of Figure 4.13. (B) Diagram showing the position and spacing of each motif in each organism in relation to the translation start site. Genus/species abbreviations are defined in Appendix 4.2.
4.4 Discussion

In this chapter, a high stringency statistical approach was used to identify and characterize 322 ancient non-coding elements which have remained conserved over large evolutionary distances. The bulk of the conserved sequences identified are specific to Hymenoptera (299/322), but nevertheless have been conserved in position for at least 180 million years of insect evolution (which occurs at a faster pace than vertebrate evolution (Wyder et al. 2007)). A small proportion of the CNEs (20) identified were at least 350 million years old, with three stretching back further still. Two CNEs are found conserved in a range of the most basal animal clades across a wide variety of both vertebrates and invertebrates and are likely over 670 (Peterson et al. 2008) and 700 million years old (Hedges, Dudley & Kumar 2006), the oldest CNEs described to date.

These two ancient CNEs are located in the 5′ UTRs of two genes that are known to interact with one another, RPLP1 and RPLP2. The two protein products of these ubiquitously expressed genes, P1 and P2, form a heterodimer; two copies of which bind to the 60s acidic ribosomal protein P0 (coded by the gene RPLP0) to form the ribosomal stalk. The ribosomal stalk is involved in translational fine tuning and is crucial for the correct folding of many proteins (Perucho et al. 2014). The depth and breadth of conservation of these sequences is indicative of a fundamental regulatory role. Indeed, the 5′ UTR of RPLP2 has already been shown to have a regulatory role in Drosophila (Patel, Jacobs-Lorena 1992), being sufficient to mark RPLP2 as a gene which is not to be translated in the early embryo, but not previously known to be conserved among animals. The fact that this CNE has been previously studied and identified as a regulatory element helps to validate the idea that other CNEs identified here are also functional regulatory elements. In
Drosophila, this CNE essentially spans the entirety of the RPLP2 5' UTR, whereas in other organisms it is only a constituent part. In this chapter, I have characterized the motifs likely to be important for the function of these regulatory elements and examined their evolution over time.

Most of the conserved regions identified in this analysis were found to be located within gene bodies as opposed to intergenic space, providing potential insights into a poorly understood class of regulatory elements. The analysis revealed conserved secondary structures in the 5' UTRs of several genes, examples including hairpin loops upstream of the Ferritin gene (an IRE), the Paramyosin gene, and the vital (Maillet et al. 2000) regulatory (Collart, Panasenko 2012) gene Not1. This class of regulatory elements is far less well understood than regulatory elements found upstream of genes, so the identification of such elements will be important for future investigations into 5' UTR regulation.

Secondary RNA structures such as hairpins can have important regulatory consequences, having the capacity to both enhance and inhibit translation. The effect of a hairpin on translation differs depending on the stability of the hairpin, its distance from the mRNA cap, and GC content (Babendure et al. 2006). These three hairpins have different fundamental characteristics, and thus likely perform their putative regulatory functions in different ways. The strong, in terms of its likely effect on translation at -52.60 kcal/mol (Babendure et al. 2006), and GC-rich (66%) Paramyosin hairpin is likely able to present a significant barrier to translation at any distance from the cap, whereas the potential function of the weaker (~15.00 kcal/mol) Not1 hairpin is less obvious. The Not1 hairpin exhibits a complex conservation pattern, with near-perfect stem complementarity, tight positional
conservation, and an associated conserved upstream AUG (uAUG), itself a tightly-suppressed (Iacono, Mignone & Pesole 2005) class of regulatory element.

The CNEs identified here, confirming similar observations in other organisms (Woolfe et al. 2005), are associated with regulatory and developmental genes. This observation is consistent with the idea of regulatory gene cascades, that genes involved in regulation are themselves tightly regulated, allowing master regulators to exert overall control of gene regulation ‘programs’ to reprogram cells (Lee, Young 2013). This result is augmented by the more specific observation that the deeply conserved set of 20 CNEs (≥ 350 Myr), which are likely to be post-transcriptional regulatory elements, are associated with genes themselves involved in post-transcriptional regulation.

Fundamental differences were also identified in basic sequence properties of the CNEs when compared with control sequences. GC content in CNEs is generally elevated (Fig S8); sharply peaking within the CNE itself but also raised in the flanking regions. This result is informative as GC content is known to be important for regulation; it is associated with regulatory mechanisms such as nucleosome occupancy (Tillo, Hughes 2009), aspects of secondary structure stability and effects on translation, and for example in chicken, variation in GC content in the 5′ UTR of genes can explain 10% of the variation in expression level (Rao et al. 2013).

The conserved elements identified were not GC rich nor GC poor, suggesting that these sequences are complex. This use of complex sequences as regulatory elements is intuitive, especially given that some of the sequences I identified have been conserved for hundreds of millions of years. One would imagine that sequences with a simple regulatory function, for example a binary expression switch, would be relatively plastic due to the many ways that a binary
signal could potentially be encoded. Changing a single nucleotide would therefore not be likely to interfere with function. These sequences would thus change easily over time and rapidly become undetectable by the methods used here. Regulatory sequences with more multi-faceted functions would require sequences to encode a lot of information. Changing a single nucleotide in a sequence such as this would thus likely have consequences for many regulatory pathways. This lack of plasticity would make these sequences detectable by the methods employed here. The fact that the CNEs identified here generally appear to be made up of several conserved ‘footprints’ also lends support to this hypothesis.

One important feature of many of the CNEs discovered here is that their positions relative to the translation start site are conserved, i.e. that the position of the CNE is conserved as well as its sequence (Figure 4.9). When more reliable transcription start site data are available, it will be possible to examine whether putative transcriptional regulatory mechanisms that I identified are conserved relative to the transcription start site, or whether some of the translational regulatory mechanisms identified are more closely associated with the mRNA cap position than the translation start site. This positional information could be useful in detecting CNEs over large evolutionary distances, under the assumption that evolution sometimes proceeds by modifying the sequences of existing cis-regulatory CNEs without significantly changing their relative position (Cande, Goltsev & Levine 2009). The analysis presented here provides a resource for other researchers who wish to investigate regulatory elements in Nasonia, and are available at WaspAtlas (http://waspatlas.com/cns_temp). The next chapter describes this gene-based Nasonia resource which includes the data presented here alongside a wealth of published and unpublished data for Nasonia genes.
5 Genomic database development

5.1 Introduction

Many model organisms have dedicated databases which greatly facilitate bioinformatic analyses (see Introduction 1.4). An important reason for this is that a large number of assays aim to reduce a whole genome down to a few candidate genes. For example, genes differentially expressed/methylated between two conditions, genes which are likely to be involved in a particular phenotype (e.g. QTL mapping), or proteins bound to other proteins in a complex. When investigating these candidate genes in an organism such as Drosophila, the integration of various types of data in databases such as FlyBase (Attrill et al. 2016) and FlyAtlas (Robinson et al. 2013) allows a researcher to quickly and easily browse all the information available about their gene of interest, potentially informing future experiments. When analysing larger sets of genes, tools such as GO term overrepresentation are useful to reveal groups of genes involved in a common function likely to be pertinent to the analysis at hand. For Drosophila, this tool is available through DAVID (Huang da et al. 2007), or found pre-installed in other tools such as BiNGO (Maere, Heymans & Kuiper 2005).

5.1.2 Nasonia genome annotations

Fewer resources exist for Nasonia. Since the original publication of the Nasonia genome (Werren et al. 2010), the genome assembly has been improved and detailed annotation projects are ongoing (Table 5.1).

---

1 The work presented in this chapter has been published in Davies & Tauber, 2015, Database
Table 5.1. *Nasonia vitripennis* genome annotations, along with associated genome build, and status of development.

<table>
<thead>
<tr>
<th>Annotation identifier</th>
<th>Genome build</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCBI annotation release 101</td>
<td>Nvit 2.1</td>
<td>Current</td>
</tr>
<tr>
<td>Official gene set v1.2 (OGS v1.2)</td>
<td>Nvit 2.1</td>
<td>Current</td>
</tr>
<tr>
<td>EvidentialGene dataset</td>
<td>Nvit 1.0</td>
<td>Current</td>
</tr>
<tr>
<td>NCBI annotation (archived release)</td>
<td>Nvit 2.0</td>
<td>Frozen</td>
</tr>
</tbody>
</table>

The level of annotation between assemblies and annotations varies. For example the EvidentialGene dataset (Munoz-Torres et al. 2011) (mapped to the first genome build) contains UTR annotation for 97% of gene models and has a significant amount of associated GO (gene ontology) (Ashburner et al. 2000) annotation, whereas OGS v1.2 (mapped to the latest genome build, adopted by Ensembl) only has UTR annotation for 37% of gene models and has relatively little GO annotation. Each gene annotation project has its own set of gene identifiers, making it difficult to compare studies that use different reference annotations, as there is no existing method for converting identifiers in batch.

5.1.3 Published datasets

Reflecting its position as an important model organism, several RNA-seq datasets have been produced for *Nasonia vitripennis*. Together, these datasets provide gene expression information for both male and female wasps, various tissues, important developmental stages, and different experimental conditions. Unfortunately, this data is currently scattered throughout different publications and has been mapped...
using various reference annotations. Given the number of experiments being carried out with *Nasonia*, it would be useful to have this data in one place to be able to find out where and when a particular gene is most highly expressed or to compare the expression patterns of groups of genes.

### 5.1.4 Databases and tools

Few tools currently exist for working with *Nasonia*, and user-aided interface access to the genome annotations varies significantly between projects, so trade-offs exist when selecting the appropriate reference annotation for a given project. *NasoniaBase* (Munoz-Torres et al. 2011) hosts the OGS v1.2 dataset and the EvidentialGene dataset. BLAST functionality is provided, as is microRNA-gene interaction scoring, a genome browser for the Nvit_2.0 version of the genome assembly, and a set of plain text downloads for the two datasets. The two annotation datasets NCBI v2.0 and NCBI v2.1 are hosted on NCBI, and therefore benefit from the NCBI user interface, providing BLAST functionality, a gene view for each gene with basic sequence properties, and genome browser access.

GO term overrepresentation must be performed manually using scripts to import annotation data into a tool such as BiNGO (Maere, Heymans & Kuiper 2005), as performed in chapter 4 of this thesis. RNAi is widely used in *Nasonia* (Lynch 2015), but there is currently no tool for performing off-target prediction. Overall, accessing gene annotation in *Nasonia* generally requires significant manual effort, or proficiency in bioinformatics techniques.
5.1.5 WaspAtlas

The goal of this part of the project was to produce a database combining data from all *Nasonia vitripennis* annotation projects, original annotation works and analyses, all currently available RNA-seq transcriptome libraries/microarray data, and DNA methylation data. The aim was to have each *Nasonia vitripennis* gene complete with, where possible, GO annotations, PFAM domain predictions (Finn et al. 2014), orthologs in other important model species, expression data comparing sexes, tissues, developmental stages, and experimental conditions, and data from as yet unpublished analyses.

The goals of WaspAtlas can be split into three main areas. Firstly, WaspAtlas is a database combining all available *Nasonia vitripennis* gene data. Secondly, WaspAtlas provides an easily navigable web interface for accessing this data, along with detailed illustrations and expansive search functionality. Thirdly and finally, WaspAtlas provides an analysis platform, helping users to perform common bioinformatic tasks for any of the available gene annotations. To demonstrate the utility of the database, a housekeeping gene analysis based on the integrated expression data was performed.

5.1.6 DNA methylation and RRBS

In order to add to the data available in WaspAtlas, the analysis and integration of an RRBS (reduced-representation bisulphite sequencing) (Gu et al. 2011) dataset, published by Pegoraro *et al.* (Pegoraro et al. 2016), is also presented in this section. DNA methylation is an important epigenetic mark which is essential for normal embryonic development (Li et al, 1992), and plays a crucial role in processes such as X-chromosome inactivation (Panning & Jaenisch, 1996) and gene imprinting.
Abnormal DNA methylation patterns have been also been associated with several diseases including cancer (Kulis & Esteller, 2010). Cytosines in CpG sites (i.e. cytosine residues followed by a guanine) are the main targets of DNA methylation (Zemach et al, 2010; Law & Jacobsen, 2010). In mammals, methylation of promoter regions has the effect of silencing transcription of the associated gene (Weber et al, 2007). In insects, gene body methylation is the primary mode of methylation (Gladstad et al, 2011), but the function of this is less well understood.

RRBS (reduced-representation bisulphite sequencing) (Gu et al. 2011) is a method that uses bisulphite sequencing to identify methylated nucleotides in combination with an enzymatic method to enrich the sample for 5′ - CCGG - 3′ sequences, with the aim of maximising the number of reads containing CpG dinucleotides and thus achieving greater coverage. The dataset used here was generated by the Tauber lab with the aim of identifying genes differentially methylated between long and short photoperiods in Nasonia (Pegoraro et al. 2016). This chapter presents the analysis that I carried out of this dataset, and its integration into WaspAtlas as a resource for Nasonia researchers providing a snapshot of DNA methylation over the Nasonia genome.
5.2 Methods

5.2.1 Implementation of WaspAtlas

WaspAtlas was implemented in Perl using the Catalyst development framework, and runs on an Apache server with a MySQL database. Catalyst was chosen due to its core language being Perl, easy integration with CPAN, and flexible MVC framework. Template toolkit was used for front-end development. The website itself is hosted at the University of Leicester, and the high-throughput analyses performed before integration into the website used the SPECTRE High Performance Computing Facility at the University of Leicester.

5.2.2 Inter-annotation mapping of gene identifiers

In order to provide a complete annotation for Nasonia vitripennis and to create a complete mapping between all gene annotations, gene models from four different annotation projects were inter-mapped using a combination of all extant gene equivalency mappings. Where an equivalency between IDs was specified in a particular dataset, these IDs were collapsed into a single locus. This was performed on an iterative basis for all annotations until the genes were collapsed into as few loci as possible based on the information available. Gene models in the EvidentialGene dataset were mapped to OGS v1.2 equivalents using the OGS1 equivalence values in the GFF. EvidentialGene gene models were further mapped to the archived NCBI annotation release (Nvit 2.0) using the refseq2 (quality1) value in the GFF. The NASONIABASE: detail in each of the NCBI datasets (annotation 101 and the archived version) was used to map these annotation releases to their equivalent models in OGS v1.2. The NCBI datasets were mapped to one another using equivalences in the gene name attribute. A reciprocal best blast hit (RBH)
approach (Tatusov, Koonin & Lipman 1997) between datasets was attempted in order to obtain more complete mappings, but the improvement obtained was negligible. The final mapping thus used only extant mappings in order to avoid potential errors introduced by original mapping analyses. The results of the inter-annotation mapping were made available on the downloads page for use by other research groups.

### 5.2.3 GO term annotation

The comprehensive loci resulting from Methods 2.27 was used to collate GO annotation data from several sources. GO term associations were parsed from NasoniaBase (blast2go format) (Munoz-Torres et al. 2011) and ensemble (plain text) (Flicek et al. 2014), combining annotations from two datasets into one. These GO terms were then expanded by traversing the GO annotation tree using the relationships ‘synonym’, ‘is_a’, and ‘part_of’ in order to include both synonymous, and more general terms. For example, a gene involved in a ‘protein metabolic process’ would also be involved in ‘macromolecule metabolic process’ and ‘primary metabolic process’. This resulted in a more complete GO term annotation.

### 5.2.4 PFAM domain prediction

To assign protein domains to amino acid sequences, PFAM protein domain prediction was performed using HMMscan version 3.1b1 (Finn, Clements & Eddy 2011) against the Pfam-A database of hidden markov models (release 27.0) using the gathering thresholds for each model. In the case of overlapping PFAM hits, the most significant non-overlapping hits were kept, as per PFAM standard protocol (Figure 5.1).
5.2.5 Calculation of orthologs in other species

In order to facilitate easy comparisons between *N. vitripennis* genes and those of other more well established model organisms orthologs were calculated using the *Nasonia vitripennis* Ensembl OGS v1.2 protein dataset in a reciprocal best blast hit approach against the corresponding protein sequences for 11 species downloaded from Ensembl (release 21 for *Aedes aegypti, Apis mellifera, Atta cephalotes, Bombyx mori, Drosophila melanogaster, Solenopsis Invicta*, and *Tribolium castaneum*, release 25 for *Amphimedon queenslandica, Caenorhabditis elegans*, and *Amphimedon queenslandica*).
Mus musculus, and Homo sapiens), combined with the Ensembl (Flicek et al. 2014) calculated orthologs where available. GO annotations found through Drosophila melanogaster orthology were subsequently included in the Nasonia vitripennis WaspAtlas annotation. More complex methods such as orthoMCL (Methods 2.7) or orthoDB were not used here, as the aim was to provide a small set of high-confidence orthologs. RBH was judged appropriate for these purposes, as the majority of orthologs are picked up by this method (2.7), and the inclusion of links to more comprehensive databases will be possible in the future.

5.2.6 Gene sequence construction

Gene co-ordinates, including transcript, exon, and CDS locations were extracted for all protein-coding genes. Full gene sequences were extracted from the genome, and these sequences subsetted based on the described co-ordinates to produce transcript sequences for each isoform. Coding sequences were extracted for each transcript and translated to form the protein sequences.

5.2.7 Analysis of transcript expression datasets

RNA-seq libraries comprising 43 samples across 7 experiments (Appendix 5.1) (Wang et al. 2013a, Akbari et al. 2013, Hoedjes et al. 2015, Sackton, Werren & Clark 2013, Wang, Werren & Clark 2015) were mapped to the NCBI Nvit 2.1 annotation using the tophat 2 v2.1.0 (Trapnell, Pachter & Salzberg 2009). Although missing some sequences when compared to other annotations such as OGS2, the NCBI Nvit 2.1 annotation was used here due to lack of fragmentary/spurious sequences and other potential confounding factors. Novel junction discovery was
disabled during this analysis to avoid the addition of potentially spurious gene sequences.

Cufflinks and cuffnorm were used with geometric (DEseq) normalisation within datasets and tissues to calculate the FPKM values for each transcript. Data was not normalised across tissues or sexes, as the underlying principle of DEseq normalisation is that the majority of genes are not differentially expressed (Anders, Huber 2010). The hugely variant transcription programs seen between sexes in Nasonia (Wang, Werren & Clark 2015) and between tissues in other organisms (Sudmant, Alexis & Burge 2015), would likely render this normalisation method meaningless. Means and standard deviations were thus calculated based on these inter-tissue, inter-sex normalised values. Enrichment values were calculated for each transcript, separately for males and females, using whole 24-hour old adult wasp whole-body samples as a baseline value. (Wang, Werren & Clark 2015). Tiling microarray data showing gene expression during several crucial stages of female development (Wang et al. 2013a) and tissues was also downloaded and integrated.

5.2.8 RRBS methylation data

The RRBS dataset consisted of two samples, both generated from 10 day old adult female wasps - one generated from wasps exposed to a short day photoperiod, and one generated from wasps exposed to a long day photoperiod. DNA was MspI digested and bisulphite-treated and the resulting fragments were sequenced on the Illumina HiSeq 2000 platform. Both samples were sequenced on one lane and multiplexed manually with a 5bp sequence at the start of the read. In total, 117,646,358 single-end 50bp reads were obtained. To address the lack of biological
replicates in this study, the statistical approaches implemented in methylKit were used (Akalin et al. 2012), and qPCR was used to validate results resulting from this analysis (Pegoraro et al. 2016).

5.2.9 RRBS data processing

Sequence quality information was analysed with FastQC version 0.10.1 (Andrews 2010). For quality control, trim galore was used to adapter and quality trim the reads. Trimming was performed with the –RRBS option enabled, as the library preparation for RRBS uses unmethylated cytosines to fill in the overhang left by the restriction enzyme at the 3’ end of each read (Gu et al. 2011). The –RRBS option in trim galore trims an extra 2bp off the end of each read to ensure that this ‘filled-in’ position is not used for methylation calling. FastX toolkit 0.0.13 (Hannon Lab 2015) was then used to remove reads with low qualities (those below a mean of 20 phred score), and reads likely to constitute noise - reads which did not begin with YGG were judged likely to be noise as due to enzymatic digestion all reads should in theory begin with CGG, and due to bisulphite treatment some of these reads should become TGG. After quality control and de-multiplexing (accomplished with a custom Perl script), there remained 30,818,609 reads from the short day sample, and 27,255,357 reads from the long day sample.

After removal of low-quality reads, 27,255,357 reads (1.06 Gbp) were obtained from the long-day sample (LD) and 30,818,609 reads (1.25 Gbp) from the short-day sample (SD). Reads were then mapped to NCBI Nvit_2.0 with Bismark (Krueger, Andrews 2011), a piece of bisulphite-sequencing specific read alignment software. High mapping efficiencies (>83%) were achieved for both samples.
Samtools (Li et al. 2009) was used for intermediate file processing (e.g. sorting mapped reads).

5.2.10 Identifying methylated cytosines

Initially, ‘non-methylated’ cytosines were identified by mapping C to T conversions relative to the genome. Methylated cytosines are immune to the bisulphite conversion, therefore cytosines present in the reads were judged to be ‘methylated’. An internal control was used to estimate the methylation calling error rate caused by incomplete conversion and/or sequencing error. As the ‘filled-in’ cytosine position present in the 3’ restriction site always contains an unmethylated cytosine (Figure 5.2), any Cs in this position either represent unsuccessful conversion or sequencing error.
Figure 5.2. The MspI digestion and filling-in protocol. (A) MspI cuts at CCGG sites (cleavage site in blue). (B) Two base pair overhangs are left on each side of the double stranded fragment (red). (C) These positions are filled in with unmethylated nucleotides (cytosine and guanine, blue). (D) Creation of adenine overhangs (green) is necessary for illumina adapter ligation.

For the purpose of estimating error, previously discarded reads that contained the illumina barcode at their 3’ end were used, taking only those reads with an 11bp overlap or more with the illumina adaptor. The estimated error rate $E$ (i.e. proportion of unmethylated Cs likely to be called as methylated) is given by the equation $E = \frac{nC^f}{nC^f + nT^f}$, where $nC^f$ is the number of cytosines at the ‘filled-in’ position, and $nT^f$ is the number of thymines at the ‘filled-in’ position. The estimated error rates were low for each sample: 0.00622 for the short day sample and 0.0068 for the long day sample.

These error rates were used to calculate the probability that each covered cytosine was truly methylated. Truly methylated cytosines are designated mC. I used the binomial distribution to calculate the probability of seeing K ‘methylated’ reads for a cytosine in N trials (read depth), using $E$ (estimated error rate) as the probability of a cytosine by chance. This was performed for every site covered at a read depth of $\geq 10$ using methylKit (Akalin et al. 2012) in R. The p-values generated from this step were FDR corrected using the Benjamini-Hochberg method, and methylated cytosines called at a threshold of 0.01.

5.2.11 Integration into WaspAtlas

All analysed CpG sites from this dataset were parsed and added into the database as an experiment dataset. CpG sites were split by gene and the methylation call was
recorded for each site giving, for each gene, a set of methylated and non-methylated CpG sites. Differential methylation between the two conditions (Pegoraro et al. 2016) was also included in the WaspAtlas dataset for each gene.

5.2.12 Search functionality in WaspAtlas
Searching in WaspAtlas is divided into two sections: quicksearch, accessed through a text input box in the top right corner of every page, and the advanced search, which has its own page on the website. The quicksearch function was designed to provide relevant results quickly, and will, where possible, guess the kind of information provided by the user. The quicksearch tool was designed to first scan the search string against a list of recognised ID formats, including gene IDs from different annotation projects (e.g. NV*, Nasvi2EG*), PFAM domain IDs, and GO terms. If the string does not match any of these recognised formats, quicksearch was designed to perform a full text search against various fields in the database (gene ID, gene name, gene description, etc). The advanced search was designed to allow users to more strictly specify the targets of their searches (e.g. specify that their search is for a gene name, not for an ID), as well as to provide functionality not fulfilled through quicksearch (e.g. matching several GO terms at once, or matching an orthologous gene in another species).

5.2.13 Illustrations in WaspAtlas
Visual aids are provided where possible to facilitate understanding of tool results, and to visualise transcripts and protein domain locations during primer design, etc. Illustrations are provided for each transcript, showing exon and intron boundaries,
as well as coding sequence boundaries. Illustrations are also provided for each protein, showing the locations of protein domains along the protein with a colour-blind friendly palette. Illustrations are generated on-the-fly using template toolkit combined with HTML canvas, so that new data added to WaspAtlas will be visualisable.

5.2.14 RNAi off-target prediction in WaspAtlas

RNAi off-target prediction splits the user-provided dsRNA sequence into all possible 19-mers, and matches these against the transcriptome for the chosen assembly and its (pre-generated) complement. The tool returns all potential targets of this dsRNA sequence (i.e. all transcripts with at least one 19-mer match), allowing users to identify potential unintended targets of their dsRNA sequence.

5.2.15 GO term overrepresentation in WaspAtlas

GO term overrepresentation takes all GO term annotation integrated into WaspAtlas and, for a set of user-provided genes, performs hypergeometric tests for each annotated GO term in turn, comparing the proportion of genes in the user-provided set annotated with this particular term with the proportion of genes in the reference set (all annotated genes in the selected annotation) positive for that term. The p-values for all GO terms are then corrected using the Benjamini-Hochberg FDR method. The test also performs the same method for annotated PFAM domains, allowing users to test if there are any overrepresented protein families in their dataset. Other tools available include a tool to retrieve basic information in batch for large sets of genes, and orthologs for sets of genes in supported WaspAtlas species.
5.2.16 Job submission system in WaspAtlas

When a job is submitted, the user is given a 10-character job ID. On the server side, a file is made for each submitted job containing the results of the submitted job (if any). When the user checks for their job status, the web page will return an ‘in progress’ status if the file is empty, an error message if an error file has been generated, a ‘404 not found’ status if no file exists, and a job completed status along with the job results if the job has completed successfully. All job results are deleted after 24 hours. All tools are provided with a javascript autofill function containing sample data to test the tool.

As well as making the full raw result data available in text form ready to be imported into programs such as excel, all tools also provide a quick summary of the most important aspects of the result. The GO/PFAM overrepresentation tool also produces a graphical summary of the top 10 significant results. As with other illustrations in WaspAtlas, this graphical summary is produced on the fly using HTML canvas.

5.2.17 Genome browser

The Genoverse (http://wtsi-web.github.io/Genoverse/) genome browser was used for the front-end of the genome browser in WaspAtlas. Nucleotide data is fed into the genome browser from FASTA files contained on the server. A dedicated back-end page was created to feed data into the genome browser upon request, including gene and transcript co-ordinates, descriptions, and names. All objects within the genome browser contain links to relevant pages on the website.
5.3 Results

5.3.1 Features and usage

Access to the WaspAtlas database is provided through a web-based interface. The interface can be conceptually divided into four main components which give access to the database: gene summaries, custom searches, genome browser, and tools. Other pages give access to pre-prepared downloads, or to information about the website and its updates.

5.3.1.1 Gene summary

The gene summary page for each gene (Figure 5.3) is divided into five sections: I) a brief summary describing the gene identifiers associated with this gene in different annotation releases and their locations on the various genome builds. Also detailed in this section are the annotated GO terms and orthology data, II) a transcripts section showing all annotation splice variants for the gene in question, containing detailed information, illustrations and downloadable sequences for all annotated transcripts, selectable by annotation release, III) a protein annotation section containing, for each splice variant, a schematic diagram of predicted PFAM protein domains and their locations within the protein (for each splice variant), again selectable by annotation release, IV) a gene expression section, showing the levels of gene expression in both sexes in various conditions and developmental stages, V) a literature section, containing data from published studies relating to the gene being browsed.
5.3.1.2 Searches

Searching for genes (Figure 5.4A) is straightforward, and can be carried out using the quick search or the advanced search. The quick search box in the upper right hand corner of every page will scan the database for genes with a certain name, symbol, or identifier, or genes annotated with a given GO term or PFAM domain. The advanced search gives users more control over their search terms, and allows users to search for groups of GO terms/PFAM domains, and gives a greater range of fields to search (e.g. by orthologous genes in other species). Searches will return a
list of genes matching the specified parameters, allowing the user to view the gene summaries for each of the returned genes.

Figure 5.4. Analysis with WaspAtlas. (A) Use case of the advanced search function, performing a search for transcription factors involved in immune response. (B) Use case of the GO overrepresentation tool, showing input and output.

5.3.1.3 Tools

WaspAtlas also functions as an analysis platform for performing common analyses of *Nasonia* genes using the latest and most complete functional annotation available (Figure 5.4B). Current tools hosted on the website include a GO term/PFAM domain overrepresentation test using clusters of genes, utilising hypergeometric tests and multiple correction to allow users to see which GO terms/PFAM domains are overrepresented in their cluster. Also included is a tool to predict potential RNAi off-targets for a given double-stranded RNA (dsRNA) fragment to make sure
that a given dsRNA is specific to one gene (or transcript) only. Users are also able to retrieve detailed functional annotation and inter-mapping for groups of genes at a time, facilitating comparisons between studies. Finally, a tool for retrieving all orthologues for a list of genes in supported WaspAtlas species has been included. Users receive a 10-character job ID with which to check the status of their job. Data from all jobs is kept for 24 hours.

5.3.3 Genome browser

A genome browser was included into WaspAtlas (see 5.2.17), split by chromosome/scaffold. All four annotations are viewable. The nucleotide assembly of each chromosome/scaffold is overlaid with gene model co-ordinates. Zooming in further allows the viewing of individual splice variants, with their exons and introns annotated. Each gene and transcript is linked to the appropriate WaspAtlas gene information page.

5.3.4 Transcriptome analysis

Using the RNA-seq data integrated into WaspAtlas, I performed an analysis of gene expression profiles and potential “house-keeping” genes, with an aim to identify those genes which exhibit constant expression within, and perhaps even across, tissues. The genes identified from such an analysis would be suitable for use in normalization procedures (e.g. in qPCR).

To examine how similar gene expression profiles are between tissues, I calculated the mean expression values of each transcript in all samples within each dataset, performing this separately for sex (where both sexes were sequenced within a single experiment). I then calculated the correlation coefficient of the transcript
expression values between all pairs of datasets and used these values to perform hierarchical clustering in R (Team 2008). The results of this analysis are presented in Table 5.2.

**Table 5.2.** *Nasonia vitripennis* RNA-seq expression correlation values for various tissues. Tissue abbreviations are: FH (Female head), FB (Female body), MT (Male testes), MH (Male head), MB (Male body). The numbers in brackets after the tissue name represent the datasets from which they were taken according to the number in Appendix 5.1.

<table>
<thead>
<tr>
<th></th>
<th>FH (1)</th>
<th>MH (1)</th>
<th>FB (2,4)</th>
<th>FH (2)</th>
<th>MT (1,2)</th>
<th>FB (1)</th>
<th>FB (3)</th>
<th>MB (1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FH (1)</td>
<td>1.00</td>
<td>0.84</td>
<td>0.22</td>
<td>0.82</td>
<td>0.09</td>
<td>0.35</td>
<td>0.35</td>
<td>0.33</td>
</tr>
<tr>
<td>MH (1)</td>
<td></td>
<td>1.00</td>
<td>1.00</td>
<td>0.75</td>
<td>0.32</td>
<td>0.54</td>
<td>0.22</td>
<td>0.37</td>
</tr>
<tr>
<td>FB (2,4)</td>
<td></td>
<td></td>
<td>1.00</td>
<td>0.18</td>
<td>0.06</td>
<td>0.54</td>
<td>0.29</td>
<td>0.30</td>
</tr>
<tr>
<td>FH (2)</td>
<td></td>
<td></td>
<td></td>
<td>1.00</td>
<td>1.00</td>
<td>0.29</td>
<td>0.19</td>
<td>0.37</td>
</tr>
<tr>
<td>MT (1,2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.00</td>
<td>0.19</td>
<td>0.18</td>
<td>0.08</td>
</tr>
<tr>
<td>FB (1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.00</td>
<td>0.97</td>
<td>0.45</td>
</tr>
<tr>
<td>FB (3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.00</td>
<td>0.45</td>
</tr>
<tr>
<td>MB (1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.00</td>
</tr>
</tbody>
</table>
Interestingly, although whole-body gene expression profiles differ extensively between males and females (Cor < 0.5, previously reported (Wang, Werren & Clark 2015)), the male and female head transcription program is highly similar (Max cor > 0.84), which would perhaps enable direct comparisons between male and female heads in differential expression studies (Figure 5.5).

To look for potential house-keeping genes, I first found those genes with the lowest variance within datasets. For each transcript in each dataset, the coefficient of variance of its expression across all samples was calculated. The transcripts with a coefficient of variance in the bottom 10% of values (i.e. the most stably expressed) were tested for overrepresented GO terms using the WaspAtlas

**Figure 5.5.** Hierarchical clustering of *Nasonia* tissue expression profiles. Performed in R using hclust, based on a distance matrix of gene expression correlation values.
overrepresentation tool. GO terms commonly significantly overrepresented (q < 0.01) in these house-keeping sets of genes included functions to do with ribosomes, organelles and mitochondria, suggesting a true fundamental housekeeping role for these genes. Although a few of these housekeeping genes were common between all datasets, the ratios between these genes were unstable, suggesting that there is no obvious set of housekeeping genes suitable for normalization across all tissues and conditions. Lists of all of these tissue-specific housekeeping genes are available to download from WaspAtlas for use in expression normalization.

5.3.5 RRBS analysis

After quality control (see 5.2.9), reads were mapped to the Nasonia genome. 4.94% (1,385,067) of genomic CpGs were covered in the short day sample (SD), and 4.65% (1,304,087) of genomic CpGs were covered in the long day sample (LD). 4.53% (1,270,716) of genomic CpG sites were covered by both samples, at an average read depth of >88x. As with in other studies in insects, a small proportion of CpGs were found to be methylated overall; 0.146% of analysed CpGs in SD and 0.124% of analysed CpGs in LD were found to be methylated, corresponding to total figures of 2,018 methylated CpG sites (mCpGs) in SD and 1,618 in LD.

The majority of methylated cytosines detected were found to be in the CpG context, with less than 6.7% of methylated cytosines found to be in the CHG or CHH context (where H = C, T, or A, Figure 6.6A). The majority of mCpGs detected were found to be methylated in both samples (where covered), whereas mCHGs and mCHHs showed either a negligible degree of overlap (mCHG, 1/125 sites) or no overlap at all (mCHH, 0/26 sites) (Figure 5.6B). The vast majority (83.6%) of

---

2 This analysis was published in Pegoraro et al., 2016, Genome res.
mCpGs were found within exons, a larger majority still in either exons or introns (i.e. the gene body, 93.4%), with only a small proportion (6.6%) of mCpGs located in either promoter regions (2 kbp upstream of each gene) or intergenic regions. Histograms of the percentage of methylated cytosines detected at each methylated CpG site reveals a bimodal distribution of methylation (Figure 5.6C), also observed in the silkmoth and honeybee (Zemach et al. 2010).

Figure 5.6. (A) Chart showing the methylated sites detected in long-day (left) and short-day (right), coloured by context. (B) Euler diagrams showing overlap between cytosines methylated in different contexts in long-day (left) and short-day (right). (C) Distribution of methylation profiles of methylated cytosines (mC) in long-day (left) and short-day (right), showing bimodal distribution of methylation status (i.e. percentage of reads for each mC which resisted bisulphite conversion).
5.4 Discussion

WaspAtlas is the most comprehensive *Nasonia vitripennis* resource developed to date, providing an easy to explore interface for accessing the most detailed *Nasonia* gene annotation available as well as the most current -omic data produced by the *Nasonia* community. The database acts as a hub for *Nasonia vitripennis* genome annotation, including all four major annotations including the latest most comprehensive annotation, OGS2 (Rago et al. 2016). The inclusion of all annotations is important, given that different research projects utilise different annotations for different purposes.

Several datasets were analysed to provide, in a fashion similar to FlyAtlas, expression levels of each mRNA transcript for all *Nasonia vitripennis* protein-coding genes. RNA-seq datasets from 7 studies (Appendix 2.3) and a developmental microarray dataset (Wang et al. 2013b) were all analysed and included in WaspAtlas. The inclusion of these into a single resource allows a researcher to, for a particular gene, see the expression levels of the gene in various tissues and to formulate hypotheses about gene function or plan experiments accordingly. For example if a particular gene is found to be lowly expressed in the tissue of interest, this may affect the required read depth for an RNA-seq study in order to appropriate quantify this gene. The identification, analysis, and inclusion of housekeeping genes in WaspAtlas will also be useful for experimental design, when selecting appropriate housekeeping genes for a given tissue.

WaspAtlas also provides an analysis platform for working with this data, and allows the WaspAtlas data to be downloaded in batch for genome-wide analyses. This is useful for researchers without proficiency in bioinformatics, as previously even relatively simple analyses such as GO term overrepresentation
required custom scripts to parse various files. The integration of several datasets into one comprehensive database increases the amount of annotation data available for such analyses, increasing the potential to spot interesting results.

The analysis of the RRBS dataset was, at the time, the first single base pair snapshot of the *Nasonia* methylome. It revealed for the first time in *Nasonia* that the majority of methylation in Nasonia is in the CpG context, that DNA methylation in Nasonia generally occurs in exons, that only a minority of CpGs are methylated, that there is a bimodal distribution of CpG methylation, and that DNA methylation differs between long and short photoperiods in *Nasonia* (Pegoraro et al. 2016). The presence of cytosine methylation in other contexts was shown to have very little overlap between samples, raising several possibilities. The first possibility is that DNA methylation in other contexts is generally caused by experimental noise, and that our statistical thresholds for identifying methylation also picked up some of this noise. This seems likely, as the vast majority of the identified CHH/CHG methylated sites were lightly (<50%) methylated and did not show the bimodal distribution seen with CpG sites. The solution to this would therefore to be to tighten the false discovery threshold used, although this comes with the possibility of excluding some truly methylated CpG sites. Similar levels of CHH/CHG methylation have been identified in other *Nasonia* methylation studies (Beeler et al. 2014).

If, however, CHG and CHH methylation are a real biological phenomenon in *Nasonia*, then the lack of overlap between samples raises two distinct possibilities. Firstly, this methylation may represent a real biological difference between the samples, perhaps of functional consequence. This would need to be confirmed with further biological replicates. Secondly, this may represent a pattern
of random, sporadic methylation at non-CpG sites. Further experiments would be necessary to investigate whether or not non-CpG methylation truly exists in *Nasonia*.

DNA methylation in *Nasonia* is of particular interest as other model insect, such as *Drosophila*, do not show widespread methylation as in other insects (Takayama et al. 2014). The picture in *Nasonia* and other Hymenoptera is more similar to mammals, in that CpG nucleotides are generally the targets of methylation (Zemach et al. 2010). There also exist differences; *Nasonia* and other insects methylate primarily in gene bodies (Glastad et al. 2011) whereas mammals methylate primarily at promoter regions (Zemach et al. 2010). It has since been proposed that methylation in *Nasonia* provides a signal for constitutive gene expression (Wang, Werren & Clark 2015).

Overall, WaspAtlas provides a platform to help with the investigation of questions such as this, integrating several types of data into one comprehensive resource.
6 General discussion

This thesis set out to carry out initial investigations into the *Nasonia* circadian clock and to advance the wasp as a model organism for chronobiology. The aim was to accomplish this through three phases. Firstly, through investigations into the *Nasonia* clock itself, its behavioural phenotypes and transcriptional outputs. Secondly, to shed light onto the mechanisms of transcriptional and translational control by comparing *Nasonia* with other species to identify common sequences likely involved in regulation. Finally, through the development of database and web interface to facilitate such (and further) analyses, bringing together different types of datasets including the examination of DNA methylation in the wasp. Figure 1.2 (Introduction) is reproduced below (Figure 6.1), along with notes on where the work presented in this thesis contributes to our understanding of the clock and its outputs. Aspects of the thesis such as the conserved region analysis and WaspAtlas, whilst relevant to the clock, are also of interest to a wider audience.
Aspects of the workings of the molecular clock were studied through temperature compensation studies and the profiling of clock gene expression. Gene regulatory networks were investigated through the CCG transcription analysis and the conserved regulatory element analysis. The identification of groups of genes involved in various biological processes (including metabolism and various neural processes) sheds light on the outputs of the clock. Finally, the creation of WaspAtlas aided this research and will aid future genetic work in *Nasonia*.

**Figure 6.1.** Illustration of the clock and its output mechanisms, annotated with where the work presented in this thesis adds knowledge.
6.1 The diversity of insect molecular clocks

The work presented in this thesis advances the case of *Nasonia* as a model organism for chronobiology and provides data upon which other studies can build. A primary reason that a new model organism is necessary in insect chronobiology is that *Drosophila* is, in many ways, esoteric. In terms of clock genes, the evidence for this comes from genomic data. Presence or absence of the genes *timeless*, *timeout*, *cryptochrome-1* and *cryptochrome-2* represent the primary differences between insect clocks. An analysis of all species in the NR protein database along with 100 transcriptomes in the 1kite database were analysed for presence or absence of these genes using a phylogenetic tree/hidden Markov model method. The results of this analysis are summarised in Figure 6.2.
Figure 6.2: Results of an analysis of >200 insect genomes and transcriptomes (including 100 1kise transcriptomes), along with a dendrogram showing inter-order relationships (Misof et al. 2014). Orders with at least three organisms analysed are shown. Candidate sequences, determined through BLAST, were classified through being scored against hidden Markov models trained based on seed alignments, produced using phylogenetic trees of sequences from representative insect genomes. The number of insects analysed per order is shown in brackets on the right.

The gene *cryptochrome*-2, which acts on the repressive arm of the clock (Yuan et al. 2007), is present in the large majority of insect orders (23/31 insect orders analysed), but not in any sequenced *Drosophila* species - the primary insect clock model. *cryptochrome*-1 is present in 12/31 insect orders, suggesting that a large number of insects, perhaps a majority, use a *cry1*-independent light input.
pathway. This may mean that the cry1-independent input mechanism into the coleopteran clock in beetles such as Tribolium castaneum (Yuan et al. 2007) arose within Coleoptera, and that at least two distinct types of beetle molecular clockwork may exist. The genes timeless and timeout were also found in the majority of insect orders (21/31, and 25/31 respectively).

It is clear from the available genomic data that there are a wide variety of clock architectures in insects, and that Drosophila alone does not represent this diversity. The dispensation of cryptochrome-2 from the repressive arm of the clock by some insects clearly represents a divergence from the ancestral insect clock (Reppert 2007). The dispensation of cryptochrome-1 as a light sensor also represents a divergence, but has clearly occurred multiple times within insects and may represent a more mammalian-like clock mechanism. The Hymenoptera represent this type of clock within insects. By extension, Nasonia as a model fills in this missing diversity within insects and provides a new model for the mammalian clock.

6.2 The Nasonia clock in context
The work presented in this thesis reveals similarities between the Nasonia clock and the clocks of other animals. Temperature compensation was shown to be well within the range of values previously shown by other animals (Brady 1979), and the clock was shown to be more tightly compensated in constant darkness than in constant light. The functions fulfilled by clock-controlled genes were found to be similar to those fulfilled by clock-controlled genes in other animals, for example metabolism (Rodriguez et al. 2013), phototransduction (Ueda et al. 2002, Rodriguez
et al. 2013) and synaptic/nervous functions (McDonald, Rosbash 2001, Ceriani et al. 2002, Claridge-Chang et al. 2001)

The conserved regulatory elements identified were generally themselves associated with genes involved in regulation, suggesting that regulatory cascades identified in other insects are likely to also be present in much the same form in *Nasonia*. Regulatory cascades are used as ‘switches’ - modifying the expression level of a single gene (or small group of genes) can ultimately produce expression changes for large groups of genes, dramatically altering the expression profile of the entire proteome. Regulatory cascades such as these are important for circadian outputs (Li et al. 2013).

One difference between *Nasonia* and other animals presented in this thesis is the discovery of widespread low-amplitude mRNA cycling. So far, a lack of high amplitude cyclers has not been demonstrated in any animal other than *Nasonia*, except in a study of nascent RNA cycling in Drosophila (Rodriguez et al. 2013). Rodriguez *et al.* quantified the circadian expression of both nascent RNA and mature mRNA, revealing that a large proportion of cycling amplitude is imposed post-transcriptionally. This perhaps suggests differences in post-transcriptional regulation between *Drosophila* and *Nasonia*, or may lend itself to other explanations such as high amplitude cycling in a small subset of cells or desynchrony between clock neurons (see Discussion 4.4).

This low cycling amplitude presents difficulties which underlie the challenge of compiling an authoritative list of clock-controlled genes in *Drosophila*. While the profiles of high amplitude genes are easy to replicate across studies in *Drosophila*, the profiles of other genes are more susceptible to noise (Keegan et al. 2007). That this phenomenon presents itself in putative core clock genes in *Nasonia*
may be the cause of disparity across studies using methods such as qPCR. For example, *period* was shown to be rhythmic in DD by Bertossa et al. (Bertossa et al. 2014), but no significant oscillation in DD was detected by Mukai and Goto (Mukai, Goto 2016). The lack of a clear set of high amplitude cyclers makes the identification of a set of putative clock genes difficult. The identification of such a set will likely require data from other experiments.

Together, these results demonstrate that while *Nasonia* may share many similarities with other animal clocks (in many cases more than *Drosophila*), there are also differences which may present unique challenges but may ultimately result in the discovery of novel clock mechanisms.

### 6.3 Beyond the clock: wider implications

As well as the applicability of the results presented in this thesis to the circadian clock, some of the discoveries presented also have wider importance. The conserved region analysis, for example, reported the identification of the two oldest conserved regulatory elements in animals. Given their wide distribution, this discovery is likely to have implications outside of *Nasonia* and outside of the insect kingdom in general. These identified elements were present in some of the earliest animals, perhaps over a billion years ago at the dawn of multicellular animals, and are thus present in almost as wide a range of animals as possible. The presence of the RPLP1 conserved regulatory element in humans means that this element in particular is likely to receive considerable further attention.

The identification of a set of conserved regulatory regions in RNA will be useful for further investigation into what is a poorly understood class of regulatory elements. Expressed non-coding portions of RNA have recently received some
attention in *Nasonia* (Rago et al. 2016) and are thus likely to feature in future research. Overall, it should be the case that each and every one of these regions will make interesting candidates for experimental analysis, helping to increase our understanding of regulation of gene expression.

The creation of and data included within WaspAtlas is also likely to have implications outside of the field of chronobiology. Since publication, WaspAtlas has received several citations and has served over 1,000 users through 2016. The integration of all four annotations is useful for cross comparison between studies. Even within this thesis all four annotations have been used for different purposes; NCBI v2.1 was used in the circadian transcription analysis due to it being the latest and most curated NCBI annotation, OGS v1.2 and OGS v2.0 were used in the conserved region analysis due to their inclusion in Ensembl and complete UTR annotation, respectively, and NCBI v2.0 was used in the reduced-representation bisulphite sequencing analysis due to it being the most complete annotation available at the time. The creation of WaspAtlas therefore allows the ability to compare genes across analyses such as these along with all the advantages provided by each different annotation dataset.

WaspAtlas also functions to provide tools to facilitate common bioinformatics tasks. These tools were used throughout this thesis, and are frequently used by other research groups. The aim of this is to cut down on the time required to carry out these analyses and to allow reproducibility. The tools section of WaspAtlas was coded in a modular fashion which has allowed tools to be easily added since the publication of the WaspAtlas paper and will allow tools to be added in the future.
6.4 Future work

Complementary approaches may be useful to build on the discoveries presented in this thesis. Rhythmic protein expression does not directly imply rhythmic transcription as several layers of regulatory mechanisms can contribute to the rhythmic expression of proteins derived from steady-state mRNAs (Kojima et al. 2012). This also results in a time-lag between rhythmic mRNA and rhythmic protein expression (Kojima et al. 2012). Further experiments should therefore seek to characterise circadian expression of proteins, as this represents one further step in the gene regulatory network responsible for circadian phenotypes., which could confound understanding rhythmic functions when examined only on the transcriptional level.

Another approach that would be complementary to this work would be the use of in-situ hybridisation or immunocytochemistry to identify and quantify co-expressed proteins. This would provide spatial expression information likely to be useful for identifying structures and neurons important to the clock. These methods would also allow for the visualisation of the rhythmic expression of the CCGs identified in Chapter 3. In-situ experiments were attempted using RNA complementary to transcripts of *period*, *Clock*, and *rhodopsin* (as a highly-expressed positive control), but were ultimately unsuccessful due the time required to successfully adapt the protocol to *Nasonia* brains.

Further investigation of both the clock and of regulatory mechanisms in *Nasonia* is likely to require a mechanistic approach. RNAi knockdown experiments have already been attempted of *period* (as a putative core clock gene) and *dyschronic* (as a putative clock output gene (Jepson et al. 2012)), but these were unsuccessful. RNAi experiments in *Nasonia* have primarily been used to knock
down developmental genes in early developmental stages, resulting in permanent developmental defects. For example, larval RNAi has been used to knock down the eye pigment gene *cinnabar*, resulting in a red-eye phenotype mutants (Werren, Loehlin & Giebel 2009). Given the time given between injection and behavioural monitoring used in this protocol, it is possible that the transient RNAi effect will have worn off by the time its effects are to be tested. The development in *Nasonia* of gene editing methods such as CRISPR CAS9 (Lynch 2015) will be useful for knocking out candidate genes and identifying phenotypic effects, especially given the potential pitfalls and weaker effect of transient RNAi knockdown. Mutagenesis screens may play a part in identifying these candidate genes, as with early circadian research in *Drosophila*.

Gene editing approaches may also prove useful in the investigation of gene regulation in *Nasonia*. It would be informative in the future to select a subset of conserved sequences, delete or edit these regions, and quantify expression (perhaps temporally and spatially) or test for other phenotypes.

Without a putative function in mind, testing the phenotypic effects of a particular conserved region can be difficult. This has been shown to be the case with UCEs. A significant portion of ultraconserved elements (UCEs) can be deleted without any clear functional consequence (Ahituv et al. 2007). However, it is possible that these conserved regions fulfill a very specific phenotype, perhaps only presenting themselves in a very specific evolutionary niche that would not be present in a laboratory setting. Most of the CNEs that I discovered exhibit footprints of conservation as opposed to the near-perfect conservation of UCEs, and are found conserved over longer periods of time. These facts together, along with the observed conservation of position, suggest that these CNEs are likely to be involved
in more crucial functions than UCEs, and are thus likely to be more straightforward to investigate.

For example, are indications (or at least possible hypotheses) for the phenotypes of some the conserved regions. The *RPLP1* conserved sequence has already been implicated in (although not identified as conserved) allowing distinction between genes which should and should not be transcribed in the *Drosophila* early embryo. It would be informative to test this phenotype in organisms other than *Drosophila* for both *RPLP1* and *RPLP2*. If this same phenotype is indeed observed in other organisms then this raises other questions: why is such a complex sequence with several footprints necessary for this function? What changes in function (if any) are associated with inter-species variation in these sequences (e.g. the mammalian-specific insertion in the *RPLP1 CNE*)? Why is there so little redundancy in this sequence, i.e. why do we observe such strong conservation? What other functions (if any) do these sequences accomplish?

Discussions with *Nasonia* research groups have already highlighted areas to which WaspAtlas could and should expand. The recent publication of a set of in-situ hybridisation experiments comes with additional data which is to be made available at WaspAtlas, providing spatial expression data for a set of genes (Pers et al. 2016). This publication, and others recently released (e.g. (Sim, Wheeler 2016)), also include RNA-seq datasets which would supplement the expression data section in WaspAtlas. As whole-genome methylation datasets are now available for *Nasonia*, the inclusion of these into WaspAtlas would allow researchers to compare expression levels, spatial expression, differential expression, etc, with methylation levels and patterns.
More flexibility in the tools would be useful, for example allowing users to modify alpha values in the GO term overrepresentation tool or allowing modification of the kmer length in the RNAi off-target prediction tool. Documentation has been prepared to allow the next administrator to continue to manage the website and make these and other necessary updates.

6.5 Concluding remarks

Overall, the research presented in this thesis has increased the understanding of the *Nasonia* circadian clock, its outputs, and regulatory mechanisms, including epigenetic mechanisms. These results also have implications outside of the field of chronobiology, particularly relating to gene expression and *Nasonia* as a model organism in general. A resource for *Nasonia vitripennis* research has been created, acting as a hub for future investigations in what is likely to become a crucial organism for our future understanding of the circadian clock.
Appendices

Appendix 2.1. Number of RBH orthologs calculated between *Nasonia* and each species.

<table>
<thead>
<tr>
<th>Organism</th>
<th>Total annotated genes</th>
<th>No. genes with RBH</th>
<th>% annotated genes with RBH</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Acyrthosiphon pisum</em></td>
<td>36939</td>
<td>6623</td>
<td>17.93%</td>
</tr>
<tr>
<td><em>Aedes aegypti</em></td>
<td>17391</td>
<td>6738</td>
<td>38.74%</td>
</tr>
<tr>
<td><em>Apis mellifera</em></td>
<td>10675</td>
<td>7434</td>
<td>69.64%</td>
</tr>
<tr>
<td><em>Atta cephalotes</em></td>
<td>18534</td>
<td>8008</td>
<td>43.21%</td>
</tr>
<tr>
<td><em>Bombyx mori</em></td>
<td>14623</td>
<td>6472</td>
<td>44.26%</td>
</tr>
<tr>
<td><em>Danaus plexippus</em></td>
<td>16260</td>
<td>6842</td>
<td>42.08%</td>
</tr>
<tr>
<td><em>Dendroctonus ponderosae</em></td>
<td>13407</td>
<td>6426</td>
<td>47.93%</td>
</tr>
<tr>
<td><em>Drosophila melanogaster</em></td>
<td>15682</td>
<td>6760</td>
<td>43.14%</td>
</tr>
<tr>
<td><em>Heliconius melpomene</em></td>
<td>16203</td>
<td>6325</td>
<td>39.04%</td>
</tr>
<tr>
<td><em>Megaselia scalaris</em></td>
<td>11784</td>
<td>4262</td>
<td>36.17%</td>
</tr>
<tr>
<td><em>Solenopsis invicta</em></td>
<td>17377</td>
<td>7249</td>
<td>41.72%</td>
</tr>
<tr>
<td><em>Tribolium castaneum</em></td>
<td>16541</td>
<td>7091</td>
<td>42.87%</td>
</tr>
</tbody>
</table>

Appendix 3.1. Hypergeometric q-values for overrepresentations of rhythmic genes in DD.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Cluster</th>
<th>q-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>DD</td>
<td>1</td>
<td>1.23E-22</td>
</tr>
<tr>
<td>DD</td>
<td>2</td>
<td>2.51E-07</td>
</tr>
<tr>
<td>DD</td>
<td>3</td>
<td>1.04E-18</td>
</tr>
<tr>
<td>DD</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>DD</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>DD</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>DD</td>
<td>7</td>
<td>1.25E-08</td>
</tr>
<tr>
<td>DD</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>DD</td>
<td>9</td>
<td>1.46E-01</td>
</tr>
<tr>
<td>DD</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>DD</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>DD</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>DD</td>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td>DD</td>
<td>14</td>
<td>1</td>
</tr>
<tr>
<td>DD</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>DD</td>
<td>16</td>
<td>1</td>
</tr>
</tbody>
</table>
Appendix 3.2. Hypergeometric q-values for overrepresentations of rhythmic genes in LL.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Cluster</th>
<th>q-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>LL</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>LL</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>LL</td>
<td>3</td>
<td>0.4201848</td>
</tr>
<tr>
<td>LL</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>LL</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>LL</td>
<td>6</td>
<td>0.0222986</td>
</tr>
<tr>
<td>LL</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>LL</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>LL</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>LL</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>LL</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>LL</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>LL</td>
<td>13</td>
<td>0.5121081</td>
</tr>
<tr>
<td>LL</td>
<td>14</td>
<td>0.7594139</td>
</tr>
<tr>
<td>LL</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>LL</td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td>LL</td>
<td>17</td>
<td>1</td>
</tr>
<tr>
<td>LL</td>
<td>18</td>
<td>1</td>
</tr>
<tr>
<td>LL</td>
<td>19</td>
<td>1.26951e-12</td>
</tr>
<tr>
<td>LL</td>
<td>20</td>
<td>9.835358e-28</td>
</tr>
<tr>
<td>LL</td>
<td>21</td>
<td>1</td>
</tr>
<tr>
<td>LL</td>
<td>22</td>
<td>2.551139e-71</td>
</tr>
</tbody>
</table>
Appendix 3.3. Genes found to be rhythmic in both *Drosophila* and *Nasonia*.

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Gene name</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOC100119121</td>
<td>ADAM 17 protease-like. putative</td>
</tr>
<tr>
<td>LOC100120927</td>
<td>LOC100120927</td>
</tr>
<tr>
<td>LOC100680296</td>
<td>Integral membrane protein 2C</td>
</tr>
<tr>
<td>LOC100121808</td>
<td>15-hydroxyprostaglandin dehydrogenase</td>
</tr>
<tr>
<td>LOC100119395</td>
<td>4-coumarate--CoA ligase 2</td>
</tr>
<tr>
<td>LOC100120655</td>
<td>Aspartate aminotransferase</td>
</tr>
<tr>
<td>LOC100123739</td>
<td>alanine--glyoxylate aminotransferase</td>
</tr>
<tr>
<td>LOC100119902</td>
<td>Ubiquitin carboxyl-terminal hydrolase 8</td>
</tr>
<tr>
<td>LOC100121036</td>
<td>Opsin. blue-sensitive</td>
</tr>
<tr>
<td>LOC100114667</td>
<td>Thrombospondin-4</td>
</tr>
<tr>
<td>LOC100122663</td>
<td>potassium voltage-gated channel protein Shal</td>
</tr>
<tr>
<td>LOC100123943</td>
<td>Cordon-bleu protein</td>
</tr>
<tr>
<td>LOC100116192</td>
<td>luciferin 4-monoxygenase. putative</td>
</tr>
<tr>
<td>LOC100114999</td>
<td>Elongation of very long chain fatty acids protein AAEL008004</td>
</tr>
<tr>
<td>LOC100116789</td>
<td>MORN repeat-containing protein 4</td>
</tr>
<tr>
<td>LOC100116430</td>
<td>4-hydroxyphenylpyruvate dioxygenase. putative</td>
</tr>
<tr>
<td>LOC100119014</td>
<td>hepatic leukemia factor</td>
</tr>
<tr>
<td>LOC100120455</td>
<td>singed protein</td>
</tr>
<tr>
<td>LOC100121681</td>
<td>Hexamerin</td>
</tr>
<tr>
<td>LOC100115086</td>
<td>TBC1 domain family member 4</td>
</tr>
<tr>
<td>LOC100118468</td>
<td>ATP-binding cassette sub-family G member 4</td>
</tr>
<tr>
<td>LOC100119658</td>
<td>Proton-coupled folate transporter</td>
</tr>
<tr>
<td>LOC100123462</td>
<td>Follistatin</td>
</tr>
<tr>
<td>LOC100679486</td>
<td>Prohormone-1</td>
</tr>
</tbody>
</table>
Appendix 4.1. All genes and organisms used in the Not1 CNE characterisation analysis.

<table>
<thead>
<tr>
<th>Organism</th>
<th>Gene identifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acyrthosiphon pisum</td>
<td>ACYPI000441</td>
</tr>
<tr>
<td>Acyrthosiphon pisum</td>
<td>ACYPI067726</td>
</tr>
<tr>
<td>Aedes aegypti</td>
<td>AAEL007326</td>
</tr>
<tr>
<td>Aedes aegypti</td>
<td>AAEL013767</td>
</tr>
<tr>
<td>Amphimedeson queenslandica</td>
<td>Aqu1.229996</td>
</tr>
<tr>
<td>Anopheles darlingi</td>
<td>ADAR005399</td>
</tr>
<tr>
<td>Anopheles gambiae</td>
<td>AGAP009057</td>
</tr>
<tr>
<td>Apis mellifera</td>
<td>GB18988</td>
</tr>
<tr>
<td>Atta cephalotes</td>
<td>ACEP10610</td>
</tr>
<tr>
<td>Bombyx mori</td>
<td>BGIBMGA012230</td>
</tr>
<tr>
<td>Brugia malay</td>
<td>Bm4512</td>
</tr>
<tr>
<td>Caenorhabditis brenneri</td>
<td>CBN01277</td>
</tr>
<tr>
<td>Caenorhabditis briggsae</td>
<td>CBG16483</td>
</tr>
<tr>
<td>Caenorhabditis elegans</td>
<td>WBGene000002845</td>
</tr>
<tr>
<td>Caenorhabditis japonica</td>
<td>CJA32020</td>
</tr>
<tr>
<td>Caenorhabditis remanei</td>
<td>CRE20280</td>
</tr>
<tr>
<td>Caenorhabditis remanei</td>
<td>CRE20281</td>
</tr>
<tr>
<td>Capitella teleta</td>
<td>CapteG157774</td>
</tr>
<tr>
<td>Crassostrea gigas</td>
<td>CGI_10028250</td>
</tr>
<tr>
<td>Crassostrea gigas</td>
<td>CGI_10023793</td>
</tr>
<tr>
<td>Culex quinquefasciatus</td>
<td>CPIJ007821</td>
</tr>
<tr>
<td>Danaus plexippus</td>
<td>KGM_03919</td>
</tr>
<tr>
<td>Daphnia pulex</td>
<td>DAPPUDRAFT_324499</td>
</tr>
<tr>
<td>Species</td>
<td>Abbreviation</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>--------------------------------------------</td>
</tr>
<tr>
<td><em>Daphnia pulex</em></td>
<td>DAPPUDRAFT_117361</td>
</tr>
<tr>
<td><em>Daphnia pulex</em></td>
<td>DAPPUDRAFT_244230</td>
</tr>
<tr>
<td><em>Dendroctonus ponderosae</em></td>
<td>YQE_03828</td>
</tr>
<tr>
<td><em>Drosophila ananassae</em></td>
<td>FBgn0089535</td>
</tr>
<tr>
<td><em>Drosophila erecta</em></td>
<td>FBgn0117400</td>
</tr>
<tr>
<td><em>Drosophila grimshawi</em></td>
<td>FBgn0129420</td>
</tr>
<tr>
<td><em>Drosophila melanogaster</em></td>
<td>FBgn0085436</td>
</tr>
<tr>
<td><em>Drosophila mojavensis</em></td>
<td>FBgn0142403</td>
</tr>
<tr>
<td><em>Drosophila persimilis</em></td>
<td>FBgn0154914</td>
</tr>
<tr>
<td><em>Drosophila persimilis</em></td>
<td>FBgn0154915</td>
</tr>
<tr>
<td><em>Drosophila pseudoobscura</em></td>
<td>FBgn0264582</td>
</tr>
<tr>
<td><em>Drosophila sechellia</em></td>
<td>FBgn0175479</td>
</tr>
<tr>
<td><em>Drosophila simulans</em></td>
<td>FBgn0181845</td>
</tr>
<tr>
<td><em>Drosophila virilis</em></td>
<td>FBgn0202229</td>
</tr>
<tr>
<td><em>Drosophila willistoni</em></td>
<td>FBgn0223484</td>
</tr>
<tr>
<td><em>Drosophila yakuba</em></td>
<td>FBgn0239473</td>
</tr>
<tr>
<td>Euphausia superba</td>
<td>N/A</td>
</tr>
<tr>
<td><em>Heliconius melpomene</em></td>
<td>HMEL007237</td>
</tr>
<tr>
<td><em>Helobdella robusta</em></td>
<td>HelroG157038</td>
</tr>
<tr>
<td><em>Ixodes scapularis</em></td>
<td>ISCW022114</td>
</tr>
<tr>
<td><em>Loa loa</em></td>
<td>LOAG_17788</td>
</tr>
<tr>
<td>Lottia gigantea</td>
<td>LotgiG239532</td>
</tr>
<tr>
<td><em>Megaselia scalaris</em></td>
<td>MESCA003458</td>
</tr>
<tr>
<td><em>Nasonia vitripennis</em></td>
<td>NV18106</td>
</tr>
<tr>
<td><em>Nematostella vectensis</em></td>
<td>NEMVEDRAFT_v1g90783</td>
</tr>
<tr>
<td><em>Onchocerca volvulus</em></td>
<td>WBGene00245503</td>
</tr>
<tr>
<td><em>Pediculus humanus</em></td>
<td>PHUMS64470</td>
</tr>
<tr>
<td><em>Pristionchus pacificus</em></td>
<td>PPA15398</td>
</tr>
<tr>
<td><em>Rhodnius prolixus</em></td>
<td>RPRC005691</td>
</tr>
<tr>
<td><em>Schistosoma mansoni</em></td>
<td>Smp_133320</td>
</tr>
<tr>
<td><em>Solenopsis invicta</em></td>
<td>SINV10644</td>
</tr>
<tr>
<td><em>Strigamia maritima</em></td>
<td>SMAR004204</td>
</tr>
<tr>
<td><em>Strongylocentrotus purpuratus</em></td>
<td>SPU_028421</td>
</tr>
<tr>
<td><em>Tetranychus urticae</em></td>
<td>tetur10g04580</td>
</tr>
<tr>
<td><em>Tribolium castaneum</em></td>
<td>TCOGS2:TC003500</td>
</tr>
<tr>
<td><em>Trichinella spiralis</em></td>
<td>EFV52447</td>
</tr>
<tr>
<td><em>Trichoplax adhaerens</em></td>
<td>TriadG23606</td>
</tr>
</tbody>
</table>

**Appendix 4.2.** Species abbreviations used for brevity in Figures 4.12 and 4.13.
<table>
<thead>
<tr>
<th>Species abbreviation</th>
<th>Full species name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nvi</td>
<td><em>Nasonia vitripennis</em></td>
</tr>
<tr>
<td>Ame</td>
<td><em>Apis mellifera</em></td>
</tr>
<tr>
<td>Sin</td>
<td><em>Solenopsis invicta</em></td>
</tr>
<tr>
<td>Ace</td>
<td><em>Atta cephalotes</em></td>
</tr>
<tr>
<td>Dme</td>
<td><em>Drosophila melanogaster</em></td>
</tr>
<tr>
<td>Aae</td>
<td><em>Aedes aegypti</em></td>
</tr>
<tr>
<td>Bmo</td>
<td><em>Bombyx mori</em></td>
</tr>
<tr>
<td>Dpl</td>
<td><em>Danaus plexippus</em></td>
</tr>
<tr>
<td>Hme</td>
<td><em>Heliconius melpomene</em></td>
</tr>
<tr>
<td>Tca</td>
<td><em>Tribolium castaneum</em></td>
</tr>
<tr>
<td>Dpo</td>
<td><em>Dendroctonus ponderosae</em></td>
</tr>
<tr>
<td>Rpr</td>
<td><em>Rhodnius prolixus</em></td>
</tr>
<tr>
<td>Phu</td>
<td><em>Pediculus humanus</em></td>
</tr>
<tr>
<td>Dpu</td>
<td><em>Daphnia pulex</em></td>
</tr>
<tr>
<td>Isc</td>
<td><em>Ixodes scapularis</em></td>
</tr>
<tr>
<td>Lgi</td>
<td><em>Lottia gigantea</em></td>
</tr>
<tr>
<td>Hro</td>
<td><em>Helobdella robusta</em></td>
</tr>
<tr>
<td>Tru</td>
<td><em>Takifugu rubripes</em></td>
</tr>
<tr>
<td>Tni</td>
<td><em>Tetraodon nigroviridis</em></td>
</tr>
<tr>
<td>Spu</td>
<td><em>Strongylocentrotus purpur</em></td>
</tr>
<tr>
<td>Nve</td>
<td><em>Nematostella vectensis</em></td>
</tr>
<tr>
<td>Mle</td>
<td><em>Mnemiopsis leidy</em></td>
</tr>
<tr>
<td>Aqu</td>
<td><em>Amphimedon queenslandica</em></td>
</tr>
</tbody>
</table>

**Appendix 5.1.** RNA-seq libraries analysed for WaspAtlas.

<table>
<thead>
<tr>
<th>Sample name</th>
<th>Strain</th>
<th>Age</th>
<th>Additional information</th>
<th>Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female whole-body (1)</td>
<td>AsymCX</td>
<td>Adult</td>
<td>24 hr</td>
<td>(Wang, et al, 2015)</td>
</tr>
<tr>
<td>Female whole-body (2)</td>
<td>AsymC</td>
<td></td>
<td></td>
<td>(Sackton, et al, 2013)</td>
</tr>
<tr>
<td>Female whole-body (3)</td>
<td>AsymCX</td>
<td>Adult</td>
<td>24 hr</td>
<td>(Wang, et al, 2013)</td>
</tr>
<tr>
<td>Female whole-body (4)</td>
<td>AsymC</td>
<td></td>
<td>Bacterial-cocktail infected wasps</td>
<td>(Sackton, et al, 2013)</td>
</tr>
<tr>
<td>Female</td>
<td>AsymCX</td>
<td>Pavlovian</td>
<td></td>
<td>(Hoedjes, et al, 2015)</td>
</tr>
<tr>
<td>Sample Type</td>
<td>Locus</td>
<td>Phenotype</td>
<td>Note</td>
<td></td>
</tr>
<tr>
<td>-------------------</td>
<td>-------</td>
<td>------------------------------------------</td>
<td>-------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Female head (2)</td>
<td>AsymC</td>
<td>conditioned &amp; non-conditioned wasps</td>
<td>Long-day and short-day exposed wasps</td>
<td></td>
</tr>
<tr>
<td>Male whole-body (1)</td>
<td>AsymCX</td>
<td>Adult, 24 hr</td>
<td>(Wang, et al, 2015)</td>
<td></td>
</tr>
<tr>
<td>Male head (1)</td>
<td>AsymC</td>
<td>Circadian collection, constant light</td>
<td>Unpublished (Nathaniel Davies, Tauber lab)</td>
<td></td>
</tr>
<tr>
<td>Male testis (1)</td>
<td>AsymC</td>
<td>Yellow-body red-eye pupae</td>
<td>(Akbari, et al, 2013)</td>
<td></td>
</tr>
<tr>
<td>Male testis (2)</td>
<td>AsymC</td>
<td>Yellow-body red-eye pupae</td>
<td>(Akbari, et al, 2013)</td>
<td></td>
</tr>
</tbody>
</table>


of the mouse liver transcriptome, as revealed by cDNA microarray, is driven by the Suprachiasmatic Nucleus", *Current Biology*, vol. 12, no. 7, pp. 540-550.


Bembom, O. 2007, "seqLogo: An R package for plotting DNA sequence logos".


Bertossa, R.C., van Dijk, J., Beersma, D.G. & Beukeboom, L.W. 2010, "Circadian rhythms of adult emergence and activity but not eclosion in males of the


Brady, J. 1979, Biological Clocks, Edward Arnold, London.


M., Sheppard, D., Taylor, K., Thormann, A., Trehanion, S.J., Vullo, A.,
Wilder, S.P., Wilson, M., Zadissa, A., Aken, B.L., Birney, E., Cunningham, F.,
Harrow, J., Herrero, J., Hubbard, T.J., Kinsella, R., Muffato, M., Parker, A.,


Friedrich, M., Chen, R., Daines, B., Bao, R., Caravas, J., Rai, P.K., Zagmajster, M.
& Peck, S.B. 2011, "Phototransduction and clock gene expression in the
troglobiont beetle Ptomaphagus hirtus of Mammoth cave", The Journal of

Fuchikawa, T. & Shimizu, I. 2007, "Effects of temperature on circadian rhythm in
the Japanese honeybee, Apis cerana japonica", Journal of insect physiology,
vol. 53, no. 11, pp. 1179-1187.

mammalian circadian timing system: from gene expression to physiology",

Gallo, S.M., Gerrard, D.T., Miner, D., Simich, M., Des Soye, B., Bergman, C.M.
& Halfon, M.S. 2011, "REDfly v3.0: toward a comprehensive database of
gregarious locusts differ in circadian rhythmicity of a visual output neuron",
*Journal of Biological Rhythms*, vol. 27, no. 3, pp. 196-205.

Gekakis, N., Saez, L., Delahaye-Brown, A.M., Myers, M.P., Sehgal, A., Young,
M.W. & Weitz, C.J. 1995, "Isolation of timeless by PER protein interaction:
defective interaction between timeless protein and long-period mutant PERL",

Gekakis, N., Staknis, D., Nguyen, H.B., Davis, F.C., Wilsbacher, L.D., King, D.P.,
Takahashi, J.S. & Weitz, C.J. 1998, "Role of the CLOCK protein in the
5369, pp. 1564-1569.

Giannoni-Guzman, M.A., Avalos, A., Marrero Perez, J., Otero Loperena, E.J.,
Kayim, M., Medina, J.A., Massey, S.E., Kence, M., Kence, A., Giray, T. &
Agosto-Rivera, J.L. 2014, "Measuring individual locomotor rhythms in honey
bees, paper wasps and other similar-sized insects", *The Journal of experimental
biology*, vol. 217, no. Pt 8, pp. 1307-1315.

20, no. 5, pp. 553-565.

Glossop, N.R.J., Lyons, L.C. & Hardin, P.E. 1999, "Interlocked feedback loops
766-768.


algorithm to functionally analyze large gene lists", *Genome biology*, vol. 8, no. 9, pp. R183.


Developmental Programs across Highly Diverged Insects", *Genome biology and evolution*, vol. 6, no. 9, pp. 2301-2320.


Kureck, A. 1979, "Two Circadian Eclosion Times in Chironomus thummi (Diptera),
Alternately Selected with Different Temperatures", Oecologia, vol. 40, no. 3,
pp. 311-323.

Langmead, B. & Salzberg, S.L. 2012, "Fast gapped-read alignment with Bowtie 2",

Lee, C., Bae, K. & Edery, I. 1999, "PER and TIM inhibit the DNA binding activity
of a Drosophila CLOCK-CYC/dBMAL1 heterodimer without disrupting
formation of the heterodimer: a basis for circadian transcription", Molecular

Lee, I., Ajay, S.S., Yook, J.I., Kim, H.S., Hong, S.H., Kim, N.H., Dhanasekaran,
S.M., Chinnaiyan, A.M. & Athey, B.D. 2009, "New class of microRNA targets
containing simultaneous 5'-UTR and 3'-UTR interaction sites", Genome

Lee, T.I. & Young, R.A. 2013, "Transcriptional regulation and its misregulation in
disease", Cell, vol. 152, no. 6, pp. 1237-1251.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G.,
Abecasis, G., Durbin, R. & 1000 Genome Project Data Processing Subgroup
2009, "The Sequence Alignment/Map format and SAMtools", Bioinformatics

Li, J., Grant, G.R., Hogenesch, J.B. & Hughes, M.E. 2015, "Chapter Sixteen -
Considerations for RNA-seq Analysis of Circadian Rhythms" in Methods in


Mito, T., Kobayashi, C., Sarashina, I., Zhang, H., Shinahara, W., Miyawaki, K., Shinmyo, Y., Ohuchi, H. & Noji, S. 2007, "even-skipped has gap-like, pair-rule-like, and segmental functions in the cricket Gryllus bimaculatus, a basal,


Nasonia reveals extensive incorporation of novelty in a regulatory network",  


R Development Core Team 2010, *R: A Language and Environment for Statistical Computing*, 2.10.1, Vienna, Austria.


Sahar, Saurabh & Sassone-Corsi, Paolo "Metabolism and cancer: the circadian clock connection", .


Saifullah, A.S.M. & Tomioka, K. 2003, "Pigment-dispersing factor sets the night state of the medulla bilateral neurons in the optic lobe of the cricket, Gryllus bimaculatus", *Journal of insect physiology*, vol. 49, no. 3, pp. 231-239.


Sauman, I., Briscoe, A.D., Zhu, H., Shi, D., Froy, O., Stalleicken, J., Yuan, Q., Casselman, A. & Reppert, S.M. 2005, "Connecting the navigational clock to


Sokolove, P.G. & Loher, W. 1975, "Rôle of eyes, optic lobes, and pars intercerebralis in locomotory and stridulatory circadian rhythms of
Teleogryllus commodus", *Journal of insect physiology*, vol. 21, no. 4, pp. 785-799.


Stelzer, R.J. & Chittka, L. 2010, *Bumblebee foraging rhythms under the midnight sun measured with radiofrequency identification.*


Yoshii, T., Funada, Y., Ibuki-Ishibashi, T., Matsumoto, A., Tanimura, T. & Tomioka, K. 2004, "Drosophila cryb mutation reveals two circadian clocks that drive locomotor rhythm and have different responsiveness to light", *Journal of insect physiology*, vol. 50, no. 6, pp. 479-488.


