Advances in studying the role of genetic divergence and recombination in adaptation in non-model species

Ramprasad Neethiraj

Academic dissertation for the Degree of Doctor of Philosophy in Population Genetics at Stockholm University to be publicly defended on Friday 22 February 2019 at 13.30 in Vivi Täckholmsalen (Q-salen), NPQ-huset, Svante Arrhenius väg 20.

Abstract
Understanding the role of genetic divergence and recombination in adaptation is crucial to understanding the evolutionary potential of species since they can directly affect the levels of genetic variation present within populations or species. Genetic variation in the functional parts of the genome such as exons or regulatory regions is the raw material for evolution, because natural selection can only operate on phenotypic variation already present in the population. When natural selection acts on a phenotype, it usually results in reduction in the levels of genetic variation at the causal loci, and the surrounding linked loci, due to recombination dynamics (i.e. linkage); the degree to which natural selection influences the genetic differentiation in the linked regions depends on the local recombination rates. Studies investigating the role of genetic divergence and recombination are common in model species such as *Drosophila melanogaster*. Only recently have genomic tools allowed us to start investigating their role in shaping genetic variation in non-model species. This thesis adds to the growing research in that domain. In this thesis, I have asked a diverse set of questions to understand the role of genetic divergence and recombination in adaptation in non-model species, with a focus on Lepidoptera.

First, how do we identify causal genetic variation causing adaptive phenotypes? This question is fundamental to evolutionary biology and addressing it requires a well-assembled genome, the generation of which is a cost, labor, and time intensive task. In paper I, I present a tool, MESPA, that stitches together exonic sequences in fragmented assemblies to produce high-quality gene models. These high-quality gene models can be used by researchers in the downstream analyses, providing genomic insights for a fraction of cost of a high quality genome.

Second, what does the pattern of recombination rate look like in chromosomes that lack centromeres (i.e. holocentric chromosomes)? In paper II, I compare the recombination landscape and the patterns of nucleotide diversity in three Lepidotera with holocentric chromosomes, *Pieris napi*, *Bombyx mandarina*, and *Bombyx mori*, with a monocentric species. Our results show that on average these three Lepidoptera have high rates of recombination across the vast majority of their genome. Our results also suggest that given similar effective population sizes, these species are likely to harbor more genetic diversity compared to monocentric species, which has important evolutionary consequences for these species.

Third, what is the potential for parallelism at the genetic level in convergent melanic phenotypes? In paper III, I investigated the genetic basis of the female-limited melanic phenotype in the green-veined white (*Pieris napi*) butterfly, and found a 20kb region, approximately 50kb from the gene *cortex*, associated with this trait. This gene has been implicated in melanic phenotypes in other Lepidoptera that diverged from *Pieris* approximately 100my, indicating very high predictability for this trait.

Finally, what is the role of cis-regulatory variation in local adaptation? In paper IV, I analyzed the relationship between allele specific expression (ASE) and genetic divergence (*F*$_{ST}$) in the F1 hybrids of *Pieris napi napi* and *Pieris napi adulwinda*. I show that intersecting results from ASE with *F*$_{ST}$ is a powerful approach to identify genes involved in local adaptation.

Keywords: Recombination, Genetic divergence, Lepidoptera, Genetic variation, Genetic diversity, Holocentric chromosomes, cis-regulatory variation, Allele-specific expression.

Stockholm 2019

http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-163416

ISBN 978-91-7797-574-8

Department of Zoology

Stockholm University, 106 91 Stockholm
ADVANCES IN STUDYING THE ROLE OF GENETIC DIVERGENCE AND RECOMBINATION IN ADAPTATION IN NON-MODEL SPECIES

Ramprasad Neethiraj
Advances in studying the role of genetic divergence and recombination in adaptation in non-model species

Ramprasad Neethiraj
To my family.
Table of contents

Papers in this thesis ............................................................................................................. 1
Candidate’s contribution ................................................................................................... 3
Introduction .............................................................................................................................. 5
  Identification of causal genetic variants ..................................................................... 6
  Recombination dynamics in holocentric chromosomes ........................................ 7
  Parallelism in convergent phenotypes .................................................................... 9
  Role of cis-regulatory variation in local adaptation ............................................. 10
Methods ............................................................................................................................... 11
Major findings ...................................................................................................................... 13
  Paper I ............................................................................................................................. 13
  Paper II ............................................................................................................................ 14
  Paper III .......................................................................................................................... 15
  Paper IV .......................................................................................................................... 16
Concluding remarks ........................................................................................................... 19
References ........................................................................................................................... 21
Svensk sammanfattning ..................................................................................................... 23
Acknowledgements ............................................................................................................ 25
Papers in this thesis

The thesis is based on the following articles, which are referred to in the text by their Roman numerals:


Paper I is reprinted with permission from the publisher.
Candidate’s contribution

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conceived the study</td>
<td>Minor</td>
<td>Significant</td>
<td>Minor</td>
<td>Minor</td>
</tr>
<tr>
<td>Designed the study</td>
<td>Minor</td>
<td>Significant</td>
<td>Minor</td>
<td>Minor</td>
</tr>
<tr>
<td>Collected the data</td>
<td>Substantial</td>
<td>Substantial</td>
<td>Substantial</td>
<td>Significant</td>
</tr>
<tr>
<td>Analysed the data</td>
<td>Substantial</td>
<td>Substantial</td>
<td>Substantial</td>
<td>Substantial</td>
</tr>
<tr>
<td>Manuscript preparation</td>
<td>Significant</td>
<td>Substantial</td>
<td>Substantial</td>
<td>Substantial</td>
</tr>
</tbody>
</table>

* Contribution Explanation

- **Minor**: contributed in some way, but contribution was limited.
- **Significant**: provided a significant contribution to the work.
- **Substantial**: took the lead role and performed the majority of the work.
Introduction

Changes in the DNA sequences that constitute the functional part of an individual’s genome such as exons or regulatory regions can affect one or several aspects of its life. At the population or species level, this diversity in the DNA sequences creates an incredible amount of phenotypic diversity on which natural selection can act. Therefore, genetic variation is essential for a population or species to respond to different selection pressures and evolve. There are two major sources of genetic variation: mutation and recombination. Mutations are a major source of genetic novelty. However, there is one major limitation to creating genetic variation through mutations; it is a slow process. For instance, human genome is $3.2 \times 10^9$ bp in size, and it has been estimated that approximately 60 new mutations are added with every new generation with a vast majority of these mutations likely to be in non-functional parts of the genome (Drake et al. 1998). Sexually reproducing organisms overcome this limitation by creating genetic variation through the process of recombination.

Recombination occurs when homologous chromosomes exchange pieces of genetic material with each other, and by doing so it shuffles variation generated by mutation at different loci to produce an array of chromosomes with diverse combination of mutations. During meiosis, this exchange can happen at random places along the chromosomes in the germ cells resulting in unique combinations of chromosomes in every gamete. The process of creating variation through recombination has two important consequences for evolution. First, it allows a mutation to be evaluated against different genetic backgrounds, which is crucial as a mutation can result in different outcomes depending on its interactions with other variation in the genome i.e., epistatic interactions. Second, variation under selection in regions of high recombination can break their association with the flanking alleles, so less variation is lost as an allele sweeps through or gets eliminated from the population.

With that in mind, in this thesis we have tried to understand the nature of genetic variation and the factors influencing it by,

- Identifying candidate genetic variation behind adaptive phenotypes (papers I, III, & IV)
- Understanding the mechanisms that govern the process of evolution (paper II).

More specifically, the aims of this thesis were as follows:

1. Generate a pipeline to identify causal genetic variation behind adaptive phenotypes quicker and cheaper by using Pool-Seq data instead of resequencing data.
2. Investigate the role of centromeres in recombination by evaluating the patterns of recombination in the chromosomes of species that lack centromeres, such as
Bombyx mori, and comparing these patterns with recombination patterns in Drosophila melanogaster, a species that contains chromosomes with centromeres.

3. Investigate the potential for parallelism at the genetic level in convergent melanic phenotypes.

4. Explore the role of cis-regulatory variation in local adaptation

Identification of causal genetic variants

What is the casual genetic variation behind a phenotype? This question is fundamental to evolutionary biology and addressing it requires a well-assembled genome. However, building a high quality genome is both time consuming and expensive, therefore genomic insights are out of reach for many researchers. In contrast, one can assemble a fragmented low quality genome with a small N50*, quickly and cheaply. Unfortunately, low quality genomes contain broken gene models making their use in conducting genomic investigations a challenge.

Here in paper I, we present an approach called Mining Exons for Scaffolding Poor Assemblies (MESPA) where a protein set from the same, or a moderately divergent species (10 to 60 MY**), can be used to improve the assembly around gene models in a highly fragmented assembly. We validated this approach using a de novo assembled genome from Pool-Seq*** data of a single wild population sample of Drosophila melanogaster (Fabian et al. 2012), and a protein dataset from a close congener that diverged 60 mya, D. virilis.

Genomes generated via Pool-Seq data are highly fragmented due to the high levels of heterozygosity, which makes it challenging for the assemblers to produce long contiguous sequences. Therefore, if MESPA was effective on Pool-Seq based assemblies it would facilitate population genomic studies in non-model species for a low cost. To demonstrate the usefulness of this pipeline in identifying the causal variation in non-model species, I then re-investigated the genetic basis of color polymorphisms in two non-model systems, a butterfly and a bird, for which the causal loci are known (Kunte et al. 2014; Poelstra et al. 2014).

In butterflies, I investigated the genetic basis of the mimicry phenotypes in the common mormon butterfly Papilio polytes. In this species females exhibit two color morphs, a male-like non-mimetic pattern (cyrus form) or one of the several mimetic patterns that resembles toxic species of the genus Pachliopta (Figure 1). A study showed that a single gene, doublesex (dsx), controls this mimicry phenotype (Kunte et al. 2014).

---

*N50 – a length statistic, indicating the size of contigs for which half the assembly is contained in contigs of length above this value

**MY – million years

***Pool-Seq - whole genome sequencing data from pools of individuals
In birds, I investigated the plumage divergence in a young European crow species pair, carrion crows (*Corvus corone corone*) and hooded crows (*Corvus corone cornix*). These species differ in their plumage color, with *C. c. corone* having a black head and torso whereas *C. c. cornix* have a back head and gray torso (Figure 2). A whole genome study revealed that between the morphs, genes involved in regulating melanogenesis and visual perception were found to be located in a single and large genomic region of high divergence (Poelstra et al. 2014).

Recombination dynamics in holocentric chromosomes

Selection at one site shapes the genetic variation patterns at the nearby sites i.e., linked sites, and the degree of this association between sites depends on the local recombination rates. In regions of low recombination, linked selection affects physically larger regions due to stronger association between alleles. In regions of high recombination, footprint of linked selection is small due to weaker association between alleles. Thus, recombination is a crucial driver of nucleotide diversity* patterns found in nature. For example, a positive relationship

*Nucleotide diversity – a measure of genetic variation in a population
has been observed between recombination rate and nucleotide diversity in several species, such as flies (*D. melanogaster*) (Figure 3), humans (*H. sapiens*) and yeast (*Saccharomyces cerevisiae*) (Mackay et al. 2012; Lohmueller et al. 2011; Cutter & Moses 2011). Interestingly in these species, low recombination rates in centromeric regions is observed. Understanding the relationship between the centromere and recombination rate is crucial as it affects the amount of standing genetic variation in the population and therefore, has important implications for the evolution of the species (Barrett & Schluter 2008). Studies that have investigated this dynamic have been primarily conducted in species with monocentric* chromosomes.

Conversely, not much is known about the relationship between recombination rate and nucleotide diversity in species that have holocentric** chromosomes. Holocentric chromosomes are found in diverse taxa, such as insects, plants, arachnids and nematodes, and present an opportunity to learn more about the effect centromeres, or its lack thereof, have on the recombination landscape of the genome (Melters et al. 2012). This will also allow us to explain the patterns of genetic diversity found in these species. If one extends the results from monocentric chromosomes one would predict that holocentric chromosomes would show a consistent recombination rate across all but the telomeric*** regions in the chromosome, and that the nucleotide diversity will exhibit a similar pattern. In paper II, we tested this prediction in three holocentric species with access to high-quality genomic resources and linkage maps: silk moth (*Bombyx mori*), wild silk moth (*Bombyx mandarina*), and green-veined white (*Pieris napi*) (Shimomura et al. 2009; Hill et al. 2017; Xia et al. 2009; Yamamoto et al. 2008; The International Silkworm Genome Consortium 2008).

---

*monocentric chromosomes – chromosomes with a single centromeres

**holocentric chromosomes – chromosomes without centromeres

***telomeres – sections of the repetitive nucleotide sequences found at each end of the chromosomes

---

**Fig. 3.** Recombination rate and nucleotide diversity in *Drosophila melanogaster*. Red points in the plot indicate nucleotide diversity and solid curves shows the recombination rate. Image taken from Mackay et al. 2012.
Parallelism in convergent phenotypes

Is evolution predictable? And if yes, at what timescale does this predictability hold? These are two longstanding questions in evolutionary biology. To answer this question, it is imperative to identify the causal variation behind similar traits among taxa that are sufficiently divergent for phenotypes to have arisen independently. Numerous studies have investigated the potential parallelism* of convergent** melanic phenotypes at the genetic level across diverse taxa, making this an excellent phenotype for assessing predictability. For instance, in mammals, selection appears to primarily target melanin biosynthesis genes such as \textit{mc1r} and \textit{agouti} to generate melanic morphs (McRobie et al. 2009; Eizirik et al. 2003; Steiner et al. 2007; Kingsley et al. 2009). However, caution has to be observed while interpreting these results as this perspective may be result of the extensive candidate gene studies looking for this effect.

Emerging genomic level studies focused upon melanic morphs among invertebrates allows us overcome this limitation and identify the causal variation in a global context. Previous studies in the peppered moth (\textit{Biston betularia}) and the Numata longwing (\textit{Heliconius numata}) butterflies show that melanic morphs can arise as a result of variation in patterning genes (e.g. \textit{cortex}) (Nadeau et al. 2016; Hof et al. 2016). In paper III, we have focused on the genetic basis of the female-limited dark melanic phenotype observed in a subspecies of \textit{P. napi}, \textit{P. napi adalwinda} as this allows us to explore whether this convergent phenotype with \textit{B. betularia} is the result of parallel selection on the same set of genes (Figure 4).

\textbf{Fig.4.} Melanic and non-melanic phenotypes of A) \textit{Biston betularia f. carbonaria}* B) \textit{Biston betularia f. typica} * C) \textit{Pieris napi adalwinda} female** D) \textit{Pieris napi napi} female**

*© Chiswick Chap/Wikimedia Commons / CC-BY-SA-3.0 / GFDL  
**Eliasson et al. 2005

*Parallelism – is defined as the use of a shared mechanism in the evolution of similar phenotypes  
**Convergence – evolution of similar phenotypes irrespective of evolutionary divergence or genetic mechanism
Role of cis-regulatory variation in local adaptation

Genetic variation in the functional parts of the genome can either affect the amino acids of the protein made, or it can change the levels of gene expression. If we detect signatures of selection associated with amino acid changes in exons, they are easy to interpret along with tests such as McDonald-Kreitman for understanding these effects (McDonald & Kreitman 1991). If the causal variation is present outside exons, we typically use genomic divergence estimates to identify such variation. However, a main limitation of this method is that it is challenging to assign a functional relevance to the identified variation. A way forward is to assess relative expression of alleles, i.e., allele specific expression (ASE), within a diploid individual. In a diploid individual there are two copies of each gene, and if these two copies vary in their cis-regulatory sequence it will result in an increased or decreased expression of one of the alleles during transcription. By assessing ASE in an F1 hybrid who is an offspring of individuals from two different populations, we can study the functional relevance of cis-regulatory variation segregating between two populations.

In paper IV, in order to investigate the role of cis-regulatory changes in local adaptation, we have analyzed ASE in the F1 hybrids of *P. n. napi* and *P. n. adalwinda*. These two subspecies have distinct wing color and diapause phenotypes, with *P. n. napi* exhibiting non-melanic and facultative diapause phenotypes, and *P. n. adalwinda* exhibiting a melanic and obligatory diapause phenotypes (Pruisscher et al. 2017; Lorkovic 1962; Petersen 1949). Additionally, these populations occupy niches that differ in several biotic and abiotic factors. For example, *P. n. napi* is primarily found in low altitude and mid to low latitude regions characterized by warmer climates, whereas *P. n. adalwinda* is found in high latitudes dominated by colder climates. Given these differences between these two subspecies, understanding the role of cis-regulatory variation within and between these populations of *P. napi* can help elucidate the genomic mechanisms that underpin adaptation and speciation.
Methods

In paper I, we present a pipeline called MESPA that can generate high-quality gene models from fragmented genome assemblies. Spaln2, a splice aware aligner of protein to genome, is the backbone of this pipeline (Iwata & Gotoh 2012). Spaln2 takes a genome assembly of the target species and a set of protein sequences from a related species, and aligns the proteins, either partially or completely, to multiple DNA scaffolds of the assembly. MESPA uses the alignment information from Spaln2 to scaffold contigs containing exons for the same protein, in correct order and orientation. This pipeline was validated using a de novo assembled genome from Pool-Seq data of a single wild population sample of D. melanogaster, and a protein dataset from D. virilis.

Paper I also reinvestigated the genetic basis of color polymorphisms in a butterfly Papilio polytes, and in a young European crow species pair, carrion crows (Corvus corone corone) and hooded crows (Corvus corone cornix). To do this, we first assembled a genome from Pool-Seq data for one color morph from each study (P. polytes for Papilio, and grey and black C. c. cornix for Corvus). Gene models were constructed with MESPA using the proteomes of the Silkmoth (Bombyx mori) and the Zebra finch (Taeniopygia guttata) as references against the Papilio and Corvus assemblies, respectively. In both cases, MESPA generated scaffolds were combined with the remaining contigs of the assembly, to create the final assembly that was used to obtain genetic divergence (FST) estimates. We then identified contigs in the published high-quality genome that are homologous to the highly differentiated contigs to validate results from our analysis using BLAST (Camacho et al. 2009).

In paper II, we first estimated recombination rate for each B. mori and P. napi chromosome, by fitting a third degree polynomial model to the genetic distance (in centimorgan or cM) as a function of chromosomal position. We then aligned the resequencing data from 11 B. mandarina and 29 B. mori individuals to the B. mori reference genome, and Pool-Seq data from 24 P. napi individuals to the P. napi reference genome. From the resulting alignments we estimated neutral nucleotide diversity (π), synonymous nucleotide diversity (πSS) and non-synonymous nucleotide diversity (πNS). To get the repeat content, repeat regions in the genome were identified and the proportion of repeat bases in each non-overlapping 100kb window was calculated. Similarly, to calculate GC content and CDS bases proportion, the number of GC bases and the sum of the lengths of all CDS that had their midpoint in a given non-overlapping 100kb window was divided by 100,000.

\*cM – 1cM is defined as the distance between two positions on the chromosome for which the expected average number of recombination events in a single generation is 0.01.
In paper III, we used data from one male and two female informative crosses, reared from the melanic *P. n. adalwinda* and the non-melanic *P. n. napi*, in a bulk segregant analysis* to map the locus carrying the dark adalwinda allele with high confidence. In order to further narrow in on the candidate region using an independent approach, we conducted: a SNP association study using genome re-sequencing data from 6 non-melanic *P. n. napi* individuals and 6 melanic *P. n. adalwinda* individuals, and a genomic divergence study using Pool-Seq samples from a *P. n. adalwinda* population from Abisko in Sweden and a *P. n. napi* population from Barcelona in Spain.

In paper IV, we generated tissue-specific transcriptomes (head, fatbody, gut, and the integument) of F1 hybrids, from both males and females, reared under directly developing and diapausing conditions. We also generated whole genome resequencing data from the F1 hybrids and the parents, to allow for unbiased estimates of segregating SNPs and ASE identification. ASE in the transcriptomes was assessed using a Bayesian method (Skelly et al. 2011), accounting for technical variation in allelic counts using whole-genome resequencing data for each F1. By intersecting our results with their gene functional pathways, we have identified an enrichment of the genes involved in energy metabolism, both mitochondrial and nuclear, suggesting this variation is potentially involved in local adaptation. In addition to gene functional pathways, we also intersected our results from the ASE analysis with genetic divergence estimates (*F*<sub>ST</sub>) to identify candidate genes involved in local adaptation.

---

*bulk segregant analysis – a technique that involves forming two groups that show opposing phenotypes for the trait of interest*
Major findings

Paper I

Validation of MESPA

In D. melanogaster, MESPA was able to make high quality gene models for nearly all the single copy ortholog genes: 94% of the genes were found and assembled at ≥ 80% of their length and with ≥ 95% accuracy. When considering all genes identified by MESPA, including those that had < 80% of their length assembled, MESPA constructed gene models for 99% of all the genes with high accuracy (Figure 5).

Papilio butterfly genomic scan

We identified 8 scaffolds, containing 111 SNPs, with high genetic divergence (FST >= 0.9) in the final assembly. A BLAST search against the recently published P. polytes genome revealed that 4 of these 8 scaffolds mapped to the “mimetic” locus and another 2 scaffolds mapped to the “non-mimetic locus” (described in (Nishikawa et al. 2015)). An additional scaffold (28kb long) had its best hit in a different contig of the genome, but upon closer inspection this scaffold was determined to be a chimera containing a part of the mimetic locus in which all the high FST SNPs were located. Thus, in sum, we find 7 out of 8 scaffolds to be in the core region identified in the original genomic study from which the data were derived (Kunte et al. 2014), which is further validation of our approach. Within the sole scaffold that did not map to the dsx region, one protein was identified: FAS-associated factor 1-like. Unfortunately, we are unable to assess whether FAS-1 and dsx are on the same chromosome, as we can neither find both loci on a single scaffold in any Lepidopteran genome nor can we find both genes in the genome of Lepidopteran genomes that have chromosomal level information (e.g. Bombyx mori, Melitaea cinxia).

Fig. 5. Results showing the performance and accuracy of each gene constructed using MESPA (n = 3,718). Each dot is a predicted D. melanogaster protein generated by MESPA, using as inputs single-copy orthologs from D. virilis and a D. melanogaster genome generated from Pool-Seq data. Coverage = length of MESPA protein/published protein length, Identity = the % identity of those two protein sequences.
**Corvus crow genomic scan**

We identified 30 scaffolds, containing 235 SNPs, with high genetic divergence ($F_{ST} > 0.9$) in the final assembly. Using BLAST, we compared the identified scaffolds back to the published *Corvus* genome, finding that all 30 scaffolds were contained within the high $F_{ST}$ region previously identified as giving rise to the color-mediated pre-zygotic isolation between the two *Corvus* semi-species (i.e. scaffolds 68 & 70 of the genome from (Poelstra et al. 2014)). These results provide a powerful validation of our approach.

**Paper II**

We observed a positive correlation between $\pi$ and recombination rate across 100 kbp windows in the autosomes of *B. mori* (Spearman’s rho = 0.2558, P-value < 2.2e-16), *B. mandarina* (Spearman’s rho = 0.3843, P-value < 2.2e-16), *P. napi* (Spearman’s rho = 0.2814, P-value = 9.896e-12). We also observed a negative correlation between $\pi$ and the proportion of CDS bases within each window in the autosomes of both *B. mori* (Spearman’s rho = -0.31, P-value < 2.2e-16), *B. mandarina* (Spearman’s rho = -0.44, P-value < 2.2e-16), and *P. napi* (Spearman’s rho = -0.4954, P-value < 2.2e-16). While we found a few other parameters showing a significant correlation with recombination rates, they were not consistent in the direction of significance across all the species used in the analysis. We also observed that approximately 75% of *B. mori* genome and 90% of *P. napi* genome have high rates of recombination (>2cM/Mb) in comparison to approximately 65% in *D. melanogaster* (Fiston-Lavier et al. 2010) (Figure 6).

![Fig 6](image-url)  
*Fig 6.* Recombination rate density plot of *D. melanogaster*, *B. mori*, and *P. napi*. Dotted lines from left indicate the mean percentile threshold, respectively.
Paper III

Using data from bulk segregant analysis on DNA pools from one male and two male informative crosses, we identified a 3Mb region in the 3’ end of chromosome 17 in the *P. napi* genome as our candidate region harboring the adalwinda locus (Figure 7). Using genome resequencing data from 6 *P. n. adalwinda* and 6 *P. n. napi* individuals, we narrowed our candidate region from a 3Mb region to a 20kb window within that region (Figure 7). The 20kb candidate region is 50kb upstream of the gene cortex, and 135 kb downstream of the gene poly(A)-specific ribonuclease. Further investigation into the nature of variation in this region identified fragments of transposable elements insertions. Genomic divergence study using Pool-Seq data from the population samples of *P. n. adalwinda* and *P. n. napi* failed to identify regions of significantly diverged regions.

**Fig. 7. Summary of results from the bulk segregant analysis and the SNP association study.** A) Female informative cross design yielding melanic and non-melanic phenotypes. B) Distribution showing the proportion of melanin pigments among the females of family 206. C) Histogram showing the number of SNPs satisfying Mendelian expectations in family 206 across the genome. D) Male informative cross design yielding melanic and non-melanic phenotypes. E) Distribution showing the proportion of melanin pigments among the females of family 23. F) Histogram showing the number of SNPs satisfying Mendelian expectations within chromosome 17 in family 23. G) Design of the SNP association study. H) Histogram showing the number of significant SNPs within the terminal 3 Mb region of chromosome 17 in 20 kb bins. Genes flanking the window with the most SNPs are highlighted.
Paper IV

On average we assessed 3,555 genes per sex, treatment, and tissue (STT) combination for the presence (or lack thereof) of ASE. Our results showed that a vast majority of genes show only moderate levels of ASE (Figure 8). Looking across all experimental groups, significant GO terms associated with molecular function (P<0.05) identified the terms “oxidoreductase activity” and “catalytic activity” as the most common (n=14), followed by “magnesium ion binding” (n=13). We also generated KEGG pathway annotations for genes that showed ASE in at least 8 STT combinations, and placed them within the known metabolic map for Pieris rapae. Of interest are genes in Oxidative Phosphorylation pathway, as these are centrally involved in energy metabolism and possible mito-nuclear interactions (Figure 9). Finally, we observed that of the 4647 genes that showed ASE in at least one STT combination, 1024 genes had a high FST window (99-percentile threshold FST > 0.2456) at an average distance of approximately 14kb away from the gene (Figure 10). Of these 1024 genes, 10 genes showed ASE in all sex, treatment combinations from gut tissues; 7 genes showed ASE in all sex, treatment combinations from head tissues (Figure 11). Of these 17 genes, we found genes such as Opsin-3 and Chaoptin, which are involved in affecting vision and phototransduction, and might play a role in affecting circadian rhythm (Cavallari et al. 2011; Reinke et al. 1988).

---

**Fig. 8. Quality assessment of ASE analysis results.** Distribution of dispersion (A) and magnitude of ASE (B) across genes in female integument tissue under diapausing conditions. C) Scatterplot of magnitude of Dispersion vs. Magnitude, with genes with a posterior probability >= 0.95 highlighted in red. D) Scatterplot of posterior probability vs magnitude of ASE.
Fig. 9. High-confidence genes that appear in at least 8 groups visualized against the oxidative phosphorylation map of *Pieris rapae*. Boxes highlighted in red correspond to high-confidence genes.

Fig. 10. Scatterplot of Fst vs. Maximum absolute magnitude across all STT groups. Each dot represents a gene. Fst values used here is the maximum of all Fst estimates across 200 bp windows that occur near the gene. Absolute magnitude value is the maximum of absolute magnitude values across all STT groups.
**Fig. 11.** UpSetR plot showing the intersection of all genes that show ASE across treatment groups, and have a high $F_{ST}$ window in the vicinity. First, second, and third letter in the group ID corresponds to sex (F: female or M: male), treatment (D: direct developing or P: diapausing), tissue (H: head, F: fatbody, G: gut, or I: integument).
Concluding remarks

In this thesis we have approached the problem of evolution by trying to understand the nature of genetic variation behind adaptive phenotypes (papers I, III, & IV) and the factors influencing it such as recombination dynamics (paper II).

In paper I, we demonstrated the feasibility of conducting genome-wide scans in non-model organisms by using a Pool-Seq only approach. Conducting genomic scans using this approach is cheaper and quicker than the traditional approach of generating a high quality genome first. Our attempts to replicate finding from previous studies on *Papilio* and *Corvus* proved successful. While we have demonstrated encouraging results of using MESPA with the Pool-Seq only approach, there are some limitations. First, the contigs in low-quality assemblies are likely to be short, on the order of 10kb. This limits us from finding regulatory variation that are 10’s of kb away from the gene, and large regions of divergence, such as in the *Papilio* system, where the causal allele was found to be in a 300kb inverted region. Moreover, many such fragmented contigs are likely to lack genic content, and more often than not, do not lead to meaningful insights. Second, using Pool-Seq data does not contain individual level SNP information and as a consequence it is not possible to assess recombination dynamics over scales >1kb. Third, this approach is useful only when there is ongoing or recent geneflow between the two pools, such that only the regions giving rise to phenotypes of interest are significantly different. Finally, it is difficult to detect genomic regions that differ in structure (inversions) or in content (insertion or deletion polymorphism).

In paper II, I show that on average Lepidoptera have high rates of recombination across the vast majority of their genome relative to monocentric species. In monocentric species, positive correlations between recombination rate and nucleotide diversity are common, and since regions of reduced recombination can include > 10% of functional genes, a substantial fraction of the genome in these species have reduced evolutionary potential. In contrast, in the holokinetic chromosomes of *B. mori* and *P. napi* we find that recombination rate does not reduce to zero near the middle of the chromosomes. But, similar to monocentric species, these Lepidoptera have a positive correlation between recombination rate and nucleotide diversity. Furthermore, we also observed higher overall rates of recombination in these holokinetic species as compared to monocentric species. Taken together, our results suggest that holokinetic chromosomes are able to carry more standing genetic variation throughout the genome despite the negative effects of linked selection on genetic diversity.

In paper III, we identified a 20kb region, approximately 50 kb upstream of the gene cortex, associated with the dark morph of *P. n. adalwinda*. This 20kb (Eliasson et al. 2005) region is composed of transposable element fragments, which suggests a mechanism very similar to that found in the peppered moth, *B. betularia*, wherein a TE insertion located in the first intron of the gene cortex was directly associated with the black (carbonaria) form.
Interestingly, while both species appear to use the same gene and involve a TE insertion to generate its corresponding melanic phenotype, the spatial patterns and extent of melanization between the melanic forms in these species are markedly different (Figure 4). These results suggest that TE insertions altering the enhancer region of cortex might be a conserved mechanism for generating other dark melanic morphs in Lepidoptera.

In paper IV, we have attempted to assess the role and extent of gene regulation in local adaptation, by studying ASE in the F1 individuals of a cross between two subspecies of P. napi, P. n. napi and P. n. adalwinda, that occur in different niches. Our results suggest that a vast majority of the genes that show ASE in our study are involved in energy metabolism. We found genes involved in oxidative-phosphorylation which might be a result of co-adaptation between mitochondrial and nuclear genomes, and likely reflects the different energy requirements in their niche environments. We also found genes involved in lipid biosynthesis and purine metabolism, which might reflect other local selection pressures. Interesting we did not find any genes associated with circadian rhythm and melanin biosynthesis, which are two of their primary phenotypic differences between the two subspecies. We also show that intersecting genomic divergence estimates with ASE is a powerful approach to assign functional relevance to divergent regions in the regulatory parts of the genome.
References


Svensk sammanfattning


Med detta i åtanke försöker vi i den här avhandlingen att komma underfund med egenskaperna som präglar genetisk variation, och vilka faktorer som påverkar den, genom att:

- Ringa in vilka arvsmasseförändringar som ligger bakom gynnade fenotyper (kapitel I, III och IV)
- Undersöka de mekanismer som styr evolutionsprocessen (kapitel II).
I detalj är avhandlingens syfte följande:

1. Att genom användning av "Pool-Seq data" istället för omsekvansiering utveckla en snabbare och billigare metod för att identifiera arvsanlagsförändringar som ger upphov till gynnade fenotyper.

2. Att förstå centromerernas roll under rekombinationen genom att undersöka rekombinationsmönster hos arter som saknar centromerer (exempelvis *Bambyx mori*), och jämföra dessa med rekombinationsmönster hos *Drosophila melanogaster*, vars kromosomer har centromerer.

3. Undersöka de genetiska förutsättningarna för parallellism (oberoende uppkomst av gemensamma egenskaper) i förekommande fall av konvergenta (sammanstrålande) färgtäckningsmönster.

4. Undersöka vilken roll förändringar i *cis-regulativa* arvsanlag har för lokal anpassning.

*Many thanks to Rasmus Erlandsson for this translation & Simon Eckerström-Liedholm for his suggestions.*
Acknowledgements

There are many people who have I am indebted to for their contribution to my time as a PhD student. First and foremost, I would like to thank my supervisor Chris for bringing me to Stockholm and giving me this wonderful opportunity. Thank you so much for your guidance, for your efforts to improve my incoherent writing, and for pulling me out of endless procrastination. Not only have you taught me the fundamentals of conducting scientific research but you have also helped me grow as a person and given me the confidence to face life. A big thank you for that! I would also like to thank my co-supervisor Soren Nylin for always being available when I needed his help. Also thanks to the other members of my follow-up committee Jason Hill, Love Dalén, Linda Laikre, and Karl Gotthard for their great support and invaluable advice.

I would also like to thank my wonderful lab mates for their continued support through these years: Alyssa, Peter, Naomi, Jason, Maria, Christen, Lisa, and Kalle. Doing a PhD is certainly the craziest thing I have done in my life, and I could not have asked for better company to do it with. This thesis would not have been possible if it wasn’t for you guys; you all have offered me plenty of advice on my work and influenced it significantly. Alyssa, thank you for being such an amazing friend and my partner-in-crime at Zootis fikas; aah-aah-aah-chooo :D Peter and Naomi – thanks for all the nice conversations; you guys were always there to lend me your support when I needed it, be it at work or in my personal life. Jason – thank you for helping me identify and bridge gaps in my knowledge; I have learnt so much from you. Maria – thank you for doing all the heavy lifting in the lab.

I would also like to thank Alyssa, Rasmus, Christina, Kalle, Ale, and Diana for being my officemates, for varying amounts of time. A quick chit-chat shared with you guys every now and then has always been time well spent. I have also shared many great conversations with Sara, Lovisa, Mariana, Simon, Olle, Charel, Sandra, Ciaran, Philipp, Patricia, Olga, Meet, and Alexandra. Thank you to Siw, Anette, Minna, and Peter for helping me navigate all the red tape, thereby making my life infinitely easier. And en stor kram från mig to all the friendly faces who have made Zootis such a fun place to work.

Thank you to all the faculty at the Department of Biotech, PSG Tech, without whom I would have not been here. Thank you, Prof. Ramamurthy, Prof. Ananthasubramanian, Prof. Rani, Prof. Selvi Subramanian, Dr. Jayanthi S, Dr. Hema, and Dr. Vidyalakshmi for your able support and guidance. Thank you, Dr. Sabarinath Chandrasekharan, for your great supervision in my Bachelor’s project.

I have also been fortunate to have several mentors throughout my life, who have been my external eyes and ears. Vignesh anna, I do not have words to express my gratitude for everything you have done for me over the years. Krishna Kumar anna, I have plagued you with several emails, calls, and discussions, but you were always happy to indulge me and
give an unbiased perspective. Thank you very much for that! Nithya akka, thank you for being part of this journey; you have been an inspiration to me. Thanks, Vasanthan anna, for several discussions that helped me learn and understand genomics.

I would like to express my heartfelt gratitude for the "Bangalore bunch": Ashick, Abdulla, Ganesan, Habib, Kristen, Mohan, Rajesh, and Wisvesh. Bangalore has changed a lot since I moved out, but you guys have not: fun conversations, late night movies, coffee at Indian Coffee House – everything is the same. You guys are the link that connects me to my immediate past, and knowing that the connection is as strong today as it was six years ago means a lot to me.

I thank my friends in Stockholm, Lakshminarayanan, Ganesh L, Praveen, Paulina, Brandon, Saranya, Richard, Mita, Tino, Ryan, and Ganesh Phad for their love, support, and encouragement throughout. I thank them for being there, and making me laugh when I had almost forgotten to.

I would like to thank my family - who have no idea what I am doing, but nevertheless being proud of me. Amma thank you for encouraging me to do what I wanted during my studies and for not convincing me to work in IT like a million other engineers in India. Thank you!

Last, but by no means the least, I would like to thank my love, Swati, for sharing this experience with me. Your constant support and encouragement have been fundamental to finishing this thesis. I know seeing your partner for ten days in a year is not anyone’s idea of a perfect relationship, but thanks for doing this with me. Looking forward to our adventures together.