UPPSALA
UNIVERSITET

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 2225

Rates and patterns of genome evolution in Lepidoptera

KARIN NÄSVALL

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2023

ISSN 1651-6214
ISBN 978-91-513-1678-9
URN urn:nbn:se:uu:diva-490587

Dissertation presented at Uppsala University to be publicly examined in Zootis-salen, EBC, Norbyvägen 14, 752 36 Uppsala, Friday, 10 February 2023 at 10:15 for the degree of Doctor of Philosophy. The examination will be conducted in English. Faculty examiner: Dr. Simon Martin (Institute of Evolutionary Biology, University of Edinburgh).

Abstract
Näsvall, K. 2023. Rates and patterns of genome evolution in Lepidoptera. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 2225. 73 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-513-1678-9.

The evolutionary consequences of genome restructuring and chromosomal rearrangements can be significant, but the underlying mechanisms are poorly understood. This thesis explores the relationships between genome restructuring, recombination, and maintenance of genetic diversity in different butterfly species. I also investigate the genetic underpinnings of different types of adaptations in two butterfly species with distinct life-history characteristics and adaptations. In Chapter I, whole genome alignments and linkage maps were used to characterize inter- and intra-specific chromosomal rearrangements in wood white butterflies (*Leptidea spp.*), revealing extensive reorganization of the chromosomes predominantly driven by fusions and fissions. Several fusion-fission polymorphisms were found segregating within populations, confirming the dynamic and ongoing process of karyotype evolution. In Chapter II, I used the linkage map information to show that chromosomal rearrangements have had considerable effects on the recombination landscape and maintenance of genetic diversity. Chapter III presents a detailed annotation of the genome of a long-distance migrant, the painted lady butterfly (*Vanessa cardui*). The annotation was used to identify expanded gene families, providing insight into the genetic underpinnings of this unique life-history. Here, I also developed linkage maps confirming a well-preserved karyotype in this species and showed that chromosome size is a major determinant of the recombination landscape. In Chapter IV, I explored the relationship between nucleotide composition, codon usage, and substitution rates across multiple Lepidoptera species, in an attempt to disentangle the relative effects of natural selection and neutral evolutionary forces on gene sequence evolution. The final two chapters focus on understanding the genetic basis of key adaptive traits in butterflies. Specifically, in Chapter V, I investigated potential local adaption in *Leptidea sinapis* by studying how host plant switch is associated with oviposition rates, larval growth and development, gene expression, and microbiome composition in populations experiencing different environmental conditions. In Chapter VI, I assessed the differential activation of regulatory elements in *V. cardui* females in response to host plant availability, which together with the gene family analysis in Chapter III, resulted in a set of candidate genes potentially associated with a migratory lifestyle. In conclusion, I have explored the complex interplay between proximate mechanisms and evolutionary forces shaping the genome structure and levels of genetic variation. I also investigated the genetic and regulatory underpinnings of adaptive traits in different butterfly species. In addition, I provided resources including chromosome-level genome assemblies, recombination maps, and annotations that will contribute to our understanding of evolutionary processes in general.

Keywords: Chromosomal rearrangements, karyotype evolution, recombination rate, Lepidoptera, genome structure, molecular evolution

Karin Näsvall, Department of Ecology and Genetics, Evolutionary Biology, Norbyvägen 18D, Uppsala University, SE-75236 Uppsala, Sweden.

ISSN 1651-6214
ISBN 978-91-513-1678-9
URN urn:nbn:se:uu:diva-490587 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-490587)

To curiosity

List of Papers

This thesis is based on the following papers. Chapters:

I. Höök, L.*, **Näsvall, K**.*, Vila, R., Wiklund, C., and Backström, N. (2022) High-density linkage maps and chromosome level genome assemblies unveil direction and frequency of extensive structural rearrangements in wood white butterflies (*Leptidea* spp.). bioRxiv doi.org/10.1101/2022.10.10.510802. *Accepted. Chromosome Res.*

II. **Näsvall, K**., Boman, J., Höök, L., Backström, N. (2022) Nascent evolution of recombination rate differences through chromosomal rearrangements. *Manuscript*

III. Shipilina, D.*, **Näsvall, K.*,** Höök, L., Vila, R., Talavera, G., Backström, N. (2022) Linkage mapping and genome annotation give novel insights into gene family expansions and regional recombination rate variation in the painted lady (*Vanessa cardui*) butterfly. *Genomics*, 114(6):11048

IV. **Näsvall, K.,** Boman, J., Talla, V., Backström, N. (2022) Base composition, codon usage and patterns of gene sequence evolution in butterflies. *Manuscript, in review.*

V. **Näsvall, K**., Wiklund, C., Mrazek, V., Künstner, A., Talla, V., Busch, H., Vila, R., and Backström, N. (2021) Host plant diet affects growth and induces altered gene expression and microbiome composition in the wood white (*Leptidea sinapis*) butterfly. *Molecular Ecology*. 30:499–516.

VI. **Näsvall, K.,** Shipilina, D., Vila, R., Talavera, G., Backström, N. (2022) Activity profiles of regulatory elements and associations with the oogenesis-flight syndrome in a long-distance butterfly migrant. *Manuscript.*

Reprints were made with permission from the respective publishers. * Equal contribution.

Other projects not included in thesis

Lohse K, Höök L, **Näsvall K**, Backström, N. (2022) The genome sequence of the wood white butterfly, *Leptidea sinapis* (Linnaeus, 1758) [version 1; peer review: awaiting peer review]. *Wellcome Open Res.* 2022, **7**:254 (https://doi.org/10.12688/wellcomeopenres.18118.1)

Palahí i Torres A, Höök L, **Näsvall** K, Shipilina D, Wiklund C, Vila R, Pruisscher P, Backström N. (2022) The fine-scale recombination rate variation and associations with genomic features in a butterfly. *Manuscript, in review.* bioRxiv 2022.11.02.514807; doi: https://doi.org/10.1101/2022.11.02.514807

Contents

Abbreviations

DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
SW	Mutation from G/C to A/T (strong to weak)
WS	Mutation from A/T to G/C (weak to strong)
gBGC	GC-biased gene conversion
LD	Linkage disequilibrium
CO	Crossover (during recombination)
NCO	Non-crossover
DSB	Double strand break
TSS	Transcription start site
TTS	Transcription termination site
UTR	Untranslated region

Introduction

Variation is the foundation for evolutionary change and a prerequisite for the rich biodiversity we are a part of today. Genetic variation emerges through mutations, as exchange, insertions or deletions involving everything from single bases, through gene duplications and rearrangements of chromosomes, to whole genome duplications. Mutations provide the raw material for change, but as important is the rearrangement of existing variation for diversification among offspring and disconnecting mutations with opposing effects. Diversity is necessary because not only are the genomes changing, the world around us constantly produces new challenges that require different abilities compared to our ancestors. Evolutionary genetics is the study of changes in the genome and their reflection in the evolution of organisms, by using the patterns of genetic variation within populations and among species to infer evolutionary processes.

In this thesis, I have explored the genetic structure and reshuffling on different levels in the genomes of two butterfly species. Essential steps for genomic analysis are the development of high contiguity assemblies (preferably at chromosome level) in the species of interest and obtaining a thorough annotation of genomic features. The linkage and reshuffling of genetic material are central factors in many evolutionary processes acting on the genome, and information thereof is crucial for correct inferences of these processes. First, we generated annotated chromosome-level genome assemblies of three species of wood whites (*Leptidea*) and linkage maps in two populations of the common wood white (*L. sinapis*). These genomic tools were combined to characterise the extent of karyotype rearrangements, including the rate and direction of chromosomal rearrangements. Furthermore, we investigated genomic features associated with chromosomal breakpoints to better understand the causes and consequences of karyotype instability. The recombination landscape was characterised using linkage map information to increase our understanding of the relationship between rearrangements, reshuffling of genetic variation and genetic diversity.

We produced a thorough annotation and inferred gene family expansions in the painted lady butterfly (*Vanessa cardui*) to identify potential candidates involved in the extraordinary life-history of this long-distant migrant. In addition, linkage maps from pedigrees allowed us to characterise the relationships between recombination rate and genomic features in this species.

Next, I used comparative genomics to investigate the complex relationship between substitution rates and nucleotide composition across Lepidoptera. Additionally, population studies and gene copy variation of tRNAs in the common wood white butterfly were used to disentangle the effects of processes like GC-biased gene conversion and codon usage.

Finally, I focus on understanding the genetic basis of key adaptive traits involved in response to different environmental conditions these butterflies may encounter. In the common wood white, we investigated potential local adaption in female host plant preferences, oviposition, larval growth and development, including differences in gene expression and microbiome composition. In the painted lady, we assessed differential activation of regulatory elements in adult females in response to hostplant availability, as part of a larger project to understand the genetic underpinnings of the migratory behaviour in this butterfly.

In summary, I have explored the relationship between proximate mechanisms and evolutionary forces influencing the genome structure and the genetic variation. Additionally, we have provided resources in the form of chromosome-level genomes, linkage maps, recombination maps and annotations. Resources that will contribute to further genomic studies, increasing our understanding of evolutionary patterns and processes.

The genome

Organisation

The chromosomes are the highest level of organisation of the DNA and serve three important aspects; to provide an environment for compaction of the vast amount of DNA contained in the nucleus, but at the same time allow transcriptional and replication activity; to ensure an equal assortment of genetic material into gametes; and, finally, to maintain DNA integrity across generations.

Chromosomes harbour a complex system of integrated support molecules and interacting DNA-binding factors in addition to DNA, and together this constitutes the chromatin (Klug et al., 2017). The primary building blocks of chromatin are the histones, proteins that form octamers around which the DNA is coiled forming nucleosomes. The nucleosomes are efficiently packing the DNA into secondary loops while simultaneously having a dynamic function varying the amount of compaction and thereby regulating the accessibility of the DNA. Heterochromatin, which is compact, tightly packed DNA, is harder to access for transcription factors and the replication machinery than open chromatin (euchromatin). Heterochromatin is one way to silence parts of the DNA from active expression. In addition, compacted DNA is less susceptible to double-strand breaks (DSBs) (Lensing et al., 2016). The tethering of DNA to the nucleosomes is mainly controlled by enzymes catalysing

modifications of the N-terminal amino acid tails of the histones. Histone tail modifications are, together with DNA-methylation, major epigenetic markers essential in chromatin regulation. The chromatin also regulates the three-dimensional structure of the DNA within the nucleus, directing its localisation in chromosomal territories, with varying amounts of intra- and interchromosomal interactions between different regulatory elements and coding sequences.

Diploid organisms have two sets of homologous chromosomes, one paternal and one maternal. The amount of genetic information is maintained during sexual reproduction through reduction to a haploid state (one chromosomal set per gamete) by two rounds of cell division during meiosis (Klug et al., 2017). The chromosomes replicate into two copies (sister chromatids), pair up with the homologous chromosome in prophase I and exchange part of their genetic material in the homologous recombination process (will be discussed later). In metaphase I, the chromosomes line up at the centre of the cell and are attached to the meiotic spindle fibres stretching from the poles of the cell. The spindle fibres connect to the chromosomes at designated locations, the centromeres. The centromeres are parts of the chromosome that contain specific centromeric histones, ensuring the attachment of the spindle fibres by recruiting the kinetochore protein complex (Quénet & Dalal, 2012). The DNA sequences in the vicinity of the centromere are often heterochromatic and enriched in repetitive sequences. Chromosomes can have one localised centromere, called monocentric chromosomes, or dispersed centromere activity along the chromosome, referred to as holocentric or holokinetic chromosomes (Bauer, 1967). The spindle fibres separate the maternal and paternal chromosomes by drawing them to opposite cell poles. The separation, haploidisation, begins in anaphase I and completes in telophase I when the new cell membrane starts to form. The next phase is the sister chromatid separation, in which metaphase, anaphase and telophase are repeated, but this time the sister chromatids are separated. A balanced segregation should result in one copy of each chromosome in the new gametes.

A problem with DNA maintenance in linear chromosomes is the risk of information loss or damage at the ends. The telomeres are specific repeat-rich sequences that protect the ends of the double helix from damage. The length of the telomeres is usually maintained through telomerase, a specific reverse transcriptase that uses RNA guide sequences as templates (Klug et al., 2017).

Despite the vital role of the chromosomes, they are not static entities but are subject to evolutionary change. The karyotype, i.e. the size and number of chromosomes, is relatively stable within lineages and between closely related species. Most eukaryote taxa have a haploid chromosome number within the range of n=10 – 50 (Stapley et al., 2017). That said, exceptions are demonstrating the potential for a remarkable variation in karyotype among organisms (de Vos et al., 2020; Stapley et al., 2017). For example, the ant *Myrmecia pilosula* with a haploid chromosome number of n=1 (Crosland & Crozier,

1986) and the lycaenid butterfly *Polyommatus atlanticus* with n=226 (V. Lukhtanov, 2015), at present, the eukaryote with the highest number of chromosomes (not counting polyploids). The meiotic machinery is strictly regulated, and loss or gain of chromosomal parts is often lethal. However, the patterns of karyotype variation in different lineages show that karyotype evolution is happening, so individual variation in the number and size of chromosomes cannot always be severely deleterious. The mechanisms behind either maintaining chromosomal integrity over long evolutionary timescales or rapid chromosomal evolution in some clades remain unknown and will be explored in Chapter I.

Genomic features and regulation

Genes

The genes are the blueprint for protein synthesis and comprise several functional parts that all can be under evolutionary constraints. Mutations in the coding regions can truncate or alter the structural properties of the protein product, while mutations in the regulatory regions can result in an increase or decrease in gene products or shifts in temporal and spatial distributions of gene expression. Knowledge of the structure of the gene sequences is important for understanding how changes in different parts can have very different evolutionary outcomes.

The leader sequence, or the 5′ untranslated region (5′UTR), starts at the transcription start site (TSS) and ends before the first initiation codon of the protein. Major features of the 5′UTR include the 5′cap that protects the mRNA from degradation, sequences regulating the secondary structure of the mRNA and regulatory binding sites (Barrett et al., 2012). The exons correspond to the sequences maintained in the mature mRNA after posttranscriptional modifications and contain the code for the protein. The genetic code is based on codons (a combination of three nucleotides), where each codon represents a specific amino acid. There are 20 different amino acids used in various configurations for protein synthesis, but combining the four nucleotides into trios leads to 64 options. With several codons corresponding to the same amino acid, there is redundancy in the code. Most of the redundancy lies in the third codon position, where many amino acids remain the same regardless of nucleotide (4-fold redundant or 4-fold degenerate). Sites in codons where two or three nucleotides are interchangeable are referred to as 2-fold and 3-fold redundant, respectively. Since mutations in redundant sites are not affected by selection for protein structure, they are assumed to be neutral and are often referred to as synonymous or silent mutations. Exons are under strong evolutionary constraints but are interspersed with introns, non-coding sequences that are enzymatically removed from the mRNA before protein synthesis in a process called mRNA-splicing. The introns often contain regulatory

sequences and different types of non-coding RNA, but also unknown, repetitive DNA. The last part of the gene is the 3´UTR, following directly after the translation termination codon, and this region is important for the polyadenylation of the mRNA and the translation initiation. The 3`UTR also contains binding sites for regulatory proteins and non-coding RNA (Pichon et al., 2012).

Gene regulation

A complex interaction of different *cis*-acting regulatory elements and *trans*-acting factors governs the activity of the genes. The *cis*-acting regulatory elements are sequences located in the vicinity of the gene, while *trans*-acting regulators usually are transcription factors or other regulatory molecules (Klug et al., 2017). They form together with chromatin regulators, non-coding RNAs and cofactors, an intricate and precise regulation of gene expression.

Promoters are regulatory *cis*-elements containing the binding sites for transcription factors that form the RNA-polymerase II transcription initiation complex (Narlikar & Ovcharenko, 2009; Nelson & Wardle, 2013; Rojano et al., 2019). Core promoters are short sequences, 100-1,000bp long, located within 100 bp of the TSS, sometimes containing specific sequences like the initiator motif (Inf) or the TATA-box (Haberle & Stark, 2018). Enhancers are another category of *cis*-acting regulatory elements involved in fine-tuning tissue- or cell-specific gene expression (Heintzman et al., 2009). They can be located very distantly, both upstream or downstream, from the target gene TSS (H. K. Long et al., 2016). The activity of promoters and enhancers is associated with specific tail modifications of the incorporated histones. For example, acetylation of lysine 27 (H3K27ac) reduces the positive charge of the histone tail, hence its affinity to DNA, thereby increasing the accessibility for transcription factors (S. Zhang et al., 2022). In addition, histone tail modifications can recruit or block transcription factors or other DNA-interacting molecules, although the exact mechanism is not yet fully known (Howe et al., 2017). Other distinguishing modifications are the trimethylation of lysine 4 (H3K4me3), which is strongly associated with active promoters and increased transcription (Heintzman et al., 2007; Santos-Rosa et al., 2002), especially together with H3K27ac (W. Zhao et al., 2021). In contrast, enhancers are enriched in mono-methylation of lysine 4 (H3K4me1) but require additional enrichment of H3K27ac for active transcription (Heintzman et al., 2009). Inactive, so-called poised enhancers in tissues and cell types where transcription of the specific gene is low are characterised by the absence of H3K27ac and instead show enrichment of the silencing marker H3K27me3 (Creyghton et al., 2010; Rada-Iglesias et al., 2011). The patterns of histone mark enrichment give a signature of the character and activity of different regulatory elements and the transcriptional activity of associated genes. Promoters and enhancers are important targets of evolution, likely more so than *trans*-acting regulators (Metzger et al., 2017; Verta & Jones, 2019). Mutations in the binding motif

for transcription factors or sequence copy number variants can influence the timing and magnitude of gene expression, directly affecting the phenotype (McGregor et al., 2007). Gene expression differences can also evolve rapidly between different populations in diverse environments (Hamann et al., 2021).

Posttranscriptional regulation encompasses the steps after the gene has been expressed. These steps are moderators of the amount of mRNA available for translation and protein synthesis and include mRNA-splicing and degradation (Klug et al., 2017). The level of the final protein product is also regulated by translational efficacy, protein folding and degradation. Alterations at any of these steps can lead to structural modifications or change the amount of the product and subsequently affect the phenotype.

The non-coding regulatory RNAs, which are transcribed sequences not translated to proteins, constitute a versatile group of regulatory factors. They are involved in gene regulation on many levels, including transcriptional regulation, alternative splicing, translation and degradation of mRNA. They are sparsely characterised, and annotations are lacking in many taxa outside of model organisms.

Transposable elements

Transposable elements (TE) are a diverse group of "selfish", repetitive sequences. TEs are either capable of self-replication or utilisation of the host machinery for proliferation within the genome, and their transposition activities are, by definition, causing structural variation. (Kidwell & Lisch, 1997; McClintock, 1956). The high sequence similarity in copies of TEs in different locations of the genome can induce mistakes in the DNA-repair machinery, so-called non-allelic (ectopic) homologous recombination (Shaw & Lupski, 2004). Both deletions and insertions of the host sequences can be mediated by TEs, as well as more extensive rearrangements (Deininger & Batzer, 1999; Hedges & Deininger, 2007). The proliferation of TEs can be substantial, and they constitute a large part of the total genome size in some organisms (Lander et al., 2001; Talla et al., 2017). The activity is deleterious mainly because proliferation can disrupt functional genes or regulatory elements. However, the domestication of TEs has resulted in important evolutionary innovations, either through effects on gene regulation or by recruiting the function of the TEs (Kazazian, 2004; Van't Hof et al., 2016). The primary defence mechanisms against transposon activity are transcriptional silencing through DNA-methylation or silencing histone marks (mainly H3K9me3 and H3k27me) and post-transcriptional degradation (Luo et al., 2020; Shalgi et al., 2010; Walter et al., 2016). These defences are mediated by different classes of small non-coding RNAs. The evolutionary arms race between TEs and the host genome defence may have contributed to innovations of mechanisms for gene regulation, especially the regulatory RNAs (Jordan & Miller, 2009). The TEs are also degraded by mutations disrupting their ability to transposition, and the evolutionary distance between transposon activity bursts can be determined by their

number of nucleotide substitutions (Kimura, 1980). The uncharacterised genome, often referred to as intergenic DNA, likely consists of a large part of degraded TEs. The relationship between TEs and rearrangements led us to investigate the enrichment of specific TE classes in the vicinity of breakpoints in the wood whites. We also explored the association between TEs and recombination rate in the painted lady.

Genetic variation

Point mutations

Mutations are the fundamental source of genetic variation. Point mutations are single nucleotide changes that entail the replacement of any of the nucleotide bases, usually caused by errors by the DNA-replication machinery during cell division. Several proofreading and repair mechanisms have evolved to increase the fidelity of the copying, but as millions and millions of bases are to be copied, even a minimal error rate results in a considerable amount of mutations. Other mutagenic factors can be external influences like oxidative damage, radiation exposure, and thermal treatment. The mutation rate for single nucleotides can be estimated either by deep sequencing of parent-offspring trios or by mutation accumulation lines. The mutation rate has been estimated to 2.0 - 3.5 $\times 10^{-9}$ per site per generation in invertebrates (Denver et al., 2009; Keightley et al., 2009, 2015), and slightly higher in vertebrates, 4.6 $\times 10^{-9}$ - 1.2 $\times 10^{-8}$ per site per generation (Milholland et al., 2017; Smeds, Qvarnström, et al., 2016).

The mutation rate is not entirely random across the genome. The transition rate, i.e. mutations between purines (A – G) or between pyrimidines (C – T), is usually higher than the transversion rate, i.e. the mutation from a purine to a pyrimidine or vice versa, which needs to be considered when building models for inferring mutation rates through time. One of the most common causes of mutations is spontaneous deamination of cytosine to uracil, which is interpreted as T, incurring a C/G to A/T mutation. This is even more common if the cytosine is methylated to 5-methylcytosine, typically at sites where C is followed by a G (CpG), because it deaminates directly to thymine which has a less efficient mismatch repair (Coulondre et al., 1978; Ehrlich & Wang, 1981). The mutation rate can therefore vary across the genome depending on GC- and CpG-content. However, GC-rich regions are also more stable, partly counteracting the increased mutation rate due to CpG (Fryxell & Zuckerkandl, 2000). Another common mutation is the oxidation of guanine to 8-oxo-guanine that have the ability to form Watson-Crick pairing with A. Hence, the incorporation of an 8-oxo-guanine during replication can lead to the mutation of a G/C to an A/T (Lynch, 2007).

Mutations from G/C to A/T are sometimes referred to as strong to weak (SW) mutations, and A/T to G/C as weak to strong (WS) since C:G has three hydrogen bonds while A:T only has two. The SW mutation types are more prevalent than WS, which leads to the generally observed A/T-biased mutation rate. If no other forces affected the occurring mutations in a population, the genome would end up in a mutation equilibrium corresponding to the mutation rate weighted by the local nucleotide content as the source of the mutations. The expected equilibrium AT-content would likely be higher than what we observe in many organisms, suggesting the presence of evolutionary forces that reduces the effect of the mutation bias (Lynch, 2007).

Structural variation

Structural variation describes mutations that lead to either copy number deviations or positional changes in a part of the genome caused by insertions, deletions or translocations of chromosomal segments. Inversions are mutations where segments are inserted in the opposite direction, whereas translocations are segments inserted in different positions. Finally, structural variations can also result from chromosomal fusion and fissions. Structural variations influence many evolutionary processes and are likely as common to affect phenotype, if not more, as single nucleotide mutations (Wellenreuther et al., 2019).

The rearrangements can originate through several mechanisms and can involve one or a few nucleotides, a whole gene, up to an entire chromosome. Rearrangements can result in tandem duplications, where multiple copies of DNA segments are located in proximity to the ancestral sequence, or dispersed duplications, with the copies located more distantly or even on different chromosomes. Mistakes during DNA-replication can cause the same sequence to be replicated multiple times or skip one segment resulting in deletion (Ohye et al., 2014). Rearrangements can also occur during DSB-repair through ectopic recombination, especially in connection to TEs (Lehrman et al., 1987; Morales et al., 2015; Rüdiger et al., 1995; Witherspoon et al., 2009; L. Zhang, 2003), or if a DSB is repaired via an alternative repair pathway using non-homologous end joining (Gorbunova & Levy, 1997; Gu et al., 2008). Retrotransposons can "copy and paste" host sequences either by extended transcription over downstream segments during transposition or by reverse transcription of the mRNA of host genes and subsequent insertion elsewhere (Hu et al., 2022).

Loss of the telomere or DSBs in the telomeric region can lead to structural rearrangements if the repair mechanism fails and triggers repair by recombination or non-homologous end joining (IJdo et al., 1991; Lo et al., 2002). If telomeric sequences from another chromosome are used for the repair, this could lead to end-to-end chromosomal fusion leaving remnants of telomeric sequences, so-called interstitial telomeric sequences (ITS) at the breakpoint (IJdo et al., 1991). The transposition of telomeric repeats can also result in

ITS, which is associated with complex rearrangements and fissions (Aksenova et al., 2013; Letsolo et al., 2010). Fissions and fusion can also be the result of ectopic recombination creating unequal fragments. In addition, DSBs can directly produce chromosome fissions if specific telomere repeats are added, protecting the new chromosome end (Fujiwara et al., 2000).

Chromosomal rearrangements were reported already a century ago by Sturtevant in *Drosophila* (Sturtevant, 1921). Inversions of chromosomal segments have since then been repeatedly identified as contributing to local adaptation and speciation (Hooper et al., 2019; Koch et al., 2022; Yan et al., 2020). A possible mechanism for the effects of inversions on evolutionary processes could be restricted recombination due to non-homology, which maintains linkage between beneficial coadapted loci. Similar effects have been observed after chromosomal fusions when whole chromosomes or chromosome arms are joined together, although the mechanism is different (Guerrero & Kirkpatrick, 2014; Wellband et al., 2019). Chromosome fusions physically link previously independently assorting loci, thereby reducing the reshuffling between them (Castiglia & Capanna, 2002; Cicconardi et al., 2021; Dumas & Britton-Davidian, 2002). A positive association between the number of chromosomes and rapid speciation rates has been observed in some taxonomic groups, indicating a connection between chromosomal rearrangements and speciation (Bush et al., 1977; Leaché et al., 2016). This could be mediated by reduced gene flow due to lower recombination or hybrid breakdown (Giménez et al., 2013; Martin et al., 2019).

Gene duplications are commonly observed in tandem clusters, although dispersed copies also occur (Hu et al., 2022; Katju & Lynch, 2003; L. Zhang, 2003). The gene duplication rate likely exceeds the nucleotide substitution rate, though most gene copies are lost by drift or purifying selection (Lipinski et al., 2011; Lynch & Conery, 2000). Multiple gene copies could disturb the delicate regulation of gene dose. These potentially deleterious effects may lead to selection for deletion or, perhaps more commonly, to the accumulation of loss of function mutations (Fischer et al., 2001). Albeit, gene family expansions have been connected with adaptation and innovation in many taxa (S. Chen et al., 2013; Moore & Purugganan, 2003; Prunier et al., 2019). A possible mechanism could be that multiple copies give redundancy and relaxation from purifying selection, which could allow new mutations to change the function of the new gene copy, so-called neofunctionalisation (Innan & Kondrashov, 2010; Lynch, 2007; Steiner et al., 2017). This is likely rare since most mutations are non-functional. Another mechanism by which the duplicated gene can be maintained is by sub-functionalisation, partitioning the subfunctions of the gene in the ancestral and the new copy, mediated by deleterious mutations, which are more frequent than beneficial mutations (Lynch, 2007). We used the assumption that gene families with rapid expansion are likely advantageous, since most gene duplications not maintained by selection

would be lost or degraded, to gain insight into functional categories possibly advantageous in the painted lady.

Forces acting on the genome

Drift and the effective population size

Evolution can be described as the change in allele frequency in a population through time. Two main forces affecting the fate of a novel mutation are genetic drift and natural selection, although other factors can also influence the rate of fixation (Nielsen & Slatkin, 2013). The ultimate fate of all mutations is either a loss from the population or a rise in the frequency to fixation. When a mutation is fixed, all members of the population carry the mutation, and the ancestral variant is substituted. When a new allele enters a population, the frequency is low since only one individual carries the new mutation on a single chromosome. Hence, the ultimate fate of novel mutations is mainly dependent on stochastic events. The main random event in sexually reproducing diploids is the independent segregation of chromosomes at meiosis that leads to an offspring inheriting only one of two alleles from the parent. The number of reproducing offspring produced by each individual varies due to differences in fertility but also due to random extrinsic factors. The stochastic element contributing to allele frequency change in finite populations in each generation is termed genetic drift. In a small population, this randomness can have a large effect on the total allele frequency change per generation, and a new mutation can, therefore, either be lost or reach fixation rapidly. A larger population is less affected by random fluctuations in reproductive success and more responsive to the effects of deterministic factors like natural selection. The population size reflected here is not the actual number of individuals in a population, but the effective population size, N_e (S. Wright, 1931). This is an idealised theoretical population with the same allele frequency changes as the real population. However, the concept of N_e is a useful approximation for building models to estimate the strength of the different forces acting on alleles, but violations of the assumptions have to be considered when interpreting results based on the model.

One of the models to understand molecular evolution is the neutral theory, which postulates that most mutations observed are selectively neutral and that evolutionary change mainly reflects the interplay between random mutations and genetic drift (Kimura, 1968). The theory predicts that the probability of fixation of a new neutral mutation is the same as its frequency in the population, while mutations under selection will go so quickly to fixation that they will rarely be observed as segregating polymorphisms in the population. If a diploid population size is N, the number of mutations arising per generation is $2N\times\mu$, where μ is the mutation rate per site per generation. The frequency of

the novel neutral allele will then be 1/(2N). The substitution rate denotes the total rate of fixed mutations per site and is the number of mutations per site times the probability of fixation $2N \times \mu \times 1/(2N) = \mu$. This means that the probability of fixation and the number of segregating mutations is proportional to the population size, but the substitution rate of neutral mutations is independent of the population size. The average time to fixation is proportional to the effective population size and is estimated to $4N_e$ generations (Kimura & Ohta, 1969). Thus, population size affects the number of segregating polymorphisms which can be observed as lower heterozygosity or nucleotide diversity in small populations and in regions of the genome with low recombination rates (Galtier, 2016). However, the population size does not affect the rate of neutral substitutions per site, so synonymous substitutions between species, which are assumed to be silent or neutral, should accumulate in proportion to the mutation rate.

Selection and the nearly neutral theory

If an allele possesses properties that affect an individual's genetic contribution to the next generation, the frequency of that allele will increase or decrease through the generations, thereby affecting the probability of fixation. The strength of selection, the selection coefficient (s), is calculated as the difference in relative fitness between different genotypes. The fixation time is determined by the selection coefficient and the initial allele frequency and will be much shorter for advantageous than for neutral alleles. Similarly, deleterious alleles are quickly removed from the population through purifying selection unless drift is strong. Hence, both beneficial and deleterious alleles will be segregating for a short time and are rarely found as segregating variants in observed data in sufficiently large populations.

A complicating factor is that the selection coefficient is not dichotomous. There is a distribution of fitness effects, from negative to positive, and many mutations are weakly deleterious or nearly neutral. The nearly neutral theory predicts that if selection is weaker than $|s| < 1/2N$, or scaled by effective population size $N_e s << 1$, the allele will be strongly affected by drift (Ohta, 1976, 1992). This will lead to the accumulation of slightly deleterious mutations in small populations. Support for the nearly neutral theory is the observed increase in the ratio of non-synonymous/synonymous polymorphisms in regions with low neutral diversity (putatively low N_e) (Castellano et al., 2018).

In summary, as with many other things in nature, there are gradients to forces acting on the genome. Evolution proceeds through gradual changes of frequencies in populations guided by random drift and variable selection coefficients. Selection is acting on the phenotype, and mutations can be selectively neutral in one environment and demographic situation but impose adaptive properties in the next.

Linked selection

One important aspect of genome structure that greatly impacts population genetic processes is the non-independence among loci. The physical linkage between loci along the genome can result in the fixation or loss of large identical haplotypes. Strong selection on one locus that rapidly goes to fixation can drag along adjacent neutral loci in a selective sweep through so-called hitchhiking (Maynard Smith and Haigh, 1974). A significant reduction in diversity was observed in proximity to regions under selection in Drosophila, supporting the hitchhiking theory (Begun & Aquadro, 1991). In extreme cases, rapid selective sweeps can form a trough in the diversity landscape. The depth and width of the trough are determined by the strength of selection, the recombination rate and the time since the fixation of the sweep (Barton, 1998; Kaplan et al., 1989). Background selection is the continuous removal of deleterious mutations along with linked neutral loci (Charlesworth et al., 1993). This process is perhaps more ubiquitous than hitchhiking since most mutations are either neutral or deleterious and can have a profound effect on genome-wide diversity (Cai et al., 2009; Lohmueller et al., 2011). Linked variants with counteracting selective pressures can result in less efficient selection due to Hill-Robertson interference (Felsenstein, 1974; Hill & Robertson, 1966). The realised selection coefficient is the sum of the selection coefficients for all linked loci, which could lead to a slower fixation time of an advantageous polymorphism due to linked deleterious variants (Betancourt & Presgraves, 2002). Furthermore, less efficient removal by purifying selection could lead to the accumulation of deleterious mutations (Rodgers-Melnick et al., 2015).

Recombination

The reshuffling of genetic material to the next generation mixes ancestral variants and creates novel combinations of alleles. This reshuffling of genetic variation is mainly produced by two very different mechanisms in diploid organisms; the independent assortment of chromosomes and homologous recombination.

Independent assortment accounts for most of the diversification of genetic variants to the offspring. Splitting up the genetic material into smaller chromosomes that can be randomly distributed to each gamete increases the probability of pairwise reshuffling between loci on different chromosomes. However, increasing the number of chromosomes renders a diminishing return as the probability of pairwise reshuffling goes towards its maximum of 0.5 (Veller et al., 2019). Splitting the chromosomes into smaller parts also decreases the probability that loci with synergistic or beneficial epistatic interactions are inherited together. One way to maintain the physical connection between the loci, still allowing some exchange, is by homologous recombination.

Homologous recombination is the exchange of genetic material between paternal and maternal chromosomes during meiosis. In prophase I of meiosis, the replicated chromosomes, also referred to as sister chromatids, pair up with their homologues to form tetrads. The chromatin forms loops held together along a central axis by structural proteins beginning to form the synaptonemal complex. Recombination is initiated within the loops by DSBs, which can be induced by endonucleases together with multiple cofactors (Keeney et al., 1997). DSBs are devastatingly harmful, and the repair is closely governed by numerous enzymes. The broken ends are first resected to single-strand DNA. The single-strand ends are protected by protein complexes and perform strand invasion by searching for a homologous sequence. This search is biased towards the homologous chromosome over the sister chromatid in meiosis (Schwacha & Kleckner, 1997). Once the homologous sequence is found, several pathways can resolve the DSB. Simplified, they result in either a crossover with reciprocal exchange of genetic material (CO) or a non-crossover (NCO or gene conversion) when the homologous chromosome is used as a template, but no reciprocal exchange occurs (J.-M. Chen et al., 2007). The process can result in more complex outcomes, and there is always a small gene conversion tract adjacent to the reciprocal crossover (Wall et al., 2022; Williams et al., 2015). The crossover event forms a physical connection between the homologues, called chiasma, visible with cytogenetic techniques (Tease & Jones, 1978). This physical connection ensures the correct orientation of the chromosomes relative to the meiotic spindle and the accurate segregation of the chromosomes when the homologues separate during anaphase (Schvarzstein et al., 2010). In some taxa with holocentric chromosomes, the orientation of the bivalent can vary dependent on the location of the CO (Hughes-Schrader, 1948). This could lead to inverted meiosis, where sister chromatid separation occurs before haploidisation.

The overall recombination rate is highly variable across taxa. A comparison of 353 eukaryotes revealed a variation spanning several orders of magnitude, from 0.03 to 119.90 cM/Mb (Stapley et al., 2017). In many organisms, at least one crossover per chromosome pair seems to be necessary for correct segregation during meiosis, although there are exceptions, like in some male Diptera and female Lepidoptera, where no recombination occurs (Morgan, 1912; Suomalainen et al., 1973; Turner & Sheppard, 1975). The necessity of one crossover per chromosome leads to a higher recombination rate per base in shorter chromosomes since there is a higher probability of a crossover event between two loci when there are fewer base pairs. This could explain the commonly observed negative association with chromosome size (Kawakami et al., 2014; Schield et al., 2020). Assuming that the genome size is maintained, this should lead to a positive correlation between haploid chromosome number and recombination rate. However, there is weak support for this, likely due to the formation of multiple chiasmata per chromosome pair in larger chromosomes, resulting in double crossovers (Stapley et al., 2017).

The distribution of crossover relative to non-crossovers is very variable among taxa. For example, 20 - 60% of detected recombination events are crossovers in yeast (Malkova et al., 2004; Mancera et al., 2008) and 3-5 times as many crossovers to non-crossovers are observed in bees and bumble bees (Kawakami et al., 2019) In contrast, there are at least ten times more NCO than CO in mice (Cole et al., 2010). The CO to NCO frequency difference between organisms could be biological, with different mechanisms determining the resolution pathway. However, non-crossovers are more challenging to detect, with usually much shorter tracts of around 100-1,000 bp, although potentially longer tracts >10 kb have been described, and the precise estimates in many taxa are limited (Wall et al., 2022).

The recombination landscape shows considerable spatial variation along chromosomes. There seem to be three dominating patterns: i) a decrease in the centre and increase at telomeres, ii) an increase at the centre and decrease at telomeres and iii) a bimodal pattern with a decrease in the centre, increase in subtelomeric regions and a decrease at the telomeres (Brazier & Glémin, 2022; Haenel et al., 2018; Vara et al., 2021). If the location of recombination would be optimally placed for maximum pairwise reshuffling, the theory predicts that centrally located COs result in more pairwise reshuffling compared to peripherally located COs if there is only one CO. In the event of multiple COs, the optimal reshuffling suggests that they should be evenly spaced along the chromosome since closely located double COs mitigate the pairwise reshuffling (Veller et al., 2019). The patterns are likely guided in part by the general chromatin state since DSBs are repressed in heterochromatin and occur more frequently in open chromatin (Jin et al., 2021). However, the number and location of recombination events occurring during meiosis are also limited by recombination interference, a mechanism reducing the probability of adjacent crossing over. For example, *C. elegans* shows total interference and is limited to only one CO per chromosome pair (Barnes et al., 1995). In contrast, interference only stretches 60 - 150 kb in *Saccharomyces cerevisiae* (Malkova et al., 2004). Gene conversion does not appear to have the same limitations (Bishop & Zickler, 2004; Malkova et al., 2004). The specific mechanism involved in interference is not known in most taxa.

The recombination landscape on a local scale can be either homogenous (Lynch et al., 2022) or very heterogeneous, with peaks of high recombination activity where most of the CO occur, so-called hotspots (Baudat et al., 2010; Myers et al., 2010). Many mammals have specific sequence motifs associated with hotspots that bind the enzyme PRDM9, which modifies H3K4m3, thereby recruiting the endonuclease Spo11 for DSB initiation (Baudat et al., 2010; Myers et al., 2010). In *C. elegans,* other enzymes have been associated with recombination rate variation and DSB initiation, but the specific mechanism is unknown (Chung et al., 2015). Chromosomal rearrangements, such as fusions, can alter the recombination landscape by decreasing the total number of crossovers but also alter the positions of the COs (Vara et al., 2021).

The recombination process is potentially harmful, including a DSB in the DNA and brings disruption of the obviously reasonably successful parental haplotypes to the offspring. The fact that homologous recombination is ubiquitously present and mostly conserved in nearly all sexually reproducing taxa, despite the potentially lethal side effect, says something of the importance of the process. There are numerous reports on the positive association between recombination and nucleotide diversity (Lercher & Hurst, 2002; Moriyama & Powell, 1996; Nachman, 1997; Stephan & Langley, 1998). Recombination breaks the linked selection and increases nucleotide diversity by allowing a higher number of segregating haplotypes in the population (Begun & Aquadro, 1992; Burri et al., 2015; Cutter et al., 2006). Neutral polymorphism originating from other haplotypes and not only mutations arising on the haplotype under selection are allowed to segregate. This is especially prominent in the case of selective sweeps, where the width of the trough in diversity is reduced by recombination breaking up the linked haplotype (Barton, 1998; Kaplan et al., 1989). However, perhaps more important for the overall diversity is the release of adjacent neutral polymorphisms from background selection (Charlesworth et al., 1993). Another contributing explanation for the association between diversity and recombination rate could be that the recombination process is mutagenic, although the evidence for this is limited (Arbeithuber et al., 2015; Halldorsson et al., 2019).

An inverse association between the ratio of non-synonymous/synonymous polymorphisms and recombination rate indicates increased efficacy of selection (Castellano et al., 2018), suggesting a release from Hill-Robertson interference. For the same reasons, there could also be an increase in genes with more efficient positive selection in highly recombining regions, potentially increasing the number of non-synonymous substitutions in a subset of genes (Betancourt & Presgraves, 2002). The rate of adaptive amino acid substitution exhibits a positive linear relationship with the recombination rate in low recombining regions but shows an asymptote in the relationship as the recombination rate increases (Castellano et al., 2016). This suggests that the effect of Hill-Robertson interference is substantial at low recombination rates but that only a small amount of linkage disruption is required to break the association between loci. The relationship between chromosomal rearrangements, recombination rate shift and the effect on molecular evolution will be explored in Chapter II.

GC-biased gene conversion

As we noted earlier, most investigated organisms have an A/T-biased mutation rate (Lynch, 2007). If mutations were the only force acting on the genome, the neutral AT-content would correspond to a mutation equilibrium proportional to the level of A/T mutations compared to the GC-changing mutation rate. The source for G/C->A/T mutations is the GC-content of the genome,

and when the GC-content is reduced, the number of G/C->A/T mutations will be reduced and end up at equilibrium (Lynch, 2007). If the observed GC-content is higher than the predicted equilibrium level, this could either mean that equilibrium is not reached yet or that there are other forces acting on these mutations before they reach fixation (H. Long et al., 2018). Heterogeneity in the GC-content along the genome suggests a variable strength of GC increasing forces along the genome (Figuet et al., 2015; Galtier et al., 2006) and evidence of selection on synonymous and non-coding sites indicates that mutational bias is not the only force affecting nucleotide composition (Eyre-Walker, 1999).

GC-biased gene conversion (gBGC) is one presumably neutral force that can counteract the A/T-mutation bias and increase the GC-content (Marais, 2003). Gene conversion is one outcome after DSB repair during recombination and refers to the unilateral repair using the homologous strand as a template without reciprocal exchange. Heterozygote sites within the conversion tract must be resolved for the DNA to pair with its original strand. If the heterozygote site is a G/C to A/T polymorphism, the A/T to a G/C mismatch repair is overrepresented, hence the term GC-biased gene conversion (Birdsell, 2002; Brown & Jiricny, 1987).

A positive correlation between GC-content and recombination rate has been observed among several eukaryotes, suggesting the presence of gBGC (Pessia et al., 2012). However, selection for high GC-content would behave similarly since recombination increases selection efficiency, and correlation does not imply causation. In support of gBGC, evidence of GC-favouring transmission bias during gene conversion has been observed (Kawakami et al., 2019; Malkova et al., 2004; Mancera et al., 2008; Smeds, Mugal, et al., 2016; Wall et al., 2022; Williams et al., 2015). Although gBGC appears in many taxa, there are contrasting studies where no biased gene conversion could be detected (H. Liu et al., 2018; Pokusaeva et al., 2022). The strength of gBGC can be inferred, unveiling a wide variability among taxa (Backström et al., 2013; Galtier et al., 2006; Muyle et al., 2011). High levels of gBGC have been observed in bees and bumble bees with a high recombination rate compared to Lepidopterans showing moderate levels (Boman et al., 2021; Kawakami et al., 2019).

The process of gBGC increases the fixation probability of the G/C variant over the A/T variant, similar to the effect of positive selection. The potential confounding signature of gBGC can therefore influence selection and demography inferences if not considered (Kostka et al., 2012; Pouyet et al., 2018; Rousselle et al., 2019). Another factor influencing synonymous mutation that potentially impacts GC-content is codon usage. How gBGC relates to codon usage, mutation bias and substitution rates is discussed in chapter IV.

Codon usage

Codon usage bias, the asymmetrical usage of synonymous codons, has been described in most investigated taxa and can have different underlying mechanisms (Galtier et al., 2018). If synonymous mutations are evolving under neutrality, the nucleotide composition of these positions should reflect the mutation bias and the composition of the non-coding genome. In line with this, a higher frequency of codons with A/T in the third codon position has been observed in many taxa, a likely consequence of the widespread SW mutation bias (Behura & Severson, 2012; Dennis et al., 2020). However, this assumption is violated in other cases, and the synonymous sites show a codon usage bias towards GC, despite the apparently ubiquitous AT-mutation bias (Vicario et al., 2007). Other neutral forces affecting the nucleotide composition, such as gBGC, could also influence codon usage as a by-product of compositional constraints (Knight et al., 2001; Lamolle et al., 2019; Palidwor et al., 2010).

Translational selection for specific codons could be an explanation for codon usage bias. Increasing the translational efficiency of one gene in slow-growing multicellular eukaryotes is likely a weak force. However, this does not exclude selection for specific codon usage. The potential mechanisms are diverse, including selection for translation efficiency, but also tRNA abundance, mRNA secondary structure and stability, and protein structure and stability, potentially conveying a more general increase of the efficiency across multiple genes and stages (Garel, 1974; Y. Liu, 2020; F. Zhao et al., 2017). The association between tRNA abundance and codon usage bias suggests that specific codons are more efficient in recruiting specific tRNAs (Behura & Severson, 2011; Duret, 2000; Garel, 1974; Moriyama & Powell, 1997). Another support for translational selection is that codons at high prevalence in highly expressed genes segregate at higher frequencies in the population (Galtier et al., 2018). There are also indications that GC-content in general, and not only specific codons, can influence mRNA stability and correct splicing and affect the secondary structure of proteins (Mordstein et al., 2020; Peng et al., 2016; Presnyak et al., 2015).

Codon usage bias can have profound effects on nucleotide composition but also on the inference of selection. Codon usage could result in a reduced substitution rate compared to the neutral expectation and lead to biased estimates of selection. The knowledge of the causes and consequences of codon usage bias is limited, and I will explore this in Chapter IV.

Research aims

Chapter I: The extent and character of karyotype evolution are largely unknown in many organisms, and our understanding of the underlying mechanisms behind chromosomal instability is limited. The aim of Chapter I was to characterise the extent and direction of intra- and inter-specific karyotype evolution in the wood white cryptic species complex of *L. sinapis*, *L. reali* and *L. juvernica*. In addition, we investigated genomic features associated with chromosomal fissions and fusion breakpoints.

Chapter II: The influence of chromosomal rearrangements on the recombination landscape and maintenance of genetic diversity can be substantial. However, studies of the effect on natural systems are rare. Here I assessed the effect of extensive karyotype rearrangements on the recombination rate and pairwise reshuffling rates, including the effects on nucleotide diversity and efficacy of selection in two populations with divergent karyotypes in *L. sinapis*.

Chapter III: Structural rearrangements in the form of gene duplications have been associated with evolutionary innovation and adaptation. The interplay between recombination rate variation, TE activity and structural changes are keys to our understanding of these processes. In Chapter III, we thoroughly annotated the genome of the migratory painted lady butterfly (*V. cardui*), aiming to investigate gene family expansion and the association with genomic features and the recombination landscape.

Chapter IV: Gene sequence evolution is also affected by forces not directly involved in protein evolution, such as mutation bias, codon usage, and GC-biased gene conversion (gBGC). Quantifying these forces and how they influence substitution patterns are essential to understand the strength and direction of natural selection. The aim of Chapter IV was to explore the effect of codon usage and nucleotide composition on substitution rates across Lepidoptera, including an in-depth exploration of patterns and processes such as fixation bias, mutation bias and tRNA-gene set in *L. sinapis*.

Chapter V: The potential to adapt to local conditions can be crucial in changing environments. Here we aimed to investigate the potential local adaption to

distinct host plant use in two populations of *L. sinapis*. We examined the behavioural and developmental differences between two karyotype extremes in *L. sinapis* in response to hostplant shift, including differential gene expression and gut microbiome composition.

Chapter VI: Environmental cues initiating and directing migratory behaviour are still unknown in the painted lady, and the underpinning genetic pathways are also yet to be determined. In Chapter VI, we investigated the differential activation of regulatory elements in response to hostplant availability in adult painted lady females. The aim was to identify differentially activated genes, which, together with the gene set identified in Chapter III, make potential candidate genes for further research on the underlying pathways associated with migration.

Study organisms

Lepidoptera

The order Lepidoptera includes butterflies and moths and comprises over 175,000 species worldwide (Shields, 1989). The general characteristics of lepidopterans are two pairs of wings covered in scales (gr. *lepids*). They carry a bi-partitioned proboscis for nectar feeding, which can be coiled up or extended. Butterflies and moths have holometabolic lifecycles, meaning they undergo a complete metamorphosis during development. Butterflies and moths have been a popular topic for naturalists and amateur collectors throughout history and are widespread study organisms in evolution and ecology research (Boggs et al. 2003). Many lepidopterans are also of direct economic importance, like the silk-producing moth *Bombyx mori*, the subject of intense studies in cell biology, functional genetics and biochemistry, including the characterisation of the gut microbiome (B. Chen et al., 2018; Fujiwara et al., 2000). Other lepidopterans are important pollinators but also pests of agricultural crops.

Lepidoptera has a ZW-sex chromosome system, and the females are achiasmatic (no recombination) (Suomalainen et al., 1973). Lepidoptera, in general, has maintained a remarkably stable karyotype (Ahola et al., 2014; Pringle et al., 2007), with many taxa showing a chromosome number close to the inferred ancestral karyotype of n=31 (de Vos et al., 2020). There are, however, exceptions from the conserved karyotype. Some groups show extraordinary variation in chromosome number, ranging from n=5 up to n=226 (de Vos et al., 2020). Butterflies and moths have holocentric chromosomes, which theoretically could facilitate the correct segregation of fusion/fissions heterozygotes since the spindle fibres attach along the whole chromosome and are not restricted by one centromere. They also display the possibility of inverted meiosis introducing flexibility in the orientation of the bivalents, thereby reducing the risk of unbalanced gametes (V. A. Lukhtanov et al., 2018; Melters et al., 2012). Holocentricity and inverted meiosis are probably not the only explanations since holocentric taxa have not been shown to have higher chromosome number evolution in general (Ruckman et al., 2020). The contrast between highly stable karyotypes in most lineages and extreme rearrangements in some groups makes Lepidoptera an interesting order for chromosomal rearrangement studies.

Butterflies and moths are ectotherms which means that they depend on the environment for body temperature regulation, leading to different strategies to cope with seasonal resource limitations and temperature drops, like diapause or migration to more suitable landscapes. The environmental cues and genetic underpinnings of these strategies have not been fully understood. Photoperiod appears to be a major environmental signal for diapause induction (Green & Kronforst, 2019; Leal et al., 2018; Pruisscher et al., 2018; Söderlind & Nylin, 2011; Wiklund et al., 2019)). The environmental cues guiding migration appear more complex, involving the initiation of movement, flight direction, and the duration of powered flight (Chapman et al., 2015; Chowdhury et al., 2021). A migratory life history involves trade-offs between energy relocation to flight muscles and fat accumulation for fuel, versus investment in reproduction. Substantial knowledge has been gained from the iconic Monarch butterfly (*Danaus plexippus*), which engages in a north American migratory route and diapause in Mexico. In this butterfly, several environmental cues have been connected to the behaviour, including magnetic field properties, photoperiod, temperature, hostplant quality and abundance (Brower et al., 2006; Iiams et al., 2019; Zhan et al., 2011, 2014; Zhu et al., 2009). However, there are still many questions regarding the evolution of this remarkable trait, and the understanding of the genetic underpinnings is still limited.

Leptidea spp.

For our purposes, *Leptidea* (Pieridae) is an excellent study system for investigating the mechanistic basis and evolutionary consequences of karyotype rearrangements. The *Leptidea* genus represents the only Palaearctic genus of the otherwise South American subfamily *Dismorphinae*, the mimetic sulphurs. In contrast to its species-rich south American relatives with outstanding wing colour varieties, the *Leptidea* genus contains only twelve Palaearctic taxa so far, almost morphologically indistinguishable with white body and white wings with a black apical spot. Three taxa, the common wood white (*L. sinapis*), Real´s wood white (*L. reali*) and the cryptic wood white (*L. juvernica*), are part of a cryptic species complex that can only be taxonomically identified by differences in genital morphology and genetic markers. They were separated into three species as late as 1988 (Boggs et al., 2003) and 2008 (Dincă et al., 2011). The ecology and behaviour of the wood whites are well known regarding habitat choice, mating behaviour and larval development (Friberg et al., 2008, 2008; Friberg & Wiklund, 2009). The genus exhibits a variable karyotype structure with haploid chromosome number 2n=52-54 in *L. reali*, 2n=76-88 in *L. juvernica* and, most extreme of them all, *L. sinapis* with a within-species karyotype cline from 2n=56 in the north to 2n=106 in the southwest (Dincă et al., 2011; V. A. Lukhtanov et al., 2011). The species also differ in genome size and show a dynamic evolution of multiple sex-chromosomes (Talla et al., 2017; Yoshido et al., 2020).

Leptidea is an excellent study system for detailed analysis of the evolution of chromosomal rearrangements, both between closely related species and within species. *L. sinapis* represents a unique system for investigating the recombination landscape in holocentric chromosomes in general, but specifically how recent chromosomal fusions and fissions affect the recombination landscape.

Vanessa cardui

The painted lady (Vanessa cardui) is a long-distance migratory nymphalid butterfly with an almost worldwide distribution. It is absent from South America and Antarctica (Shields, 1992). Recent studies have revealed that this butterfly can cross the Mediterranean sea and the vast Saharan desert to reach the rich resources in the sub-Saharan savannah before returning northbound when these resources dry out in the spring (Stefanescu et al., 2016; Talavera & Vila, 2016). This long circuit was confirmed with stable isotope data validating the origin of spring migrants in the Mediterranean area to tropical Africa (Talavera et al., 2018). The painted lady is highly polyphagous, which is advantageous for a migratory lifestyle (Celorio-Mancera et al., 2016). Unlike many close relatives, the painted lady does not hibernate and is highly fecund, with up to 10 generations per year (Menchetti et al., 2019; Wiklund & Friberg, 2022). The high genetic variation detected in this species is likely associated with its migratory lifestyle (García-Berro et al., 2022). The versatility of the painted lady has led to its use in research investigating genetic pathways involved in colour patterns and development (Connahs et al., 2016; Otaki, 2007; L. Zhang et al., 2021). The painted lady has retained the ancestral karyotype number (n=31), hence a good candidate for characterising the recombination landscape and small-scale rearrangements in an overall conserved karyotype. The genetic underpinnings and associated regulatory pathways involved in the exceptional migratory lifestyle are yet to be unveiled. Here, we analysed gene family expansion to identify candidate genes potentially involved in the painted lady's unique lifestyle. We also investigate the behavioural response to different environmental cues potentially involved in triggering migratory behaviour.

Main methods

Field and greenhouse work

Selection is acting on the phenotype, so to study evolution, we have to consider the effect of genetic changes on the phenotypic level. Knowledge of the system, behaviour and ecology is essential for asking the right questions and correctly interpreting genomic findings. Four projects (Chapters I, III, V, VI) entailed fieldwork and rearing butterflies in common garden settings. The sample locations for the fieldwork were determined by earlier efforts from collaborators and butterfly sightings reported in public databases (https://www.artportalen.se/). The public databases were valuable for current phenological information for appropriate sampling timing and geographic location. The dedicated butterfly lab in our institution was developed during my PhD, so some of the experiments were performed at the Zoology department at Stockholm University. The butterfly lab maintains a constant temperature and direct daylight lamps with automated timing to simulate constant environmental conditions. The captured females were kept in custom-made rearing flasks with double space in the bottom to accommodate plant cuttings for oviposition or in cages of different sizes, depending on the experiment. Since the species we have worked with are nectar feeders as adults, we provided them with 20% sugar water on cotton as an energy source. The larvae were kept in the rearing flasks described above and required a constant supply of greenhouse-grown hostplant cuttings and daily cleaning.

Genital dissection was performed on adult butterflies to ensure correct taxonomical assignment in the wood whites, and for the pedigrees, we confirmed single mating by checking the number of spermatophores (Dincă et al., 2011).

DNA- and RNA-extraction

To perform genomic analyses, we first have to translate the information in the living cells to human readable sequences. The first step is to extract the nucleic acids from the cells using extraction protocols suitable for the downstream analysis. Regardless of the method, the main steps are cell lysis, removal of proteins, capturing of nucleic acid and purification.

For the linkage maps, we used restriction site associated DNA-sequencing (RAD-seq). RAD-seq is a reduced representation technique, where a

restriction enzyme fragments the genome and only fragments containing the enzyme cut site are sequenced. To extract the DNA, we used a combination of cell lysis, proteinase K digestion, and phenol-chloroform protocols (Sambrook & Russell, 2006), which usually give high yields of high-quality DNA from butterflies. The downside is that it includes handling hazardous chemicals. A fast and economical alternative was a high-salt protocol (Aljanabi, 1997) with less harmful chemicals used for the pedigree of the painted lady.

We used silica-membrane columns to extract the RNA for the gene expression analysis. In principle, the tissue was lysed and homogenised before mixing with ethanol. The ethanol facilitated RNA binding to the silica membrane while the contaminants were removed by repeated washing. The RNA was finally released from the membrane by eluding it with water (RNeasy Kit, Qiagen). For the bacterial DNA, we used phenol-chloroform protocols as above.

Chromatin immunoprecipitation (ChIP)-seq entails using specific antibodies targeting histone tail modifications and precipitating them while the histones still are attached to the DNA. Sequencing of the DNA allows detection of the genomic location of specific histone modification enrichment peaks and characterisation of associated genomic features. For ChIP-seq, it is essential that the chromatin remains intact. First, we fixated the DNA-protein binding using formaldehyde before digesting the chromatin with a nuclease, and then we treated the fragmented chromatin with antibodies targeting specific histone modifications. In this case, we used H3K4me3 and H3K27ac to detect active regulatory elements. The antibodies were precipitated using magnetic beads covered with protein G with high antibody affinity, washed to remove non-target chromatin and, finally, the chromatin was released with specific buffers. The DNA was extracted using silica columns (Simple ChIP Enzymatic Chromatin IP KIT, Cell Signaling Technology).

We used the facilities provided by the National Genomic Infrastructure at SciLifeLab in Stockholm for library preparation and sequencing on Illumina platforms.

Gene order and gene family expansion

For an overview of the extent of rearrangements, we compared pairwise gene order between genomes in synteny analyses. For comparing distantly related taxa, it is suitable to use relatively conserved sequences like genes and preferably single-copy orthologues. For the synteny analyses, we used the Basic Local Alignment Search Tool (BLASTP) to find homologous genes (Altschul et al., 1990) and MCScanX to sort the genes after genomic position and compute collinear blocks (Wang 2012). Collinear blocks, or synteny blocks, are chromosomal regions sharing the same order of homologous genes from a common

ancestor. The software performs progressive alignment of multiple collinear blocks against the reference, successively using each genome as a reference.

To investigate gene copy number variation in the painted lady, we used gene annotations from several butterflies in the same family (Nymphalidae) to compare the number of gene copies in each orthogroup (gene family). Orthogroups were inferred with Orthofinder, which uses different methods, including BLASTP, to find orthologous gene families in a phylogenetic framework (Emms & Kelly, 2019). The orthogroups can then be used with a phylogenetic tree to infer the rate of gene gains and losses for the included taxa, implementing maximum likelihood models with Badirate (Librado et al., 2012).

In order to identify potential functions of genes of interest, we indirectly inferred the pathways involved. Gene ontology information is a hierarchical system assigning functional annotations to each gene if known, and the genes are clustered in higher-order functions and pathways. The descriptions are from three different gene ontologies (GO); biological function, molecular process and cellular compartment of the function (Alexa & Rahnenfuhrer, 2021). The enrichment test describes if a specific function is overrepresented in the focal gene set compared to a reference set of genes. One limitation of GO-enrichment analysis is that the information is derived from functional studies only performed in a small set of organisms. Although some functions are conserved, many functions, especially for rapidly evolving genes, are not classified and will be missed. The overrepresented genes are potential genes of interest for later functional experiments by knockdown or knockout of the specific genes, verifying their function and their involvement in the specific phenotype in the focal taxon.

Estimation of nucleotide diversity

The amount of genetic variation in a population can be described by several different characteristics. The variable associated with genetic variation, theta, is equal to the number of mutations in a population scaled by the effective population size, theta=$4N_e\mu$, where μ is the mutation rate (Nielsen & Slatkin, 2013). This model is under the assumption of neutrality and allows no recurrent mutations at the same site. The diversity landscape is shaped by different selection pressures, recombination rate variation and admixture. Since diversity is linked to population size, population bottlenecks or expansions will also leave signatures in the patterns of diversity. Pairwise nucleotide difference, π, is an estimator of theta widely used to describe genetic variation. This statistic compares all sites in an alignment of multiple individuals and counts the pairwise differences relative to the total number of comparisons. Here we used the software *pixy,* that only uses callable sites to calculate pairwise differences

(Korunes & Samuk, 2021). Restricting the analysis to callable sites avoids the assumption that missing sites are the same as in the reference, which can lead to an underestimation of the total diversity.

Detecting signals of selection

In this thesis, I inferred substitutions between species to explore the association between substitution rates and nucleotide composition in Lepidoptera. First, single-copy orthologous gene sequences were extracted from gene annotations of each species and aligned to identify changes in homologous sites. The substitution rate between species was inferred with a codon model, describing the probabilities of nucleotide substitutions within the codons. A phylogenetic tree was constructed, presenting the evolutionary relationship between the species to correctly place the substitutions along the branches (Dutheil & Boussau, 2008; Z. Yang, 2007).

The rate of synonymous substitutions is assumed to be equal to the mutation rate and represents the total substitution rate under neutrality. Evaluating the ratio of non-synonymous substitutions per non-synonymous site (d_N) to synonymous substitutions per synonymous site (d_S) can inform about the amount of selection on coding regions. Most non-synonymous mutations are believed to be deleterious, and the removal of deleterious mutations, purifying selection, would result in $d_N / d_S < 1$. If the non-synonymous and synonymous substitution rates are equal, $d_N / d_S = 1$, this suggests that the non-synonymous mutations are effectively neutral, representing relaxed selection pressure, for example, in a pseudogene. However, it could also represent a high level of both neutral and advantageous substitutions. A d_N / d_S -ratio > 1 suggests positive selection, with repeated advantageous mutations with a higher fixation rate than neutral substitutions.

Signals of selection can also be detected using polymorphisms since strongly advantageous or deleterious mutations will only be in a polymorphic state over comparatively short timeframes and are, therefore, underrepresented in a population sample. The ratio of non-synonymous polymorphisms to synonymous polymorphisms can inform about the strength of direct selection in coding sequences. The polymorphism ratio represents purifying selection when less than one and relaxed selection when equal to one. However, if the ratio is above one, which is rare, it most likely represents balancing selection since polymorphic sites under strong positive selection would quickly go to fixation (Hahn, 2018; Lynch, 2007). We used this method to assess the association between recombination rate and efficacy of selection in Chapter II.

Recombination rate estimation

Linkage map

The recombination rate can be estimated either per generation or by population data averaging over many generations. Per generation data gives unbiased, direct estimates not affected by demography or selection. For my projects, I constructed linkage maps from pedigrees, a commonly used method to estimate the recombination rate (Apuli et al., 2020; Coop et al., 2008; Kawakami et al., 2014; Smeds, Mugal, et al., 2016).

Recombination is the shift in parental haplotype or genealogy along the chromosome and can be detected in the offspring as an alternative combination of markers compared to the parental genotypes. In the simplest case, there are two loci, each with two alleles, A/a and B/b, giving four possible haplotypes AB, Ab, aB, and ab. The classical way of estimating the recombinant fraction is to cross two homozygous groups, where all have either AB or ab. The offspring (F_1) will have one haplotype AB and one haplotype ab. The F_1 generation is then crossed with each other or back-crossed. The resultant offspring (F_2) should also get the haplotypes AB and ab, unless a recombination event occurred in one of the parents. The haplotypes in the F_2 generation are counted, and the recombinant fraction is calculated.

The fraction of recombinants can also be interpreted as the probability of a recombination event occurring between the markers. The probability of recombination is proportional to the physical distance between the markers. If two markers are closely located, they are less likely to have a crossover event than two more distant markers (Sturtevant, 1913). If we observe the same number of all haplotypes, the recombinant fraction would be 50%. In this case, we can assume that the loci are independent, either on different chromosomes or distantly located on the same chromosome. Repeating the process for more markers across the genome creates a map of the physical relations between markers along each chromosome. The map distance, or genetic distance, is usually denoted Morgan, or centiMorgan (cM), and reflects the probability of a recombination event. If the total map length of a chromosome exceeds 50 cM, this is indicative of double crossovers. One problem with double crossovers is that they can go unnoticed if they are relatively close, as the markers on either side of the two crossovers will be in the same parental haplotype state. Map functions, such as the Kosambi and the Haldane map functions, have been developed to account for this at larger crossover distances when converting recombination fractions to map distances (Kosambi, 1944).

Extending this process to hundreds or even millions of markers using whole genome sequence information allows the combination of information from the surrounding markers. Several algorithms have been developed for estimating the likelihood of recombination events given the data. I used LepMap3 to obtain maximum likelihood estimates of recombination distances based on

genotype likelihoods (Rastas, 2017; Rastas et al., 2016). LepMap3 is developed for high marker density and uses the information from the pedigree to infer and reconstruct parental markers. The tool is also designed to accommodate heterochiasmy or achiasmy in one parent. The first module converted the input genotype likelihoods into posterior probabilities of the ten possible genotypes for each informative variant. Informative variants are heterozygous in at least one of the parents and were used to infer the probability of the four different haplotype states and to calculate the probability of two markers having identical haplotypes. The logarithm of the odds (LOD) that two loci are inherited together is the ratio of the probability of linkage to the probability of not being linked. We used the probabilities of identical haplotypes to calculate the LOD-score, which was used to separate markers into linkage groups. The next step was to order the markers along the linkage group. For this, a maximum likelihood hidden Markov model is used with two parameters; the recombination probability and the probability of interference. Every change in state means a recombination event and gives the genetic position. The likelihood was obtained by finding the path through the model, giving the highest probability for the data. The module runs with alternative phasing to improve the likelihood. The recommendation is to iterate the model due to the stochasticity in the method and use the iteration with the highest likelihood.

The genetic maps were subsequently used to verify and correct genome assemblies and to estimate the recombination rate. We calculated the recombination rate as the genetic distance between the markers divided by the physical distance, usually denoted in centiMorgan per Megabase (cM/Mb). We also estimated the recombination rate as the slope of a function with genetic distance and physical distance as variables, either by local linear least square regression or window-based estimates of the slope (Rezvoy et al., 2007).

The drawbacks of recombination estimates using pedigree-based linkage maps are that the resolution of the map is limited by the number of informative meiosis analysed, i.e. the number of offspring included. In addition, linkage maps give the recombination rate in one individual, or family, at one time point. Other methods to estimate recombination rate are direct cytological evidence of recombination occurring by counting the number of chiasmata formed during meiosis (Anderson et al., 2003; Herickhoff et al., 1993; Holm & Rasmussen, 1980; Pollock & Fechheimer, 1978). Genotyping of sperms results in high-density fine-scale maps with the possibility to record thousands of meiotic events per individual (Meier et al., 2021; Xu et al., 2015; L. Yang et al., 2022). Sperms are haploid, which means that the data is already phased, eliminating one source of error in estimations of recombination rate.

The patterns of recombination rate through generations render a time-averaged impact on sequence evolution. This can be used to infer the population-averaged recombination rate. The population recombination parameter ρ is under the standard neutral model defined as $\rho = 4N_e c$, where N_e is the effective

population and c is the recombination rate per site. There are several methods to infer ρ, but the most commonly used methods are indirect inference based on the non-random association of alleles at two different loci linkage disequilibrium (LD) (Chan et al., 2012; Hudson, 2001; McVean et al., 2002). LD-based estimates are valuable but inherently biased by their dependence on allele frequencies (Lynch, 2007).

Estimation of total genetic reshuffling

Pairwise reshuffling probability estimates the probability that two random loci are inherited from different parental haplotypes. The difference between estimates of recombination rate and pairwise reshuffling probability is that total reshuffling also takes the recombination by independent assortment into account. In addition, while calculating the intrachromosomal contribution of pairwise reshuffling, the locations of the homologous recombination events are also considered. For example, the pairwise reshuffling rate is lower if two recombination events happen in close proximity to each other and a recombination event occurring towards the end of a chromosome renders fewer opportunities for pairwise reshuffling. Interchromosomal pairwise reshuffling can be calculated as the probability of two random loci located on different chromosomes having different parental origins in the gamete. This is equal to the probability of being located on different chromosomes and the probability of independent assortment of the chromosomes: (1 - sum((chromosome length/assembly length) ^2))×0.5 (Veller et al., 2019). Intrachromosomal pairwise reshuffling is the probability of two random markers on the same chromosome having different parental origins, taking into account the effect of the location and genetic distance between the markers. The sum of pairwise genetic distances was divided by the total number of comparisons for each chromosome, where the sum of the contribution of each chromosome was weighted by the probability of two loci being located on the same chromosome (chromosome length/assembly length)^2 (Veller et al., 2019).

Inferring signal of GC-biased gene conversion

The strength of gBGC can be inferred with b=cnr, where c represents the G/C-transmission bias from heterozygotes (i.e. the deviation from the expected segregation rate, 0.5), n the length of the conversion tract and r the probability of having a conversion event. If $c > 0$, the change is positive (increasing the GC-content), while if $c < 0$, the expression is negative (increasing the AT-content). If the population size is very small compared to the timescale of conversion ($N_e < 1/nr$), the heterozygous sites will quickly be lost or fixed due to drift, so they are marginally affected by gBGC. The effect of gBGC can be compared

to the effect of selection since it affects the probability of fixation, where b is analogous to the selection coefficient (s). This also gives us a limit for the detection of gBGC. If the strength of gBGC is smaller than the coalescent time, there will be no effect ($b << 1/2N_e$). To estimate the strength of gBGC directly is non-trivial and requires large sample sizes of high-density genetic maps to accurately estimate conversion tract length and transmission bias. One alternative, which we used in *L. sinapis*, is to use population genetic models to infer the population scaled strength of deviation from neutral expectations ($B= 4N_eb$) in population data, where b represents the intensity of the bias. The models take advantage of the effect of gBGC on the site frequency spectrum, which shifts to medium-high frequencies for WS mutations (right shift) and more low frequency alleles for SW (left shift) (Duret & Galtier, 2009; Lachance & Tishkoff, 2014). This model uses the derived allele frequency spectrum in the different mutation categories SW, WS and GC-conservative mutations to obtain maximum likelihood estimates of the intensity of a GC favouring fixation bias (Glémin et al., 2015). This method does not explicitly state the conversion bias; it only states if there is a fixation bias in either direction. The mechanism behind the fixation bias could be biased gene conversion, but also other forces like selection could affect the estimate.

Codon usage

We first estimated the codon usage frequencies, basically counting the number of occurrences of a codon per 1,000 bp with the software *cusp* from the European Molecular Biology Open Software Suit (EMBOSS) (Rice et al., 2000). We were interested in the differential usage of synonymous codons and calculated the relative synonymous codon usage (RSCU). RSCU uses codon counts and their relative distribution per amino acid, normalised by the expected counts per synonymous codon if all codons are used at equal frequencies (RSCU= 1). Overrepresented codons have RSCU > 1, and underrepresented codons have RSCU < 1 (Sharp & Li, 1986). Another way to estimate the codon usage we also applied is to calculate the 'effective number of codons' (ENC). This statistic ranges from 20 – 61, stop codons excluded, and is based on the number of amino acids in each degeneracy class and the relative frequencies of the synonymous usage for each codon (F. Wright, 1990). Similar to the effective population size, ENC can be seen as a representative number of the distribution of codon usage frequencies in a gene. Lower values mean fewer codons are used, hence stronger codon usage bias. The observed ENC can be compared to the expected number of codons based on nucleotide composition, $ENCexp = 2 + GC3 + 29 / (GC3^2 + (1\text{-}GC3)^2)$ (Wright 1990). If codon usage were determined by GC-content alone, ENC would be similar to ENCexp (Li et al. 2018). ENC is a composite of the codon usage in the entire gene, so information about specific codons is lost.

Summary of papers

Chapter I – **High-density linkage maps and chromosome level genome assemblies unveil direction and frequency of extensive structural rearrangements in wood white butterflies (*Leptidea* spp.).**

In this chapter, we characterised the rate and direction of chromosomal rearrangement in three species in the *Leptidea* cryptic species complex. Karyotypes are, in general, highly conserved between closely related species. Large chromosome rearrangements can cause unbalanced gametes due to problems in chromosome segregation during meiosis, or produce negative fitness consequences in heterozygotes. In the order Lepidoptera, most studied taxa have retained the ancestral karyotype and gene synteny is often conserved across deep divergences. However, many clades demonstrate highly variable karyotypes and a few examples of extensive genome reshuffling have recently been demonstrated. The underlying reasons for the rapid karyotype evolution in these lineages have not been determined. The genus *Leptidea* has an extreme level of chromosome number variation and rearranged sex chromosomes, but a detailed characterisation of restructuring across the genome has not been done before.

First, we constructed eight chromosome-level genome assemblies for males and females from each species in the wood white (*Leptidea*) species complex, including two populations with distinct karyotypes within *L. sinapis*. We used a combination of 10X linked reads and HiC contact data to generate assemblies, and improved them using linkage maps for the two populations of *L. sinapis*. The genome sizes ranged between 655 – 685 Mb, with pseudochromosome numbers of *L. juvernica* n=42,43, *L. reali* n=26, Swedish *L. sinapis* n=28,29 and Catalan *L. sinapis* n=52, which is within the expected ranges. To evaluate the extent of reorganisation compared to the ancestral karyotype, we performed gene order analyses with genomes from the silkmoth, *Bombyx mori* (n=28) and the nymphalid butterfly *Melitea cinxia* (n=31), which both have conserved gene synteny, with the exception of three fusions in *B. mori*. The synteny analysis revealed extensive rearrangements in *Leptidea* compared to the ancestral karyotype. The average lengths of homologous tracts were on the level of 1 Mb, and each chromosome contained, on average, 12.5 - 14.5 of those blocks.

We also observed substantial rearrangements within the *Leptidea* clade. In *L. juvernica*, which we used as an outgroup to polarise the direction of the rearrangements the other species, we found 44 breakpoints. We found 20 fusions and six fissions in *L. reali*, 11 fusions in the Swedish, and two fusions and 14 fissions in the Catalan population of *L. sinapis*. One of the three Z-chromosomes was conserved across the species, while the other two were rearranged. We further detected several rearrangement polymorphisms between the male and female assemblies in both *L. sinapis* and *L. juvernica*. Using the linkage maps, we confirmed five segregating fusion/fission polymorphisms in the Swedish and three in the Catalan population of *L. sinapis*. To understand more of the processes behind the karyotype instability, we characterised the type of rearrangements and the genetic features associated with the breakpoints. Most of the restructuring could be explained by fissions and fusions, while translocations appear relatively rare in this clade. LINEs and LTR transposable elements were enriched in breakpoints, suggesting that ectopic recombination might be an important mechanism behind the chromosomal rearrangements.

In conclusion, we found extensive chromosomal rearrangements within *Leptidea*, mainly driven by chromosome fusions and fissions. The segregating rearrangement polymorphisms support a highly dynamic ongoing genome structure evolution in this clade. The considerable rearrangements, on both species- and population-level, show that chromosome count alone may underestimate the extent of genome restructuring.

Chapter II – **Nascent evolution of recombination rate differences through chromosomal rearrangements.**

Here we explored the effect of chromosomal rearrangements on the recombination rate, pairwise reshuffling and subsequent effects on diversity and efficacy of selection. Reshuffling existing variation is essential to create novel combinations of alleles and break the linkage between advantageous and deleterious variants, increasing the efficacy of selection. Two different mechanisms mainly produce the reshuffling of genetic variation in diploid organisms; i) independent assortment, the pairwise reshuffling between loci situated on different chromosomes, and ii) homologous recombination, the exchange of genetic material between homologous chromosomes breaking the association between loci located on the same chromosome. Many organisms require at least one crossover per chromosome pair for correct segregation during meiosis. This leads to a positive association between the number of chromosomes, the probability of pairwise reshuffling and the recombination rate. Despite the evolutionary importance of reshuffling, studies of how chromosomal rearrangements affect reshuffling and subsequent evolutionary processes are limited in natural systems.

Here we characterized the recombination rate variation in the wood white butterfly (*Leptidea sinapis*), a species with extreme intraspecific karyotype variation. We used pedigree-based linkage maps from a Swedish (2n=52-58) and a Catalan (2n=100-108) population to estimate the recombination rate within the two populations and the extent of inter- and intrachromosomal pairwise reshuffling. In addition, we assessed the effects of recombination rate changes on diversity and the efficacy of selection.

The recombination rate was higher in the population with more chromosomes, and the recombination rate had a strong negative association with chromosome size. However, when removing the effect of at least one crossover per chromosome, the chromosome length was positively correlated with the excess map length. The total probability of pairwise reshuffling was slightly higher in the population with a higher chromosome count. The diminishing return of increasing the number of chromosomes, together with the increased frequency of double crossovers in the larger chromosomes, reduced the difference between the populations. We observed an overall positive association between recombination rate and neutral nucleotide diversity, and a negative association with the ratio of non-synonymous to synonymous polymorphisms in both populations. Both associations displayed an asymptote at higher recombination rates. There was a positive effect of recombination rate variation on the change in intergenic diversity in chromosomes involved in derived fusions or fissions. There was no difference in estimates of direct selection between the populations, possibly due to high enough recombination rates to remove slightly deleterious mutations in both populations.

We concluded that extensive karyotype rearrangements resulting in a twofold difference in karyotype in *L. sinapis* significantly affected the recombination rate, but had limited effect on pairwise reshuffling. We observed a positive effect of recombination rate on the diversity in rearranged chromosomes, likely due to the reduction of linked selection as an effect of a higher recombination rate. Chromosomal rearrangements are an important driver of recombination rate variation, subsequently shaping the patterns of genetic diversity.

Chapter III - **Linkage mapping and genome annotation give novel insights into gene family expansions and regional recombination rate variation in the painted lady (*Vanessa cardui*) butterfly.**

Structural changes such as gene duplications can provide genetic variation for natural selection to act upon, with potentially profound evolutionary consequences. Common causes of rearrangements are non-homologous repair of double-strand breaks during recombination and the duplication of host sequences mediated by transposable elements. Characterisation of the interplay between gene family expansions, transposable element activity and recombination rate variation are keys to our understanding of these processes.

Here, we developed a linkage map and detailed genome annotation of the painted lady butterfly (*Vanessa cardui*). The painted lady is a non-diapausing, highly polyphagous species famous for its long-distance migratory behaviour and almost cosmopolitan distribution. Migratory behaviour is a complex phenotype where multiple approaches are necessary to unveil the genetic underpinnings and microevolutionary processes involved. We used the annotation to explore gene family expansions, transposon distribution across the genome, and the relationship of the genomic features with variation in the recombination landscape.

The painted lady assembly showed conserved large-scale synteny with the ancestral Lepidopteran karyotype. Several lineage-specific gene family expansions were identified, most spatially clustered, suggesting that the primary mechanism behind expansion has been tandem duplication. The expanded families were involved in protein and fat metabolism functions, detoxification, and defence against infection, which are vital for the painted lady's life-history. We also found enrichment of several TE classes in association with regions of gene expansion. The recombination maps allowed us to characterise a heterogeneous regional recombination landscape, with a strong effect of chromosome size on the recombination rate. Furthermore, short interspersed elements (SINEs) showed a positive association with recombination rate, suggesting that SINEs use the recombination machinery for transposition, or initiate the recombination events, although other forces such as open chromatin affecting both recombination rate and SINEs cannot be excluded.

In conclusion, we developed a thorough annotation and a linkage map of the painted lady genome as a resource for further research. We found that the interaction of recombination, transposable element activity, and gene duplications have shaped the genome structure. In addition, we provide a first candidate gene set tentatively involved in the evolution of the migratory lifestyle of the painted lady.

Chapter IV - **Base composition, codon usage and patterns of gene sequence evolution in butterflies.**

Sequence evolution is influenced by natural selection and neutral evolutionary forces, and the classification of genomic features evolving under specific forces is key to understanding microevolutionary processes. The rates and patterns of protein changing (non-synonymous) and putatively neutral (synonymous) substitutions have been used in coding sequences to infer the rate of evolution. However, the effects of other forces not involved in protein evolution per se, such as mutation bias, codon usage, and GC-biased gene conversion (gBGC), on gene sequence evolution have not been explored in many species. Quantifying how these forces shape the substitution patterns is vital to understand the strength and direction of natural selection.

Here, we used comparative genomics of eight taxa to investigate the association between base composition and codon usage on gene sequence evolution in butterflies and moths (Lepidoptera). In addition, we inferred the amount of mutation bias and the strength of fixation bias, as well as characterising the tRNA-gene set, to further investigate the underlying patterns and processes in one species, *Leptidea sinapis*.

We found variable GC-content in third codon positions among taxa, and the overall estimates of dN/dS using all substitution classes were lower compared to estimates of GC-conservative substitution classes. There was a G/C to A/T substitution bias at synonymous sites in most butterfly lineages with some variation in strength. However, the substitution bias was lower than expected from previously estimated mutation rate ratios. Population models in *L. sinapis* showed an A/T-mutation bias close to three and fixation bias in favour of higher GC-content (B=0.46), the magnitude of fixation bias suggests a notable deviation from neutrality at 4-fold degenerative sites. We explored the presence of codon usage bias, and the analysis showed an overrepresentation of A/T-ending codons in most species and a negative association between the magnitude of codon usage bias and GC-content in third codon positions, which is expected considering the A/T mutation bias. Another cause of codon usage bias could be translational selection, so we examined the tRNA-gene population in *L. sinapis* as a proxy for tRNA abundance. There was a strong association between the number of tRNA gene copies and the frequency of amino acids used in coding sequences, but a weaker association in the relative usage of specific codons. In addition, there was higher GC-content at third codon positions in the tRNA gene set compared to the coding sequences and less overrepresentation of A/T-ending codons, suggesting translational selection in favour of increased GC-content. We observed an inverse relationship between synonymous substitutions and codon usage bias, indicating selection on synonymous sites. This further supports that codon usage is affected by forces other than mutation bias.

We conclude that the evolutionary rates in Lepidoptera are affected by a complex interaction between an underlying G/C -> A/T mutation bias and partly counteracting fixation biases, predominantly conferred by overall purifying selection, gBGC and selection on codon usage.

Chapter V - **Host plant diet affects growth and induces altered gene expression and microbiome composition in the wood white (*Leptidea sinapis*) butterfly.**

The importance of herbivore-plant interactions in shaping biodiversity cannot be overstated, and a detailed understanding of specific resource utilization strategies is necessary to understand the underlying processes. Host switches in phytophagous insects are common, but the physiological and behavioural

responses are poorly understood. The potential for utilization of novel host plants depends on the similitude of plant chemical composition, its recognition by adult females for oviposition, and the fitness of the larvae feeding on the plant. The common wood white butterfly (*Leptidea sinapis*) has different host plant preferences in the latitudinal extremes of the distribution range. The northern (Swedish) population prefers *Lotus corniculatus*, while the Iberian (Catalan) population preferentially uses *Lotus dorycnium* in the wild. The two populations are also representing two extremes of the karyotype cline. Here we explored female host plant preference differences between the two populations, the effect of host plant diet on larval growth, population-specific differential gene expression and gut microbiome composition.

Female Catalonian wood whites lacked preference for either host plant in the greenhouse trials, while Swedish females strongly preferred *L. corniculatus*. The larvae from both populations exhibited longer developmental times and gained a smaller adult body size, feeding on *L. dorycnium*. The Catalonian larvae showed a more pronounced response to host plant treatment in gene expression profiles and microbiome community.

Our conclusion is that local conditions and population-specific plasticity determine the choice to use a specific host plant in wood whites.

Chapter VI – **Activity profiles of regulatory elements and associations with the oogenesis-flight syndrome in a long-distance butterfly migrant.**

Migratory behaviour is a complex phenotype comprising several behavioural, physiological and morphological modifications. A migratory life history involves trade-offs between energy allocation to flight muscles and fat accumulation for fuel versus investment in reproduction. The oogenesis-flight syndrome in insects, where reproduction is delayed until after the migratory phase, is one example of the realisation of those trade-offs. The environmental cues involved in the initiation of flight or reproduction and the molecular pathways eliciting the response to those cues are likely lineage-specific and poorly understood in many taxa.

In this study, we investigated the differential activation of regulatory elements in the painted lady (*V. cardui*) in response to host plant availability, since the scarcity of host plants for oviposition could be a potential cue for migration. We applied two treatments, with and without access to the host plant, *M. sylvestris*, and two replicates in each group. The females were sampled five days after emerging as adults, and the head and antennae were processed for chromatin immunoprecipitation (ChIP). We used ChIP-sequencing to detect the overrepresentation of histone tail modifications, specifically H3K4m3 and H3K27ac, indicators of active gene transcription, whose enrichment is associated with promoters and enhancers.

There were apparent behavioural differences between the two treatment groups. The females with access to host plants laid considerably more eggs, >850 and >500 in total, compared to the females without hostplant (0 and 90). We identified 14,100 H3K27ac and 9,846 H3K4me3 enriched regions, where a majority of the H3K27ac peaks were intergenic, while more than half of the H3K4me3 peaks were detected in the proximity of an annotated TSS. The overall differential activation levels between the groups were minor, and we found only a small number of differentially activated H3K27ac regions. These peaks were located in the vicinity of genes associated with DNA-replication, oogenesis, glucose metabolism, mating behaviour and muscle activity, including one odorant-binding protein gene and a putative juvenile hormone esterase-like gene.

The genes in proximity to the differentially activated peaks give an initial insight into pathways possibly involved in regulating oviposition and host plant recognition in the painted lady, and represent potential candidates for further investigation of the migratory behaviour.

Conclusions and future prospects

In this thesis, I wanted to investigate the causes and consequences of structural rearrangements and how different evolutionary forces shapes the genome. We found that the interaction of structural rearrangements, transposable element activity, and recombination has shaped the genome structure and influenced genetic variation. Furthermore, nucleotide composition and substitution rates are affected by both selective and neutral forces acting on the genome, which should be considered when inferring and interpreting evolutionary rates. We have also investigated the link between genotype and phenotype, and how the genomes interact with their environments. We used gene family expansions and experimental setups to obtain candidate genetic pathways underlying behavioural responses and local adaption to different environmental conditions. Each of the projects has given rise to many new questions that hopefully will be answered in the future. The resources we have developed in the form of chromosome-level genomes, annotations, linkage and recombination maps will be an asset for further studies within evolutionary genomics and molecular evolution, as well as ecological genomics.

We found that the chromosomal rearrangements in *Leptidea* were mainly driven by fusions/fissions, and the degree of rearrangements superseded the expected rearrangements using only the chromosome numbers. The segregating fission/fusion polymorphisms suggest an ongoing genome structure evolution in the *Leptidea* clade, especially within *L. sinapis*. One approach to investigate the ongoing karyotype rearrangements is to extend the sampling within *L. sinapis* across the karyotype gradient in Europe to see how the gradient relates to the karyotype extreme populations. This could answer if the populations in central Europe result from secondary contact between the northeast and the southwest after the last glaciation or if they represent other populations with their own karyotype dynamics. Extending the sampling across the cline could inform on the effects of rearrangements on the potential for hybridisation and population divergence, and can complement ongoing studies on hybrid incompatibilities within *L. sinapis*.

Another route would be to extend the phylogenetic sampling over the whole genus and include sister clades of their closest relatives, the mimetic sulphurs, to quantify the variation in rearrangement rate. Or an even more comprehensive comparative genomic analysis of karyotype instability with other clades or subfamilies of butterflies. A macroevolutionary approach

including both conserved and unstable karyotypes to infer rates and patterns of karyotype evolution on a larger scale and unravel associations between karyotype evolution and species divergence and speciation rate.

One outstanding question is why most investigated butterflies have remarkably well-preserved karyotypes over more than 130 million years, while some lineages display substantial rearrangements. Large-scale comparative genomic studies could indicate if there are any common denominators behind these chromosomal instabilities. Damage and breaks to the DNA can occur anywhere, and most of the time, it is strongly deleterious. By investigating the relationship between genomic characters and breakpoints, we could learn which structures could be involved in the actual breakpoint and which structures are more important to keep intact. We found indications that specific TEs were enriched in fusion breakpoints. These could be remnants of the repeat-rich telomeric regions of the fused chromosome ends, but it is a possible cause of initiation of ectopic recombination, leading to fusions. Do these TEs have specific characteristics more prone to fusion and fissions? Finding those or other specific TEs associated with breakpoints in other taxa, would support the action of specific TE-classes as the causal mechanism behind the rapid chromosome evolution.

Nucleotide composition is highly variable across the genome, and we do not fully understand the mechanisms involved in shaping the composition or the consequences of variation. We concluded that gBGC affects the nucleotide composition in butterflies, but other mechanisms also increase or maintain the GC-content in coding regions of the genome. To correctly infer the strength of evolutionary processes, there is a need for a null model incorporating neutral processes such as gBGC, but perhaps other mechanisms as well, which requires further studies on the nature of these processes in different taxa. Comparative genomics on multiple taxa of intermediate divergence would be helpful to further examine the evolutionary consequences of the interplay between mutation bias and different fixation biases in butterflies. This would include sister pair comparisons to infer substitution rates and associations with population estimates of the strength and direction of fixation bias and codon usage bias. In addition, evaluating the fitness effects of these processes in the populations would further increase our understanding of the causes and consequences of this intricate interaction between A/T mutation bias on one side and gBGC and other GC-increasing forces on the other.

Migration is a complex phenotype, and a combination of different approaches will be required to unravel the underlying molecular mechanisms. By analysing gene duplications and differential activation, we uncovered an initial set of candidate genes. These genes, together with the result of ongoing studies using differential gene expression and DNA-methylation in the responses to environmental cues, could be used in further studies of the link between genotype and phenotype. The assessment of phenotypic response using activity monitoring or flight mills, will be necessary for evaluating the

effects of manipulations. In addition to physiological examinations, to assess changes in muscle mass, fat deposition and egg maturation in response to environmental cues, perhaps by 3D-imaging using computed tomography of the abdomen and thorax of the females. Other ongoing projects use comparative genomics to detect genes with a signal of positive selection and specific gene family dynamics in sedentary versus migratory sister taxa. Besides coding sequence evolution, we are also interested in understanding the level of conservation of the non-coding parts, to identify associations between specific life history strategies and non-coding genetic variation. The ChIP-seq experiment was a part of a larger project to identify and characterise regulatory regions in the genome of the painted lady, complementing computational predictions. The aim was to identify conserved regulatory elements in other closely related species for comparative analyses using sedentary and migratory sister pairs to understand the genetic basis of migratory behaviour in that clade.

During the course of my PhD, the number of published chromosome-level lepidopteran genomes available on NCBI (www.ncbi.nlm.nih.gov/) has gone from two to 267, and more are coming with initiatives like Darwin Tree of Life and The Earth BioGenome Project, aspiring to sequence the biodiversity on Earth. The high throughput of high-quality chromosome-level genomes will enable large-scale studies of chromosome rearrangements, evolutionary rate variation and the effects of gBGC and codon usage, non-coding regulatory element evolution, gene expansion and loss and the enigmatic TEs. The future for evolutionary genomics is promising and will answer many of the questions above, and likely produce as many new questions waiting to be answered.

Svensk sammanfattning

Evolutionär genetik handlar om att undersöka och karaktärisera processer på molekylär nivå som ligger till grund för evolutionära förändringar i populationer och arter. Genetisk variation mellan individer är en förutsättning för evolutionen. Dessa variationer kan vara enstaka basmutationer eller strukturella kromosomavvikelser som involverar hela eller delar av kromosomen. Det som styr hur genetiska varianter av en gen, så kallade alleler, ökar eller minskar i frekvens i populationen är olika evolutionära krafter, de viktigaste är mutationsfrekvensen, genetisk drift, naturligt urval (också kallat selektion) och rekombination av olika genetiska varianter. I min avhandling har jag studerat hur strukturen på arvsmassan och rekombination i samverkan med olika evolutionära krafter påverkat arvsmassan hos olika fjärilsarter samt hur de svarat på olika miljöbetingelser.

Genetisk rekombination innebär att genetiskt material blandas och fördelas till nästa generation, och det sker framför allt på två sätt. Varje kromosompar fördelas oberoende av varandra till nästa generation så att den nya könscellen får en blandning av kromosomer med olika ursprung. Det resulterar i en blandning av gener som ligger på olika kromosomer, men gener och funktionella delar som ligger på samma kromosom ärvs då alltid tillsammans. Det kan vara ett problem om det finns varianter med nackdelar och fördelar på samma kromosom, då kommer inte selektionen att kunna verka lika effektivt eftersom selektionstrycket är summan av effekterna av länkade delar. DNA-rekombination är en lösning på problemet, det innebär ett utbyte av genetiskt material mellan kromosomerna inom varje kromosompar, på så sätt bryts länkade varianter hos föräldrakromosomerna, och nya kombinationer av alleler uppkommer.

Vi har karakteriserat omfattningen av kromosomförändringar hos vitvingar, ett kryptiskt artkomplex där det nyligen upptäcktes att de var tre olika arter, skogsvitvinge (*Leptidea sinapis*), ängsvitvinge (*L. juvernica*) och Reals vitvinge (*L. reali*). Inom gruppen fjärilar och malar har de flesta undersökta arter välbevarad kromosomuppsättning och genordning över väldigt långa evolutionära avstånd, över 130 miljoner år. *Leptidea* är ett av undantagen, de varierar kraftigt i kromosomantal mellan arterna och skogsvitvingen har den största variationen i kromosomtal som hittills upptäckts inom en art. Vi visade att endast mycket korta segment av genordningen var bevarad och att de omfattande kromosomförändringarna mellan och inom arterna framför allt utgörs av multipla fusioner och fissioner av kromosomerna. Vi fann också flera fall

av kromosomavvikelser hos individer inom arterna, vilket tyder på att kromosominstabiliteten är en pågående process. Antalet kromosomer kan ha betydelse för rekombinationen och påverka den genetiska mångfalden i populationen. Vi undersökte hur detta påverkades hos två populationer av skogsvitvinge med skillnad i kromosomantal, en svensk population med ca 29 kromosompar och en katalansk population med cirka 52 kromosompar. Vi fann högre rekombinationshastighet i populationen med högre kromosomtal, men skillnaden var mindre än förväntat. Det beror på att populationen med få kromosomer också har större kromosomer med större sannolikhet för dubbla överkorsningar. Hos de kromosomer som genomgått omstruktureringar såg vi att ökningen i rekombinationshastighet hade ett samband med ökad genetisk variation.

Vi var intresserade av hur arvsmassans baskomposition, det vill säga innehållet av de olika kvävebaserna i DNA; adenin (A), tymin (T), guanin (G) och cytosin (C), påverkas och påverkar olika evolutionära processer. Basparsmutationer sker oftare från G eller C till A eller T, vilket ger en högre andel A och T hos de flesta organismer, men finns också mekanismer som verkar åt andra hållet, som alltså ökar G/C-innehållet. Vi undersökte hur dessa mekanismer påverkar arvsmassan i olika fjärilar och fann att det är att intrikat samspel mellan mutationsbias, fixeringsbias och andra evolutionära krafter.

Förmågan att använda flera värdväxter eller att byta värdväxt kan vara avgörande för om en art ska kunna överleva förändringar i miljön eller hur framgångsrikt den kan utöka sitt utbredningsområde. Vi tog reda på hur två populationer av skogsvitvinge från Sverige och Katalonien svarade på användandet av två olika värdväxter, en med generell utbredning och en som enbart finns vid Medelhavet. Det visade sig att de svenska honorna i huvudsak lade ägg på den växt som också finns i Sverige medan de katalanska inte visade någon preferens. Larverna från båda populationerna hade längre utvecklingstid och visade en mindre vuxenstorlek när de utfodrades med den växt som inte finns i Sverige. Det var en liten skillnad i larvernas genuttryck och kompositionen hos tarmmikrobiomet mellan värdväxtgrupperna, men skillnaden var olika beroende på var larven hade sin ursprungspopulation. Användningen av en viss värdväxt är följaktligen en kombination av miljöförhållanden och lokal anpassning.

Många arter flyttar, migrerar, som ett sätt att minska trycket från konkurrenter, parasiter och för att öka möjligheten att hitta resurser. Det är en framgångsrik strategi för många men det innebär också en avvägning mellan fördelar och kostnader i form av energiförbrukning, risker under flytten och att de ofta måste skjuta upp fortplantningen tills flytten är avklarad. Här har vi undersökt en riktig långflygare, tistelfjärilen (*Vanessa cardui*), som gör en av djurvärldens längsta migrationer från söder om Sahara upp till norra Skandinavien. Migrationen tar flera generationer på vägen norrut på våren och något färre på återfärden på hösten. Vi vet inte vad som sätter igång migrationen på våren eller hur varje ny generation vet om de ska stanna eller i vilken riktning

de ska fortsätta, men vi har gläntat lite på dörren till de genetiska mekanismerna som ligger bakom. Till att börja med så gjorde vi en kartläggning av de funktionella generna i tistelfjärilens genom, och vi fann kandidatgener genom att titta på vilka genfamiljer som genomgått högre expansion i tistelfjärilen jämfört med andra arter från samma familj (praktfjärilar). De duplicerade generna var bland annat involverade i fett- och proteinmetabolism samt immunförsvar, alla viktiga funktioner hos en migrerande fjäril. De flesta av de duplicerade generna låg grupperade tillsammans inom varje genfamilj, vilket indikerar att de troligen skapats genom upprepade så kallade tandemduplikationer. Vi undersökte också överkorsningsmönstret och fann ett starkt omvänt samband med kromosomstorlek. Förekomsten av ett transposabelt element, en hoppande gen, SINE, var kopplat till variationer i överkorsningshastigheten, och det använder troligtvis rekombinationsmaskineriet för att multiplicera sig i arvsmassan. Som en del i att öka vår förståelse för hur olika miljöfaktorer påverkar beteenden som reproduktion och migration, och vilka gener och regulatoriska nätverk som är involverade, undersökte vi hur tillgången på värdväxter påverkar äggläggningen hos vuxna tistelfjärilshonor. Vi fann att honorna med värdväxt lade betydligt fler ägg än de som var utan värdväxt. En molekylär undersökning av genaktiveringsgraden visade en viss skillnad i gener som kan vara kandidater för fortsatt forskning för att för att förstå de genetiska mekanismerna bakom tistelfjärilens reproduktiva beteende och migratoriska livscykel.

Sammantaget så har vi visat hur kromosomavvikelser kan vara mer omfattande än vad som kromosomantalet visar, samt att kromosomförändringar påverkar genetisk variation genom en förändrad överkorsningshastighet. Vi har också visat på värdet av att karaktärisera strukturella variationer, överkorsningshastighet och transposabla element för att undersöka olika evolutionära processer. I flera av projekten har vi utvecklat resurser som genkopplingskartor och funktionell kartläggning för att besvara frågor hur strukturella förändringar och olika evolutionära krafter samverkar och påverkar arvsmassan, som förhoppningsvis kan bidra till fortsatt forskning.

Acknowledgements

A PhD is supposed to teach you to become an independent scientist, but more importantly, you learn how to cooperate. Ideas need to be honed and trimmed in the heat of arguments, and the most rewarding parts of science are the discussions. I want to thank all the people at all departments at EBC for making these years so wonderful, all of you who light up my day when I meet you in the corridor or at the coffee machine!

Thank you, Niclas, for giving me the opportunity to work with these exciting questions. By providing an open environment and always being encouraging and supportive, you have let me explore my own paths, but at the same time, kept me on track, which is not an easy task, I know. I have learnt so much from you during these years. There have been some wonderful field trips and excursions; we should always have our lab-meetings outdoors!

Thanks to Martin for being co-supervisor, the always invigorating political discussions, sharing the tolls of commuting, and the book-club discussions providing much more than the actual text.

Thanks, Venkat, for helping out at the start, for all the fun adventures, for meeting your family and cousins, and for the great pleasure of getting to know Manu! Thanks, Jesper, so full of genius and knowledge. It has been great fun arguing and discussing science and life with you. Thank you, Lars, for your hard work and wise and well-thought-out insights in the field, lab and discussions. Don't forget to go to your cabin, especially during your final year. Dasha, the dream team leader, thank you for all your support and for being a fantastic friend. You are a great scientist and a role model, for work and life (and thank you, Max, and my lab-assistants Marfa and Tisha). Thanks, Linnéa, for being a kind and caring person (and biologist, yes you are), for your organisational skills and friendly energy. Veronika, thanks for the excellent cooperation and for being an enthusiastic friend. Thanks to Alessandro, Luis, Per, Peter, Laura, Snow, AnneMieke, Orazio, Melih, Yishu, Elenia, and everyone who joined the lab for fun field trips and great discussions.

Thanks, Gerard, for delivering so much energy in the form of great ideas and candy, and Aurora, keep up your good work and amazing lifestyle. To Roger Vila, it has been a privilege working with you, and Vlad Dincă, your enthusiasm in the field is contagious. Thank you, Christer Wiklund, for the commitment to the butterflies in Stockholm, you are such a great inspiration and motivator! To Axel Künstner, for our brief but fun cooperation!

Philipp, I am so happy you are my office mate. You have been a great friend and support throughout this journey. I would not have survived without our well-needed coffee breaks and sour candy. Madee, I have such respect for your intelligent opinions on science and society, giving me great hope for the future. Erik, thanks for literally sustaining the world, you are not just thinking about saving the world, you are actually doing it. David, you are so positive, I always end up in a good mood when talking to you. Thanks, Gabriel, for your interesting stories and discussions on life. Good luck with your PhD, Milena, you have come to a great place. Thank you, Aleix, glad you came back and for taking over my teaching.

Thank you to everyone else I have had the privilege to meet at EBC. To the PIs in the Evolutionary Biology programme, Anna, Elina, Arild, Hans and Vaishali, for creating the scientific environment; I admire your achievements. For engaging and fun lunchtimes Veera, Martyna, Ioana, Axel, Sara, Thad, McKenna and all former colleagues Merce, Agnese, Shadi, Carina, Arvid, Zaenab, Marisol, Berrit, Alex, Vale, Mahwash, Homa, Willian, Ghazal, Marcin, Kerri, Robert, Taki, and everyone I have not mentioned.

Thank you, Ingrid, for showing how an excellent teacher cares for the students. Thanks, Martin, for sharing your sharp scientific thinking and for all the nice chats! Thanks, Julian, for the great company at Klubban! To Ronny, Tommy, Malin, Andreas, Peter, Banafsheh, Afshaneh and Mats for helping me during teaching at Klubban and EBC. Thank you, Doug, Ylva, Malin, Mattias, Yvonne, Gunilla, Christoffer, and the admin, Frida, Annette and Jenny, for assisting me over the years.

I also want to thank the funding agencies Anna-Maria Lundin, Tullbergs f biologisk forskning, Thelins G, Bjurzons, Ellen och Tage Westins, W Stenholms, for travel grants allowing me to participate in conferences and field work. Thanks to NGI at SciLifeLab, Stockholm and especially Elísabet Einarsdóttir, for patiently handling our sequencing orders.

I want to thank my friends, Verena and Robin, for sharing home office in the first part of the plague, making the lockdown endurable. Thank you, Evgenia and Torunn, Anna-Karin and Henrik, and many more at HHAB. Thanks to Love Dalén and Erik, Rasa, Patricia, Johanna and David at SU for introducing me to evolutionary genomics in my master project.

Many, many thanks to my family for their unconditional support for all my crazy ideas. Thank you, Anna, for being the best big sister and for keeping an eye on my house. To Håkan, Linus, Johanna and Fredrik with families. Thanks to Anton for keeping me sane during home office writing with regular walks and coffee breaks. Thanks to Sara for being the best little sister and a support in sickness and in health. And to Sezar, Henning, Lukas, Axel and Arvid, for being there and giving my life so much life. And thank you, Kent, for everything and more. Ett stort tack till min pappa, Lars-Eric, för all uppmuntran och för att du lärt oss att allt är möjligt. Till minne av mamma, Birgitta, du finns kvar, alltid.

References

Ahola, V., Lehtonen, R., Somervuo, P., Salmela, L., Koskinen, P., … Hanski, I. (2014). The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera. *Nature Communications*, *5*(1), Article 1. https://doi.org/10.1038/ncomms5737

Aksenova, A. Y., Greenwell, P. W., Dominska, M., Shishkin, A. A., Kim, J. C., … Mirkin, S. M. (2013). Genome rearrangements caused by interstitial telomeric sequences in yeast. *Proceedings of the National Academy of Sciences*, *110*(49), 19866–19871. https://doi.org/10.1073/pnas.1319313110

Alexa, A., & Rahnenfuhrer, J. (2021). *topGO: Enrichment analysis for gene ontology* (2.44.0). Bioconductor version: Release (3.13). https://doi.org/10.18129/B9.bioc.topGO

Aljanabi, S. (1997). Universal and rapid salt-extraction of high quality genomic DNA for PCR- based techniques. *Nucleic Acids Research*, *25*(22), 4692–4693. https://doi.org/10.1093/nar/25.22.4692

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Anderson, L. K., Doyle, G. G., Brigham, B., Carter, J., Hooker, K. D., … Stack, S. M. (2003). High-resolution crossover maps for each bivalent of Zea mays using recombination nodules. *Genetics*, *165*(2), 849–865.

Apuli, R.-P., Bernhardsson, C., Schiffthaler, B., Robinson, K. M., Jansson, S., … Ingvarsson, P. K. (2020). Inferring the genomic landscape of recombination rate variation in European aspen (*Populus tremula*). *G3;Genes|Genomes|Genetics*, *10*(1), 299–309. https://doi.org/10.1534/g3.119.400504

Arbeithuber, B., Betancourt, A. J., Ebner, T., & Tiemann-Boege, I. (2015). Crossovers are associated with mutation and biased gene conversion at recombination hotspots. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(7), 2109–2114. https://doi.org/10.1073/pnas.1416622112

Backström, N., Zhang, Q., & Edwards, S. V. (2013). Evidence from a house finch (*Haemorhous mexicanus*) spleen transcriptome for adaptive evolution and biased gene conversion in passerine birds. *Molecular Biology and Evolution*, *30*(5), 1046–1050. https://doi.org/10.1093/molbev/mst033

Barnes, T. M., Kohara, Y., Coulson, A., & Hekimi, S. (1995). Meiotic recombination, noncoding DNA and genomic organization in *Caenorhabditis elegans*. *Genetics*, *141*(1), 159–179. https://doi.org/10.1093/genetics/141.1.159

Barrett, L. W., Fletcher, S., & Wilton, S. D. (2012). Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding

elements. *Cellular and Molecular Life Sciences*, *69*(21), 3613–3634. https://doi.org/10.1007/s00018-012-0990-9
Barton, N. H. (1998). The effect of hitch-hiking on neutral genealogies. *Genetics Research*, *72*(2), 123–133. https://doi.org/10.1017/S0016672398003462
Baudat, F., Buard, J., Grey, C., Fledel-Alon, A., Ober, C., … de Massy, B. (2010). PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science (New York, N.Y.)*, *327*(5967), 836–840. https://doi.org/10.1126/science.1183439
Bauer, H. (1967). Die kinetische Organisation der Lepidopteren-Chromosomen. *Chromosoma*, *22*(2), 101–125. https://doi.org/10.1007/BF00326724
Begun, D. J., & Aquadro, C. F. (1991). Molecular population genetics of the distal portion of the X chromosome in *Drosophila*: Evidence for genetic hitchhiking of the yellow-achaete region. *Genetics*, *129*(4), 1147–1158. https://doi.org/10.1093/genetics/129.4.1147
Begun, D. J., & Aquadro, C. F. (1992). Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature*, *356*(6369), Article 6369. https://doi.org/10.1038/356519a0
Behura, S. K., & Severson, D. W. (2011). Coadaptation of isoacceptor tRNA genes and codon usage bias for translation efficiency in *Aedes aegypti* and *Anopheles gambiae*. *Insect Molecular Biology*, *20*(2), 177–187. https://doi.org/10.1111/j.1365-2583.2010.01055.x
Behura, S. K., & Severson, D. W. (2012). Comparative analysis of codon usage bias and codon context patterns between dipteran and hymenopteran sequenced genomes. *PLoS ONE*, *7*(8), e43111. https://doi.org/10.1371/journal.pone.0043111
Betancourt, A. J., & Presgraves, D. C. (2002). Linkage limits the power of natural selection in *Drosophila*. *Proceedings of the National Academy of Sciences*, *99*(21), 13616–13620. https://doi.org/10.1073/pnas.212277199
Birdsell, J. A. (2002). Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. *Molecular Biology and Evolution*, *19*(7), 1181–1197. https://doi.org/10.1093/oxfordjournals.molbev.a004176
Bishop, D. K., & Zickler, D. (2004). Early decision: Meiotic crossover interference prior to stable strand exchange and synapsis. *Cell*, *117*(1), 9–15. https://doi.org/10.1016/S0092-8674(04)00297-1
Boggs, C. L., Watt, W. B., & Ehrlich, P. R. (Eds.). (2003). *Butterflies: Ecology and Evolution Taking Flight*. University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/B/bo3644192.html
Boman, J., Mugal, C. F., & Backström, N. (2021). The effects of GC-biased gene conversion on patterns of genetic diversity among and across butterfly genomes. *Genome Biology and Evolution*, *13*(5), evab064. https://doi.org/10.1093/gbe/evab064
Brazier, T., & Glémin, S. (2022). Diversity and determinants of recombination landscapes in flowering plants. *PLOS Genetics*, *18*(8), e1010141. https://doi.org/10.1371/journal.pgen.1010141
Brower, L. P., Fink, L. S., & Walford, P. (2006). Fueling the fall migration of the monarch butterfly. *Integrative and Comparative Biology*, *46*(6), 1123–1142. https://doi.org/10.1093/icb/icl029
Brown, T. C., & Jiricny, J. (1987). A specific mismatch repair event protects

mammalian cells from loss of 5-methylcytosine. *Cell*, *50*(6), 945–950. https://doi.org/10.1016/0092-8674(87)90521-6
Burri, R., Nater, A., Kawakami, T., Mugal, C. F., Olason, P. I., … Ellegren, H. (2015). Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of *Ficedula* flycatchers. *Genome Research*, *25*(11), 1656–1665. https://doi.org/10.1101/gr.196485.115
Bush, G. L., Case, S. M., Wilson, A. C., & Patton, J. L. (1977). Rapid speciation and chromosomal evolution in mammals. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(9), 3942–3946. https://doi.org/10.1073/pnas.74.9.3942
Cai, J. J., Macpherson, J. M., Sella, G., & Petrov, D. A. (2009). Pervasive hitchhiking at coding and regulatory sites in humans. *PLOS Genetics*, *5*(1), e1000336. https://doi.org/10.1371/journal.pgen.1000336
Castellano, D., Coronado-Zamora, M., Campos, J. L., Barbadilla, A., & Eyre-Walker, A. (2016). Adaptive evolution is substantially impeded by Hill–Robertson interference in *Drosophila*. *Molecular Biology and Evolution*, *33*(2), 442–455. https://doi.org/10.1093/molbev/msv236
Castellano, D., James, J., & Eyre-Walker, A. (2018). Nearly neutral evolution across the *Drosophila melanogaster* genome. *Molecular Biology and Evolution*, *35*(11), 2685–2694. https://doi.org/10.1093/molbev/msy164
Castiglia, R., & Capanna, E. (2002). Chiasma repatterning across a chromosomal hybrid zone between chromosomal races of *Mus musculus domesticus*. *Genetica*, *114*(1), 35–40. https://doi.org/10.1023/A:1014626330022
Celorio-Mancera, M. de la P., Wheat, C. W., Huss, M., Vezzi, F., Neethiraj, R., … Janz, N. (2016). Evolutionary history of host use, rather than plant phylogeny, determines gene expression in a generalist butterfly. *BMC Evolutionary Biology*, *16*(1), 59. https://doi.org/10.1186/s12862-016-0627-y
Chan, A. H., Jenkins, P. A., & Song, Y. S. (2012). Genome-wide fine-scale recombination rate variation in *Drosophila melanogaster*. *PLOS Genetics*, *8*(12), e1003090. https://doi.org/10.1371/journal.pgen.1003090
Chapman, J. W., Reynolds, D. R., & Wilson, K. (2015). Long-range seasonal migration in insects: Mechanisms, evolutionary drivers and ecological consequences. *Ecology Letters*, *18*(3), 287–302. https://doi.org/10.1111/ele.12407
Chen, B., Du, K., Sun, C., Vimalanathan, A., Liang, X., … Shao, Y. (2018). Gut bacterial and fungal communities of the domesticated silkworm (Bombyx mori) and wild mulberry-feeding relatives. *The ISME Journal*, *12*(9), Article 9. https://doi.org/10.1038/s41396-018-0174-1
Chen, J.-M., Cooper, D. N., Chuzhanova, N., Férec, C., & Patrinos, G. P. (2007). Gene conversion: Mechanisms, evolution and human disease. *Nature Reviews Genetics*, *8*(10), Article 10. https://doi.org/10.1038/nrg2193
Chen, S., Krinsky, B. H., & Long, M. (2013). New genes as drivers of phenotypic evolution. *Nature Reviews Genetics*, *14*(9), Article 9. https://doi.org/10.1038/nrg3521
Chowdhury, S., Fuller, R. A., Dingle, H., Chapman, J. W., & Zalucki, M. P. (2021). Migration in butterflies: A global overview. *Biological Reviews*, *96*(4), 1462–1483. https://doi.org/10.1111/brv.12714
Chung, G., Rose, A. M., Petalcorin, M. I. R., Martin, J. S., Kessler, Z., …

Boulton, S. J. (2015). REC-1 and HIM-5 distribute meiotic crossovers and function redundantly in meiotic double-strand break formation in *Caenorhabditis elegans*. *Genes & Development*, *29*(18), 1969–1979. https://doi.org/10.1101/gad.266056.115

Cicconardi, F., Lewis, J. J., Martin, S. H., Reed, R. D., Danko, C. G., & Montgomery, S. H. (2021). Chromosome fusion affects genetic diversity and evolutionary turnover of functional loci but consistently depends on chromosome size. *Molecular Biology and Evolution*, *38*(10), 4449–4462. https://doi.org/10.1093/molbev/msab185

Cole, F., Keeney, S., & Jasin, M. (2010). Comprehensive, fine-scale dissection of homologous recombination outcomes at a hot spot in mouse meiosis. *Molecular Cell*, *39*(5), 700–710. https://doi.org/10.1016/j.molcel.2010.08.017

Connahs, H., Rhen, T., & Simmons, R. B. (2016). Transcriptome analysis of the painted lady butterfly, *Vanessa cardui* during wing color pattern development. *BMC Genomics*, *17*(1), 270. https://doi.org/10.1186/s12864-016-2586-5

Coop, G., Wen, X., Ober, C., Pritchard, J. K., & Przeworski, M. (2008). High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. *Science (New York, N.Y.)*, *319*(5868), 1395–1398. https://doi.org/10.1126/science.1151851

Coulondre, C., Miller, J. H., Farabaugh, P. J., & Gilbert, W. (1978). Molecular basis of base substitution hotspots in Escherichia coli. *Nature*, *274*(5673), Article 5673. https://doi.org/10.1038/274775a0

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., … Jaenisch, R. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences*, *107*(50), 21931–21936. https://doi.org/10.1073/pnas.1016071107

Crosland, M. W. J., & Crozier, R. H. (1986). *Myrmecia pilosula*, an ant with only one pair of chromosomes. *Science*, *231*(4743), 1278–1278. https://doi.org/10.1126/science.231.4743.1278

Cutter, A. D., Baird, S. E., & Charlesworth, D. (2006). High nucleotide polymorphism and rapid decay of linkage disequilibrium in wild populations of *Caenorhabditis remanei*. *Genetics*, *174*(2), 901–913. https://doi.org/10.1534/genetics.106.061879

de Vos, J. M., Augustijnen, H., Bätscher, L., & Lucek, K. (2020). Speciation through chromosomal fusion and fission in Lepidoptera. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1806), 20190539. https://doi.org/10.1098/rstb.2019.0539

Deininger, P. L., & Batzer, M. A. (1999). Alu repeats and human disease. *Molecular Genetics and Metabolism*, *67*(3), 183–193. https://doi.org/10.1006/mgme.1999.2864

Dennis, A. B., Ballesteros, G. I., Robin, S., Schrader, L., Bast, J., … Gadau, J. (2020). Functional insights from the GC-poor genomes of two aphid parasitoids, *Aphidius ervi* and *Lysiphlebus fabarum*. *BMC Genomics*, *21*(1), 376. https://doi.org/10.1186/s12864-020-6764-0

Denver, D. R., Dolan, P. C., Wilhelm, L. J., Sung, W., Lucas-Lledó, J. I., … Baer, C. F. (2009). A genome-wide view of *Caenorhabditis elegans* base-substitution mutation processes. *Proceedings of the National Academy of Sciences*, *106*(38), 16310–16314. https://doi.org/10.1073/pnas.0904895106

Dincă, V., Lukhtanov, V. A., Talavera, G., & Vila, R. (2011). Unexpected layers of cryptic diversity in wood white *Leptidea* butterflies. *Nature Communications*, *2*(1), Article 1. https://doi.org/10.1038/ncomms1329

Dumas, D., & Britton-Davidian, J. (2002). Chromosomal rearrangements and evolution of recombination: Comparison of chiasma distribution patterns in standard and Robertsonian populations of the house mouse. *Genetics*, *162*(3), 1355–1366. https://doi.org/10.1093/genetics/162.3.1355

Duret, L. (2000). TRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes. *Trends in Genetics*, *16*(7), 287–289. https://doi.org/10.1016/S0168-9525(00)02041-2

Duret, L., & Galtier, N. (2009). Biased gene conversion and the evolution of mammalian genomic landscapes. *Annual Review of Genomics and Human Genetics*, *10*(1), 285–311. https://doi.org/10.1146/annurev-genom-082908-150001

Dutheil, J., & Boussau, B. (2008). Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. *BMC Evolutionary Biology*, *8*(1), 255. https://doi.org/10.1186/1471-2148-8-255

Ehrlich, M., & Wang, R. Y. (1981). 5-Methylcytosine in eukaryotic DNA. *Science (New York, N.Y.)*, *212*(4501), 1350–1357. https://doi.org/10.1126/science.6262918

Emms, D. M., & Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*(1), 238. https://doi.org/10.1186/s13059-019-1832-y

Eyre-Walker, A. (1999). Evidence of selection on silent site base composition in mammals: Potential implications for the evolution of isochores and junk DNA. *Genetics*, *152*(2), 675–683. https://doi.org/10.1093/genetics/152.2.675

Felsenstein, J. (1974). The evolutionary advantage of recombination. *Genetics*, *78*, 737–756.

Figuet, E., Ballenghien, M., Romiguier, J., & Galtier, N. (2015). Biased gene conversion and gc-content evolution in the coding sequences of reptiles and vertebrates. *Genome Biology and Evolution*, *7*(1), 240–250. https://doi.org/10.1093/gbe/evu277

Fischer, G., Neuvéglise, C., Durrens, P., Gaillardin, C., & Dujon, B. (2001). Evolution of gene order in the genomes of two related yeast species. *Genome Research*, *11*(12), 2009–2019. https://doi.org/10.1101/gr.212701

Friberg, M., Bergman, M., Kullberg, J., Wahlberg, N., & Wiklund, C. (2008). Niche separation in space and time between two sympatric sister species—A case of ecological pleiotropy. *Evolutionary Ecology*, *22*(1), 1–18. https://doi.org/10.1007/s10682-007-9155-y

Friberg, M., & Wiklund, C. (2009). Host plant preference and performance of the sibling species of butterflies *Leptidea sinalis* and *Leptidea reali*: A test of the trade-off hypothesis for food specialisation. *Oecologia*, *159*(1), 127–137. https://doi.org/10.1007/s00442-008-1206-8

Fryxell, K. J., & Zuckerkandl, E. (2000). Cytosine deamination plays a primary role in the evolution of mammalian isochores. *Molecular Biology and Evolution*, *17*(9), 1371–1383. https://doi.org/10.1093/oxfordjournals.molbev.a026420

Fujiwara, H., Nakazato, Y., Okazaki, S., & Ninaki, O. (2000). Stability and telomere structure of chromosomal fragments in two different mosaic

strains of the Silkworm, *Bombyx mori*. *Zoological Science*, *17*(6), 743–750. https://doi.org/10.2108/zsj.17.743
Galtier, N. (2016). Adaptive protein evolution in animals and the effective population size hypothesis. *PLOS Genetics*, *12*(1), e1005774. https://doi.org/10.1371/journal.pgen.1005774
Galtier, N., Bazin, E., & Bierne, N. (2006). GC-biased segregation of noncoding polymorphisms in *Drosophila*. *Genetics*, *172*(1), 221–228. https://doi.org/10.1534/genetics.105.046524
Galtier, N., Roux, C., Rousselle, M., Romiguier, J., Figuet, E., … Duret, L. (2018). Codon usage bias in animals: Disentangling the effects of natural selection, effective population size, and GC-biased gene conversion. *Molecular Biology and Evolution*, *35*(5), 1092–1103. https://doi.org/10.1093/molbev/msy015
García-Berro, A., Talla, V., Vila, R., Wai, H. K., Shipilina, D., … Talavera, G. (2022). Migratory behaviour is positively associated with genetic diversity in butterflies. *Molecular Ecology*. https://doi.org/10.1111/mec.16770
Garel, J.-P. (1974). Functional adaptation of tRNA population. *Journal of Theoretical Biology*, *43*, 211–225. https://doi.org/doi.org/10.1016/S0022-5193(74)80054-8
Giménez, M. D., White, T. A., Hauffe, H. C., Panithanarak, T., & Searle, J. B. (2013). Understanding the basis of diminished gene flow between hybridizing chromosome races of the house mouse. *Evolution*, *67*(5), 1446–1462. https://doi.org/10.1111/evo.12054
Glémin, S., Arndt, P. F., Messer, P. W., Petrov, D., Galtier, N., & Duret, L. (2015). Quantification of GC-biased gene conversion in the human genome. *Genome Research*, *25*(8), 1215–1228. https://doi.org/10.1101/gr.185488.114
Gorbunova, V., & Levy, A. A. (1997). Non-homologous DNA end joining in plant cells is associated with deletions and filler DNA insertions. *Nucleic Acids Research*, *25*(22), 4650–4657. https://doi.org/10.1093/nar/25.22.4650
Green, D. A., & Kronforst, M. R. (2019). Monarch butterflies use an environmentally sensitive, internal timer to control overwintering dynamics. *Molecular Ecology*, *28*(16), 3642–3655. https://doi.org/10.1111/mec.15178
Gu, W., Zhang, F., & Lupski, J. R. (2008). Mechanisms for human genomic rearrangements. *PathoGenetics*, *1*(1), 4. https://doi.org/10.1186/1755-8417-1-4
Guerrero, R. F., & Kirkpatrick, M. (2014). Local adaptation and the evolution of chromosome fusions. *Evolution*, *68*(10), 2747–2756. https://doi.org/10.1111/evo.12481
Haberle, V., & Stark, A. (2018). Eukaryotic core promoters and the functional basis of transcription initiation. *Nature Reviews Molecular Cell Biology*, *19*(10), Article 10. https://doi.org/10.1038/s41580-018-0028-8
Haenel, Q., Laurentino, T. G., Roesti, M., & Berner, D. (2018). Meta-analysis of chromosome-scale crossover rate variation in eukaryotes and its significance to evolutionary genomics. *Molecular Ecology*, *27*(11), 2477–2497. https://doi.org/10.1111/mec.14699
Hahn, M. W. (2018). *Molecular population genetics*. Oxford University Press ; Sinauer Associates.
Halldorsson, B. V., Palsson, G., Stefansson, O. A., Jonsson, H., Hardarson,

M. T., … Stefansson, K. (2019). Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science*, *363*(6425), eaau1043. https://doi.org/10.1126/science.aau1043

Hamann, E., Pauli, C. S., Joly-Lopez, Z., Groen, S. C., Rest, J. S., … Franks, S. J. (2021). Rapid evolutionary changes in gene expression in response to climate fluctuations. *Molecular Ecology*, *30*(1), 193–206. https://doi.org/10.1111/mec.15583

Hedges, D. J., & Deininger, P. L. (2007). Inviting instability: Transposable elements, double-strand breaks, and the maintenance of genome integrity. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, *616*(1–2), 46–59. https://doi.org/10.1016/j.mrfmmm.2006.11.021

Heintzman, N. D., Hon, G. C., Hawkins, R. D., Kheradpour, P., Stark, A., … Ren, B. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, *459*(7243), Article 7243. https://doi.org/10.1038/nature07829

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., … Ren, B. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics*, *39*(3), Article 3. https://doi.org/10.1038/ng1966

Herickhoff, L., Stephen Stack, & Sherman, J. (1993). The relationship between synapsis, recombination nodules and chiasmata in tomato translocation heterozygotes. *Heredity*, *71*. https://doi.org/10.1038/hdy.1993.152

Hill, W. G., & Robertson, A. (1966). The effect of linkage on limits to artificial selection. *Genetical Research*, *8*, 269–294. https://doi.org/10.1017/S001667230800949X

Holm, P. B., & Rasmussen, S. W. (1980). Chromosome pairing, recombination nodules and chiasma formation in diploid Bombyx males. *Carlsberg Research Communications*, *45*(6), 483. https://doi.org/10.1007/BF02932921

Hooper, D. M., Griffith, S. C., & Price, T. D. (2019). Sex chromosome inversions enforce reproductive isolation across an avian hybrid zone. *Molecular Ecology*, *28*(6), 1246–1262. https://doi.org/10.1111/mec.14874

Howe, F. S., Fischl, H., Murray, S. C., & Mellor, J. (2017). Is H3K4me3 instructive for transcription activation? *BioEssays*, *39*(1), e201600095. https://doi.org/10.1002/bies.201600095

Hu, Y., Wu, X., Jin, G., Peng, J., Leng, R., … Zhang, C. (2022). Rapid genome evolution and adaptation of *Thlaspi arvense* mediated by recurrent RNA-based and tandem gene duplications. *Frontiers in Plant Science*, *12*. https://www.frontiersin.org/articles/10.3389/fpls.2021.772655

Hudson, R. R. (2001). Two-locus sampling distributions and their application. *Genetics*, *159*(4), 1805–1817. https://doi.org/10.1093/genetics/159.4.1805

Hughes-Schrader, S. (1948). Cytology of Coccids (Coccoïdea-Homoptera). In M. Demerec (Ed.), *Advances in Genetics* (Vol. 2, pp. 127–203). Academic Press. https://doi.org/10.1016/S0065-2660(08)60468-X

Iiams, S. E., Lugena, A. B., Zhang, Y., Hayden, A. N., & Merlin, C. (2019). Photoperiodic and clock regulation of the vitamin A pathway in the brain mediates seasonal responsiveness in the monarch butterfly. *Proceedings of the National Academy of Sciences*, *116*(50), 25214–25221. https://doi.org/10.1073/pnas.1913915116

IJdo, J. W., Baldini, A., Ward, D. C., Reeders, S. T., & Wells, R. A. (1991). Origin of human chromosome 2: An ancestral telomere-telomere fusion. *Proceedings of the National Academy of Sciences*, *88*(20), 9051–9055. https://doi.org/10.1073/pnas.88.20.9051

Innan, H., & Kondrashov, F. (2010). The evolution of gene duplications: Classifying and distinguishing between models. *Nature Reviews Genetics*, *11*(2), Article 2. https://doi.org/10.1038/nrg2689

Jin, X., Fudenberg, G., & Pollard, K. S. (2021). Genome-wide variability in recombination activity is associated with meiotic chromatin organization. *Genome Research*, *31*(9), 1561–1572. https://doi.org/10.1101/gr.275358.121

Jordan, I. K., & Miller, W. J. (2009). Genome defense against Transposable Elements and the origins of Regulatory RNA. In D.-H. Lankenau & J.-N. Volff (Eds.), *Transposons and the Dynamic Genome* (pp. 77–94). Springer. https://doi.org/10.1007/7050_2008_041

Kaplan, N. L., Hudson, R. R., & Langley, C. H. (1989). The "hitchhiking effect" revisited. *Genetics*, *123*(4), 887–899. https://doi.org/10.1093/genetics/123.4.887

Katju, V., & Lynch, M. (2003). The structure and early evolution of recently arisen gene duplicates in the *Caenorhabditis elegans* genome. *Genetics*, *165*(4), 1793–1803. https://doi.org/10.1093/genetics/165.4.1793

Kawakami, T., Smeds, L., Backström, N., Husby, A., Qvarnström, A., … Ellegren, H. (2014). A high-density linkage map enables a second-generation collared flycatcher genome assembly and reveals the patterns of avian recombination rate variation and chromosomal evolution. *Molecular Ecology*, *23*(16), 4035–4058. https://doi.org/10.1111/mec.12810

Kawakami, T., Wallberg, A., Olsson, A., Wintermantel, D., de Miranda, J. R., … Webster, M. T. (2019). Substantial heritable variation in recombination rate on multiple scales in honeybees and bumblebees. *Genetics*, *212*(4), 1101–1119. https://doi.org/10.1534/genetics.119.302008

Kazazian, H. H. (2004). Mobile elements: Drivers of genome evolution. *Science*, *303*(5664), 1626–1632. https://doi.org/10.1126/science.1089670

Keeney, S., Giroux, C. N., & Kleckner, N. (1997). Meiosis-specific DNA double-strand breaks are catalyzed by Spo11, a member of a widely conserved protein family. *Cell*, *88*(3), 375–384. https://doi.org/10.1016/S0092-8674(00)81876-0

Keightley, P. D., Pinharanda, A., Ness, R. W., Simpson, F., Dasmahapatra, K. K., … Jiggins, C. D. (2015). Estimation of the spontaneous mutation rate in *Heliconius melpomene*. *Molecular Biology and Evolution*, *32*(1), 239–243. https://doi.org/10.1093/molbev/msu302

Keightley, P. D., Trivedi, U., Thomson, M., Oliver, F., Kumar, S., & Blaxter, M. L. (2009). Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. *Genome Research*, *19*(7), 1195–1201. https://doi.org/10.1101/gr.091231.109

Kidwell, M. G., & Lisch, D. (1997). Transposable elements as sources of variation in animals and plants. *Proceedings of the National Academy of Sciences*, *94*(15), 7704–7711. https://doi.org/10.1073/pnas.94.15.7704

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*, *217*(5129), Article 5129. https://doi.org/10.1038/217624a0

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, *16*(2), 111–120.

https://doi.org/10.1007/BF01731581
Kimura, M., & Ohta, T. (1969). The average number of generations until fixation of a mutant gene in a finite population. *Genetics*, *61*(3), 763–771. https://doi.org/10.1093/genetics/61.3.763
Klug, W. S., Cummings, M. R., Spencer, C. A., & Palladino, M. A. (2017). Chromosome structure and DNA sequence organization. In *Essentials of Genetics* (9th edition, pp. 215–230). Pearson Education Limited.
Knight, R. D., Freeland, S. J., & Landweber, L. F. (2001). A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biology*, *2*(4), research0010.1. https://doi.org/10.1186/gb-2001-2-4-research0010
Koch, E. L., Ravinet, M., Westram, A. M., Johannesson, K., & Butlin, R. K. (2022). Genetic architecture of repeated phenotypic divergence in Littorina saxatilis ecotype evolution. *Evolution*, *76*(10), 2332–2346. https://doi.org/10.1111/evo.14602
Korunes, K. L., & Samuk, K. (2021). pixy: Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Molecular Ecology Resources*, *21*(4), 1359–1368. https://doi.org/10.1111/1755-0998.13326
Kosambi, D. (1944). The etimation of map distances from recombination values. *Annals of Eugenics*, *12*, 172–175.
Kostka, D., Hubisz, M. J., Siepel, A., & Pollard, K. S. (2012). The role of GC-biased gene conversion in shaping the fastest evolving regions of the human genome. *Molecular Biology and Evolution*, *29*(3), 1047–1057. https://doi.org/10.1093/molbev/msr279
Lachance, J., & Tishkoff, S. A. (2014). Biased gene conversion skews allele frequencies in human populations, increasing the disease burden of recessive alleles. *The American Journal of Human Genetics*, *95*(4), 408–420. https://doi.org/10.1016/j.ajhg.2014.09.008
Lamolle, G., Fontenla, S., Rijo, G., Tort, J. F., & Smircich, P. (2019). Compositional analysis of flatworm genomes shows strong codon usage biases across all classes. *Frontiers in Genetics*, *10*, 771. https://doi.org/10.3389/fgene.2019.00771
Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., … The Wellcome Trust: (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), Article 6822. https://doi.org/10.1038/35057062
Leaché, A. D., Banbury, B. L., Linkem, C. W., & de Oca, A. N.-M. (2016). Phylogenomics of a rapid radiation: Is chromosomal evolution linked to increased diversification in north american spiny lizards (Genus Sceloporus)? *BMC Evolutionary Biology*, *16*(1), 63. https://doi.org/10.1186/s12862-016-0628-x
Leal, L., Talla, V., Källman, T., Friberg, M., Wiklund, C., … Backström, N. (2018). Gene expression profiling across ontogenetic stages in the wood white (*Leptidea sinapis*) reveals pathways linked to butterfly diapause regulation. *Molecular Ecology*, *27*(4), 935–948. https://doi.org/10.1111/mec.14501
Lehrman, M. A., Russell, D. W., Goldstein, J. L., & Brown, M. S. (1987). Alu-Alu recombination deletes splice acceptor sites and produces secreted low density lipoprotein receptor in a subject with familial hypercholesterolemia. *The Journal of Biological Chemistry*, *262*(7), 3354–

3361.
Lensing, S. V., Marsico, G., Hänsel-Hertsch, R., Lam, E. Y., Tannahill, D., & Balasubramanian, S. (2016). DSBCapture: In situ capture and sequencing of DNA breaks. *Nature Methods*, *13*(10), Article 10. https://doi.org/10.1038/nmeth.3960
Lercher, M. J., & Hurst, L. D. (2002). Human SNP variability and mutation rate are higher in regions of high recombination. *Trends in Genetics*, *18*(7), 337–340. https://doi.org/10.1016/S0168-9525(02)02669-0
Letsolo, B. T., Rowson, J., & Baird, D. M. (2010). Fusion of short telomeres in human cells is characterized by extensive deletion and microhomology, and can result in complex rearrangements. *Nucleic Acids Research*, *38*(6), 1841–1852. https://doi.org/10.1093/nar/gkp1183
Librado, P., Vieira, F. G., & Rozas, J. (2012). BadiRate: Estimating family turnover rates by likelihood-based methods. *Bioinformatics*, *28*(2), 279–281. https://doi.org/10.1093/bioinformatics/btr623
Lipinski, K. J., Farslow, J. C., Fitzpatrick, K. A., Lynch, M., Katju, V., & Bergthorsson, U. (2011). High spontaneous rate of gene duplication in *Caenorhabditis elegans*. *Current Biology : CB*, *21*(4), 306–310. https://doi.org/10.1016/j.cub.2011.01.026
Liu, H., Huang, J., Sun, X., Li, J., Hu, Y., … Yang, S. (2018). Tetrad analysis in plants and fungi finds large differences in gene conversion rates but no GC bias. *Nature Ecology & Evolution*, *2*(1), 164–173. https://doi.org/10.1038/s41559-017-0372-7
Liu, Y. (2020). A code within the genetic code: Codon usage regulates co-translational protein folding. *Cell Communication and Signaling*, *18*(1), 145. https://doi.org/10.1186/s12964-020-00642-6
Lo, A. W. l., Sabatier, L., Fouladi, B., Pottier, G., Ricoul, M., & Mumane, J. P. (2002). DNA amplification by breakage/fusion/bridge cycles initiated by spontaneous telomere loss in a human cancer cell line. *Neoplasia*, *4*(6), 531–538. https://doi.org/10.1038/sj.neo.7900267
Lohmueller, K. E., Albrechtsen, A., Li, Y., Kim, S. Y., Korneliussen, T., … Nielsen, R. (2011). Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. *PLOS Genetics*, *7*(10), e1002326. https://doi.org/10.1371/journal.pgen.1002326
Long, H. K., Prescott, S. L., & Wysocka, J. (2016). Ever-changing landscapes: Transcriptional enhancers in development and evolution. *Cell*, *167*(5), 1170–1187. https://doi.org/10.1016/j.cell.2016.09.018
Long, H., Sung, W., Kucukyildirim, S., Williams, E., Miller, S. F., … Lynch, M. (2018). Evolutionary determinants of genome-wide nucleotide composition. *Nature Ecology & Evolution*, *2*(2), 237–240. https://doi.org/10.1038/s41559-017-0425-y
Lukhtanov, V. (2015). The blue butterfly *Polyommatus (Plebicula) atlanticus* (Lepidoptera, Lycaenidae) holds the record of the highest number of chromosomes in the non-polyploid eukaryotic organisms. *Comparative Cytogenetics*, *9*(4), Article 4. https://doi.org/10.3897/CompCytogen.v9i4.5760
Lukhtanov, V. A., Dincă, V., Friberg, M., Šíchová, J., Olofsson, M., … Wiklund, C. (2018). Versatility of multivalent orientation, inverted meiosis, and rescued fitness in holocentric chromosomal hybrids. *Proceedings of the National Academy of Sciences*, *115*(41), E9610–E9619. https://doi.org/10.1073/pnas.1802610115

Lukhtanov, V. A., Dincă, V., Talavera, G., & Vila, R. (2011). Unprecedented within-species chromosome number cline in the Wood White butterfly *Leptidea sinapis* and its significance for karyotype evolution and speciation. *BMC Evolutionary Biology*, *11*(1), 109. https://doi.org/10.1186/1471-2148-11-109

Luo, S., Zhang, H., Duan, Y., Yao, X., Clark, A. G., & Lu, J. (2020). The evolutionary arms race between transposable elements and piRNAs in Drosophila melanogaster. *BMC Evolutionary Biology*, *20*(1), 14. https://doi.org/10.1186/s12862-020-1580-3

Lynch, M. (2007). *The Origins of Genome Architecture*. Sinauer.

Lynch, M., & Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science*, *290*(5494), 1151–1155. https://doi.org/10.1126/science.290.5494.1151

Lynch, M., Ye, Z., Urban, L., Maruki, T., & Wei, W. (2022). The linkage-disequilibrium and recombinational landscape in *Daphnia pulex*. *Genome Biology and Evolution*, *14*(11), evac145. https://doi.org/10.1093/gbe/evac145

Malkova, A., Swanson, J., German, M., McCusker, J. H., Housworth, E. A., … Haber, J. E. (2004). Gene conversion and crossing over along the 405-kb left arm of *Saccharomyces cerevisiae* Chromosome VII. *Genetics*, *168*(1), 49–63. https://doi.org/10.1534/genetics.104.027961

Mancera, E., Bourgon, R., Brozzi, A., Huber, W., & Steinmetz, L. M. (2008). High-resolution mapping of meiotic crossovers and noncrossovers in yeast. *Nature*, *454*(7203), 479–485. https://doi.org/10.1038/nature07135

Marais, G. (2003). Biased gene conversion: Implications for genome and sex evolution. *Trends in Genetics*, *19*(6), 330–338. https://doi.org/10.1016/S0168-9525(03)00116-1

Martin, S. H., Davey, J. W., Salazar, C., & Jiggins, C. D. (2019). Recombination rate variation shapes barriers to introgression across butterfly genomes. *PLOS Biology*, *17*(2), e2006288. https://doi.org/10.1371/journal.pbio.2006288

McClintock, B. (1956). Controlling Elements and the Gene. *Cold Spring Harbor Symposia on Quantitative Biology*, *21*(0), 197–216. https://doi.org/10.1101/SQB.1956.021.01.017

McGregor, A. P., Orgogozo, V., Delon, I., Zanet, J., Srinivasan, D. G., … Stern, D. L. (2007). Morphological evolution through multiple cis-regulatory mutations at a single gene. *Nature*, *448*(7153), 587–590. https://doi.org/10.1038/nature05988

McVean, G., Awadalla, P., & Fearnhead, P. (2002). A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics*, *160*(3), 1231–1241. https://doi.org/10.1093/genetics/160.3.1231

Meier, J. I., Salazar, P. A., Kučka, M., Davies, R. W., Dréau, A., … Chan, Y. F. (2021). Haplotype tagging reveals parallel formation of hybrid races in two butterfly species. *Proceedings of the National Academy of Sciences*, *118*(25), e2015005118. https://doi.org/10.1073/pnas.2015005118

Melters, D. P., Paliulis, L. V., Korf, I. F., & Chan, S. W. L. (2012). Holocentric chromosomes: Convergent evolution, meiotic adaptations, and genomic analysis. *Chromosome Research: An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome*

Biology, *20*(5), 579–593. https://doi.org/10.1007/s10577-012-9292-1
Menchetti, M., Guéguen, M., & Talavera, G. (2019). Spatio-temporal ecological niche modelling of multigenerational insect migrations. *Proceedings of the Royal Society B: Biological Sciences*, *286*(1910), 20191583. https://doi.org/10.1098/rspb.2019.1583
Metzger, B. P. H., Wittkopp, P. J., & Coolon, Joseph. D. (2017). Evolutionary dynamics of regulatory changes underlying gene expression divergence among *Saccharomyces* Species. *Genome Biology and Evolution*, *9*(4), 843–854. https://doi.org/10.1093/gbe/evx035
Milholland, B., Dong, X., Zhang, L., Hao, X., Suh, Y., & Vijg, J. (2017). Differences between germline and somatic mutation rates in humans and mice. *Nature Communications*, *8*(1), Article 1. https://doi.org/10.1038/ncomms15183
Moore, R. C., & Purugganan, M. D. (2003). The early stages of duplicate gene evolution. *Proceedings of the National Academy of Sciences*, *100*(26), 15682–15687. https://doi.org/10.1073/pnas.2535513100
Morales, M. E., White, T. B., Streva, V. A., DeFreece, C. B., Hedges, D. J., & Deininger, P. L. (2015). The contribution of Alu elements to mutagenic DNA double-strand break repair. *PLoS Genetics*, *11*(3), e1005016. https://doi.org/10.1371/journal.pgen.1005016
Mordstein, C., Savisaar, R., Young, R. S., Bazile, J., Talmane, L., … Kudla, G. (2020). Codon usage and splicing jointly influence mRNA localization. *Cell Systems*, *10*(4), 351-362.e8. https://doi.org/10.1016/j.cels.2020.03.001
Morgan, T. H. (1912). Complete linkage in the second chromosome of the male of *Drosophila*. *Science*, *36*(934), 719–720. https://doi.org/10.1126/science.36.934.719
Moriyama, E. N., & Powell, J. R. (1996). Intraspecific nuclear DNA variation in *Drosophila*. *Molecular Biology and Evolution*, *13*(1), 261–277. https://doi.org/10.1093/oxfordjournals.molbev.a025563
Moriyama, E. N., & Powell, J. R. (1997). Codon usage bias and tRNA abundance in *Drosophila*. *Journal of Molecular Evolution*, *45*(5), 514–523. https://doi.org/10.1007/PL00006256
Muyle, A., Serres-Giardi, L., Ressayre, A., Escobar, J., & Glémin, S. (2011). GC-biased gene conversion and selection affect GC content in the *Oryza* genus (rice). *Molecular Biology and Evolution*, *28*(9), 2695–2706. https://doi.org/10.1093/molbev/msr104
Myers, S., Bowden, R., Tumian, A., Bontrop, R. E., Freeman, C., … Donnelly, P. (2010). Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science (New York, N.Y.)*, *327*(5967), 876–879. https://doi.org/10.1126/science.1182363
Nachman, M. W. (1997). Patterns of DNA variability at X-linked loci in *Mus domesticus*. *Genetics*, *147*(3), 1303–1316. https://doi.org/10.1093/genetics/147.3.1303
Narlikar, L., & Ovcharenko, I. (2009). Identifying regulatory elements in eukaryotic genomes. *Briefings in Functional Genomics and Proteomics*, *8*(4), 215–230. https://doi.org/10.1093/bfgp/elp014
Nelson, A. C., & Wardle, F. C. (2013). Conserved non-coding elements and cis regulation: Actions speak louder than words. *Development*, *140*(7), 1385–1395. https://doi.org/10.1242/dev.084459
Nielsen, R., & Slatkin, M. (2013). *An introduction to population genetics: Theory and applications*. Sinauer Associates.

Ohta, T. (1976). Role of very slightly deleterious mutations in molecular evolution and polymorphism. *Theoretical Population Biology*, *10*(3), 254–275. https://doi.org/10.1016/0040-5809(76)90019-8
Ohta, T. (1992). The nearly neutral theory of molecular evolution. *Annual Review of Ecology and Systematics*, *23*(1), 263–286. https://doi.org/10.1146/annurev.es.23.110192.001403
Ohye, T., Inagaki, H., Ozaki, M., Ikeda, T., & Kurahashi, H. (2014). Signature of backward replication slippage at the copy number variation junction. *Journal of Human Genetics*, *59*(5), Article 5. https://doi.org/10.1038/jhg.2014.20
Otaki, J. M. (2007). Stress-induced color-pattern modifications and evolution of the painted lady butterflies *Vanessa cardui* and *Vanessa kershawi*. *Zoological Science*, *24*(8), 811–819. https://doi.org/10.2108/zsj.24.811
Palidwor, G. A., Perkins, T. J., & Xia, X. (2010). A general model of codon bias due to GC mutational bias. *PLoS ONE*, *5*(10), e13431. https://doi.org/10.1371/journal.pone.0013431
Peng, Z., Uversky, V. N., & Kurgan, L. (2016). Genes encoding intrinsic disorder in Eukaryota have high GC content. *Intrinsically Disordered Proteins*, *4*(1), e1262225. https://doi.org/10.1080/21690707.2016.1262225
Pessia, E., Popa, A., Mousset, S., Rezvoy, C., Duret, L., & Marais, G. A. B. (2012). Evidence for Widespread GC-biased Gene Conversion in Eukaryotes. *Genome Biology and Evolution*, *4*(7), 675–682. https://doi.org/10.1093/gbe/evs052
Pichon, X., Wilson, L. A., Stoneley, M., Bastide, A., King, H. A., … Willis, A. E. (2012). RNA binding protein/RNA element interactions and the control of translation. *Current Protein & Peptide Science*, *13*(4), 294–304. https://doi.org/10.2174/138920312801619475
Pokusaeva, V. O., Diez, A. R., Espinar, L., Pérez, A. T., & Filion, G. J. (2022). Strand asymmetry influences mismatch resolution during single-strand annealing. *Genome Biology*, *23*(1), 93. https://doi.org/10.1186/s13059-022-02665-3
Pollock, D. L., & Fechheimer, N. S. (1978). The chromosomes of cockerels (*Gallus domesticus*) during meiosis. *Cytogenetics and Cell Genetics*, *21*(5), 267–281. https://doi.org/10.1159/000130903
Pouyet, F., Aeschbacher, S., Thiéry, A., & Excoffier, L. (2018). Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences. *ELife*, *7*, e36317. https://doi.org/10.7554/eLife.36317
Presnyak, V., Alhusaini, N., Chen, Y.-H., Martin, S., Morris, N., … Coller, J. (2015). Codon optimality is a major determinant of mRNA stability. *Cell*, *160*(6), 1111–1124. https://doi.org/10.1016/j.cell.2015.02.029
Pringle, E. G., Baxter, S. W., Webster, C. L., Papanicolaou, A., Lee, S. F., & Jiggins, C. D. (2007). Synteny and chromosome evolution in the lepidoptera: Evidence from mapping in *Heliconius melpomene*. *Genetics*, *177*(1), 417–426. https://doi.org/10.1534/genetics.107.073122
Pruisscher, P., Nylin, S., Gotthard, K., & Wheat, C. W. (2018). Genetic variation underlying local adaptation of diapause induction along a cline in a butterfly. *Molecular Ecology*, *27*(18), 3613–3626. https://doi.org/10.1111/mec.14829
Prunier, J., Giguère, I., Ryan, N., Guy, R., Soolanayakanahally, R., … Porth, I. (2019). Gene copy number variations involved in balsam poplar (*Populus balsamifera L.*) adaptive variations. *Molecular Ecology*,

28(6), 1476–1490. https://doi.org/10.1111/mec.14836
Quénet, D., & Dalal, Y. (2012). The CENP-A nucleosome: A dynamic structure and role at the centromere. *Chromosome Research*, *20*(5), 465–479. https://doi.org/10.1007/s10577-012-9301-4
Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S. A., Flynn, R. A., & Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, *470*(7333), 279–283. https://doi.org/10.1038/nature09692
Rastas, P. (2017). Lep-MAP3: Robust linkage mapping even for low-coverage whole genome sequencing data. *Bioinformatics*, *33*(23), 3726–3732. https://doi.org/10.1093/bioinformatics/btx494
Rastas, P., Calboli, F. C. F., Guo, B., Shikano, T., & Merilä, J. (2016). Construction of ultradense linkage maps with Lep-MAP2: Stickleback F_2 recombinant crosses as an example. *Genome Biology and Evolution*, *8*(1), 78–93. https://doi.org/10.1093/gbe/evv250
Rezvoy, C., Charif, D., Gueguen, L., & Marais, G. A. B. (2007). MareyMap: An R-based tool with graphical interface for estimating recombination rates. *Bioinformatics*, *23*(16), 2188–2189. https://doi.org/10.1093/bioinformatics/btm315
Rice, P., Longden, I., & Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, *16*(6), 276–277. https://doi.org/10.1016/S0168-9525(00)02024-2
Rodgers-Melnick, E., Bradbury, P. J., Elshire, R. J., Glaubitz, J. C., Acharya, C. B., … Buckler, E. S. (2015). Recombination in diverse maize is stable, predictable, and associated with genetic load. *Proceedings of the National Academy of Sciences*, *112*(12), 3823–3828. https://doi.org/10.1073/pnas.1413864112
Rojano, E., Seoane, P., Ranea, J. A. G., & Perkins, J. R. (2019). Regulatory variants: From detection to predicting impact. *Briefings in Bioinformatics*, *20*(5), 1639–1654. https://doi.org/10.1093/bib/bby039
Rousselle, M., Laverré, A., Figuet, E., Nabholz, B., & Galtier, N. (2019). Influence of recombination and gc-biased gene conversion on the adaptive and nonadaptive substitution rate in mammals versus birds. *Molecular Biology and Evolution*, *36*(3), 458–471. https://doi.org/10.1093/molbev/msy243
Ruckman, S. N., Jonika, M. M., Casola, C., & Blackmon, H. (2020). Chromosome number evolves at equal rates in holocentric and monocentric clades. *PLOS Genetics*, *16*(10), e1009076. https://doi.org/10.1371/journal.pgen.1009076
Rüdiger, N. S., Gregersen, N., & Kielland-Brandt, M. C. (1995). One short well conserved region of Alu-sequences is involved in human gene rearrangements and has homology with prokaryotic chi. *Nucleic Acids Research*, *23*(2), 256–260.
Sambrook, J., & Russell, D. W. (2006). Purification of nucleic acids by extraction with Phenol:Chloroform. *Cold Spring Harbor Protocols*, *2006*(1), pdb.prot4455. https://doi.org/10.1101/pdb.prot4455
Santos-Rosa, H., Schneider, R., Bannister, A. J., Sherriff, J., Bernstein, B. E., … Kouzarides, T. (2002). Active genes are tri-methylated at K4 of histone H3. *Nature*, *419*(6905), Article 6905. https://doi.org/10.1038/nature01080
Schield, D. R., Pasquesi, G. I. M., Perry, B. W., Adams, R. H., Nikolakis, Z. L., … Castoe, T. A. (2020). Snake recombination landscapes are

concentrated in functional regions despite PRDM9. *Molecular Biology and Evolution*, *37*(5), 1272–1294. https://doi.org/10.1093/molbev/msaa003
Schvarzstein, M., Wignall, S. M., & Villeneuve, A. M. (2010). Coordinating cohesion, co-orientation, and congression during meiosis: Lessons from holocentric chromosomes. *Genes & Development*, *24*(3), 219–228. https://doi.org/10.1101/gad.1863610
Schwacha, A., & Kleckner, N. (1997). Interhomolog bias during meiotic recombination: Meiotic functions promote a highly differentiated interhomolog-only pathway. *Cell*, *90*(6), 1123–1135. https://doi.org/10.1016/S0092-8674(00)80378-5
Shalgi, R., Pilpel, Y., & Oren, M. (2010). Repression of transposable-elements – a microRNA anti-cancer defense mechanism? *Trends in Genetics*, *26*(6), 253–259. https://doi.org/10.1016/j.tig.2010.03.006
Sharp, P. M., & Li, W. H. (1986). Codon usage in regulatory genes in Escherichia coli does not reflect selection for "rare" codons. *Nucleic Acids Research*, *14*(19), 7737–7749. https://doi.org/10.1093/nar/14.19.7737
Shaw, C. J., & Lupski, J. R. (2004). Implications of human genome architecture for rearrangement-based disorders: The genomic basis of disease. *Human Molecular Genetics*, *13*(suppl_1), R57–R64. https://doi.org/10.1093/hmg/ddh073
Shields, O. (1989). World numbers of butterflies. *Journal of the Lepidopterist's Society*, *43*, 178–183.
Shields, O. (1992). *Journal of the Lepidopterists' Society.* (Vol. 46, pp. 1–376). Lepidopterists' Society,. https://www.biodiversitylibrary.org/item/128069
Smeds, L., Mugal, C. F., Qvarnström, A., & Ellegren, H. (2016). High-resolution mapping of crossover and non-crossover recombination events by whole-genome re-sequencing of an avian pedigree. *PLOS Genetics*, *12*(5), e1006044. https://doi.org/10.1371/journal.pgen.1006044
Smeds, L., Qvarnström, A., & Ellegren, H. (2016). Direct estimate of the rate of germline mutation in a bird. *Genome Research*, *26*(9), 1211–1218. https://doi.org/10.1101/gr.204669.116
Söderlind, L., & Nylin, S. (2011). Genetics of diapause in the comma butterfly *Polygonia c-album*. *Physiological Entomology*, *36*(1), 8–13. https://doi.org/10.1111/j.1365-3032.2010.00756.x
Stapley, J., Feulner, P. G. D., Johnston, S. E., Santure, A. W., & Smadja, C. M. (2017). Variation in recombination frequency and distribution across eukaryotes: Patterns and processes. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *372*(1736), 20160455. https://doi.org/10.1098/rstb.2016.0455
Stefanescu, C., Soto, D. X., Talavera, G., Vila, R., & 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*(10), 20160561. https://doi.org/10.1098/rsbl.2016.0561
Steiner, C., Bozzolan, F., Montagné, N., Maïbèche, M., & Chertemps, T. (2017). Neofunctionalization of "Juvenile Hormone Esterase Duplication" in Drosophila as an odorant-degrading enzyme towards food odorants. *Scientific Reports*, *7*(1), Article 1. https://doi.org/10.1038/s41598-017-13015-w
Stephan, W., & Langley, C. H. (1998). DNA polymorphism in Lycopersicon and crossing-over per physical length. *Genetics*, *150*(4), 1585–1593.

https://doi.org/10.1093/genetics/150.4.1585
Sturtevant, A. H. (1913). The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. *Journal of Experimental Zoology*, *14*(1), 43–59. https://doi.org/10.1002/jez.1400140104
Sturtevant, A. H. (1921). A case of rearrangement of genes in *Drosophila*. *Proceedings of the National Academy of Sciences of the United States of America*, *7*(8), 235–237.
Suomalainen, E., Cook, L. M., & Turner, J. R. G. (1973). Achiasmatic oogenesis in the Heliconiine butterflies. *Hereditas*, *1973*(74), 302–304. https://doi.org/doi/10.1111/j.1601-5223.1973.tb01134.x
Talavera, G., Bataille, C., Benyamini, D., Gascoigne-Pees, M., & Vila, R. (2018). Round-trip across the Sahara: Afrotropical Painted Lady butterflies recolonize the Mediterranean in early spring. *Biology Letters*, *14*(6), 20180274. https://doi.org/10.1098/rsbl.2018.0274
Talavera, G., & Vila, R. (2016). Discovery of mass migration and breeding of the painted lady butterfly *Vanessa cardui* in the Sub-Sahara: The Europe–Africa migration revisited. *Biological Journal of the Linnean Society*, *2016*(n/a). https://doi.org/10.1111/bij.12873
Talla, V., Suh, A., Kalsoom, F., Dincă, V., Vila, R., … Backström, N. (2017). Rapid increase in genome size as a consequence of transposable element hyperactivity in wood-white (*Leptidea*) butterflies. *Genome Biology and Evolution*, *9*(10), 2491–2505. https://doi.org/10.1093/gbe/evx163
Tease, C., & Jones, G. H. (1978). Analysis of exchanges in differentially stained meiotic chromosomes of *Locusta migratoria* after BrdU-substitution and FPG staining. *Chromosoma*, *69*(2), 163–178. https://doi.org/10.1007/BF00329915
Turner, J. R. G., & Sheppard, P. M. (1975). Absence of crossing-over in female butterflies (*Heliconius*). *Heredity*, *1975*(34 (2)), 265–269.
Van't Hof, A. E., Campagne, P., Rigden, D. J., Yung, C. J., Lingley, J., … Saccheri, I. J. (2016). The industrial melanism mutation in British peppered moths is a transposable element. *Nature*, *534*(7605), 102–105. https://doi.org/10.1038/nature17951
Vara, C., Paytuví-Gallart, A., Cuartero, Y., Álvarez-González, L., Marín-Gual, L., … Ruiz-Herrera, A. (2021). The impact of chromosomal fusions on 3D genome folding and recombination in the germ line. *Nature Communications*, *12*(1), Article 1. https://doi.org/10.1038/s41467-021-23270-1
Veller, C., Kleckner, N., & Nowak, M. A. (2019). A rigorous measure of genome-wide genetic shuffling that takes into account crossover positions and Mendel's second law. *Proceedings of the National Academy of Sciences*, *116*(5), 1659–1668. https://doi.org/10.1073/pnas.1817482116
Verta, J.-P., & Jones, F. C. (2019). Predominance of cis-regulatory changes in parallel expression divergence of sticklebacks. *ELife*, *8*, e43785. https://doi.org/10.7554/eLife.43785
Vicario, S., Moriyama, E. N., & Powell, J. R. (2007). Codon usage in twelve species of *Drosophila*. *BMC Evolutionary Biology*, *7*(1), 226. https://doi.org/10.1186/1471-2148-7-226
Wall, J. D., Robinson, J. A., & Cox, L. A. (2022). High-resolution estimates of crossover and noncrossover recombination from a captive baboon colony. *Genome Biology and Evolution*, *14*(4), evac040. https://doi.org/10.1093/gbe/evac040

Walter, M., Teissandier, A., Pérez-Palacios, R., & Bourc'his, D. (2016). An epigenetic switch ensures transposon repression upon dynamic loss of DNA methylation in embryonic stem cells. *ELife*, *5*, e11418. https://doi.org/10.7554/eLife.11418

Wellband, K., Mérot, C., Linnansaari, T., Elliott, J. a. K., Curry, R. A., & Bernatchez, L. (2019). Chromosomal fusion and life history-associated genomic variation contribute to within-river local adaptation of Atlantic salmon. *Molecular Ecology*, *28*(6), 1439–1459. https://doi.org/10.1111/mec.14965

Wellenreuther, M., Mérot, C., Berdan, E., & Bernatchez, L. (2019). Going beyond SNPs: The role of structural genomic variants in adaptive evolution and species diversification. *Molecular Ecology*, *28*(6), 1203–1209. https://doi.org/10.1111/mec.15066

Wiklund, C., & Friberg, M. (2022). Testing the migration syndrome: Comparative fecundity of migratory and non-migratory nymphaline butterflies. *Ecological Entomology*, *47*(6), 1061–1067. https://doi.org/10.1111/een.13195

Wiklund, C., Lehmann, P., & Friberg, M. (2019). Diapause decision in the small tortoiseshell butterfly, *Aglais urticae*. *Entomologia Experimentalis et Applicata*, *167*(5), 433–441. https://doi.org/10.1111/eea.12780

Williams, A. L., Genovese, G., Dyer, T., Altemose, N., Truax, K., … Przeworski, M. (2015). Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. *ELife*, *4*, e04637. https://doi.org/10.7554/eLife.04637

Witherspoon, D. J., Watkins, W. S., Zhang, Y., Xing, J., Tolpinrud, W. L., … Jorde, L. B. (2009). Alu repeats increase local recombination rates. *BMC Genomics*, *10*(1), 530. https://doi.org/10.1186/1471-2164-10-530

Wright, F. (1990). The 'effective number of codons' used in a gene. *Gene*, *87*(1), 23–29. https://doi.org/10.1016/0378-1119(90)90491-9

Wright, S. (1931). Evolution in Mendelian Populations. *Genetics*, *16*(2), 97–159.

Xu, S., Ackerman, M. S., Long, H., Bright, L., Spitze, K., … Lynch, M. (2015). A male-specific genetic map of the microcrustacean *Daphnia pulex* based on single-sperm whole-genome sequencing. *Genetics*, *201*(1), 31–38. https://doi.org/10.1534/genetics.115.179028

Yan, Z., Martin, S. H., Gotzek, D., Arsenault, S. V., Duchen, P., … Keller, L. (2020). Evolution of a supergene that regulates a trans-species social polymorphism. *Nature Ecology & Evolution*, *4*(2), Article 2. https://doi.org/10.1038/s41559-019-1081-1

Yang, L., Gao, Y., Li, M., Park, K.-E., Liu, S., … Liu, G. E. (2022). Genome-wide recombination map construction from single sperm sequencing in cattle. *BMC Genomics*, *23*(1), 181. https://doi.org/10.1186/s12864-022-08415-w

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, *24*(8), 1586–1591. https://doi.org/10.1093/molbev/msm088

Yoshido, A., Šíchová, J., Pospíšilová, K., Nguyen, P., Voleníková, A., … Marec, F. (2020). Evolution of multiple sex-chromosomes associated with dynamic genome reshuffling in *Leptidea* wood-white butterflies. *Heredity*, *125*(3), 138–154. https://doi.org/10.1038/s41437-020-0325-9

Zhan, S., Merlin, C., Boore, J. L., & Reppert, S. M. (2011). The monarch butterfly genome yields insights into long-distance migration. *Cell*, *147*(5),

1171–1185. https://doi.org/10.1016/j.cell.2011.09.052
Zhan, S., Zhang, W., Niitepõld, K., Hsu, J., Haeger, J. F., … Kronforst, M. R. (2014). The genetics of monarch butterfly migration and warning colouration. *Nature*, *514*(7522), Article 7522. https://doi.org/10.1038/nature13812
Zhang, L. (2003). Does recombination shape the distribution and evolution of tandemly arrayed genes (TAGs) in the *Arabidopsis thaliana* genome? *Genome Research*, *13*(12), 2533–2540. https://doi.org/10.1101/gr.1318503
Zhang, L., Steward, R. A., Wheat, C. W., & Reed, R. D. (2021). High-Quality Genome Assembly and Comprehensive Transcriptome of the Painted Lady Butterfly *Vanessa cardui*. *Genome Biology and Evolution*, *13*(7). https://doi.org/10.1093/gbe/evab145
Zhang, S., Postnikov, Y., Lobanov, A., Furusawa, T., Deng, T., & Bustin, M. (2022). H3K27ac nucleosomes facilitate HMGN localization at regulatory sites to modulate chromatin binding of transcription factors. *Communications Biology*, *5*(1), Article 1. https://doi.org/10.1038/s42003-022-03099-0
Zhao, F., Yu, C., & Liu, Y. (2017). Codon usage regulates protein structure and function by affecting translation elongation speed in *Drosophila* cells. *Nucleic Acids Research*, *45*(14), 8484–8492. https://doi.org/10.1093/nar/gkx501
Zhao, W., Xu, Y., Wang, Y., Gao, D., King, J., … Liang, F.-S. (2021). Investigating crosstalk between H3K27 acetylation and H3K4 trimethylation in CRISPR/dCas-based epigenome editing and gene activation. *Scientific Reports*, *11*(1), Article 1. https://doi.org/10.1038/s41598-021-95398-5
Zhu, H., Gegear, R. J., Casselman, A., Kanginakudru, S., & Reppert, S. M. (2009). Defining behavioral and molecular differences between summer and migratory monarch butterflies. *BMC Biology*, *7*(1), 14. https://doi.org/10.1186/1741-7007-7-14

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 2225

Editor: The Dean of the Faculty of Science and Technology

A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology. (Prior to January, 2005, the series was published under the title "Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology".)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-490587

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2023