Reflection (598 words)

The proposed project aim was to investigate the genetic basis of migratory behaviour in butterflies as a deeper understanding of migration can be used to create strategies for mitigating the consequences of climate change. This aligned with the modules I had chosen in Year 2 and piqued my interest as a keen environmentalist.

Upon my arrival, we did preliminary dissections in the lab in order to determine which parts of each developmental stage of *V. cardui,* was most appropriate for RNA extraction which was exciting, as planning an investigation from the very beginning was novel to me. I enjoyed the privilege of designing aspects of the research myself, being involved in conversations about ordering equipment and making compromises when our aims didn't align with the possible methods. This was also an enlightening experience to the expense of professional research.

After initial demonstrations from PhD students, I conducted the majority of my lab work independently. This was exacerbated by restrictions as limited number of staff in the lab meant there was reduced opportunity to consult others when faced with an obstacle. I feel I responded appropriately to issues and contacted my supervisors where appropriate but overall gained confidence in my own lab skills and problem solving. I extracted RNA following the QIAGEN RNeasy Mini Kit and evaluated the quality of these results using NanoDrop, Qubit, Bioanalyzer and gel electrophoresis. RNA sequencing has experienced delays worldwide so it was disappointing but not surprising that the RNA-seq data was not available to me before my placement ended.

I dedicated time to familiarising myself with using UPPMAX, a tool for bioinformatic analysis of large data sets, and continued research into migration and its genetic basis. This involved learning about quality control and quality filtering of RNA reads, transcriptome assembly, read mapping, expression quantification and gene annotation. This was completed at home due to safety restrictions but I had weekly meetings with my research group where there was opportunity, and we were encouraged, to discuss our work. I really struggled to grasp these techniques at first as I wasn't accustomed to using code but I soon felt more comfortable.

Once weekly the team would also host a Journal Club, where we would explore a suggested paper. This was often relevant to one or more project within the research group and would offer a chance to ask questions to develop understanding, critique the methods/approach of a study and explore potential links to our own work. I particularly struggled to engage with these sessions as I found many of the papers difficult to comprehend (especially if they were unrelated to my field) and didn't want to slow the flow of ideas by questioning each point. This was a personal issue as my colleagues and supervisors always created a welcoming environment to ask questions. I felt I could speak to them openly about not understanding a concept without fear of judgement.

In general the group I joined, led by Niclas Backström, was the most friendly and accepting academic group I have ever been a part of. I was respected despite being the least advanced in my studies and actively encouraged to pitch ideas/comments for work produced by people in much senior academic positions. I was also valued as a native English speaker in an international community so enjoyed having a skill I could easily use to help others. I especially thrived in hosting social activities for the group including weekly quizzes and a bonfire night celebration to teach others about English traditions. I would recommend Uppsala University and Sweden in general to any prospective PTY students.

A differential gene expression study using *Vanessa cardui* as a model species for migratory behaviour

Elenia Parkes

Biochemistry BSc with Professional Training Year

Vocational Supervisor: Niclas Backström

Evolutionary Biology, Department of Ecology and Genetics, Uppsala University

5879 words (excluding Abstract + Acknowledgments)

Abstract

Populations of the migratory butterfly, the painted lady (*Vanessa cardui*) can be spotted in most continents, from Sweden to South Africa, on their multi-generational annual migration. Migration can be an adaptive response to changes in habitat and host plant availability. *Vanessa cardui* is the most widely distributed migratory butterfly and populations have displayed variation in their behaviour correlating to environmental changes, making it a model organism for studying factors affecting migration. However, some populations of the painted lady, mostly located in tropical regions where the climate allows for constant availability of host plants, rarely migrate which indicates food availability could be an important predictor of migratory behaviour. The aim of this project is to use varied resource allocation and population density to investigate the genetic basis of migratory behaviour. By comparing gene expression profiles of adult females in treatments that encourage migration (high density [HD], limited resources [LI]) or reproduction (low density [LD], *ad libitium* [AL]) we can explore the effects of environmental factors on migration. The impacts across developmental stages were studied with a set-up including multiple biological replicates of three treatment groups (HDAL, HDLI & LDAL). RNA samples were extracted from the head and abdomen of individuals from each treatment at four developmental stages (Instar III larva, Instar V larva, Pupa [male and female] and Adult [male and female]) in each treatment group using the Qiagen RNeasy Mini Kit and samples were sequenced using Illumina paired-end sequencing.

It is predicted that there will be a difference between transcriptome profiles of butterflies in the treatments with *ad libitum* resources and those with limited resources. Results should reveal potential correlations between population density/host plant availability and expression profiles of specific functional categories. This information can be used to infer the specific pathways underlying induction of migration or investment in reproduction. Timing of energy and resource allocation is key in migratory species, so we speculate that a large fraction of differentially expressed genes will be involved in metabolic processes.

1 | Introduction:

1.1 | Background

Dispersal and migration are both key factors of an insect species' role in an ecosystem, as a pollinator of important crops (TEEB 2010), as a predatory/parasitic species and carriers of disease (Altizer et al., 2011). Unlike dispersal, which can be arbitrary or influenced by unplanned events, migration appears to be an evolutionary response to environmental changes. The painted lady butterfly, *Vanessa cardui,* has long been used as a model for migratory behaviour due to its long migration route and abundance in varied climates across multiple continents (Talavera et al., 2018). Some popular study systems for

migratory insect behaviour include the monarch butterfly (*Danaus plexippus*) (Zhan et al., 2011) and globe wonderer dragonfly *(Pantala flavescens)* (Hobson et al., 2012). Despite study of these organisms, the genetic basis of migration and what specific genes are responsible for the underlying mechanisms, are still poorly understood. In addition, migratory behaviour has evolved many times across insects in general, and the order of Lepidoptera in particular. This indicates that the specific genetic underpinnings of particular migratory strategies in different lineages, may vary. Hence, to draw general conclusions about functional gene categories and pathways involved in migratory behaviour, independent investigations involving different migratory species have to be made.

In August 2021, the International Panel on Climate Change (IPCC) published a damning report highlighting how severe the climate crisis is. The effects of capitalist human behaviours (including deforestation, demand for infrastructure and overwhelming greenhouse emissions) has consequences that can be seen across all life forms, as the delicate interspecific interactions that ecosystems rely on are being disrupted (IPCC 2021). Apart from the reduction in biodiversity, climate change causes shift in the phenology and movement of animals and, in particular, can affect their migration patterns (Seebacher and Post 2015). Migrations are often characterized by seasonal changes to escape unfavourable environmental conditions but can vary largely in length and distance between species (Cresswell et al. 2011, Merlin and Liedvogel 2019). The time and length of migration is affected by temperature and humidity (Kristensen et al. 2015) meaning that for a specific study system, these characteristic are not fixed but variable, and that a deeper understanding of the mechanisms that determine the migratory response, can prove important to predict and mitigate the impacts of climate change. Migratory patterns are likely to be impacted from climate change as once stable habitats become disrupted, which could have devastating effects on pollination and pest control of agricultural crops and predator/prey interspecific relationships in ecosystems where migratory insects can be found.

Climate change can affect species which already show migratory strategies. This has been seen in multiple animal systems already. Pacific pink and sockeye salmon faced an increased risk of upstream migration failure when held in water at a higher temperature, and experienced higher mortality than those kept in cooler water (Eliason et al. 2011). Harsh winters and warm summers drive the phenology and extent of migration in aphids, a common agricultural pest which could pose a threat to food security as the routes and extent of migration become unpredictable (Bell et al., 2015). Global warming studied over 27 years showed the migration range of the cosmopolitan cutworm (*Agrotis ipsilon*), which feeds on a variety of vegetables and many important grains, increased by 708 km (Zeng et al. 2020). In all cases, the trigger of the shifts in migratory trends can be attributed to changes in habitat as a result of climate change.

The displacement of organisms toward the Poles is a repeated pattern and is estimated to occur at a median rate of 17 km per decade in terrestrial species (Chen et al. 2011) and 72 km per decade in marine ones (Poloczanska et al. 2013). Additionally, this rate has been gradually increasing during the last decades (Lowing & Polly 2011) and is highly likely to keep increasing and affect a greater number of species in the coming years (Pecl et al. 2017). Temporal shifts in one species of a food chain also has knock-on impacts on the ecosystem and can negatively impact other organisms interacting with that species. Wild bee pollination is responsible for 85% of food crops (Zattara and Aizen 2021) and the recent drastic decline in populations has posed real and immediate threat to food security. Human populations will not be the only organisms impacted by changes in pollinator population levels due to altered migratory patterns of insects.

Therefore, further studies are essential to be able to detect and understand the onset of changes in migration patterns of different species and apprehend how the situation may change in the future. Assisted migration for non-migratory species has been proposed and applied as a method to maintain biodiversity in systems whose habitats are severely impacted by climate change (Butt et al. 2020); understanding the underlying mechanisms for migration is key to this approach. A more thorough characterization and deeper understanding of the genetic and molecular regulation of the migratory response in organisms with an already established strategy (either with total or partial migration, or populations with differing migration behaviours), is essential for making more accurate predictions. This goes beyond the influence of ecological factors and requires gaining comprehension of the genetic and metabolic mechanisms different organisms use to determine their migratory strategy.

1.2 | Genetics of migration:

Understanding migration has been a challenge as there are multiple factors that can potentially influence the ‘decision’ to migrate. Some research has been performed to elucidate the environmental influences on migratory behaviour, but the understanding of the genetic underpinning of migration is still in its infancy. Temperature, light quality and food availability of a population’s habitat, in addition to population density, are all factors that impact whether a population will migrate to or from a specific location (Taylor and Norris 2007).

The North American Monarch butterfly (*Danaus plexippus*), one of the most well-studied migratory insects, responds to changes in photoperiod and temperature to trigger their seasonal migration (Merlin and Liedvogel 2019). A single migration cycle occurs over multiple generations (whose individuals will therefore have similar genetic make-up) indicating the decision to initiate migration cannot be entirely genetic: responses to external environmental stimuli must have an impact. Re-migrants, so called because they travel in the opposite orientation to the individuals that originally migrated earlier in the cycle, locate their predecessors original breeding sites with extreme accuracy,

suggesting epigenetic control of migration rather than genetic influence alone (Merlin and Liedvogel 2019). The Monarch butterfly is one of the best studied migratory insects but a recent extensive study showed no genomic difference between the eastern monarchs, who undergo a massive migration from east North America to Oyamel forest in Mexico, to spend the winter months, and western monarchs, which have a distribution range close to the Rocky Mountains and perform a much shorter annual migration circuit along the Pacific coast. These two populations were previously thought to be divergent and distinct species, but the lack of genomic differentiation suggests that the difference in migratory behaviour is not encoded in the genome, but rather an effect of different epigenetic modifications and/or gene expression cascades triggered by environmental cues (Talla et al., 2020).

In 2014, a spearheading genomic study of 101 Monarch individuals, representing both migratory and sedentary populations, was aimed at investigating genes under positive selection in migratory populations. The authors found a correlation between migratory behaviour and positive selection on genes associated with flight muscles (Zhan et al., 2014). Such an association is rather expected, since long-distance migration involves a much more active lifestyle. However, this difference in selection pressure does not reveal the specific mechanisms triggering the migratory response/behaviour. Hence, comparative genomics studies must be complemented by in-depth investigations of epigenetic and/or gene expression differences in individuals that are predisposed to migrate and compare to individuals who are predisposed to reproduce. Here, we try to investigate such differences in another migratory butterfly species, the painted lady (*Vanessa cardui*) to get information about different gene expression profiles in different treatments groups.

In addition, although both comparative genomics studies (e.g. Zhan et al., 2014) and in-depth epigenetic and gene expression profiling could reveal differences between migratory and sedentary species/populations, such investigations have to be complemented with functional studies (e.g. RNAi-editing and/or CrisprCas9-modification) to verify the causal relationship between genetic specificities and migratory behaviour.

1.3| *Vanessa cardui* as a model species:

Lepidoptera (butterflies and moths) represent a perfect system to be used as a case study in a multitude of projects in evolutionary biology, given the wide array of habitats they occupy, and adaptations they display. Butterflies and moths have also long been tractable study organisms for naturalists all over the world and detailed knowledge about their life-history and ecology has been gathered (Boggs et al., 2003). Among these adaptations, migration is one that has evolved repeatedly across the butterfly tree of life (Wahlberg & Rubinoff, 2011). Sometimes these massive movements of individuals to new habitats happen in a regular interval (e.g., one full migratory cycle per year) in a never-ending pursuit of suitable habitats and greater food availability (Alerstam et al., 2003), while in

other cases migration appears as a temporary response to the lack of resources in the environment the individuals are inhabiting compared to other available niches elsewhere (Stefanescu et al.,2016).

Vanessa sp. represents a taxon with a high variability in migratory strategies, with variation between closely related species and even populations within the same species. The low genetic differentiation between the different species, highlights how important gene expression profile studies with regards to external environment can be, as individuals with apparently highly similar genetic make-ups behave in different ways. Specifically, the painted lady (*V. cardui*) presents a majority of populations with a spectacular migratory behaviour, and a reduced number of populations that are virtually sedentary and remain constantly in the same locations (most of them concentrated in tropical islands where the homogeneous meteorological conditions make it possible to stay permanently).

Vanessa cardui ranks among the most cosmopolitan of all butterflies in the world. Its distribution covers most continents, with a reduced presence in South America and absence in Antarctica (Savela 1999). In the northern parts of the distribution range, winter temperatures are too low to guarantee the individuals' survival and in the southern parts it is likely that the lack of precipitation in summer makes the availability of host plants limited. As opposed to most species in the Lepidoptera order, the painted lady does not enter a stage of dormancy (i.e., diapause) when conditions are not favourable, but continue reproducing uninterruptedly throughout the year while migrating to more suitable environments. In the West Palearctic region (including Northern Africa, Europe and parts of the Arabian Peninsula and Eastern Asia), a circular migratory route has recently been discovered. The annual migratory cycle includes up to 10 generations of butterflies, with some individuals flying > 4,000 km during their lifetime (Suchan *et al.* 2019, Talavera et al. 2018) and the entire cycle can sum up to 15,000 km (Menchetti et al., 2019).

Migration time for *V. cardui* can occur anytime over the year as they do not enter diapause like other insects during the winter (Menchetti et al., 2019). Adult *V. cardui* can be spotted in Southern Europe and North Africa as early as February/March during their migration North, which is concluded in the temperate Scandinavia and north-east Europe. Once in these higher latitudes, the population remains between May to August taking advantage of the warm temperatures. The direction of travel is reversed during autumn as populations travel down through Europe to spend the wintering months (between November and March) settled in the tropical conditions of Sub-Saharan Africa (Menchetti et al., 2019). The painted lady embarks upon a continual cycle of migration and reproduction where mated females interrupt their migration for some time to lay eggs on a host plant, so it appears that investment in either strategy is dependent on host plant availability. When the eggs hatch, the larvae feed on the plant that was selected by their mother. After sufficient growth and reaching Instar V, they pupate and remain in this state for a few days/weeks, before emerging from the pupa and immediately

continuing with the migration route (Stefanescu et al., 2013).*Vanessa cardui* migration, occurs in the pre-reproductive period meaning that the majority of females in active migration are non-mated (Stefanescu et al. 2021). Moreover, it has been shown that, host plant abundance positively impacted mating frequency/timing when modelled and it was suggested that mated females can detect suitable breeding areas (Stefanescu et al. 2021). Hence, *V. cardui* represents a classical example of the oogenesis-flight syndrome (Stefanescu et al., 2020), where females invest either in dispersal/migration or reproduction, probably dependent on the availability of host plants where they emerge (and/or population density at the breeding site).

As mentioned above, timing of mating has been shown to be affected by host plant availability so that females invest earlier in reproduction if host plants are abundant (Stefanescu et al. 2021). Another study has shown that a full fat body could power a Monarch for 40 hours of continuous migratory flight (Gibo and McGurdy 1993), and it is possible that the painted lady could behave in a similar way. This provides the motivation for our investigation into genes associated with migratory behaviour linked to host plant abundance and population density. These could be linked to a range of gene families with functions covering, for example, overall metabolic rate and fat storage, wing muscle development or egg maturation.

1.4 | Aims of the project

In this project we will investigate the difference in transcriptome profiles between populations of *V. cardui* reared under different treatments aimed to either induce or inhibit a migratory response, which can be used as a tool for mitigating the consequences of climate change and promote conservation of biodiversity (Bauer and Hoye 2014). Analysis the differential profile of RNA at different developmental stages where individuals are either predisposed to migrate or remain stationary to reproduce has the potential to reveal the genetic pathways involved in this classical oogenesis-flight trade-off. Our results will be compared to sets of candidate genes that have been linked to migration in other species to draw conclusions on the generality of the genetic underpinnings of migratory behaviour in butterflies. Since migratory behaviour has evolved independently in many butterfly lineages, and because migratory strategies in other systems in general do include a diapause stage (e.g. as in Monarchs, discussed above), we predict that the genetic pathways underlying this complex trait may vary across lineages.

2 | Materials and methods:

2.1 | Preliminary investigations

Initially, it was planned to isolate the gonads of each of the four developmental stages (Instar III, Instar V, Pupa and Adult), in order to investigate the affects the environmental conditions play on the gene

expression in these organs. Unfortunately, and as opposed to other Lepidoptera species, gonads were difficult to identify and impossible to isolate from accompanying tissues in all stages. Consequently, it was decided to dissect a larger section of the abdomen to make sure the gonads were included in it. This approach makes it more difficult to detect gonad specific expression patterns since the signal gets diluted, but the procedure allowed for easier dissection, which was preferable to prevent RNA degradation with the material thawing.

Two kits for RNA extraction were preliminarily tested with samples not included in the experimental design: RNeasy Mini Kit (Qiagen) & RNeasy Lipid Tissue Mini Kit (Qiagen) for efficacy and practicality. The RNeasy Lipid Tissue Mini Kit (Qiagen) was tested as *V. cardui* pupae (and some adults) have a substantial proportion of body fat. This kit however produced substantially lower yields overall, insufficient according to the sequencing service requirements. Consequently, the RNeasy Mini Kit (Qiagen) was selected to proceed with the RNA extractions.

2.2 | Different Resource Allocation Treatments – Developmental Stages

Wild *V. cardui* adult females were collected during field research trips. The females were kept in individual cages in a common garden with specific settings (18:6 light:dark regime, +25°C, access to a host plant (*Malva sylvestris*) and 10% sugar water). Eggs from each female were split into three different treatment groups intended to induce either migratory behaviour or investment in reproduction (i.e. mating). Individual females can lay up to 500 eggs (Hammad and Raafat 1972) but as nearly all the females caught in the wild are already mated, the yield is usually substantially lower than this. Once hatched, larvae were fed with sprouts of the host plant *M. sylvestris* under a light:dark regime of 18:6 hours.

To induce a migration response, some treatments were established to generate stress to the larvae, as it has been shown that unfavourable conditions (lack of host plants or high population density during larval development, or both) trigger the need to migrate to a new location (Stefanescu et al. 2021). Two parameters were used to create this stress: high population density and limited host plant resources (HDLI). To encourage stationary behaviour and investment in reproduction, another treatment was set at a high population density and an abundance (*ad libitum*) of host plants (HDAL). Finally, to assess if population density is an important factor regardless of food resources, another set of eggs was kept at conditions of low population density and abundant resources (LDAL). These three treatments should allow us to discriminate between population density (HDAL vs. LDAL comparison) and food availability (HDAL vs HDLI) as the main factor affecting differences in gene expression profiles (Figure 1).

Figure 1

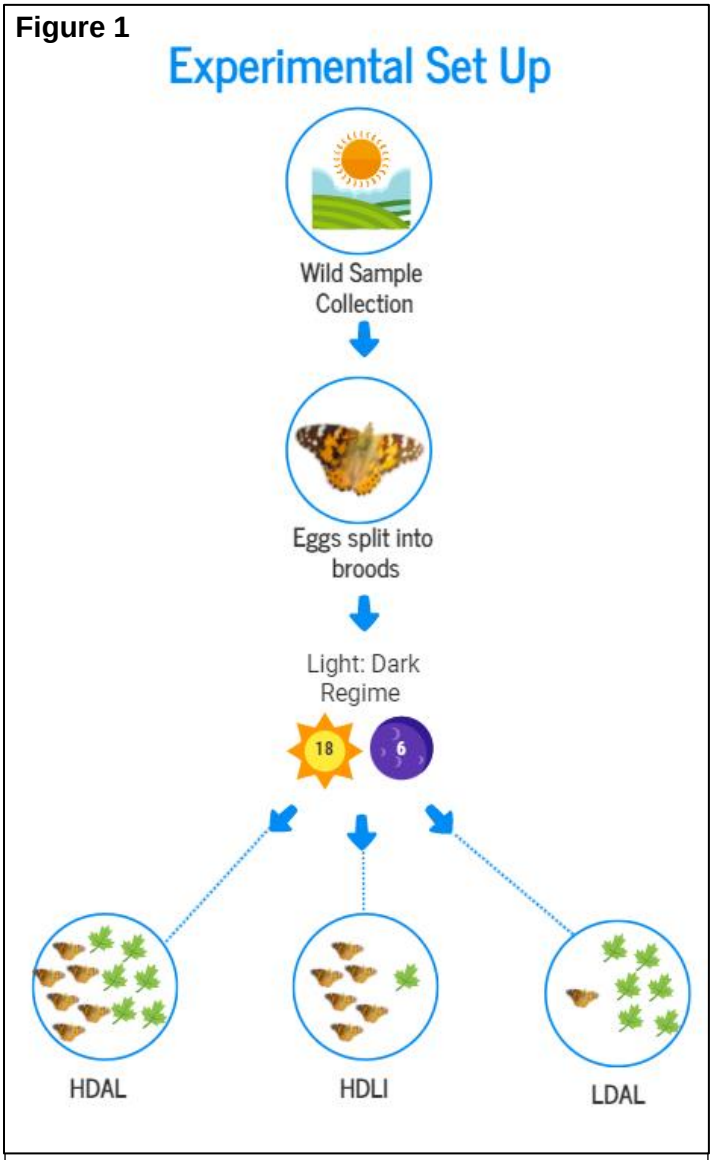

Figure 1: The experimental set up to investigate gene expression changes between different treatments of developing *Vanessa cardui* butterflies. Wild caught females were allowed to lay eggs on an abundance on host plant (*Malva sylvestris)* and eggs were transferred to 1 of 3 environments to observe the impact of population density and food availability on the transcriptome. HDAL = high density population, *ad libitum* resources, HDLI = high density population, limited resources, LDAL = low density, ad libitum resources. A light regime of 18hs light and 6h darkness was applied to the cages. The number of individuals sampled from each developmental group + sex are shown. Infographic created using: Venngage Inc.

Figure 2

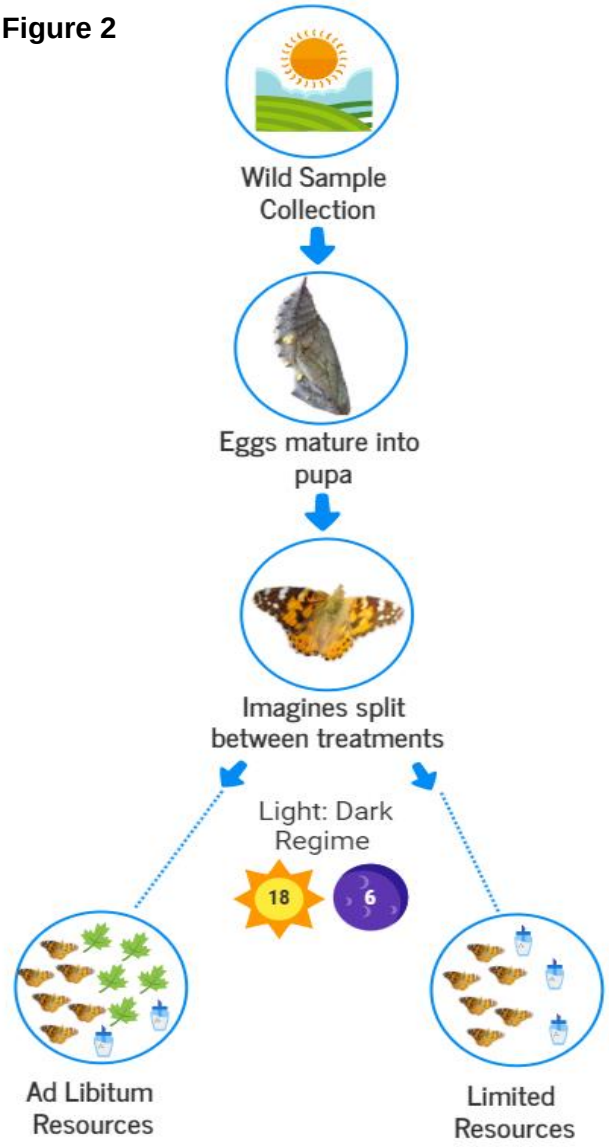

Figure 2: The experimental set up to investigate gene expression changes between different treatments of adult female *Vanessa cardui* butterflies. Eggs, laid by wild caught females, were raised in the lab and emergent adult females were placed in 80x80x50cm cages with males in one of two treatments. Ab libitum resources was an abundance of host plant (*Malva sylvestris)* and sugar water, limited resources was sugar water alone. A light regime of 18hs light and 6h darkness was applied to the cages. The number of individuals sampled from each treatment was 5 adult females. Infographic created using: Venngage Inc.

Individuals were reared under the different treatments and sampled at different developmental stages

as genes could be differentially expressed at any stage: larvae at instar III and instar V harvested at the day they entered the larval stage in question, female and male pupae harvested one day after pupation, and female and male adults harvested at the day of emergence from the chrysalis. Note: sex cannot be determined based on morphology in *Vanessa cardui* larvae and therefore we could not sample males and females separately for those developmental stages. All samples were flash frozen with liquid nitrogen and stored at –80°C until RNA extraction.

2.3 | Different Resource Allocation Treatments – Adults

We have yet to obtain sufficient evidence to suggest the trade-off of migratory behaviour vs investment in reproduction occurs only during the development of *V*. *cardui.* Therefore, *a* simultaneous set up was prepared with adult males and females in two food availability conditions: *ad libitum* and no resources, to analyze the influence that environmental conditions may have on investment in migration or reproduction in imagines. Again, a light:dark regime of 18:6 hours was set throughout the whole experiment. Here, newly emerged female imagines were individually marked with a unique ID (with sharpie under left hindwing) and released in one of two different large (80*80*50 cm) cages with free flying males. One cage contained an abundance of host plants (9 pots with *M. sylvestris*) and sugar water and the other cage contained only sugar water. Females were collected exactly five days after emergence, when the predisposition to either mate or migrate should be set (Stefanescu et al. 2021), flash frozen in kryo tubes in liquid nitrogen and stored in –80°C until the RNA extraction was performed.

2.4 | RNA Extraction Library Preparation and Sequencing

Samples were dissected into head and abdominal subsections on a cold dissection plate placed on ice. Keeping temperature as low as possible throughout the extraction protocol was key to slow RNA degradation as much as possible. Once a sample was dissected, it was disrupted and homogenized using a guanidine-isothiocyanate lysis buffer and micro-pestle, followed by QiaShredder. In the samples collected at the larval stages, the gut contained abundant plant material. This was easily removed without damaging the surrounding tissue to prevent contamination of the insect RNA. Given the difficulty dissecting the gonads specifically, the abdomen was sampled between the 6^{th}-8^{th} segments, inclusive. The head was dissected from the first segment of the thorax. In pupae, the section equivalent to the head was dissected, as well as the lowest segments of the abdomen, which include the region where the gonads are found. Pupal internal anatomy undergoes drastic reorganizations, which makes it impossible to identify the gonads of the individuals. Hence, sex determination was performed by phenotyping external morphological characteristics on the pupae and adults. The RNeasy Mini Kit (Qiagen) was used to extract RNA following the general instructions in the Kit handbook and resuspending RNA extractions in nuclease free water.

RNA integrity and fragment length were evaluated by 1% agarose gel electrophoresis at 90V for 60 minutes. NanoDrop (ThermoFisher) was used to evaluate the purity of the samples (lack of contamination from reagents and/or proteins) by measuring the light absorbance at different wavelengths and calculating the 260/230 nm and 260/280 nm ratios. Additionally, NanoDrop was used to provide a preliminary idea about the RNA concentration of the samples.

Concentrations were more precisely measured immediately after extraction with Qubit (ThermoFisher). Many samples had concentrations above the range the kit could measure (600 ug/ml upper threshold), making dilutions necessary. All extractions were preserved at -80°C and, where required, were diluted and remeasured with NanoDrop and Qubit right before library preparation and sequencing to ensure that the samples had the requested concentration for library preparation and sequencing.

RNA solutions at the appropriate concentration were sent to the National Genomics Infrastructure (NGI), Science for Life Laboratory (SciLife) for library preparation (RNA sequencing (RNA-seq) with poly-A tail enrichment) and sequenced in multiplex on two Illumina S4-300 v1.5 lanes with 150 base pairs (bp) paired-end reads.

2.5 I Alternative approaches

Gene expression was the most logical route for this investigation as different individuals within the same population of *V. cardui* displayed different behaviours despite having similar genetic make-up. However, we could have taken a theory-based approach, similar to what was proposed in a study conducted by Martin et al. (2015), who, while working with Chinook salmon, modelled the difference in metabolic fitness of migration in response to temperature (Martin et al. 2015). This approach limits the scope of research to one category of genetic processes as the model must be specifically designed, whereas a whole transcriptome sequencing approach offers the opportunity to observe changes in multiple pathways of the genome and correlate population density or resource allocation with processes including reproduction, fat storage and muscle development.

A gene expression study had already been conducted by the same lab on host plant effects in *Leptidea sinapis* (Näsvall et al., 2020), which made this set of techniques even more appropriate, facilitated by the experience, resources and pipelines previously developed for RNA analysis that were available.

3 | Research Plan

3.1 I Raw data (RNA-sequencing reads) quality assessment and filtering

Genomic analysis has been vastly improved with the aid of high-throughput sequencing (Schmieder and Edwards 2011) but can be streamlined by quality control of the raw (in this case RNA-seq) reads. Here the plan for processing the RNA-seq data is detailed, with each bioinformatic tool in CAPITALS for clarity.

In order to quickly assess the overall quality of the RNA-seq reads, FASTQC version 0.11.5 (Andrews 2016) will be used to produce graphical representations of the quality scores per base, quality scores per sequence, per base sequence content (expected to be identical across all sequences), per base GC content (expected to be a little uneven at start of RNA sequences due to remaining Illumina primer sequences) and GC content per base sequence (presented as a relationship to an expected normal curve). The plots are useful for estimating the necessary quality trimming and adjusting the parameters of filtering software.

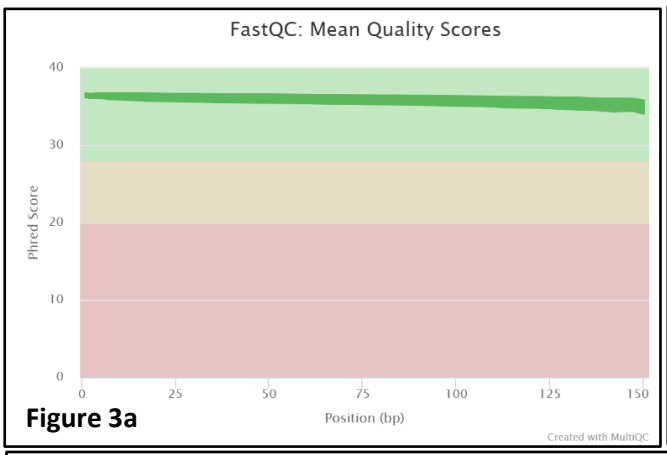

Figure 3a

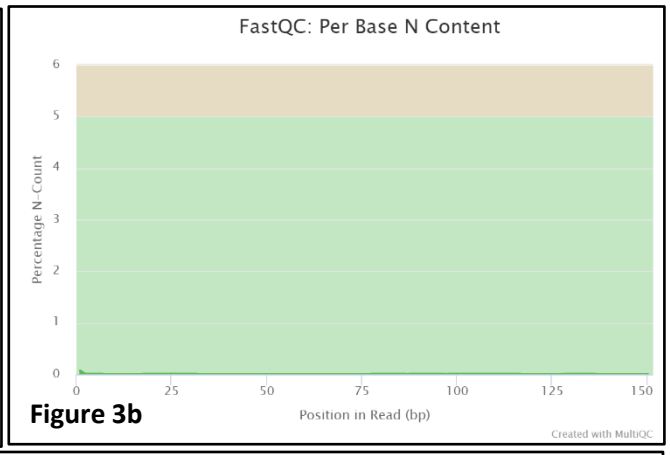

Figure 3b

Figure 3: Figure 3a demonstrates the mean quality of bases (Phred score) called at each position in RNA-sequence data produced by Illumina paired-end sequencing. A score of >34, as all bp here have, is equivalent to an error rate of 0.0004. Figure 3b displays the per base N content of the same reads. Ns are called when the quality of the sequence is too low to accurate determine which base is at the position. Here the majority of bp have a score of 0.00 while the highest percentage is 0.07.

After an initial view at FASTQC, it is evident that high quality RNA-sequencing reads were achieved as the mean Phred score across all base pairs was > 34, corresponding to an error rate of 0.0004 (Ewing et al.1998) meaning there is a lower than 1 in 4000 chance of a base pair being called incorrectly (Figure 3a). There is also a very low proportion of low quality reads that were unable to be identified and therefore masked N (Figure 3b); the majority of base pairs show 0.00 for proportion of N reads and the highest at Base 1 is only 0.07. This indicates that the RNA reads are of very high quality and are a good starting point for gene expression analysis.

Known adapters used in the Illumina sequencing process will be removed using TRIMGALORE version 0.6.1 (Krueger 2017); this software also filters out bases with a higher than 0.001 probability of incorrectly called bases, i.e. a Phred score of < 30 (Ewing et al.1998). A high stringency (min 1) is set to remove as small as 1 bp overlaps between the adapter and the target sequence. This setting is strict to ensure minimal contamination that would otherwise lead to potential mis-mapping with the reference genome, in subsequent analytical steps. It is important to note that the sequence content across all bases should be stable, but after adapter trimming there is a decrease of A in the final position as even a tail of 1 A will be removed from the 3' end to minimize the risk of including any post-transcriptional modifications (for example the poly-A-tail in the case of eukaryotic mRNA). Trimming of sequences can sometimes leave very short reads remaining, especially in RNA-seq data, so all sequences < 30 bp will also be filtered out to minimize the risk of erroneous mapping.

As paired-end sequencing was conducted, TRIMGALORE will be set to remove both reads if at least one falls below this threshold in order to be compatible with aligners requiring both reads. This is to ensure the data are suitable for software used later in the analysis that has a minimum threshold for sequence length and to make the size of the output file more manageable. The error rate, i.e., the number of errors per length of sequence, is set to 0.1 and any bases marked N will be removed.

FASTQ MASKER (http://hannonlab.cshl.edu/fastx_toolkit/; accessed 29-08-2021) will be used to mask (replace with N) nucleotides in the reads with a Phred score quality lower than 20. This corresponds to a 0.01 error rate. These masked N nucleotides will be removed using PRINSEQ 0.20.4 (Schmieder and Edwards 2011), as well as any potential remaining poly(A) tails (with a minimum of 3 bp from both 5' and 3' end). This software will only remove poly-A or poly-T sections if they are located at the end of the sequence. After this processing step, sequence ends with a lower Phred quality than 30 will be filtered out as well as reads with a high N content (maximum allowed N fraction threshold = 10%). Any remaining sequences shorter than 30 bp will again be filtered out. Complexity assessed with the DUST algorithm (modelled on the system used by BLAST) and sequences with a score of < 7 (low complexity) will also be removed (Schmeider & Edwards 2011, Morgulis et al. 2006). These filtering steps are all standard procedure in preparing raw RNA-seq data for detailed gene expression analysis (Näsvall et al. 2020).

CUTADAPT version 3.1 (Martin 2011) will be used to filter out internal poly -A/T sequences longer than 10 bp as it supports gapped alignment. It will also remove Ns from each end of the read, discard sequences with > 10% N content and sequences shorter than 30bp. CONDETRI (Smeds and Künstner 2011) uses quality scores for individual bases to assess each sequence and discards reads with < 30 Phred Score in over 80% of the read. This will save computational time and storage as the data is streamlined.

Non-coding ribosomal sequences (rRNA) will be removed from the meta-transcriptomic RNA with SORTMERNA version 3.0.3 (Kopylova et al. 2012). FASTQC version 0.11.5 (Andrews 2016) will be run again to provide an overview of the outcome of the trimmed data and identify any further necessary quality control steps.

FASTQC SCREEN 0.11.1 (Wingett 2017) with the aligner BOWTIE2 version 2.3.5 (Langmead & Salzberg, 2012) will be used to isolate and screen for foreign RNA from other species. The host plant *M. sylvestris* is a likely contaminant as well as remaining rRNA, undetected Illumina adapters and primers and genetic material from human researchers, *Drosophila melanogaster*, *Wolbachia* sp. (a parasite bacterium found in many insect species) and *Leptidea sinapis* (another butterfly species reared in the same lab). This is especially important as genes expressed in foreign transcriptomes could lead to inaccurate conclusions about the significance of this data and shared contaminant segments can lead to unrelated sequences being aligned to each other.

3.2 | Mapping reads + Evaluating gene expression

STAR version 2.7.2b (Dobin et al. 2013) will be run in two-pass mode to map the quality filtered reads to a previously published *V.cardui* genome and transcriptome (Zhang et al. 2021, Connahs et al. 2016). STAR is the preferred software to other aligners, as it has been shown to outperform in speed and precision (Dobin et al. 2013), and is especially suited to RNA data as it performs splice-aware alignment (introns that are not transcribed are taken into account when mapped to the reference genome). Another benefit of using STAR is that sequence junctions that were mapped during the first pass can be “annotated” in the second pass, resulting in a greater number of spliced regions to be mapped to novel junctions.

STRINGTIE version 2.1.4 (Pertea et al., 2015) assembles a series of reads into tangible transcripts that can be used as ‘gene sequences’ to quantify gene specific expression levels. This software is suitable as it is fast and can provide good estimates of expression levels. It also produces a competitive number of transcripts compared to other leading transcriptome reconstruction software (Pertea et al., 2015).

Differential gene expression analysis will be performed using the package DESEQ2 version 3.6 (Love et al. 2014) as implemented in R version 4.0.5 (R Core Team, 2013). Initially, genes that are not present in any sample or have only been expressed in 1 out of the 144 samples will be discarded (0 read counts for every sample or only a single >0 read count), to eliminate expression changes from single individuals that are likely to be factors for changes which do not represent the treatment group; small, yet statistically significant changes in gene expression may not be interesting for further analysis (Love et al., 2014). A count of 1 will be added to every gene to stabilize the variance for lowly

expressed genes (working with 0 makes population-scale statistics difficult). Genes with a low mean expression across all samples (>2 to account for the added 1) will be discarded to only analyze the most informative genes. DESEQ2 with default settings will allow us to normalize the counts per gene by number of reads in a specific library (library size) and perform significance testing on individual genes using the Wald test (Love et al. 2014). To account for the high number of significance tests performed (and reduce the number of false positives), false discovery rate (FDR) adjustment (the anticipated quantity of incorrectly rejected hypothesis) will be generated using the Benjamini and Hochberg (1995) method and used to adjust significance levels for each specific gene.

TOPGO version 2.38.1 (Alexa and Rahnenfuhrer 2016) within the BIOCONDUCTOR package in R version 3.4.3 (R Core Team, 2013) attributes specific function to identified genes displaying differential gene expression between experimental cohorts. This will be achieved by comparing genes in the database org.Dm.eg.db from R-package AnnotationDbi (Pagès et al., 2019**)** for orthologous genes, assuming genes with similar sequences will perform similar functions i.e., a homologous amino acid sequence behaves the same way in the focal organism as in the organism used for functional verification. Finally, we will quantify gene ontology categories (e.g., biological process, cellular component and molecular function) that are enriched for differentially expressed genes in each comparison. Such functional information will be important to understand the specific genetic pathways that are affected by the different treatments, leading to differential gene expression between individuals predisposed to reproduce or migrate.

3.3 | Expected Results

It is naturally difficult to predict the precise gene families that will be expressed differently in this investigation and the literature around the topic is limited to different approaches of genomic research.

We do anticipate differences in the transcriptome profiles of the *V. cardui* populations raised in treatments with *ad libitum* resources and limited resources (HDAL vs HDLI) revealing a potential correlation between host plant availability and expression levels of genes of a specific function. It is likely that genes associated with metabolic pathways will be impacted by changes in resource allocation, as storing of energy reserves and timing of specific energy allocations is necessary for the success of migration or reproduction. This information can be used to infer the specific genes and pathways underlying this trade – off. Monarchs *(D. plexippus)* experience reproductive diapause when undertaking a migratory flight (Zhan et al., 2011) suggesting a lower expression level for genes associated with reproductive processes in migratory populations when compared to sedentary ones. This same pattern may be expected in *V. cardui* as well, with reduced expression of genes associated with reproductive activity in those individuals reared under conditions of limited resources.

It is also expected that some genes will be differentially expressed between populations in treatments with different population densities (HDAL vs LDAL). We speculate that affected genes could be associated with reproduction and fat storage as there is a trade-off between migratory behaviour and investment in reproduction: the oogenesis-flight syndrome. Hence, we are especially interested in the abdominal samples of adults as this tissue will contain gonad RNA and could provide the most valuable insights into the genetic underpinnings of the migration-reproduction trade-off.

It is also possible that we will observe a difference in expression of genes associated with flight muscle function as selection for these genes was associated with migration in Monarch butterflies (Zhan et al., 2014). An apparent difference between haplotypes of a collagen protein and the *kettin* gene (associated with flight muscles) in migratory vs non-migratory populations suggest a selection for different flight muscle phenotypes (Zhan et al. 2014) in migratory and sedentary populations and that transcriptomic approaches like ours could advance the understanding of phenotypic differences between sub-populations of partially migratory species.

4 | Conclusion

Our study, among other epigenetic and gene expression approaches, can reveal differences between sedentary and migratory populations and indicate genes associated with migratory behaviour. The importance of individual genes is highlighted but their specific role will not yet be defined. Therefore, in order to fully verify the causal relationship between the transcriptome and phenotype, further functional studies, such as RNAi-editing or Crispr-Cas9, must be conducted.

Our results will demonstrate the genetic underpinnings of migratory behaviour in *V. cardui* but will have significant implications for migration across all insects. This information can be used to predict the influence of climate change on migratory organisms and advise assisted migration as a mitigation tool to protect endangered non-migratory species experiencing habitat loss. Results will also be important for the agricultural sector, who rely on insect pollination for crops and would benefit from being able to anticipate potential changes to pest species migration. This is especially important as the demand for food increases and food security is continually threatened by the climate crisis. Conservation groups combating the consequences of climate change could rely on migration as a tool to promote biodiversity as well.

5 | Acknowledgements

I thank Niclas Backström, my vocational supervisor, and Lars Höök, my project supervisor, for their consistent guidance and support throughout the project and for helpful comments to a previous version of this project proposal. I especially thank Aleix Palahí i Torres, for his suggestions to the work and similar support. I also thank Karin Näsvall and Veronika Mrazek for assistance with the RNA

sequence data analysis protocol and the remaining of the Backström research group: Daria Shipilina, Peter Pruisscher, Jesper Bowman, Venkat Talla, Orazioluca Paternò, Xuejing Hu and Yishu Zhu for their input and advice throughout the year.

I acknowledge support from the National Genomics Infrastructure in Stockholm and Uppsala funded by the Science for Life Laboratory, The computations were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) at Uppsala University.

6 | References

Alerstam, T., Hedenström, A. and Åkesson, S. 2003. Long-distance migration: evolution and determinants. *Oikos* 103, pp. 247-260.

Alexa, A. and Rahnenfuhrer, J. 2016. topGO: Enrichment analysis for gene ontology. [R package]. Retrieved from: https://www.bioconductor.org/packages/release/.

Altizer, S., Bartel, R. and Han, B. A. 2011. Animal migration and infectious disease risk. *Science* 331, pp. 296-302.

Andrews, S. 2010. FastQC: A Quality Control Tool for High Throughput Sequence Data [Software]. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Bauer, S. and Hoye, B. J. 2014. Migratory animals couple biodiversity and ecosystem functioning worldwide. *Science* 344, pp. 1242552.

Bell, J.R., et al. 2015. Long-term phenological trends, species accumulation rates, aphid traits and climate: five decades of change in migrating aphids. *Journal of Animal Ecology* 84, pp. 21–34.

Benjamini, Y. and Hochberg, Y. 995. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society B* 57, pp. 289 – 300.

Boggs, C. L., Watt, W. B. and Ehrlich, P. R. 2003. Butterflies: ecology and evolution taking flight. Chicago: University of Chicago Press.

Butt, N. et al. 2020. Importance of species translocation under rapid climate change. *Conservation Biology* 35, pp. 1 – 9. doi: 10.1111/cobi.13643.

Chen, I-C., Hill, J.K., Ohlemuller, R., Roy, D.B. and Thomas, C. D.
2011. Rapid range shifts of species associated with high levels of climate warming. *Science* 333, pp. 1 024–1026. doi: 10.1126/science.1206432.

Connahs, H., Rhen, T. and Simmons, R. B.
2016. Transcriptome analysis of the painted lady butterfly *Vanessa cardui* during wing color pattern de velopment. *BMC Genomics* 17, p. 270. doi: 10.1186/s12864-016-2586-5.

Cresswell, K. A., Satterthwaite, W. H. and Sword, G.
A. Understanding the evolution of migration through empirical examples. In: Milner-Gulland E.
J., Fryxell J. M. and Sinclair A. R. E. eds. *Animal migration: a synthesis*. New York:
Oxford University Press, pp. 1-16.

Dobin, A. et al. 2013. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 1pp. 5–
21. doi: 10.1093/bioinformatics/bts635.

Eliason, E.J. et al.
2011. Differences in thermal tolerance among sockeye salmon populations. *Science* 332, pp. 109–
112.

Gibo, D.L. and McCurdy, J.
1993. Lipid accumulation by migrating monarch butterflies (*Danaus plexippus L.*).
Canadian Journal of Zoology 71, pp. 76-82.

Hobson, K. A., Anderson, R. C., Soto, D. X. and Wassenaar, L. I.
2012. Isotopic evidence that dragonflies (*Pantala flavescens*) migrating through the Maldives come fro m the northern Indian subcontinent. PLoS One 7, e52594.

IPCC. 2021. Summary for Policymakers.
In: *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge:
Cambridge University Press.

Kopylova, E., Noé, L. and Touzet,
H. 2012. SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinf ormatics* 28, pp. 3211 – 3217. doi: 10.1093/bioinformatics/bts611.

Kristensen, N.P., Johansson, J., Ripa, J. and Jonzen, N.
2015. Phenology of two interdependent traits in migratory birds in response
to climate change. *Procedures of the Royal Society B* 282, pp. 20150288.

Krueger, F. 2017. Trim Galore, v. 0.4.4: a wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files. Available
at: https://github.com/FelixKrueger/TrimGalore.

Love, M. I., Huber, W. and Anders,
S. 2014. Moderated estimation of fold change and dispersion for RNA-
seq data with DESeq2. *Genome Biology* 15, p. 550. doi:10.1186/s13059-014-0550-8.

Martin, M. 2011. Cutadapt removes adapter sequences from high-
throughput sequencing reads. *Embnet Journal* 17, pp. 10–12. doi: 10.14806/ej.17.1.200.

Martin, B.T., Nisbet, R.M., Pike, A., Michel, C.J. and Danner, E.M. 2015.
Sport science for salmon and other species: ecological consequences of metabolic power constraints. *Ecology Letters* 6, pp. 535-44. doi: 10.1111/ele.12433.

Menchetti, M., Guéguen, M. and Talavera, G. 2019. Spatio-
temporal ecological niche modelling of multigenerational insect migrations. *Proceedings of the Royal S ociety B* 286, pp. 20191583. doi:10.1098/rspb.2019.1583.

Merlin, C. and Liedvogel, M. 2019. The genetics and epigenetics of
animal migration and orientation: birds, butterflies and beyond. *Journal of Experimental Biology* 6, 222. doi: 10.1242/jeb.191890.

Morgulis, A., Gertz, E.M., Schäffer, A.A. and Agarwala, R. 2006.
A fast and symmetric DUST implementation to mask low-
complexity DNA sequences. *Journal of Computational Biology* 2006, pp. 1028.

Näsvall, K. et al. 2020. Host plant diet affects growth and induces altered gene expression and microbiome composition in the wood white (*Leptidea sinapis*) but terfly. *Molecular Ecology* 30, pp. 499 – 516. doi: 10.1111/mec.15745.

Pagès, H., Carlson, M., Falcon, S. and Li, N. 2019. AnnotationDbi: manipulation of SQLite-based annotations in Bioconductor. version 1.48.0. [R package]. Retrieved from: https://bioconductor.org/packages/AnnotationD.

Pecl, G.T. et al. 2017. Biodiversity redistribution under climate change: Impacts on ecosystems and human well-being. *Science* 355, pp. eaai9214. doi: 10.1126/science.aai9214.

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T-C., Mendell, J. T. and Salzberg, S. L. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology* 33, pp. 290–295. doi: 10.1038/nbt.3122.

Poloczanska, E.S. et al. 2013. Global imprint of climate change on marine life. Nature *Climate Change* 3, pp. 919–925. doi: 10.1038/nclimate1958.

Savela, M. 1999. *Vanessa cardui* (Linnaeus, 1758). Lepidoptera and Some Other Life Forms. Available at: http://www.nic.funet.fi/pub/sci/bio/life/insecta/lepidoptera/ditrysia/papilionoidea/nymphalidae/nymphalinae/vanessa/#cardui. [Accessed on August 3, 2021].

Schmieder, R. and Edwards, R. 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27, pp. 863–864, doi:10.1093/bioinformatics/btr026.

Seebacher, F. and Post, E. 2015. Climate change impacts on animal migration. *Climate Change Responses* 2, pp. 5. doi: 10.1186/s40665-015-0013-9.

Smeds, L. and Künstner, A. 2011. ConDeTri - a content dependent read trimmer for Illumina data. *PLoS One* 6, pp. e26314. doi: 10.1371/journal.pone.0026314.

Stefanescu, C. et al. 2013. Multi-generational long-distance migration of insects: studying the painted lady butterfly in the Western Palearctic. *Ecography* 36, pp. 474 – 486.

Stefanescu, C., Soto, D.X., Talavera, G., Vila, R. and Hobson, K.A. 2016. Long-distance autumn migration across the Sahara by painted lady butterflies: exploiting resource pulses in the tropical savannah. *Biology Letters* 12, pp. 20160561.

Stefanescu, C., Ubach, A. and Wiklund, C. 2021. Timing of mating, reproductive status and resource availability in relation to migration in the pain ted lady buttefly. *Animal Behaviour* 172, pp. 145 – 153. doi:10.1016/j.anbehav.2020.12.013

Suchan, T., Talavera, G., Sáez, L., Ronikier, M. and Vila, R. 2019. Pollen metabarcoding as a tool for tracking long-distance insect migration. *Molecular Ecology Resources* 19, pp. 149-162.

Talavera, G., Bataille, C., Benyamini, D., Gascoigne-Pees, M. and Vila, R. 2018. Round-trip across the Sahara: Afrotropical painted lady butterflies recolonize the Mediterranean in early sprin g. *Biology Letters* 14.

Talla, V. et al. 2020. Genomic evidence for gene flow between monarchs with divergent migratory phenotypes and fli ght performance. *Molecular Ecology* 29, pp. 2567 – 2582. Doi: 10.1111/mec.15508.

Taylor, C.M. and Norris, D.R. 2007. Predicting conditions for migration: effects of density dependence and habitat quality. *Biology Le tters* 3, pp. 280 – 284. doi: 10.1098/rsbl.2007.0053.

TEEB. 2010. The Economics of Ecosystems and Biodiversity: Mainstreaming the Economics of Nature: A synthesis of the approach, conclusions and recommendations of TEEB.

Wahlberg, N. and Rubinoff, D. 2011. Vagility across *Vanessa* (Lepidoptera: Nymphalidae): mobility in butterfly species does not inhibi t the formation and persistence of isolated taxa. *Systematic Entomology* 36, pp. 362-370.

Zattara, E.E. and Aizen, M.A. 2021. Worldwide occurrence records suggest a global decline in bee species richness. *One Earth* 4, pp. 114 – 123. doi: 10.1016/j.oneear.2020.12.005.

Zhan, S., Merlin, C., Boore, J. L. and Reppert, S. M.
2011. The monarch butterfly genome yields insights into long-distance migration. *Cell* 147, pp. 1171-1185.

Zhan, S. et al. 2014. The genetics of monarch butterfly migration and warning coloration. *Nature* 514, pp. 317 – 321. doi: 10.1038/nature13812.

Zeng, J., Liu, Y., Zhang, H., Liu, J., Jiang, Y., Wyckhuys, K.A.G. and Wu, K. 2020.
Global warming modifies long-distance migration of an agricultural insect pest. *Journal of Pest Science* 93, 569–581 (2020). doi: 10.1007/s10340-019-01187-5.