Master of Science Thesis

De novo RNA sequencing enables transcriptome studies in non-model species: gene expression patterns involved in *Drosophila montana* reproductive diapause

Mikko Merisalo

University of Jyväskylä
Department of Biological and Environmental Science
Ecology and Evolutionary Biology
2.7.2014

JYVÄSKYLÄN YLIOPISTO, Matemaattis-luonnontieteellinen tiedekunta

Bio- ja ympäristötieteiden laitos
Ekologia ja evoluutiobiologia

Merisalo, M.:	RNA-sekvensointi mahdollistaa transkriptomitason tutkimukset ei-mallilajeilla: geenien toiminta *D. montana* -kärpäsen lisääntymislepokaudessa.
Pro Gradu –tutkielma:	36 s.
Työn ohjaajat:	Dos. Maaria Kankare, FT Tiina Salminen, MSc Darren Parker
Tarkastajat:	Prof. Anneli Hoikkala, Dos. Emily Knott
Heinäkuu 2014	

Hakusanat: RNA-sekvensointi, lisääntymislepokausi, kriittinen valojakso, *Drosophila*

TIIVISTELMÄ

Monet pohjoiset hyönteiset selviävät talvesta lisääntymislepokaudessa, joka on geneettisesti määräytyvä ja hormonaalisesti ohjattu fysiologinen lepotila. Lisääntymislepokautta on pääosin tutkittu ekologisesta ja fysiologisesta näkökulmasta, mutta sen geneettisestä taustasta tiedetään vähemmän. Geneettisen tiedon puuttumiseen ovat vaikuttaneet sopivan tutkimuslajin ja menetelmien puute, sekä toisaalta lisääntymislepokauden varsin monimutkainen luonne. Nykyaikaiset sekvensointimenetelmät mahdollistavat varsin laajat geneettiset tutkimukset lajeilla, joiden genomia ei ole vielä sekvensoitu, mutta joiden ekologiasta on paljon tutkimustietoa. Esimerkkinä kyseisenlaisesta tutkimuslajista lisääntymislepokauden geneettiseen taustaan liittyen ovat *Drosophila montana* -kärpäset, jotka talvehtivat aikuisvaiheen lisääntymislepokaudessa. Pro Gradu -tutkimukseni tarkoituksena oli tunnistaa nykyaikaisen RNA-sekvensointimenetelmän avulla geenejä, joiden toiminta muuttuu *D. montana* kärpästen lisääntymislepokauden aikana. SOLiD 5500XL sekvensointi tuotti 200 miljoonaa lyhyttä sekvenssiä, jotka laatumuokkausten jälkeen koottiin noin 32 000 pitemmäksi sekvenssiksi (engl. contig). Contigeja vastaavien geenien määräksi saatiin noin 70 % *D. viriliksen* genomin koosta, mikä viittasi siihen, että tutkimuksessa pystyttiin tuottamaan hyvälaatuinen transkriptomi tutkimuksen loppuosaa varten. Geenien toiminta contigien tasolla erosi siten, että se oli 17 %:lla contigeista merkitsevästi korkeampaa ja 19 %:lla matalampaa diapaussaavilla kärpäsillä ei-diapaussaaviin verrattuna. Toiminnantasoiltaan eroavat contigit jaettiin laajempiin geeniryhmiin. Korkeamman toiminnan geeniryhmät sisälsivät geenejä, jotka liittyvät ympäristön muutoksiin vastaamiseen, soluntukirankaan, aineenvaihduntaan, kuten esimerkiksi glukoosi- ja rasva-aineenvaihdunta, sekä (ionien) kuljetukseen. Myös matalamman geenitoiminnan geenit liittyivät aineenvaihduntaan ja kuljetukseen sekä lisäksi mitoosiin/meiosiin ja DNA/RNA:n toimintaan. Lopuksi kymmenen toiminnaltaan eniten diapaussaavien ja ei-diapaussaavien naaraiden välillä eroavaa geeniä pyrittiin nimeämään tarkemmin. Geeneistä kaksi liittyi mitä luultavimmin ympäristön havainnointiin hajuaistin avulla, kolme geeniä glukoosiaineenvaihduntaan ja yksi geeni sekä rasva- että glutationi-aineenvaihduntaan. Lisäksi kaksi myosiinigeeniä liittyi todennäköisesti solutukirangan toimintaan. Listan viimeisestä geenistä ei löytynyt sen toimintaa määrittelevää tutkimustietoa. Yhteenvetona voidaan todeta, että tutkimuksessa löydettiin useita mielenkiintoisia geenejä ja geeniryhmiä, joita voidaan käyttää tulevaisuudessa lisääntymislepokauteen liittyvissä tarkemmissa tutkimuksissa.

UNIVERSITY OF JYVÄSKYLÄ, Faculty of Mathematics and Science

Department of Biological and Environmental Science
Ecology and Evolutionary Biology

Merisalo, M.:	De novo RNA sequencing enables transcriptome studies in non-model species: gene expression patterns involved in *Drosophila montana* reproductive diapause.
Master of Science Thesis:	36 p.
Supervisors:	Doc. Maaria Kankare, PhD Tiina Salminen, MSc Darren Parker
Inspectors:	Prof. Anneli Hoikkala, Doc. Emily Knott
July 2014	

Key Words: RNA sequencing, diapause, critical photoperiod, *Drosophila*

ABSTRACT

A wide range of insects are able to survive the harsh winter conditions in the northern hemisphere by entering a genetically programmed and hormonally controlled state of dormancy known as diapause. Most of the research so far has concentrated on the ecology and physiology of diapause. The complex nature of diapause in addition to the lack of suitable genetic study species and methodology has hindered the growth of genetic knowledge on diapause. Recently developed high-throughput sequencing technology however enables large scale genetic studies also on non-model species with well-characterized ecology. An example of such a study species for diapause is a north adapted Drosophila fly*,* *Drosophila montana*, that overwinters in an adult reproductive diapause. The goal of this study was to use RNA sequencing to study differentially expressed genes in response to diapause in *D. montana* flies. SOLiD 5500XL sequencing resulted in 200 million paired-end reads, which after quality filtering were assembled into 32 000 contigs. The contigs matched close to 70 % of the genes in *D. virilis* genome indicating a good reference transcriptome, which was then be used as the basis for the rest of this study. Differential expression analyses found significant upregulation for 17 % and downregulation for 19 % of the contigs in the diapause treatment. Functional annotation clustering divided the up- and downregulated contigs into larger units. Genes in the upregulated clusters have been connected to response to external stimulus, cytoskeleton, (ion) transportation and metabolism, such as glucose and fatty acid metabolism. Downregulated clusters included genes also connected to metabolism and transportation as well as to mitosis/meiosis and DNA/RNA activity. Finally, information for the ten most upregulated genes was collected. Two of these genes most likely function in sensing the environment through olfaction, three genes have been connected to glucose metabolism and one gene in both fatty acid and glutathione metabolism. Two myosin genes could function in cytoskeleton. The last gene had limited research information about its function. In conclusion, this study introduced many interesting genes and gene clusters, which could be used in the future to study the many different and interesting aspects of diapause.

Contents

1. INTRODUCTION

Seasonally changing environmental conditions pose a great challenge for organisms inhabiting northern latitudes. An adaptation to periodically suppressing organism's normal active development into a state of dormancy enables them to survive over the hostile seasons. This has arguably enabled a large variety of insects to inhabit even the harshest of environments. A common type of dormancy in insects is diapause, a genetically programmed and hormonally controlled set of physiological events, where growth and reproduction are halted over the harsh season and synchronized to resume in a more favorable season (Denlinger 2002).

Diapause is a wide spread adaptation in insects (Nishizuka 1998) and can occur at any stage of development from embryonic, pre-pupal and larval to adult stage, but it is generally limited to only one developmental stage in a species' life-cycle (Denlinger 2002). Egg or larval stage diapause is often an obligatory diapause occurring in every generation in univoltine (only one generation per year) species with only minor environmental control (Danks 1987). A more common facultative diapause necessitates a decision whether to enter diapause or to resume normal development in multivoltine species (Denlinger 2002). Facultative diapause is often observed in adult stages, where development is suspended and reproduction is postponed to more favorable conditions (Danks 1987). In facultative diapause the decisions to enter, maintain and terminate diapause are direct responses to changing environmental conditions. The most reliable signal to track seasonal changes is the photoperiod (Tauber et al. 1986), which plays a major role in helping to coordinate diapause with seasonal changes especially in the northern latitudes (Beck 1980). Also temperature has an effect on the diapause response (Danks 1987). Photoperiodic signals are less useful close to the equator, where other cues such as temperature, drought or food availability are used instead (Denlinger 1986).

The mechanisms for measuring seasonally changing photoperiodic signals, sometimes referred as a seasonal clock, theoretically consists of four units (Saunders 2002). In the first unit (I), the input of photoperiodic signals through light receptors, such as cryptochrome (Goto & Denlinger 2002), connects the seasonal clock to the environment. In the second unit, information from the light receptors is used by the photoperiod clock (II). Currently, the mechanisms for this process are poorly understood as is the relationship between the photoperiodic clock and the more studied daily time measurement system known as circadian clock. Views of the connection of seasonal clock with the photoperiod clock ranges from being part of the same system to being completely independent systems (Saunders 2002; Emerson et al. 2009; Koštál 2011). Nonetheless, the theoretical clock system would then feed information to the third unit, photoperiodic counter mechanism (III), which accumulates information on successive photoperiods and triggers the last unit (IV), the output pathways, when a certain threshold is reached (Saunders 2002; Koštál 2011). The number of cumulative photoperiods needed to reach the threshold level varies between species and is also thought to be affected, for example, by temperature (Saunders 2002).

Photoperiodic signals that induce diapause are effective only within a species specific sensitive period in insect’s life cycle, which is occurs before the actual diapause state (Danks 1987). A critical day length represents a stationary photoperiod for a study population during their sensitive period when a diapause response is observed in half of the individuals while the other half continues normal active development (Beck 1980). Usually the width of the critical photoperiod is very narrow, and subtle changes in either direction affect the resulting diapause incidence (Xiao et al. 2010). Hence, selection favors the optimal timing of diapause initiation in different latitudes (Bradshaw 1976; Lankinen 1986), for example the effects of climate change in the form of longer growing seasons (Menzel & Fabian 1999) has already been shown to favor more southern, shorter critical day lengths (Bradshaw & Holzapfel 2001a).

Output pathways of the seasonal clock system effect the insect's endocrine system (Saunders 2002; Emerson et al. 2009) and several hormones involved in diapause regulation have been identified (Denlinger et al. 2011). One of the most studied examples of embryonic diapause comes from the silkworm *Bombyx mori*, where maternally secreted diapause hormone (DH) leads to the production of diapause destined eggs (Yamashita 1996). In larval and pupal diapause the absence of ecdysteroids (insect moulting hormones) has been connected to the diapause state (Hamel et al. 1998; Denlinger et al. 2011). Adult diapause has most often been connected to juvenile hormone (JH) (Denlinger 2011), which typically shows lower levels during diapause, but markedly higher levels shortly after diapause termination and during normal development (Schooneveld et al. 1977; Saunders et al. 1990; Readio et al. 1999). Additionally, ecdysteroid levels (Richard et al. 1998) and insulin signaling pathway have been found to behave similarly to JH in adult diapause, the latter being involved in nutrient management (Badisco et al. 2013) making it one of the key candidate for controlling adult diapause (Giannakou & Partridge 2007; Sim & Denlinger 2008; Hahn & Denlinger 2011).

Changes in hormone levels lead to the physiological characteristics observed in diapausing insects (Emerson et al. 2009). The sensitive period is followed by a preparative phase, when diapause destined individuals change their behavioral (Calvert & Brower 1986) and feeding patterns (Bowen 1992) to prepare for the adverse season. Nutrient reserves are accumulated in the insect fat body mainly as lipids (triacylglyceride) (Arrese & Soulages 2013), but also as glycogen (Zhou & Miesfeld 2009) and storage proteins (Burmester 1999; Denlinger 2002) with the expense of ovarian development in the adult reproductive diapause (Adams 1985). It is critical for the success of diapause that individuals gather enough energy reserves in advance since feeding is often minimized if not arrested over the actual diapause phase and insufficient reserves can affect diapause entry and termination (Hahn & Denlinger 2007). Also, extra reserves could enhance post-diapause development (Zhou & Miesfeld 2009). In addition to nutrient reserves, molecular chaperones and cryoprotectants, for example heat shock proteins and glycerol, are often synthesized to protect proteins and tissue from stressful conditions such as freezing or desiccation (Ishiguro et al. 2007; Rinehart et al. 2007). During the actual diapause phase, metabolic levels are suppressed with varying degree (Guppy & Withers 1999) and energy usage is transferred away from costly tissues such as the flight muscles (Kim & Denlinger 2009) to more critical systems such as the brain (Hahn & Denlinger 2011). As a result, diapausing individuals live longer than normally reproducing individuals, which is needed to survive over the long hostile season (Herman & Tatar 2001; Tatar et al. 2001).

Much of the diapause research attention has been given to economically important species, such as the silkworm (*B. mori*), potato pest Colorado potato beetle (*Leptinotarsa decemlineata*), rice pest Rice stem borer (*Chilo suppressalis*) and mosquito species acting as disease vectors (*e.g. Aedes aegypti*, *Culex pipiens*), with emphasis on the ecological and physiological aspects of diapause as described above. In contrast, the underlying molecular and genetic patterns are much less well known (Bradshaw & Holzapfel 2001b; Denlinger 2002). An impeding factor has been the complexity of the diapause phenotype, which has most likely appeared several times independently in the history (Nishizuka 1998; MacRae 2010). Moreover, diapause is considered as an alternative dynamic pathway to normal active development with various phases (Kostal 2006) and modules (Emerson et al. 2009).

Research on the molecular basis of diapause has also been complicated by a shortage of suitable genetic model study species (Denlinger 2002). The fairly recent discovery of an adult stage diapause in a genetic model species *Drosophila melanogaster* (Saunders et al. 1989) enabled more in depth research on the genetic aspects of diapause than before (Saunders et al. 1990; Williams & Sokolowski 1993; Stanewsky et al. 1998; Tatar et al. 2001; Schmidt et al. 2005; Baker & Russell 2009; Paaby & Schmidt 2009). However, the diapause response in *D. melanogaster* is shallow, young in origin and only observed under certain temperatures never reaching a full 100 percent response (Saunders & Gilbert 1990; Schmidt & Paaby 2008). Therefore, other species, e.g. those inhabiting northern latitudes, with a more robust diapause response would be better suited for studies on diapause genetics than *D. melanogaster* (Denlinger 2002).

Use of more northern non-model species to study diapause genetics has been hindered by lack of suitable methodology to attain genetic knowledge on these species, especially in a complex phenotype such as diapause. However, new technology enables leaps from studies of only few genes to many, even up to the level of transcriptome, which is the set of all RNA transcripts in a cell, tissue or whole organism in a certain physiological stage or time (Wang et al. 2009). Previously, one of the most used methods to examine change in gene expression across many genes has been hybridization-based techniques, such as microarrays (Schena et al. 1995; Heller 2002). However, this technology is limited by the pre-existing genomic knowledge on the studied organism (Wang et al. 2009), which is usually lacking for the most part for non-model organisms. Also, designing probes for a study species based on sequences from a model species can give false results via cross-hybridization (Casneuf et al. 2007). Another set of technologies often used to study transcriptomes are tag-based methods such as the serial analysis of gene expression (SAGE) (Harbers & Carninci 2005). However, these methods are mostly based on expensive and laborious Sanger sequencing technology (Sanger et al. 1977), which limits their use (Wang et al. 2009).

Recently developed high-throughput sequencing technology is changing the way of studying non-model organisms. This technology enables large scale molecular studies on species with well-characterized ecological background, but limited genetic and genomic knowledge (Ekblom & Galindo 2011). These so called next-generation sequencing methods use massively parallel sequencing that produce millions of short sequence reads from single, amplified DNA sequences in a single machine run (Shendure & Ji 2008). Several platforms of this technology (Ansorge 2009) offers various applications including genome wide de novo (Li et al. 2010) and re-sequencing (see Table 2 in Metzker 2010), ChIP-Seq to profile DNA regulatory proteins (Park 2009), metagenomics to determine microbial DNA from environmental samples (Chistoserdova 2010), targeted sequencing of individual genes or sections of DNA (Levin et al. 2009) and RNA sequencing (RNA-seq) to study gene expression variation (Wang et al. 2009).

Out of the variety of the different applications RNA-seq has been one of the most popular (Ekblom & Galindo 2011). In a standard RNA-seq protocol RNA samples are first fragmented, adapter ligated and converted to cDNA sequences (library preparation step), which are then attached separately to a solid surface or otherwise immobilized and amplified in some platforms. Next, all the attached templates are sequenced simultaneously in high-throughput manner resulting in a large amount of short sequence reads (Costa et al. 2010). Typically, in a RNA-seq data analysis pipeline, reads are mapped to a reference, either a genome or a de novo assembled transcriptome, normalized for within and between library differences, subjected for statistical testing of differential expression and finally functionally classified based on, for example, Gene Ontology (GO) searches (Oshlack et al. 2010).

RNA-seq, and next-generation sequencing technology in general, holds great advantages over other methods for genome and transcriptome wide studies. The technology has no reliance on existing genomic knowledge, which is especially suitable for non-model organisms (Wang et al. 2009). In the case of RNA-seq, the produced data contains information for example about sequence variation (e.g. SNPs and indels), transcription boundaries, transcriptome characteristics, splicing patterns, gene expression levels and genetic markers (Wang et al. 2009; Ozsolack & Milos 2010; Ekblom & Galindo 2011). There are many challenges for this still maturing technology, especially in the field of data analysis to develop efficient and reliable software to analyze the large and ever-growing data sets (Field et al. 2006; Nekrutenko & Taylor 2012). Nonetheless, the technology has already been successfully applied to many different organisms and purposes with varying genomic background knowledge (Ekblom & Galindo 2011; Qian et al. 2014).

With these developments in sequencing technology, more attention can now be paid to non-model organisms with ecologically and evolutionary interesting study systems. An excellent candidate species of such type for diapause research is a northern fruit fly *Drosophila montana* from the *Drosophila virilis* species group with a divergent time from *D. virilis* of about 9 million years (Morales-Hojas et al. 2011) and from *D. melanogaster* of about 63 million years (Tamura et al 2004). *D. montana* most likely originated in Asia from where it spread to the northern hemisphere reaching latitudes from 30°N to 70°N (Throckmorton 1982). *D. montana* females overwinter in an adult reproductive diapause, which is induced mainly by shortening photoperiod in the autumn, well in advance of the harsh winter period (Lumme 1978). Critical day length for diapause entry in *D. montana* changes along a latitudinal cline, where subtle changes in photoperiod affects the diapause incidence showing a strong photoperiodic response and local adaptation even in the presence of high gene flow (Tyukmaeva et al. 2011; Lankinen et al. 2013). *D. montana* flies are relatively cold tolerant (Vesala & Hoikkala 2011), show changes in rhythmicity adjusting them better for the long summer photoperiods (Kauranen et al. 2012) and have a strong photoperiodic diapause response (Tyukmaeva et al. 2011; Salminen & Hoikkala 2013) which together make this species well adapted to conditions in the north and an excellent target to study genetic aspects in diapause (Kankare et al. 2010).

The overall goal for this study was to use next-generation RNA sequencing data from diapausing and non-diapausing *D. montana* females collected from the critical day length to study differentially expressed genes in response to the reproductive diapause. Using flies reared in population specific critical day length enabled a novel way to eliminate the effects of varying photoperiod on the results, a factor that has been overlooked so far (e.g. Poelcahu et al. 2011). The overall goal can be divided further into four more specific aims for this study. The first aim (i) was to produce a reference transcriptome for *D. montana,* which could be used as the basis for the other three aims of this study, which were (ii) to determine gene expression differences between diapausing and non-diapausing females, (iii) to functionally classify differentially expressed genes and (iiii) to present ten most upregulated genes in diapausing females.

2. MATERIALS AND METHODS

2.1. Sample collection and RNA sequencing

Three *D. montana* isofemale strains (3OL8, 175OJ8 and 265OJ8) were used in this study. Each line was founded from the offspring of a single female fly collected from Oulanka, Finland (66ºN) in 2008. Since collection, the strains have been maintained in laboratory with overlapping generations under diapause preventing conditions of constant light, 19ºC and 60% humidity in half-pint plastic bottles containing malt medium (Lakovaara 1969).

Virgin female flies for the studies were collected from maintenance bottles within one day after eclosion using light CO2 anesthesia to help to separate the sexes. Females were put into vials with malt medium and transferred to a climate chamber (Sanyo MLR-351H). Flies were reared in a population specific critical day length, which for Oulanka population is approximately 18,5 hours of light and 5,5 hours of dark (16ºC and 60 % humidity) (Tyukmaeva et al. 2011). Three weeks (21 days) in the climate chamber in these conditions is sufficient time for the flies to develop mature ovaries or start to prepare for winter via reproductive diapause and allocate resources elsewhere than to ovarian development (Salminen et al. 2012). After 21 days, the flies were taken out and immediately snap frozen with liquid nitrogen (–196ºC) in order to preserve the state of their RNA. Frozen flies were then stored in a –85ºC freezer.

The stage of ovarian development was used to determine whether a female fly is diapausing or non-diapausing. Pre-vitellogenic small and transparent ovaries with no yolk accumulation or visible segments were classified as diapausing and large vitellogenic ovaries with visible eggs as non-diapausing (Salminen & Hoikkala 2013). Flies that had intermediate ovaries with some yolk accumulation and segments visible but no eggs were not used in the study. To be able to dissect the flies and check the developmental stages without damaging their RNA, frozen flies were submerged into RNAlater ICE (Ambion) according to company instructions and dissected under a light microscope. Ten diapausing or non-diapausing flies separated based on their ovarian development stage were pooled into each sample.

RNA extraction was performed using Tri Reagent (Sigma-Aldrich) extraction kit followed by RNeasy Mini (Qiagen) kit with DNase treatment. RNA concentration and purity was checked using NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies) and integrity using 2100 Bioanalyzer (Agilent Technologies). Based on the concentration and quality values, best three diapausing and three non-diapausing samples, one of each from the three isofemale strains used in this study, were sent for the library preparation in the Finnish Microarray and Sequencing Centre in Turku.

A large part of a total RNA sample usually consists of ribosomal RNA and only a small amount is messenger RNA. Therefore, MicroPoly(A) Purist kit (Ambion AM1919) was used to enrich for mRNA from total RNA sample. Sequencing libraries were constructed from each sample using SOLiD total RNA-seq Kit (Whole Transcriptome library) with unique barcodes (SOLiD Transcriptome Multiplexing Kit). Libraries were then sequenced with SOLiD 5500XL Genetic Analyzer (Applied Biosciences) using sequencing by ligation chemistry (Shendure et al. 2005) in a six lane flow cell with 75 base pair forward reads and 35 base pair reverse reads (paired-end reads).

The above mentioned six samples from the three different isofemale strains (Appendix 1, 3 diapausing and 3 non-diapausing samples, nos. 1-6) were sequenced together with ten additional *D. montana* samples (Appendix 1, diapausing, non-diapausing and cold acclimated samples, nos. 7-16). The full 16 sample data set was used in the first aim of this study to produce a reference transcriptome for *D. montana*. The reference transcriptome then served as the basis for the other three aims where only the six samples collected from critical day length were used.

2.2 Transcriptome assembly and annotation

2.2.1 Read quality control

The SOLiD sequencing platform does not have a built-in system to pre-filter low quality reads, but all reads with successful base calls, regardless of their quality, are added to the sequence data. The two main types of errors that can occur in a sequence data like this are polyclonal errors, when the entire read is of bad quality, and single erroneous bases, when single bases have been miss-called (Sasson & Michael 2010). Therefore, it is necessary to run the SOLiD reads through effective filtering systems to reduce these errors and ensure good quality data for the assembly stage. For these steps all the reads from 16 samples were merged into two .csfasta-files (F3all.csfasta and F5all.csfasta) and two .qual-files (F3all.qual and F5all.qual). Raw sequence reads were first trimmed using SOLiD TRIM (with run options: -p 3 -q 22 -y y -e 2 -d 10) to remove polyclonal errors from the data (Sasson & Michael 2010). Reads that passed this filter were then error corrected using SOLiD Accuracy Enhancer Tool (SAET tools) to reduce the amount of color calling errors, or erroneous bases, in the sequence. The program was run using default option and a reference length of 20 million bases. The filtered reads were then imported to CLC Genomics Workbench 5.0.1 (CLC bio) with paired-end distance between forward and reverse reads set from 80 to 300 bases. An additional filtering step was then performed using Genomics Workbench's trimming option to remove reads with quality score greater than 0.02 and with a sequence less than 20 nucleotides.

2.2.2 De novo assembly

Since *D. montana* does not currently have a reference genome available, a de novo assembly of the transcriptome was used. In the de novo assembly there is no reference sequence to assemble the reads against but only the sequenced reads themselves. Therefore, sequence reads from all the 16 samples as described before were used in the assembly. In addition, a mode of assembly was chosen where the sequence reads were mapped back to the assembled contigs right after the assembly was finished. This provides information on mapping and contig statistics, which were used to check the quality of the assembly and the contigs. The de novo assembly parameters in Genomics Workbench were left as default values (i.e. word size = 24, bubble size = 50 and minimum contig length of 200 base pairs with scaffolding option on).

2.2.3 Sequence annotation

Contigs were annotated using Blast2GO (Conesa et al. 2005), which uses online or local Blast searches from a chosen database to find similar sequences for a data set of input sequences. Assembled contigs were blasted against non-redundant protein sequence database (nr database with blastx search). Blast cut-off threshold for E-value, which is the significance score for a blast result, was set to 0.001 (default value). Contigs without a hit were blasted to non-redundant nucleotide collection (nr database with blastn search) and the contigs still without a hit were blasted against Reference Sequence (RefSeq) genomic database (blastn with default values) (Pruitt et al. 2007). Contigs with a genomic scaffold hit to RefSeq were blasted against annotated genes from the 12 *Drosophila* genomes (Appendix 2). Contigs that had a blast hit to ribosomal RNA or non-arthropod sequences were removed from the contig list.

Contigs were further annotated using *D. melanogaster* gene orthologs. Contigs with a gene blast hit were connected to their corresponding *D. melanogaster* gene homologs, if known, using ortholog information from Flybase.org (St. Pierre et al. 2014). In order to estimate the amount of genes present in the blast results, i.e. the transcriptome size, the gene annotations and *D. melanogaster* orthologs were used to pool all contigs blasting to the same gene to represent a single gene.

2.3 Differential contig expression between diapausing and non-diapausing females

2.3.1 Mapping

In order to obtain expression estimates for each contig, sequence reads were mapped to the de novo assembly on a sample by sample basis to get read counts for every contig in a sample. Read counts are simply the number of individual reads that match to a contig's sequence. However, this can be seen as the expression level of a certain contig in the studied sample (Mortazavi et al. 2008). Since the error correction and trimming stages were previously done to a combined file of all 16 samples, it was repeated again with just the six samples focused on in the latter part of this study. Default parameters were used to map the samples back to the contigs using Genomic Workbench in order to get read counts for each contig in each of the six samples.

2.3.2 Differential contig expression analysis with DESeq

Read counts for each of the six samples were loaded into DESeq package (version 1.9.4, Anders & Huber 2010) in R programming environment (2.14.2). Samples were assigned to "diapause" and "non-diapause" conditions in order to detect differential expression of contigs between these two phenotypes. Between-library normalization was used to account for variation in the library size caused by differences in the sequence depths between the samples. Normalization in DESeq is accomplished by calculating specific size factors for each sample to adjust read counts, which has been shown to be an effective between-library normalization method (Dillies et al. 2012). Often an additional normalization step is used to account for variation in contig length, because the amount of reads mapping to a contig is proportional to its length. However, when expression levels of the same contigs are compared between samples, the bias caused by contig length is usually canceled out leaving no need for within-library normalization (Oshlack et al. 2010).

A generalized linear model (GLM) with a negative binomial distribution was fitted in DESeq with diapause state as a factor. Pooled dispersion were then estimated across all samples and used in a GLM likelihood ratio test to determine if each contig was differentially expressed between each of the diapause and non-diapause states. p-values from the GLM likelihood ratio test were multiple test corrected using Benjamini and Hochberg's algorithm to control for false discovery rate (FDR) (Benjamini and Hochberg 1995) with a significance level of <5% (FDR < 0.05).

Based on these results contigs were divided into lists of upregulated and downregulated contigs. Upregulated contigs have higher and downregulated contigs have lower mean expression values in the diapausing females compared to non-diapausing ones. The division to up- and downregulation enables to study contigs more closely that have different expression levels in the two phenotypes.

2.4 Functional gene annotation

DAVID Bioinformatics Resources web program (v. 6.7) (Huang et al. 2009) was used to carry out functional clustering for the differentially expressed contigs. For DAVID to find the annotations, all gene IDs from Blast2Go results were converted to Flybase IDs, apart from the results that did not have a Flybase ID (i.e. contigs that blasted to non-flybase species), which were converted to Refseq annotation numbers. Transcriptome data usually contains multiple contigs blasting to the same gene and since DAVID contains gene specific information, all duplicate gene IDs were removed. Hence, contigs were matched to their corresponding genes, which were then used in the gene level clustering analysis. The set of all unique gene IDs were used as the background list of genes.

For the functional annotation analysis, gene IDs were arranged in the up- and down-regulated gene lists based on the multiple test corrected p-values from DESeq. However, DAVID limits the amount of IDs that can be imported in a study list to 3000. Therefore, in both of the lists a p-value of 0.01 from DESeq results was used to limit the amount of gene IDs imported to DAVID as the study list. Duplicate gene IDs were also removed from the two study gene lists.

DAVID compares a list of study gene IDs to a background list in an enrichment analysis to identify functional categories, or clusters, of annotation terms that are over-represented in the study list. Clusters are ranked based on their level of enrichment, which is defined as the geometric mean of p-values (EASE score, see Hosack et al. 2003) for each annotation term within the group (Huang et al. 2009).

Gene ontology (GO) is a structured and controlled way to annotate genes and give functions and roles to genes in an organism (The Gene Ontology Consortium 2000). These annotations, or GO terms, are divided into three different categories: biological process, molecular function and cellular component. Here, the Gene Ontology options were changed to include only biological processes and all the other options were left to default values.

From the results, enrichment score can be converted to a p-value by negative logarithmic transformation. Clusters from both of the study lists that have the corresponding p-value less than 0.05 were used in further analyses along with gene IDs falling into those clusters. The significant clusters were then divided into broader functional groups based on the annotation terms from DAVID and genes involved in the clusters. Explanations for the annotation terms used to define a cluster are accessible via the annotation term ID.

2.5 Top upregulated genes in diapausing females

The list of upregulated contigs was arranged based on the multiple test corrected p-value from DESeq and top contigs with the lowest p-value were chosen for closer investigation. Contigs were annotated to their corresponding genes and regarded as genes thereafter. Since even the *D. melanogaster* genome have not been annotated completely, only genes having a *D. melanogaster* ortholog with a specified gene name were accepted to the list, which should ensure necessary annotation information for every gene. Along with data from the previous steps in this study, additional information was searched also from literature and Flybase (for example from development and anatomy expression data, St. Pierre et al. 2014).

3. RESULTS

3.1 Transcriptome assembly and annotation

The sequencing with SOLiD 5500XL Genetic Analyzer resulted in approximately 200 million paired-end reads of 75 and 35 bases in length (Table 1). These raw reads were filtered using SOLiD TRIM and SOLiD Accuracy Enhancer Tool (SAET), which removed 3% of the reads. Trimming the reads further with Genomics Workbench corrected 82 % of the reads and the mean read length shortened from 55 bases to 45 when trimming removed bad quality bases from the end of the reads.

The trimmed reads were assembled into 31 880 contigs with 15 million bases in total (Table 1). The contig N50 value, which is a weighted median that describes the length where 50 % of the assembled contigs sorted by length are equal or longer than this value, was 527 and mean contig length 471 when the minimum contig length was set to 200 bases. Highest contig length frequency was between 220 and 260 bases (Figure 1).

Table 1. Sequencing and assembly statistics. N25, N50 and N75 values refer to the length of a contig at the corresponding percentages in a list of contigs sorted by length.

Number of original sequence reads	199 433 554
Number of reads remaining after SOLiD trim and SAET	198 793 392
Number of reads remaining after CLC trim	164 724 753
Number of contigs	31 880
N75	327
N50	527
N25	932
Minimum contig length	186
Maximum contig length	7 371
Mean contig length	471

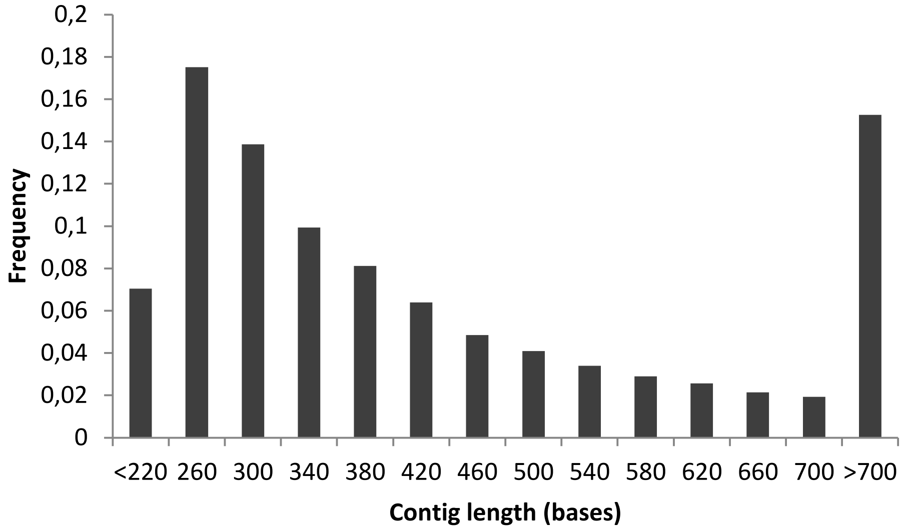

Figure 1. Histogram of contig length frequencies from 31 880 contigs. Minimum contig length was 186 bases and maximum 7371 bases. Highest contig length frequency was between 220 and 260 bases.

Blasting the assembled contigs to the nr database resulted in blast hits for 26 549 of the contigs (83 %). Blasting the remaining contigs to RefSeq genomic database produced significant blast results for 5010 of the contigs. As a result 25 644 contigs had a hit to a gene, 5000 contigs blasted only to genomic scaffolds from RefSeq, 321 contigs had no blast hit and 645 contigs were discarded as rRNA or as non-arthropod sequences. Genomic scaffolds in the RefSeq database are stretches of genomic sequence that have no annotation information yet. Performing a local blast for the contigs with genomic hits against the annotated genes of the 12 *Drosophila* genomes (Appendix 2) produced an additional 376 gene hits. The rest of the genomic scaffold hits were annotated only by the species their blasted to (Table 2).

A vast majority of the top blast hits matched to *D. virilis* sequences (Figure 2), which is expected since this species is the closest relative of *D. montana* with a sequenced genome available. Also, most of the other blast hits matched to one of the 12 sequenced *Drosophila* species, and less than 2 % of the blast hits were to other arthropod species.

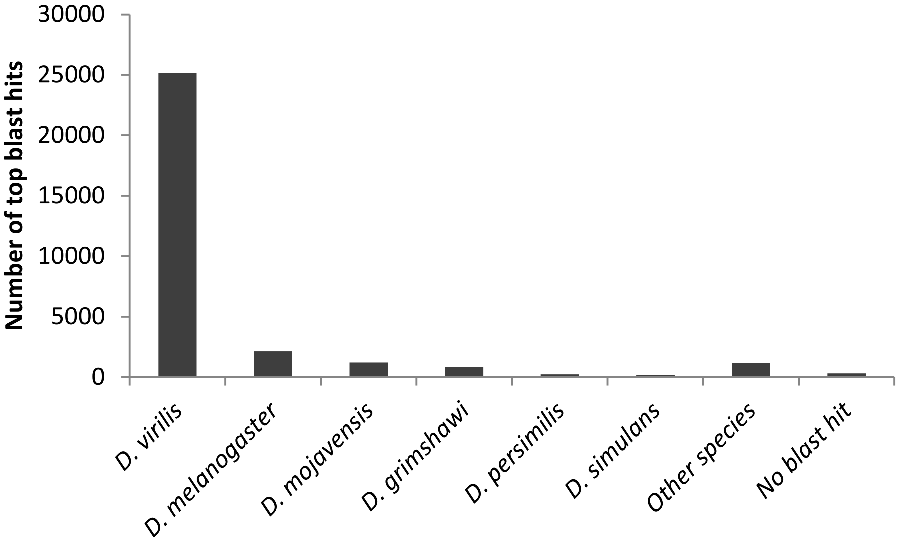

Figure 2. Top blast hit species for assembled contigs from Blast2GO. Other species include 1160 contigs, of which only 130 are non-*Drosophila* arthropod species and the rest different *Drosophila* species.

D. melanogaster gene orthologs were retrieved for contigs with a gene hit, which resulted in 23 612 contigs being annotated to *D. melanogaster* genes leaving 2408 contigs with no *D. melanogaster* ortholog. In order to estimate a transcriptome size for *D. montana*, all the contigs with a *D. melanogaster* ortholog that match the same gene were pooled to represent a single gene, which reduced the amount of contigs down to 8628 genes. The remaining contigs (2408) with no known *D. melanogaster* gene ortholog were cut down to 1538 unique gene IDs based on the Blast2GO results (Table 2).

Table 2. Contig annotation results. Contigs with a gene hit were connected to their corresponding *D. melanogaster* orthologs, if one existed. Duplicate contigs that matched to the same ortholog were removed (contigs with a unique *D. melanogaster* ortholog). Duplicates were removed also from contigs that did not have an ortholog, but had a blast hit to the same gene. Genomic scaffolds are stretches of genomic sequence in the RefSeq database that have no annotation information yet. Discarded contigs had blast hits to ribosomal RNA or to non-arthropod species.

	Number of contigs
Contigs	31880
Contigs with a blast hit	31559
Contigs with a gene blast hit	26020
Contigs with a *D. melanogaster* ortholog	23612
Contigs with a unique *D. melanogaster* ortholog	8628
Contigs with a unique gene ID but no *D. melanogaster* ortholog	1538
Contigs with only a genomic scaffold hit	4624
Discarded contigs (non-arthropod blast hits)	645

3.2 Differential contig expression between diapausing and non-diapausing females

Differential expression test between diapausing and non-diapausing *D. montana* females was performed in DESeq for the library size corrected read counts from a total of 31235 contigs. Using a 0.05 multiple-test corrected p-value, 5353 contigs were significantly upregulated and 5825 were significantly downregulated. That is, approximately 17% of the contigs had significantly higher expression level in the three diapausing samples than in the three non-diapausing samples and approximately 19% of the contigs had lower expression. The large number of differentially expressed contigs is visualized in Figure 3, which shows that even when the mean read count was as low as ten, the test was able to identify differentially expressed contigs with close to the same level of fold change as with higher mean read counts.

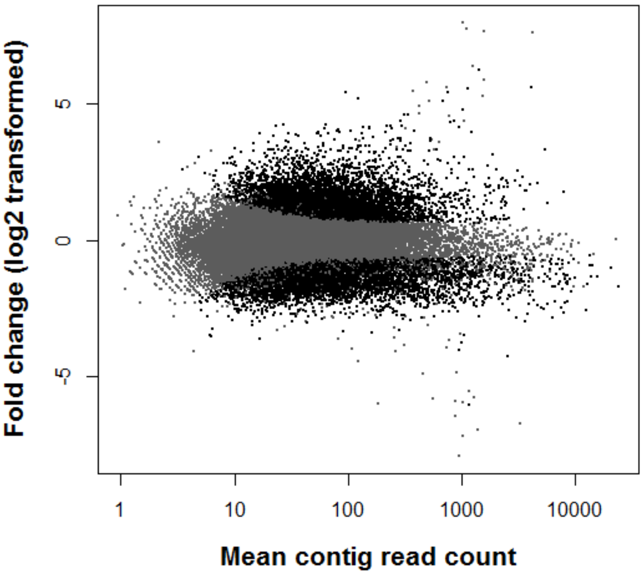

Figure 3. Scatter plot of logarithmic fold change values (non-diapausing vs. diapausing) against mean normalized read counts (Anders & Huber 2010). Black dots represent contigs with a significant differential expression when using 5% false discovery rate (FDR) significance level.

3.3 Functional gene annotation

A total of 12 649 gene IDs were imported to DAVID Bioinformatics Resources web program to be used as a background list in the enrichment analysis. The two study gene lists consisted of 2158 upregulated and 2435 downregulated genes in diapausing females. Enrichment analysis done separately for the gene lists divided upregulated genes into 159 clusters and downregulated genes into 151 clusters. Enrichment score of 1.3, which corresponds to a p-value of 0.05, was used as a cutoff value to select significant clusters. 31 and 33 clusters (Appendixes 3-4) had enrichment scores higher than 1.3 for the upregulated (marked with i) and downregulated (marked with j) genes, respectively. The significantly enriched clusters from the two gene lists were both organized into four larger cluster groups (Figure 4).

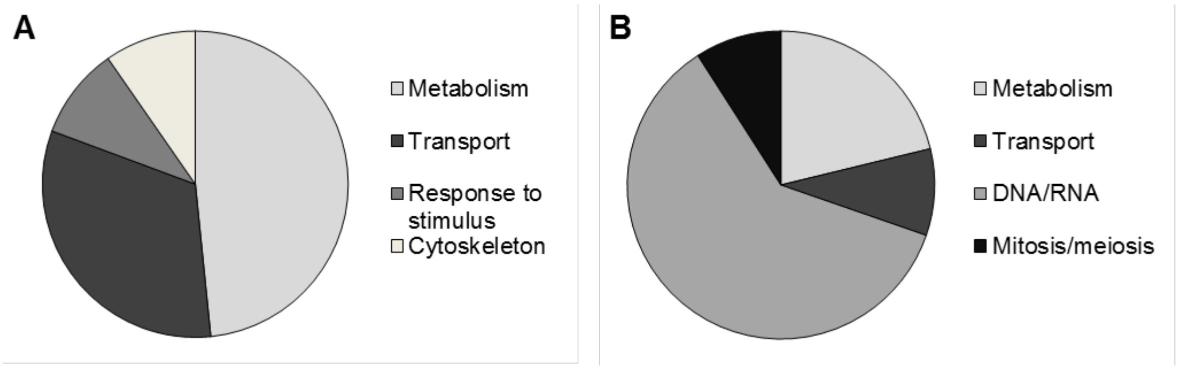

Figure 4. Annotation cluster groups for the significantly enriched upregulated (A) and downregulated (B) gene annotation clusters.

The first group (i) from the upregulated genes was named as “response to stimulus”. It included three clusters with annotation terms on heat shock proteins, sensory perception and rhodopsin, which is a visual pigment in photoreception cells. All the three stimulus clusters were quite similar with genes intertwined between the clusters. For example, the *neither inactivation nor afterpotential E* (*ninaE*) gene was assigned to all of the clusters and two heat shock proteins *Hsp22* and *Hsp67Bc* to the first two clusters.

The second group (ii) consisted of three clusters that had annotation terms spectrin repeat, immunoglobulin and actin. Spectrin functions in proteins involved in cytoskeleton structure along with actin. Immunoglobulins are protein superfamily categorized based on structural features and it includes domains that are involved in various functions including cell-surface receptors, muscle structures and the immune system. The immunoglobulin cluster included genes with functions in all of the above mentioned categories. However, most of the top differentially expressed genes, for example *Unc-89*, *bent* and *Stretchin-Mlck*, have been connected to cytoskeleton or myosin related function and myosin has also been connected to actin. Therefore, the cluster was grouped under “cytoskeleton”. Genes involved with immune response included e.g. *Relish (Rel)* and *PDGF- and VEGF-receptor related (Pvr).*

Different clusters of metabolic genes were both up- and downregulated in diapausing females. The upregulated metabolism group (iii) consisted of 15 clusters with annotation terms on, for example, oxidation reduction, chymotrypsin, cytochrome P450, peptidase and glycoside hydrolase. Chymotrypsin and peptidase function in protein catabolism with various more specific functions. Glycoside hydrolase belongs to a group of enzymes that hydrolyze glycosidic bonds in carbohydrates. Cytochrome p450 is part of an enzyme superfamily, which is found in all kingdoms of life and which functions in oxidizing multitude of substrates. Finally, oxidation reaction consists of all metabolic processes where electrons are transferred between reactants.

The next group on transport can be also found from both the up- and downregulated cluster sets. This last group from the upregulated gene list (iiii) included 10 clusters out of which four are directly connected to ion transport functions through cell membranes and one to sugar transport. The C2 membrane-targeting proteins and Munc-13 proteins are part of calcium dependent membrane targeting and also EF-Hand proteins function in calcium binding. Basic leucine zipper proteins mediate sequence-specific DNA-binding and JHBP is annotated as juvenile hormone binding protein.

The downregulated metabolism group (j) is smaller than the one in upregulated genes with 7 clusters and annotation terms on, for example, DNA repair, ribosome biogenesis, chaperonin and tetratricopeptide. Out of these clusters the latter two refer to proteins that function in protein-protein interactions to enable proper protein assembly.

There are only three clusters in the transport group in the downregulated genes (jj). The first two clusters included genes affecting chromosome organization and transporting molecules into, out of or within the nucleus. The third cluster named Armadillo is annotated as a group of genes with a specific amino acid tandem repeat, which have many functions such as intracellular signaling or cytoskeletal regulation.

The last two groups from the downregulated gene list are mitosis/meiosis group (jjj) that has only 3 clusters and the largest DNA/RNA group (jjjj) with 20 clusters. Genes in these two groups are involved in managing the cell cycle and DNA replication processes by, for example, unfolding the DNA double helix (helicases), replicating a new strand (DNA polymerases), packing the DNA into nucleosomes (e.g. histones) and locating specific sequences of DNA or protein (e.g. zinc fingers).

3.4 Top upregulated genes in diapausing females

The top ten most upregulated genes with a *D. melanogaster* ortholog are listed in Table 3 along with the expression and annotation information. All the genes are very highly differentially expressed and most of them have also very high fold change values. Only three genes are not involved in any of the upregulated annotation clusters detailed above. The rest of the genes belong to different metabolism clusters except *Odorant-binding protein 44a (Obp44a)*, which is part of a transportation cluster.

The first gene *antdh* has been connected to olfaction (Wang et al. 1999) and based on expression information (St. Pierre et al. 2014) it is almost entirely active in adult heads, or more specifically in the antennae. Also the second, a non-metabolism gene *Obp44a* has highest expression in the head, but it is also highly expressed in the central nervous system (St. Pierre et al. 2014).

The third gene, *Zwischenferment* (*Zw*, also known as *G6PD*), is a glucose metabolism gene that functions in redox reactions. The gene has the highest expression in adult crops (St. Pierre et al. 2014). Also two other genes in the top 10 list are involved in similar type of metabolic activities. *target of brain insulin (tobi)* gene acts in carbohydrate metabolism and *Maltase A1 (Mal-A1)* in glucose metabolism. Furthermore, *tobi* is involved in insulin signaling mediating balance between dietary protein and sugar (Buch et al. 2008) and *Mal-A1* is named as maltase, which refers to the hydrolysis of disaccharide maltose into glucose. *tobi* and *Mal-A1* genes are most highly expressed in the adult digestive system (St. Pierre et al. 2014).

The next metabolism gene is *Desaturase 2 (Desat2)*, which functions in fatty acid metabolism. More specifically, *Desat2* is a cuticular hydrocarbon pheromone (Wicker-Thomas 2007) responsible for *D. melanogaster* pheromone polymorphism (Takahashi et al. 2001). The last gene connected to metabolism, *Glutathione S transferase E6 (GstE6)*, is involved in glutathione metabolic activities. It is most highly expressed in the adult digestive system (St. Pierre et al. 2014).

The next two genes have annotation terms linking them to myosin activity. The *Myofilin (Mf)* functions most likely in muscle myosin assembly (Qiu et al 2005) and the *Myosin binding subunit (Mbs)* gene has many ontogenesis related annotations.

The final gene, *PFTAIRE-interacting factor 1B (Pif1B)*, has very limited annotation information thus far. It was discovered when it interacted with a *Drosophila* early development gene *Ecdysone-induced protein 63E* (Rascle et al. 2003). *Pif1B* has high expression levels throughout different adult fly tissues, but it has the highest expression in the carcass of larvae (i.e. remaining tissues after the CNS, gut, trachae and most fat body have been removed) and adult flies (i.e. remaining tissues after the gut and sexual tracts have been removed) (St. Pierre et al. 2014).

Table 3. Ten most upregulated genes in diapausing *D. montana* females with multiple test corrected p-values <0.001. *D. melanogaster* gene name refers to the *D. melanogaster* ortholog that has been connected to the blast hit gene ID. Other information include contig length in bases, fold change for the gene expression differences between diapausing and non-diapausing females, upregulated gene clusters where the gene is part of in the gene functional annotation analysis results, the number of contigs each gene have in the data, Flybase anatomy expression information (St. Pierre et al. 2014) about the tissue, organ or body part where the gene is most highly expressed in and finally annotation information about the genes.

No.	Blast hit gene ID	*D. melanogaster* gene name	Contig length	Fold change	Clusters involved in (upregulated)	Same gene contigs	Highly expressed (St. Pierre et al. 2014)	Annotation
1	FBgn0205775	*antdh*	1072	6,3	1	1	Head, antennae	Oxidation-reduction, NAPDH metabolism, olfaction
2	FBgn0207363	*Odorant-binding protein 44a*	602	6,0	6	1	Head, nervous system	Odorant binding, olfaction
3	FBgn0203146	*Zwischenferment*	704	6,2	1,8	2	Crop	Oxidation-reduction, glucose metabolism,
4	FBgn0210847	*target of brain insulin*	1163	5,0	8	1	Digestive system	Carbohydrate metabolism, insulin signalling
5	FBgn0210267	*Desaturase 2*	1259	5,3	1,4,16,19,20	1	No specific expression	Fatty acid metabolism, stress resistance
6	FBgn0068265	*Myofilin*	524	4,8	-	3	No specific expression	Muscle myosin assembly
7	FBgn0207770	*Maltase A1*	1016	4,9	8,13	2	Digestive system	Glucose metabolism, maltose metabolism
8	FBgn0200487	*Myosin binding subunit*	273	4,9	-	3	No specific expression	Myosin activity
9	FBgn0210834	*PFTAIRE-interacting factor 1B*	1502	5,0	-	3	Carcass	No specific annotation
10	FBgn0207035	*Glutathione S transferase E6*	459	5,2	12	1	Digestive system	Glutathione metabolism

4. DISCUSSION

4.1 Transcriptome assembly and annotation

A reference transcriptome was produced for *D. montana* from sequence read libraries containing 16 diapausing, non-diapausing and cold acclimated female and male samples (Appendix 1). The transcriptome was assembled from more than 160 million error corrected reads into almost 32 000 contigs with an N50 of 527 and average contig length of 471. These results are comparable with a transcriptome study using the same platform indicating a successful assembly (Everett et al. 2011). However, SOLiD sequencing platform has rarely been used in de novo transcriptome sequencing studies leaving very few studies to cross-check the results to. Therefore, the next step of annotation served also as a validation for the transcriptome's quality.

The transcriptome annotation yielded a blast result for 99 % of the contigs. Out of the results, close to 82 % of the contigs blasted to known genes and over 14 % to genomic scaffolds in the RefSeq database. As expected, most of the blast hits were to sequences from *D. virilis* (>25 000), which is the closest relative to *D. montana* with a sequenced genome. The remaining hits were almost all to other *Drosophila* species and only 645 contigs were discarded as non-arthropod sequences, which indicate good quality of the original RNA samples, but also that the results produced in the sequencing and assembly steps were of good quality.

There are about 15 000 genes in the *D. virilis* genome (Drosophila 12 Genomes Consortium 2007; St. Pierre et al. 2014), which is much less than the amount of contigs with blast hits to genes in this study. In order to get a better estimate of the transcriptome content for *D. montana* the contigs were reduced down to less than 9000 genes using *D. melanogaster* orthologs. This is about 58 % of the genes in *D. virilis* genome and when combined with contigs having a unique gene id, but no *D. melanogaster* ortholog, the amount rises into about 68 %. The observed genetic coverage and low amount of contamination in this study indicates a good reference transcriptome for *D. montana*, which then also served as a reliable reference for the rest of this study.

4.2 Differential contig expression between diapausing and non-diapausing females

Differential expression was determined for original contigs instead of genes, because there were more than 4000 contigs that matched only to general genomic sequences. It is unlikely that all of these sequences would represent a novel gene, which is also supported by the high amount of contigs that match to different genes in this study. A more likely scenario is that most of these sequences are transcripts from a known gene, but they fall outside the current gene model boundaries. In addition, many contigs that matched to a gene blasted to known intron areas of that gene instead of exons (data not shown) representing possible alternative splicing events. Consequently, combining contigs with annotation data of this kind could heavily bias the expression results.

Out of the contigs, 17 % were significantly upregulated and 19 % significantly downregulated in diapausing *D. montana* females. Altogether, more than one third of all the contigs were found to be differentially expressed, which is a high percentage compared to the few other transcriptome studies on gene expression differences in diapause (Poelchau et al. 2011; Ning et al. 2013). The main factor behind this observed difference between the studies is most probably due to sample design. Even though next generation sequencing technology has a lower price when compared to conventional Sanger sequencing methods taking into account the amount of data achieved (Hall 2013), it is still expensive to get samples sequenced with this new technology. The need for a sound experimental design has been overrun by the need to cut the costs and many of the initial next generation sequencing projects lack, for example, the use of biological replicates (Auer & Doerge 2010). Without replicates, there is no information about the biological variation between samples, which causes problems for statistical testing and the reliability of the results (Anders & Huber 2010).

Using good quality samples with three biological replicates in this study enabled to find differential expression even with low read counts for many contigs between the sample groups. Additionally, all the samples were reared in the same conditions in population specific critical day length, which could reduce unwanted variation in gene expression levels due to differences in lighting conditions. Also, a good normalization method, as the one used in the DESeq (Dillies et al. 2012), is required to remove the unwanted technical variation from the results. Even though the concentration in the original RNA samples were leveled before the sequencing step, size factors for the between-library normalization step did differ somewhat between few of the samples, which if left uncared for, would have caused variation in the samples.

Consequently, the high amount of differentially expressed contigs in this study is very likely to have arisen from substantial differences between the two phenotypes being compared. The diapause response involves the activation of many gene modules from photoperiodic and temperature signal measurement systems to hormonal control mechanisms, which leads to the physiological characteristics observed in diapausing individuals (Emerson et al. 2009). These changes include adjusting behavioral and feeding patterns, accumulating nutrient reserves, molecular chaperones and cryoprotectants before the diapause stage and suppressing metabolic levels during diapause. Also, in adult reproductive diapause such as in *D. montana*, ovaries are left at pre-vitellogenic stage with no yolk accumulation or egg development during the diapause phase. The distinction to normal active development of growth and reproduction is clear, which is seen not only in the mentioned physiological differences between the two phenotypes, but also in the genetic level as observed in this study.

More than one third of all the contigs being differentially expressed is a problematic result when trying to assess the importance of individual genes on the diapause response. Moreover, differences are so large between the two phenotypes that using non-diapausing individuals as a control treatment against diapausing individuals is perhaps not the most optimal choice. Tissue heterogeneity between diapausing and non-diapausing phenotypes, for example as differences in ovary and fat body size, could add noise to the gene expression results (Neville & Goodwin 2012). However, there are no other straightforward control phenotypes available for diapausing flies. Other potential controls in addition to non-diapausing mature flies could be, for example, to use young flies or to remove ovaries from the non-diapausing flies, but both of these options have problematic issues. Age difference could cause differential expression in developmental genes and if the young flies would be collected from critical day length, it would not be possible to know the future phenotype for the young flies as they make the decision whether to enter diapause or not approximately at the age of 4 days (Salminen & Hoikkala 2013). On the other hand, removing ovaries altogether might leave out genes that at worst could regulate some key features in diapause since also the diapausing females have ovaries but no eggs or yolk in them. A compromise would be to sequence all of the above mentioned life stages and/or samples and compare the results, or to use a limited number of genes to verify the results using, for example, qPCR technology.

4.3 Functional gene annotation

Due to the large amount of differentially expressed genes the data needed to be clustered into assemblages of contigs with similar biological function, which can then be investigated more easily. The annotated clusters from both of the up- and downregulated gene lists were further organized into 4 larger cluster groups.

The first upregulated group, response to stimulus, included clusters on sensory perception and heat resistance. High enrichment scores for genes involved in heat shock proteins (HSP) is expected amongst upregulated genes in diapause. HSPs acts as molecular chaperones for other proteins in various stressful conditions such as cold or desiccation (Rinehart et al. 2007) and many different HSPs are found to be active in diapausing individuals (Yocum et al. 1998; Rinehart et al. 2000; Vesala & Hoikkala 2011). In this study, heat shock proteins genes *Hsp22* and *Hsp67Bc* are examples of genes having higher expression levels in diapausing than in non-diapausing individuals.

For the clusters on visual perception the observed result is not so straightforward since there does not seem to be obvious benefits of having higher expression for these genes in diapausing than in non-diapausing individuals. One possible explanation could be that diapausing individuals need to follow photoperiod more carefully when entering diapause than non-diapausing individuals. Critical photoperiods are very narrow for *D. montana* populations and change along the latitudinal gradient (Tyukmaeva et al. 2011) indicating a high pressure for the correct timing of diapause. However, even after the decision to enter diapause these individuals will follow the changing conditions and will break diapause if transferred to non-diapausing conditions (Salminen & Hoikkala 2013). The advantage of a correctly timed decision to enter diapause is high. A late critical photoperiod might not leave enough time to prepare for diapause by not being able to accumulate the necessary energy reserves (Hahn & Denlinger 2007). On the other hand, early decision could prevent producing an extra generation that could still survive over the harsh period. Therefore, any gene expression changes that would enable the flies to follow the environmental signals more accurately could be seen as an important adaptation for the diapause phenotype.

Examples of genes with high involvement in the visual clusters are rhodopsins, especially a *ninaE* gene that produces the major rhodopsin Rh1, which functions as photoreception pigment in visual perception (Kiselev & Subramaniam 1994). Interestingly, the seasonal photoperiodic calendar system is not well known, neither are the proteins involved in photoreception in this system. So far, cryptochrome has been one of the most potential candidates for a photoreceptor due to its connection to circadian rhythms (Emery et al. 1998; Stanewsky et al. 1998). Also opsins, such as melanopsin and boceropsin, have been speculated as potential photoreceptors (Denlinger et al. 2011). Furthermore, the opsin gene with expression differences observed here, rhodopsin, has suitable characteristics for diapause photoreceptor, like blue-light sensitivity (Denlinger 2011; Kiselev & Subramaniam 1994), but it has not been connected to photoperiodism yet. However, rhodopsin also functions in thermal detection (Shen et al. 2011), which could affect the observed expression patterns. It could help diapause destined individuals in preparation for colder weather or even be involved in the cold tolerance of the flies (Vesala et al. 2012b).

Cytoskeleton as the main annotation of the second cluster group has many functions in the cell including cellular division, intracellular transport and cell shape. Changes in its structure have been observed in low temperatures aiding the cold tolerance and temperature sensing in plants (Abdrakhamanova et al. 2003; Pokorna et al. 2004). Also in insects, low temperatures cause a change in the distribution and structure of actin (Kim et al. 2006), a core constituent of cytoskeleton. Moreover, in diapausing individuals the change is even larger accompanied by upregulation of actin genes (Kim et al. 2006; Robich et al. 2007). In this study, several different actin, myosin and microtubule related genes in the cytoskeleton group had high expression levels in diapausing samples supporting the role of actin for being involved in the diapause phenotype as also observed by Salminen et al (submitted manuscript) for *D. montana*.

One of the clusters in the cytoskeleton group comprised of genes connected to the multifunctional immunoglobulin superfamily. Most of the top differentially expressed genes in this cluster were connected to cytoskeleton related functions, but the cluster also included immune response genes. For example, *Rel* gene is part of the Rel family of genes that functions in immune response not only in insects, but also in mammals (Dushay et al. 1996). Another gene, *Pvr*, has also been connected to immune response in *D. melanogaster* (Bond & Foley 2009). Thus, a more in-depth analysis of the genes part of the immunoglobulin cluster could provide further information on even previously unknown genes that act in *Drosophila* immune response and/or in diapause.

Clusters grouping under metabolism appear in both upregulated and downregulated genes. Large amount of metabolism genes expressed in the diapausing females compared to a smaller amount in the non-diapausing females is at first somewhat surprising since diapause destined individuals must usually gather energy reserves during the preparative phase. The actual diapause stage is characterized by metabolic suppression and decline in the stored energy reserves (Tauber et al. 1986). Whereas, the non-diapausing individuals have access to food resources throughout the summer, which female flies use, for example, to develop mature ovaries for reproduction. Nonetheless, managing the limited energy reserves during diapause most likely requires a lot of effort since it involves not only their proper use but also assessing the levels of the remaining reserves (Hahn & Denlinger 2007). Also, despite of being a successful and necessary adaptation to survive over long periods of environmental hardships, diapause is a metabolically expensive and stressful strategy with many trade-offs and fitness costs (Bradshaw et al. 1998; Ellers & Van Alphen 2002). Consequently, large amount of metabolism genes needs to be active also during the diapause stage.

Many of the GO terms in the metabolic clusters are very broad, such as oxidation reduction or carbohydrate metabolic process, making it difficult to estimate the involvement or importance of individual clusters in diapause. However, many of the upregulated genes seem to belong to clusters that are part of resource utilization metabolism such as protein catabolism or carbohydrate, fatty acid and organic acid metabolism. Metabolic clusters involved in the downregulated genes are more connected to the management or protection of DNA replication and protein synthesis. The low amount of possible dietary metabolism clusters in the non-diapausing females is surprising and could be explored further.

As an example of a potentially interesting metabolic upregulated gene group with a very high enrichment score is the cytochrome p450 cluster. These genes belong to a superfamily of genes found in a large variety of phyla and also many genes have been identified in *Drosophila* flies (Tijet et al. 2001). In general, these genes function in metabolizing many different compounds such as steroids and fatty acids (Brun et al 1996) and few genes has been associated with insecticide resistance (Liu & Scott 1995; Le Goff et al. 2003). Cytochrome p450 genes are suggested to affect stress resistance in aging flies (Pletcher et al. 2002) and be involved in larval diapause in silkmoth (*Antheraea yamamai*) (Yang et al. 2008). Thus, it would be interesting to examine closer the connection of cytochrome p450 genes with *D. montana* diapause for example using RNA interference (RNAi) technology.

Many of the clusters in the last upregulated cluster group on transportation have very broad and imprecise annotation terms. For example, most of the ten clusters are annotated with function in ion transportation, especially with calcium. Also the three clusters in the corresponding downregulated transportation group have the same problem with annotations to large groups of genes with many different functions. This raises an issue with the GO terms when some of the terms are quite precise and some others so broad that it is difficult to assess the actual function of the gene or the gene group. Since the upregulated transportation group is quite large, the data should be investigated further to enable a more precise analysis of the genes involved. As an example of possible analysis methods are tools utilizing the Kyoto Encyclopedia of Genes and Genomes (KEGG) databases (Kanehisa & Goto 2000), which could be used to find enriched pathways in which the studied genes are involved in to better evaluate their function in the studied target system (e.g. Ragland et al. 2011).

The only upregulated transportation cluster with a fairly narrow annotation is the juvenile hormone binding protein cluster (JHBP). Juvenile hormone (JH) regulates many important functions in insects (Wyatt & Davey 1996) and most of it is carried to the target tissues by JHBPs (Zalewska et al. 2009). JH has been connected to diapause regulation and usually the characteristic feature of diapause is the deficiency of JH (Flatt et al. 2005). Here, genes in the JHBP cluster are upregulated in the diapausing females. This raises few interesting questions. Does the high expression of putative JHBP in the study samples correspond to a high JH titer? And if yes, could JH behave differently in *D. montana* than in many other diapausing insects? Or could the diapause intensity still be low by the time the females were collected and hence also the JH levels would still be high? Or could the clustered JHBP genes mediate the transport of some other hormone or substrate than JH? Since the JHBP cluster has only 13 genes and there is very little annotation information available for the genes involved, it would be very interesting to use, for example, previously mentioned KEGG analysis to try to get more information about the genes and their connection to the behavior of JH.

The final two cluster groups on mitosis/meiosis and DNA/RNA management connected with the downregulated genes highlights the difference between the two studied phenotypes. These groups are lacking from the upregulated clusters, but together they comprise two thirds of all the significant downregulated clusters. Genes involved in the normal development of non-diapausing females enables them to grow, develop ovaries and reproduce using the vast amount of resources available during the summer. A clear contrast exists for diapausing females, who need to gather large energy reserves in autumn and survive over the long winter period before preparing to produce the next generation in the following spring.

4.4 Top upregulated genes in diapausing females

In the last aim of this study a list of top upregulated contigs was used as the basis to present the ten most upregulated genes in diapausing *D. montana* females. Selected contigs were annotated to corresponding genes using *D. melanogaster* ortholog information, i.e. only genes with a *D. melanogaster* ortholog were used, which should ensure enough annotation information to speculate the reason for the high upregulation for each of these genes in diapausing females.

Diapausing as well as non-diapausing females use environmental cues to follow daily and seasonal changes and act accordingly. As speculated before, the observation of sensory perception genes could indicate a greater need to follow environmental signals in diapause compared to non-diapausing phenotype. Hence, to find two genes (*Obp44a* and *antdh*) connected to olfaction amongst the top ten most upregulated genes in this study is surprising and interesting. During diapause the token stimuli, or the main signal that is used to determine whether to enter diapause or not, helps to maintain the diapause state and affects also the termination phase (Koštál 2006). In *D. montana*, the token stimuli is photoperiod, but also temperature affects the decision (Salminen & Hoikkala 2013).

Consequently, what is the reason for the high expression of the observed olfaction genes in the diapausing *D. montana* females?

Olfaction signals could also be used as pheromones for example for males to sense whether a female fly is in diapausing state or not (Outi Ala-Honkola pers. communication), or to locate, for instance, suitable places where to diapause over the winter or food resources when preparing for a diapause, which is also evident when collecting *D. montana* flies from nature by luring them with malt porridge baits. However, an interesting feature in the results is that there are altogether ten odorant binding protein (obp) genes expressed more in diapausing females and two in non-diapausing samples, but only one odorant receptor gene (*Or92a*) is even present in the data and with very low read counts (data not shown). Odorant binding proteins are suggested to be part of odorant reception system in insects along with odorant receptors (Hekmat-Scafe et al. 2002; Leal 2013). Therefore, it would be interesting to study the function of these genes in diapausing *D. montana* flies more carefully, since there are 51 "obp" genes and 60 "or" genes known in *Drosophila* (St. Pierre et al. 2014). Are these genes indeed taking part in odor recognition and how, or could they even have a role in stress resistance by protecting the olfactory organs as Wang et al (1999) shortly speculate for the *antdh* gene?

Three of the top genes (*Zw, tobi and Mal-A1*) were found to be part of various glucose metabolism activities in the adult digestive systems. Glucose can have many functions in organisms and it could also affect diapause in different ways. Firstly, glucose is an important carbohydrate and energy source, which is stored in insects mainly as glycogen, a glucose polysaccharide. However, feeding is usually arrested during diapause when overwintering insects have little access to food resources and they have to rely on their energy stores to survive (Hahn & Denlinger 2007). Therefore, it is not likely that the observed glucose metabolism genes would function in storing glucose, unless the study flies were still feeding and gathering energy reserves when they were collected. The flies used in this study were grown on top of their food resource making it possible for them to feed until the collections were made. Secondly, glucose levels have been observed to increase in the autumn in those *D. montana* flies preparing for winter, which could improve the cold tolerance of these flies (Vesala et al. 2012a). Most common cryoprotectant is glycerol, but many other polyols and sugars, such as glucose, can also function in the same purpose (Lee 1991). Also, the function of glucose-6-phosphate dehydrogenase enzyme, which is coded by *Zw (or G6PD)* gene in *Drosophila*, could affect the synthesis of cryoprotectant polyols in insect larvae (Storey et al. 1991). Thirdly, overexpression of *Zw* gene has been associated with increased life span in *D. melanogaster* (Legan et al. 2008). *Zw* is a key enzyme in NADPH syntesis where glucose 6-phosphate reacts with NADP+ to produce NADPH, which has been suggested to directly function as an antioxidant (Kirsch & De Groot 2001). However, NADPH functions also as an indirect antioxidant by controlling the generation of glutathione (Wolin et al. 2005), which itself is an important antioxidant preventing oxidative damage, but it also removes xenobiotic and electrophilic compounds from the cell (Forman et al. 2009). Interestingly, one of the top ten genes in this study, *GstE6*, also functions in glutathione metabolism. Gst stands for Glutathione S transferase, which are a vital part of a conjugation reaction to neutralize these reactants (Forman et al. 2009). Since the diapausing *D. montana* individuals supposedly live much longer than pure summer generations, these two genes could also have a role in aging in *D. montana*. Additionally, the GstE6 gene has been observed to be upregulated in an insecticide treatment possibility aiding in insecticide resistance (Alias & Clark 2010).

In addition to glucose metabolism, a gene that functions in fatty acid metabolism was discovered amongst the top ten genes. *Desat2* is a cuticular hydrocarbon that affects male courtship behavior and reproductive isolation (Coyne 1996), but also plays a role in desiccation (Rouault et al 2004), cold and starvation tolerance (Greenberg et al. 2003; but see Coyne & Elwyn 2006). A potentially similar function could be expected for this gene also in diapausing *D. montana* females. Interestingly, the gene was found to be downregulated in cold acclimated *D. virilis* flies, but not in *D. montana* flies (Vesala et al 2012b) while in this study it was heavily upregulated. Furthermore, another stearoyl-CoA 9-desaturase gene, *Desat1*, was just left out from the top most upregulated genes in this study, but in Vesala et al (2012b) it was found to be downregulated in *D. montana*. The samples in Vesala et al. (2012b) were non-diapausing flies taken from cold acclimation treatments. Thus, it would be interesting to see what kind of expression levels these genes or their different transcripts have in cold acclimated diapausing flies and how they affect *D. montana* cold tolerance.

Myosin is most often associated with muscles where it drives muscle contraction by a cyclical interaction with actin. However, even though flight muscles could store glycogen (Hahn & Denlinger 2011), most likely muscle size is decreased during diapause in order to save energetic costs (de Kort 1990). Therefore, it is likely that two of the top ten upregulated genes associated with myosin have some other function in diapausing individuals than increased muscle contraction. One potential cause for their high expression could be a structural role in cytoskeleton. As mentioned before, various actin genes have been found to be upregulated in diapause and also in this study several cytoskeleton, actin and myosin related genes were found to be enriched amongst the upregulated genes. Thus, myosin could have an effect on the cytoskeletal changes observed in diapause through its connection to actin, but whether the two myosin related genes presented here could actually have the effect on cytoskeleton needs further clarification.

The final gene in the upregulated gene list is the least annotated one of the top ten genes. *Pif1B* gene was discovered when it interacted with a *Drosophila* development gene *Eip63E* (previously *L63*) (Rascle et al. 2003). In our study, the *Eip63E* gene has only few low read count contigs, which does not correspond to the expression levels of *Pif1B* that well. Therefore, it is likely that the *Pif1B* gene has also other, yet unknown functions in *Drosophila*. Also, *Pif1B* protein is coded by the same gene as is another protein product, *PiF1A*, which has very similar expression pattern as *Pif1B* (data not shown) and it is the first gene left out from the list of top ten most upregulated genes. An interesting feature in the Flybase expression data for these two genes is that they both have high expression in larvae and adult carcass. However, due to the limited annotation information, more studies are needed to find out what kind of roles these genes might have in diapause.

In conclusion, next-generation RNA sequencing was used to study genes connected with the reproductive diapause of a north adapted fly species *D. montana*. This study is an example of the suitability of such a technology in enabling genetic studies on a species with limited genetic background, but a well-characterized ecological basis. The results presented in this study demonstrate a large genetic difference between the diapause phenotype capable of surviving the long winters and the non-diapausing summer phenotype. This is in accordance with the current view of diapause being an alternative dynamic adaptation and not just an arrest in the normal active summer development.

A good quality and comprehensive transcriptome created in this study enabled the functional annotation of the differentially expressed genes to present interesting gene clusters and genes that have various potential effects on the diapause of *D. montana*. These results can be used in the future to study the many different aspects of diapause including stress resistance, photoperiodism and photoreception, metabolism, endocrinology and aging in more details. In general, next-generation sequencing studies, as the one depicted here, should never be seen as an end product, but a starting point and a basis for many other studies to come.

ACKNOWLEDGEMENTS

I thank my supervisors Maaria Kankare, Darren Parker and Tiina Salminen for their excellent guidance and support during this project. I also thank Laura Vesala and Asta Laiho for their contribution to this work and Anneli Hoikkala for her support and discussions during the project. This work was supported by grants from Centre of Excellence in Evolutionary Research, Suomen Hyönteistieteellinen Seura ry, Suomen Biologian Seura Vanamo ry and Kuopion Luonnon Ystäväin Yhdistys ry, Betty Väänäsen rahasto.

REFERENCES

Abdrakhamanova A., Wang Q.Y., Khokhlova L. & Nick P. 2003. Is Microtubule Disassembly a Trigger for Cold Acclimation? *Plant Cell Physiol.* 44: 676–686.

Adams A.J. 1985. The critical field photoperiod inducing ovarian diapause in the cabbage whitefly, *Aleyrodes proletella* (Homoptera: Aleyrodidae). *Physiol. Entomol.* 10: 243–249.

Alias Z. & Clark A.G. 2010. Adult *Drosophila melanogaster* glutathione S-transferases: Effects of acute treatment with methyl parathion. *Pestic. Biochem. Phys.* 98: 94–98.

Anders S. & Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol.* 11: R106.

Ansorge W.J. 2009. Next-generation DNA sequencing techniques. *New Biotechnol.* 25: 195–203.

Arrese E.L. & Soulages J.L. 2013. Insect fat body: energy, metabolism and regulation. *Annu. Rev. Entomol.* 55: 207–225.

Auer P.L. & Doerge R.W. 2010. Statistical design and analysis of RNA sequencing data. *Genetics* 185: 405–416.

Badisco L., Van Wielendaele P. & Vanden Broeck J. 2013. Eat to reproduce: a key role for the insulin signaling pathway in adult insects. *Front. Physiol*. 4: 202.

Baker D.A. & Russell S. 2009. Gene expression during *Drosophila melanogaster* egg development before and after reproductive diapause. *BMC Genomics* 10: 242.

Beck S.D. 1980. *Insect photoperiodism*, 2nd edition. Academic press, New York and London, pp. 387.

Benjamini Y. & Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statis. Soc.* B 57: 289–300.

Bond D. & Foley E. 2009. A quantitative RNAi screen for JNK modifiers identifies *Pvr* as a novel regulator of *Drosophila* immune signaling. *PLoS Pathog.* 5: e1000655.

Bowen M.F. 1992. Patterns of sugar feeding in diapausing and nondiapausing *Culex pipiens* (Diptera:culicidae) females. *J. Med. Entomol.* 29: 843–849.

Bradshaw W.E. 1976. Geography of photoperiodic response in a diapausing mosquito. *Nature* 262: 384–386.

Bradshaw W.E. & Holzapfel C.M. 2001a. Genetic shift in photoperiodic response correlated with global warming. *P. Natl. Acad. Sci.* USA. 98: 14509–14511.

Bradshaw W.E. & Holzapfel C.M. 2001b. Phenotypic evolution and the genetic architecture underlying photoperiodic time measurement. *J. Insect Physiol.* 47: 809–820.

Bradshaw W.E., Armbruster P.A. & Holzapfel C.M. 1998. Fitness consequences of hibernal diapause in the Pitcher-plant Mosquito, *Wyeomyia smithii*. *Ecology* 79: 1458–1462.

Brun A., Cuany A., Le Mouel T., Berge J. & Amichot M. 1996. Inducibility of the *Drosophila melanogaster* Cytochrome P450 gene, *CYP6A2*, by phenobarbital in insecticide susceptible or resistant strains. *Insect Biochem. Molec.* 26: 697–703.

Buch S., Melcher C., Bauer M., Katzenberger J. & Pankratz M.J. 2008. Opposing effects of dietary protein and sugar regulate a transcriptional target of *Drosophila* insulin-like peptide signaling. *Cell Metab.* 7: 321–332.

Burmester T. 1999. Evolution and function of the insect hexamerins. *Eur. J. Entomol.* 96: 213–225.

Calvert W.H. & Brower L.P. 1986. The location of monarch butterfly *(Danaus plexippus L.)* overwintering colonies in Mexico in relation to topography and climate. *J. Lepid. Soc.* 40: 164–187.

Casneuf T., Van de Peer Y. & Hubert W. 2007. In situ analysis of cross-hybridisation on microarray and the inference of expression correlation. *BMC Bioinformatics* 8: 461.

Chistoserdova L. 2010. Recent progress and new challenges in metagenomics for biotechnology. *Biotechnol. Lett.* 32: 1351–1359.

Conesa A., Götz S., Garcia-Gomez J.M., Terol J., Talon M. & Robles M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21: 3674–3676.

Costa V., Angelini C., De Feis I. & Ciccodicola A. 2010. Uncovering the complexity of transcriptomes with RNA-Seq. *J. Biomed. Biotechnol.* 853916.

Coyne J.A. 1996. Genetics of Differences in Pheromonal Hydrocarbons Between *Drosophila melanogaster* and *D. simulans*. *Genetics* 143: 353–364.

Coyne J.A. & Elwyn S. 2006. Does the desaturase-2 locus in *Drosophila melanogaster* cause adaptation and sexual isolation? *Evolution* 60: 279–291.

Danks H.V. 1987. *Insect dormancy: an ecological perspective*. Biological Survey of Canada (Terrestrial Anthropods), Ottawa, pp. 439.

de Kort C.A.D. 1990. Thirty-five years of diapause research with the Colorado potato beetle. *Entomol. Exp. Appl.* 56: 1-13.

Denlinger D.L. 1986. Dormancy in tropical insects. *Annu. Rev. Entomol.* 31, 239–264.

Denlinger D.L. 2002. Regulation of diapause. *Annu. Rev. Entomol.* 47: 93–122.

Denlinger D.L., Yocum G.D. & Rinehart J.P. 2011. Hormonal control of diapause. In: Gilbert, L.I. (Ed.), *Insect Endocrinology*, Elsevier, Amsterdam. Chapter 10. pp. 430–463.

Dillies M-A., Rau A., Aubert J. et al. 2012. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinform.* 14: 671–683.

Drosophila 12 Genomes Consortium. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 450: 203–218.

Dushay M.S., Åsling B. & Hultmark D. 1996. Origins of immunity: *Relish*, a compound *Rel*-like gene in the antibacterial defense of *Drosophila*. *P. Natl. Acad. Sci. U S A.* 93: 10343–10347.

Ekblom R. & Galindo J. 2011. Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity* 107: 1–15.

Ellers J. & Van Alphen J.J.M. 2002. A trade-off between diapause duration and fitness in female parasitoids. *Ecol Entomol.* 27: 279–284.

Emerson K.J., Bradshaw W.E. & Holzapfel C.M. 2009. Complications of complexity: integrating environmental, genetic and hormonal control of insect diapause. *Trends Genet.* 25: 217–225.

Emery P., So W.V., Kaneko M., Hall J.C. & Rosbash M. 1998. CRY, a *Drosophila* clock and light-regulated cryptochrome, is a major contributor to circadian rhythm resetting and photosensitivity. *Cell* 95: 669–679.

Everett M.V., Grau E.D. & Seeb J.E. 2011. Short reads and non-model species: Exploring the complexities of next generation sequence assembly and SNP discovery in the absence of a reference genome. *Mol. Ecol. Resour.* 11 (Suppl. 1): 93–108.

Field D., Tiwari B., Booth T., Houten S., Swan D., Bertrand N. & Thurston M. 2006. Open software for biologists: from famine to feast. *Nat Biotechnol.* 24: 801–803.

Flatt T., Tu M.P. & Tatar M. 2005. Hormonal pleiotropy and the juvenile hormone regulation of *Drosophila* development and life history. *Bioessays* 27: 999–1010.

Forman H.J., Zhang H. & Rinna A. 2009. Glutathione: overview of its protective roles, measurement, and biosynthesis. *Mol. Aspects Med.* 30: 1–12.

Giannakou M.E. & Partridge L. 2007. Role of insulin-like signalling in *Drosophila* lifespan. *Trends Biochem. Sci.* 32: 180–188.

Goto S.G. & Denlinger D.L. 2002. Short-day and long-day expression patterns of genes involved in the flesh fly clock mechanism: *period, timeless*, *cycle* and *cryptochrome*. *J. Insect Physiol.*48: 803–816.

Greenberg A.J., Moran J.R., Coyne J.A. & Wu C-I. 2003. Ecological adaptation during incipient speciation revealed by precise gene replacement. *Science* 302: 1754–1757.

Guppy M. & Withers P. 1999. Metabolic depression in animals: physiological perspectives and biochemical generalizations. *Biol. Rev.* 74: 1–40.

Hahn D.A. & Denlinger D.L. 2007. Meeting the energetic demands of insect diapause: nutrient storage and utilization. *J. Insect Physiol.* 53: 760–773.

Hahn D.A. & Denlinger D.L. 2011. Energetics of Insect *Diapause*. *Annu. Rev. Entomol.* 56: 103–121.

Hall N. 2013. After the gold rush. *Genome Biol.* 14: 115.

Hamel M., Géri C. & Auger-Rozenberg M.-A. 1998. The effects of 20-hydroxyecdysone on breaking diapause of *Diprion pini*L. (Hym., Diprionidae). Physiol. Entomol. 23: 337–346.

Harbers M. & Carninci P. 2005. Tag-based approaches for transcriptome research and genome annotation. *Nat. Methods* 7: 495–502.

Hekmat-Scafe D.S., Scafe C.R., McKinney A.J. & Tanouye M.A. 2002. Genome-wide analysis of the odorant-binding protein gene family in *Drosophila melanogaster*. Genome Res. 12: 1357–1369.

Herman W.S. & Tatar M. 2001. Juvenile hormone regulation of longevity in the migratory monarch butterfly. *P. Roy. Soc. B-Biol. Sci.* 268: 2509–2514.

Heller M.J. 2002. DNA microarray technology: devices, system, and applications. *Annu. Rev. Biomed. Eng.* 4: 129–153.

Hosack D.A., Dennis G.Jr, Sherman B.T., Lane H.C. & Lempicki R.A. (2003) Identifying biological themes within lists of genes with EASE. *Genome Biol.* 4: R70.

Huang da W., Sherman B.T. & Lempicki R.A. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4: 44–57.

Ishiguro S., Li Y., Nakano K., Tsumuki H. & Goto M. 2007. Seasonal changes in glycerol content and cold hardiness in two ecotypes of the rice stem borer, *Chilo suppressalis*, exposed to the environment in the Shonai district, Japan. *J. Insect Physiol.* 53: 392–397.

Kanehisa M. & Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28: 27–30.

Kankare M., Salminen T., Laiho A., Vesala L. & Hoikkala A. 2010. Changes in gene expression linked with adult reproductive diapause in a northern malt fly species: a candidate gene microarray study. *BMC Ecol.* 10: 3.

Kauranen H., Menegazzi P., Costa R., Helfrich-Förster C., Kankainen A. & Hoikkala A. 2012. Flies in the north: Locomotor behavior and clock neuron organization of *Drosophila montana*. *J. Biol. Rhythm.* 27: 377–387.

Kim M. & Denlinger D. L. 2009. Decrease in expression of beta-tubulin and microtubule abundance in flight muscles during diapause in adults of *Culex pipiens*. *Insect. Mol. Biol.* 18: 295–302.

Kim M., Robich R.M., Rinehart J.P. & Denlinger D.L. 2006. Upregulation of two actin genes and redistribution of actin during diapause and cold stress in the northern house mosquito, *Culex pipiens*. *J. Insect Physiol.* 52: 1226–1233.

Kirsch M. & De Groot H. 2001. NAD(P)H, a directly operating antioxidant? FASEB J. 15: 1569–1574.

Kiselev A. & Subramaniam S. 1994. Activation and regeneration of rhodopsin in the insect visual cycle. *Science* 266: 1369–1373.

Koštál V. 2006. Eco-physiological phases of insect diapause. *J. Insect Physiol.* 52: 113–127.

Koštál V. 2011. Insect photoperiodic calendar and circadian clock: Independence, cooperation, or unity? *J. Insect Physiol.* 57: 538–556.

Lankinen P. 1986. Geographical variation in circadian eclosion rhythm and photoperiodic adult diapause in *Drosophila littoralis*. *J. Comp. Physiol.* A 159: 123–142.

Lankinen P., Tyukmaeva V.I. & Hoikkala A. 2013. Northern *Drosophila montana* flies show variation both within and between cline populations in the critical day length evoking reproductive diapause. *J. Insect Physiol.* 59: 745–751.

Lakovaara S. 1969. Malt as a culture medium for *Drosophila* species. *Dros. Info. Serv.* 44: 128.

Le Goff G., Boundy S., Daborn P.J., Yen J.L., Sofer L., Lind R., Sabourault C., Madi-Ravazzi L. & ffrench-Constant, R.H. 2003. Microarray analysis of cytochrome P450 mediated insecticide resistance in Drosophila. *Insect Biochem. Molec.* 33: 701–708.

Leal W.S. 2013. Odorant Reception in Insects: Roles of Receptors, Binding Proteins, and Degrading Enzymes. *Annu. Rev. Entomol.* 58: 373–391.

Lee R .E. 1991. Principles of insect low temperature tolerance. In: *Insects at Low Temperature* (eds. R. E. Lee and D. L. Denlinger), pp. 17-46. Chapman and Hall, New York.

Legan S.K., Rebrin I., Mockett R.J., Radyuk S.N., Klichko V.I., Sohal R.S. & Orr W.C. 2008. Overexpression of glucose-6-phosphate dehydrogenase extends the life span of *Drosophila melanogaster*. *J. Biol. Chem.* 283: 32492–32499.

Levin J.Z., Berger M.F., Adiconis X., et al. 2009. Targeted next-generation sequencing of a cancer transcriptome enhances detection of sequence variants and novel fusion transcripts. *Genome Biol.* 10: R115.

Li R., Fan W., Tian G., et al. 2010. The sequence and de novo assembly of the giant panda genome. *Nature* 463: 311–317.

Liu N. & Scott J.G. 1995. Genetics of Resistance to Pyrethroid Insecticides in the House Fly, *Musca domestica*. *Pestic. Biochem. Phys.* 52: 116–124.

Lumme J. 1978. Phenology and Photoperiodic diapause in Northern Populations of *Drosophila*. In: *Evolution of Insect Migration and Diapause*. Ed. Dingle H.

MacRae T.H. 2010. Gene expression, metabolic regulation and stress tolerance during diapause. Cell. Mol. Life Sci. 67: 2405-2424.

Menzel A. & Fabian P. 1999. Growing season extended in Europe. *Nature* 397: 659.

Metzker M.L. 2010. Sequencing technologies — the next generation. *Nat. Rev. Genet.* 11: 31–46.

Morales-Hojas R., Reis M., Vieira C.P. & Vieira J. 2011. Resolving the phylogenetic relationships and evolutionary history of the *Drosophila virilis* group using multilocus data. *Mol. Phylogenet. Evol.* 60: 249–258.

Mortazavi A. Williams B.A., McCue K., Schaeffer L. & Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat. Methods* 5: 621–628.

Nekrutenko A. & Taylor J. 2012. Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nat. Rev. Genet.* 13: 667-672.

Neville M. & Goodwin S.F. 2012. Genome-wide approaches to understanding behaviour in *Drosophila melanogaster*. *Brief. Funct. Genomics.* 11: 395-404.

Ning J., Wang M., Li C. & Sun S. 2013. Transcriptome sequencing and de novo analysis of the copepod *Calanus sinicus* using 454 GS FLX. *PLoS One* 8: e63741.

Nishizuka M., Azuma A. & Masaki S. 1998. Diapause response to photoperiod and temperature in *Lepisma saccharina linnaeus* (Thysanura : Lepismatidae). *Ent. Science* 1: 7-14.

Oshlack A., Robinson M.D. & Young M.D. 2010. From RNA-seq reads to differential expression results. *Genome Biol.* 11: 220.

Ozsolack F. & Milos P.M. 2010. RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.* 12: 87–98.

Paaby A.B. & Schmidt P.S. 2009. Dissecting the genetics of longevity in *Drosophila melanogaster*. *Fly* 3: 29–38.

Park P.J. 2009. ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* 10: 669–680.

Pletcher S.D., Macdonald S.J., Marguerie R., Certa U., Stearns S.C., Goldstein D.B. & Partridge, L. 2002. Genome-wide transcript profiles in aging and calorically restricted *Drosophila melanogaster*. *Curr Biol.* 12: 712–723.

Poelchau M.F., Reynolds J.A., Denlinger D.L., Elsik C.G. & Armbruster P.A. 2011. A de novo transcriptome of the Asian tiger mosquito, *Aedes albopictus*, to identify candidate transcripts for diapause preparation. *BMC Genomics* 12: 619.

Pokorna J., Schwarzerova K., Zelenkova S., Petrasek J., Janotova I., Capkova V. & Opatrny Z. 2004. Sites of actin filament initiation and reorganization in cold treated tobacco cells. *Plant Cell Environ.* 27: 641–653.

Pruitt K.D., Tatusova T. & Maglott D.R. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 35(Database issue): D61–65.

Qian X., Ba Y., Zhuang Q. & Zhong G. 2014. RNA-Seq technology and its application in fish transcriptomics. *OMICS* 18: 98-110.

Qiu F., Brendel S., Cunha P.M., Astola N., Song B., Furlong E.E., Leonard K.R. & Bullard B. 2005. Myofilin, a protein in the thick filaments of insect muscle. *J. Cell Sci.* 118(Pt 7): 1527–1536.

Ragland G.J., Egan S.P., Feder J.L., Berlocher S.H. & Hahn D.A. 2011. Developmental trajectories of gene expression reveal candidates for diapause termination: a key life-history transition in the apple maggot fly *Rhagoletis pomonella*. *J. Exp. Biol.* 214: 3948–3959.

Rascle A., Stowers R.S., Garza D., Lepesant J.A. & Hogness D.S. 2003. *L63*, the *Drosophila* PFTAIRE, interacts with two novel proteins unrelated to cyclins. *Mech. Develop.* 120: 617–628.

Readio J., Chen M.H. & Meola R. 1999. Juvenile hormone biosynthesis in diapausing and nondiapausing *Culex pipiens* (Diptera: Culicidae). *J. Med. Entomol.* 36: 355–360.

Richard D.S., Watkins N.L., Serafin R.B. & Gilbert L.I. 1998. Ecdysteroids regulate yolk protein uptake by *Drosophila melanogaster* oocytes. *J. Insect Physiol.* 44: 637–644.

Rinehart J.P., Yocum G.D. & Denlinger D.L. 2000. Developmental upregulation of inducible *hsp70* transcripts, but not the cognate form, during pupal diapause in the flesh fly, *Sarcophaga crassipalpis*. *Insect Biochem. Molec.* 30: 515–521.

Rinehart J.P., Li A,. Yocum G.D., Robich R.M., Hayward S.A.L. & Denlinger D.L. 2007. Up-regulation of heat shock proteins is essentail for cold survival during insect diapause. *P. Natl. Acad. Sci. USA.* 104: 11130–11137.

Robich R.M., Rinehart J.P., Kitchen L.J. & Denlinger D.L. 2007. Diapause-specific gene expression in the northern house mosquito, *Culex pipiens* L., identified by suppressive subtractive hybridization. J. *Insect Physiol.* 53: 235–245.

Rouault J.-D., Marican C., Wicker-Thomas C. & Jallon J.-M. 2004. Relations between cuticular hydrocarbon (HC) polymorphism, resistance against desiccation and breeding temperature; a model for HC evolution in *D. melanogaster* and *D. simulans*. *Genetica* 120: 195–212.

Sanger F., Nicklen S. & Coulson A.R. 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U S A. 74: 5463–5467.

Salminen T.S. & Hoikkala A. 2013. Effect of temperature on the duration of sensitive period and on the number of photoperiodic cycles required for the induction of reproductive diapause in *Drosophila montana*. *J. Insect Physiol.* 59: 450–457.

Salminen T. S., Vesala L. & Hoikkala A. 2012. Photoperiodic regulation of life-history traits before and after eclosion: Egg-to-adult development time, juvenile body mass and reproductive diapause in *Drosophila montana*. *J. Insect Physiol.* 58: 1541–1547.

Salminen T.S., Vesala L., Laiho A., Merisalo M., Hoikkala A. & Kankare M. Seasonal phenotypic plasticity and gene expression kinetics in northern *Drosophila virilis* group species overwintering in reproductive diapause. Submitted manuscript.

Sasson A. & Michael T.P. 2010. Filtering error from SOLiD output. *Bioinformatics* 26: 849–850.

Saunders D.S. 2000. Larval diapause duration and fat metabolism in three geographical strains of the blow fly, Calliphora vicina. *J. Insect Physiol.* 46: 509–517.

Saunders D.S. 2002. In: Steel, C.G.H., Vafopoulou, X., Lewis, R.D. (Eds.), *Insect Clocks*, 3rd ed. Elsevier Science, Amsterdam, pp. 560.

Saunders D.S., Henrich V.C. & Gilbert L.I. 1989. Induction of diapause in *Drosophila melanogaster*: photoperiodic regulation and the impact of arrhythmic clock mutations on time measurement. *P. Natl. Acad. Sci. USA.* 86: 3748–3752.

Saunders D.S. & Gilbert L.I. 1990. Regulation of ovarian diapause in *Drosophila melanogaster* by photoperiod and moderately low temperature. *J. Insect Physiol.* 36: 195–200

Saunders D.S., Richard D.S., Applebaum S.W., Ma M. & Gilbert L.I. 1990. Photoperiodic diapause in *Drosophila melanogaster* involves a block to the juvenile hormone regulation of ovarian maturation. *Gen. Comp. Endocr.* 79: 174–184.

Sim C. & Denlinger D.L. 2008. Insulin signaling and *FOXO* regulate the overwintering diapause of the mosquito *Culex pipiens*. *P. Natl. Acad. Sci. U S A.* 105: 6777–6781.

Schena M., Shalon D., Davis R.W. & Brown P.O. 1995. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science* 270: 467–470.

Schmidt P.S. & Paaby A.B. 2008. Reproductive diapause and life-history clines in North American populations of *Drosophila melanogaster*. *Evolution* 62: 1204–1215.

Schmidt P.S., Paaby A.B. & Heschel M.S. 2005. Genetic Variance for Diapause Expression and Associated Life Histories in *Drosophila melanogaster*. *Evolution* 59: 2616–2625.

Schooneveld H., Sanchez A.O. & de Wilde J. 1977. Juvenile hormone-induced break and termination of diapause in the Colorado potato beetle. *J. Insect Physiol.* 23: 689–696.

Shen W.L., Kwon Y., Adegbola A.A., Luo J., Chess A. & Montell C. 2011. Function of Rhodopsin in temperature discrimination in *Drosophila*. *Science* 331: 1333–1336.

Shendure J. & Ji H. 2008. Next-generation DNA sequencing. *Nat Biotechnol* 26: 1135–1145.

Shendure J., Porrecal G.J., Reppas N.B., Lin X., McCutcheon J.P., Rosenbaum A.M., Wang M.D., Zhang K., Mitra R.D. & Church G.M. 2005. Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome. *Science* 309: 1728–1732.

St. Pierre S.E., Ponting L., Stefancsik R., McQuilton P. & the FlyBase Consortium. 2014. FlyBase 102 - advanced approaches to interrogating FlyBase. *Nucleic Acids Res.* 42: D780-D788.

Stanewsky R., Kaneko M., Emery P., Beretta B., Wager-Smith K., Kay S.A., Rosbash M. & Hall J.C. 1998. The *cry(b)* mutation identifies cryptochrome as a circadian photoreceptor in *Drosophila*. *Cell* 95: 681–692.

Storey K.B., Keefe D., Kourtz L. & Storey J.M. 1991. Glucose-6-phosphate dehydrogenase in cold hardy insects: Kinetic properties, freezing stabilization, and control of hexose monophosphate shunt activity. *Insect Biochem.* 21: 157–164.

Takahashi A., Tsaur S-C., Coyne J.A. & Wu C-I. 2001. The nucleotide changes governing cuticular hydrocarbon variation and their evolution in *Drosophila melanogaster*. *P. Natl. Acad. Sci. U S A.* 98: 3920–3925.

Tamura K., Subramanian S. & Kumar S. 2004. Temporal Patterns of Fruit Fly (*Drosophila*) Evolution Revealed by Mutation Clocks. *Mol. Biol. Evol.* 21: 36–44.

Tatar M., Chien S.A. & Priest N.K. 2001. Negligible senescence during reproductive dormancy in *Drosophila melanogaster*. *Am. Nat.* 158: 248–258.

Tauber M.J., Tauber C.A. & Masaki S. 1986. *Seasonal adaptations of insects*. Oxford University Press, Oxford, pp. 426.

The Gene Ontology Consortium. 2000. Gene Ontology: tool for the unification of biology. *Nat. Genet.* 25: 25–29.

Throckmorton L. 1982. The virilis species group. *The Genetics and Biology of Drosophila.* Vol. 3b, ed. Ashburner M., Carson H. & Thompson J. Academic Press, London.

Tijet N., Helvig C. & Feyereisen R. 2001. The cytochrome P450 gene superfamily in *Drosophila* melanogaster: Annotation, intron-exon organization and phylogeny. *Gene* 262: 189–198.

Tyukmaeva V.I., Salminen T.S., Kankare M., Knott K.E. & Hoikkala A. 2011. Adaptation to a seasonally varying environment: a strong latitudinal cline in reproductive diapause combined with high gene flow in *Drosophila montana. Ecol. Evol.* 1: 160–168.

Vesala L. & Hoikkala A. 2011. Effects of photoperiodically induced reproductive diapause and cold hardening on the cold tolerance of *Drosophila montana*. *J. Insect Physiol.* 57: 46–51.

Vesala L., Salminen T.S., Koštál V., Zahradníčková H. & Hoikkala A. 2012a. Myo-inositol as a main metabolite in overwintering flies: seasonal metabolomic profiles and cold stress tolerance in a northern drosophilid fly. *J. Exp. Biol.* 215(Pt 16): 2891–2817.

Vesala L., Salminen T.S., Laiho A., Hoikkala A. & Kankare M. 2012b. Cold tolerance and cold-induced modulation of gene expression in two *Drosophila virilis* group species with different distributions. *Insect. Mol. Biol.* 21: 107–118.

Wang Q., Hasan G. & Pikielny C.W. 1999. Preferential expression of biotransformation enzymes in the olfactory organs of *Drosophila melanogaster*, the antennae. *J. Biol. Chem.* 274: 10309–10315.

Wang Z., Gerstein M. & Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10: 57–63.

Wicker-Thomas C. 2007. Pheromonal communication involved in courtship behavior in Diptera. *J. Insect Physiol.* 53: 1089–1100.

Williams K.D. & Sokolowski M.B. 1993. Diapause in *Drosophila melanogaster* females: a genetic analysis. *Heredity* 71: 312–317.

Wolin M.S., Ahmad M. & Gupte S.A. 2005. Oxidant and redox signaling in vascular oxygen sensing mechanisms: basic concepts, current controversies, and potential importance of cytosolic NADPH. *Am. J. Physiol.-Lung C.* 289: L159–173.

Wyatt G.R. & Davey K.G. 1996. Cellular and molecular actions of juvenile hormone. II. Roles of juvenile hormones in adult insects. *Adv. Insect Physiol.* 26: 1–155.

Xiao H-J., Mou F-C., Zhu X-F. & Xue F-S. 2010. Diapause induction, maintenance and termination in the rice stem borer *Chilo suppressalis* (Walker). *J. Insect Physiol.* 56: 1558–1564.

Yamashita O. 1996. Diapause hormone of the silkworm, *Bombyx mori*: structure, gene expression and function. *J. Insect Physiol.* 42: 669–679.

Yang P., Tanaka H., Kuwano E. & Suzuki K. 2008. A novel cytochrome P450 gene (*CYP4G25*) of the silkmoth *Antheraea yamamai*: Cloning and expression pattern in pharate first instar larvae in relation to diapause. *J. Insect Physiol.* 54: 636–643.

Yocum G.D., Joplin K.H. & Denlinger D.L. 1998. Upregulation of a 23 kDa small heat shock protein transcript during pupal diapause in the flesh fly, *Sarcophaga crassipalpis*. *Insect Biochem. Molec.* 28: 677–682.

Zalewska M., Kochman A., Estève J.P., Lopez F., Chaoui K., Susini C., Ozyhar A. & Kochman M. 2009. Juvenile hormone binding protein traffic - Interaction with ATP synthase and lipid transfer proteins. *Biochim. Biophys. Acta* 1788: 1695–705.

Zhou G. & Miesfeld R.L. 2009. Energy metabolism during diapause in *Culex pipiens* mosquitoes. *J. Insect Physiol.* 55: 40–46.

APPENDIXES

Appendix 1. Table of 16 *D. montana* RNA samples used in assembling the transcriptome. The first six samples were from diapausing (D) and non-diapausing flies (ND) from three different isofemale strains (3OL8, 175OJ8 and 265OJ8). Flies for these samples were reared in population specific critical day length (CDL) of 18.5 hours of light and 5.5 hours of dark (LD 18.5:5.5) and were the main focus of this thesis. The rest of the diapausing flies were reared in LD 16:8 and non-diapausing flies in LD 22:2. Flies in the last four non-diapausing samples (three female and one male sample, No. 13-15) received acclimation treatment and were reared in LD 22:2. Each sample comprised of ten individual flies.

No.	Sample ID	Phenotype	Strain	Light:Dark cycle	Sex
1	1_CDL_D	Diapause	3OL8	18.5:5.5	Female
2	2_CDL_D	Diapause	175OJ8	18.5:5.5	Female
3	3_CDL_D	Diapause	265OJ8	18.5:5.5	Female
4	4_CDL_ND	Non_diapause	3OL8	18.5:5.5	Female
5	5_CDL_ND	Non_diapause	175OJ8	18.5:5.5	Female
6	6_CDL_ND	Non_diapause	265OJ8	18.5:5.5	Female
7	7_D	Diapause	175OJ8	16:8	Female
8	8_D	Diapause	175OJ8	16:8	Female
9	9_D	Diapause	175OJ8	16:8	Female
10	10_ND	Non_diapause	175OJ8	22:2	Female
11	11_ND	Non_diapause	175OJ8	22:2	Female
12	12_ND	Non_diapause	175OJ8	22:2	Female
13	13_ACC	Acclimation	175OJ8	22:2	Female
14	14_ACC	Acclimation	175OJ8	22:2	Female
15	15_ACC	Acclimation	175OJ8	22:2	Female
16	16_ACC	Acclimation	175OJ8	22:2	Male

Appendix 2. Table of *Drosophila* genome versions for 12 sequenced *Drosophila* species (Drosophila 12 Genomes Consortium 2007). Version refers to the released versions of each of the genomes downloaded from Flybase on May 5th, 2012.

Species	Version
D. melanogaster	r5.44
D. simulans	r1.3
D. sechellia	r1.3
D. yakuba	r1.3
D. erecta	r1.3
D. ananassae	r1.3
D. pseudoobscura	r1.04
D. persimilis	r1.3
D. willistoni	r1.3
D. virilis	r1.2
D. mojavensis	r1.3
D. grimshawi	r1.3

Appendix 3. Table of upregulated gene annotation clusters. Enrichment score is the level of enrichment for the cluster. Annotation term and ID refer to annotation terms in different databases, which are gene ontology (GO), interpro (IPR) and SMART (SM). NA refers to SP-PIR Keywords for which there were no term IDs. Other listed information includes the amount of genes in the cluster, broader cluster annotation group and two examples of *D. melanogaster* genes (symbol) belonging to the cluster.

Cluster	Enrichment score	Cluster annotation term	Annotation term ID	No. of genes	Annotation group	D. melanogaster genes (e.g.)
1	11,3	Oxidation reduction	GO:0055114	125	Metabolism	*antdh, zw*
2	7,1	CHK kinase-like	IPR015897	21	Metabolism	*CG1561, CG13360*
3	6,0	Spectrin repeat	IPR002017	39	Cytoskeleton	*Mhc, zormin*
4	5,6	Cytochrome P450	IPR002401	51	Metabolism	*Cyp12a5, Cyp6a2*
5	4,8	Peptidase/chymotrypsin	IPR001254	111	Metabolism	*amon, gammaTry*
6	4,8	Ion transport	GO:0006811	236	Transport	*Obp44a, wtrw*
7	4,7	cation transport	GO:0006812	126	Transport	*CG6484, Tret1-1*
8	4,6	Carbohydrate metabolic process	GO:0005975	150	Metabolism	*tobi, Mal-A1*
9	4,4	Immunoglobulin	IPR013098	71	Cytoskeleton	*Unc-89, Strn-Mlck*
10	4,2	JHBP	SM00700	13	Transport	*CG5945, to*
11	3,8	C2 membrane targeting protein	IPR018029	25	Transport	*inaC, Caps*
12	3,1	Peptidase M13	IPR018497	23	Metabolism	*Pex1, GstE6*
13	2,9	Glycoside hydrolase	IPR013781	18	Metabolism	*Mal-A1, Gal*
14	2,5	Alkaline phosphatase	IPR001952	10	Metabolism	*Sulf1, Aph-4*
15	2,4	Munc13 homology 1	IPR014770	9	Transport	*Unc-13, CG34349*
16	2,4	Fatty acid biosynthetic process	GO:0006633	44	Metabolism	*norpA, desat2*
17	2,2	Sodium ion transport	GO:0006814	18	Transport	*Nhe2, NAAT1*
18	1,9	Cellular respiration	GO:0045333	40	Metabolism	*Punch, pug*
19	1,8	Carboxylic acid metabolic process	GO:0019752	79	Metabolism	*Gclc, ACC*
20	1,8	Nucleotide metabolic process	GO:0009117	235	Metabolism	*ATP citrate lyase, desat1*
21	1,8	EF-HAND	IPR018247	64	Transport	*Nhe2, para*
22	1,7	Sugar transporter	IPR005829	18	Transport	*CG6484, Tret1-1*
23	1,7	Leucine-rich repeat	IPR001611	27	Metabolism	*ltl, chp*
24	1,7	Carbohydrate kinase	IPR018485	16	Metabolism	*Gpo-1, Inos*
25	1,6	Adenylyl cyclase	IPR018297	27	Metabolism	*Adgf-D, rut*
26	1,6	Heat shock protein Hsp20	IPR002068	14	Response to stimulus	*Hsp22, ninaE*
27	1,5	Sensory perception	GO:0007600	32	Response to stimulus	*Rh6, shep*
28	1,5	Actin	IPR004001	13	Cytoskeleton	*alpha Spectrin, Actin 57B*
29	1,4	Na+ channel	IPR001696	12	Transport	*Na channel protein 60E, Irk2*
30	1,4	Basic leucine zipper	IPR011700	9	Transport	*kay, Xrp1*
31	1,4	GPCR, rhodopsin-like superfamily	IPR017452	53	Response to stimulus	*seven up, Notch*

Appendix 4. Table of downregulated gene annotation clusters. Enrichment score is the level of enrichment for the cluster. Annotation term and ID refer to annotation terms in different databases, which are gene ontology (GO), interpro (IPR) and SMART (SM). NA refers to SP-PIR Keywords for which there were no term IDs. Other listed information includes the amount of genes in the cluster, broader cluster annotation group and two examples of *D. melanogaster* genes (symbol) belonging to the cluster.

Cluster	Enrichment score	Cluster annotation term	Annotation term ID	No. of genes	Annotation group	D. melanogaster genes (e.g.)
1	12,2	Zinc finger	IPR015880	154	DNA/RNA	*lola, fs(1)Ya*
2	11,2	DNA repair	GO:0006281	110	Metabolism	*Ssl1, Pxt*
3	11,1	Cellular metabolic process	GO:0044237	726	Metabolism	*Mcm3, RnrL*
4	8,7	Chromosome organization	GO:0051276	131	Transport	*Semaphorin-2A, like-AP180*
5	7,7	RNA recognition motif	IPR000504	68	DNA/RNA	*tsunagi, MAN1*
6	7,3	DEAD-like helicase	IPR014001	105	DNA/RNA	*Iswi, Chd1*
7	6,6	DNA replication initiation	GO:0006270	29	DNA/RNA	*Mcm3, Pole2*
8	5,9	PHD	SM00249	35	DNA/RNA	*Mi-2, nej*
9	5,7	Ribosome biogenesis	GO:0042254	54	Metabolism	*Gem2, Smn*
10	5,4	WD40 repeat	IPR001680	85	DNA/RNA	*groucho, tomosyn*
11	5,1	Nuclear transport	GO:0051169	240	Transport	*Nup133, Dmel_CG6446*
12	3,8	Chromatin assembly	GO:0006333	21	DNA/RNA	*enok, HP1b*
13	2,9	Tudor domain	IPR002999	17	DNA/RNA	*MBD-R2, Pcl*
14	2,7	Transcription factor jumonji	IPR013129	16	DNA/RNA	*lid, Bap170*
15	2,7	mRNA processing	GO:0006397	19	DNA/RNA	*Gem2, U2af50*
16	2,6	Macromolecule catabolic process	GO:0044265	92	Metabolism	*Drep-4, Ubpy*
17	2,5	Bromodomain	IPR001487	16	DNA/RNA	*BRWD3, Br140*
18	2,1	DNA packaging	GO:0006323	13	DNA/RNA	*Nap1, cid*
19	2,1	Chaperonin Cpn60/TCP-1	IPR002423	14	Metabolism	*Tcp-1zeta, Droj2*
20	1,8	Transcription, DNA-dependent	GO:0006351	20	DNA/RNA	*skuld, Tbp*
21	1,8	Meiotic cell cycle	GO:0051321	22	Mitosis/meiosis	*CAP-D2, sallimus*
22	1,8	Cell cycle	GO:0007049	38	Mitosis/meiosis	*barr, Nipped-B*
23	1,7	Tetratricopeptide	IPR001440	30	Metabolism	*Utx, Nup358*
24	1,7	mRNA splicing	NA	9	DNA/RNA	*mael, Ars2*
25	1,7	Cyclin	IPR004367	22	Mitosis/meiosis	*hu li tai shao, encore*
26	1,6	SET	IPR001214	20	DNA/RNA	*egg, G9a*
27	1,6	DNA-directed DNA polymerase	IPR006134	12	DNA/RNA	*RpII215, DNApol-alpha180*
28	1,5	HAS subgroup	IPR013999	4	DNA/RNA	*brm, dom*
29	1,5	Regulation of metabolic process	GO:0019222	238	Metabolism	*Septin-1, staufen*
30	1,4	Pre-SET zinc-binding region	IPR007728	8	DNA/RNA	*MBD-R2, Art4*
31	1,4	SANT, DNA-binding	IPR001005	14	DNA/RNA	*MTA1-like, Myb*
32	1,3	Zinc-finger	NA	118	DNA/RNA	*modifier of mdg4, Sgf11*
33	1,3	Armadillo	IPR000225	9	Transport	*Kap-alpha1,Apc*