Ann. Zool. Fennici 54: 275–291 ISSN 0003-455X (print), ISSN 1797-2450 (online)
Helsinki 15 May 2017 © Finnish Zoological and Botanical Publishing Board 2017

Ilkka Hanski: The legacy of a multifaceted ecologist

Butterfly genomics: insights from the genome of *Melitaea cinxia*

Virpi Ahola[1], Niklas Wahlberg[2] & Mikko J. Frilander[3,*]

[1] *Department of Biosciences, P.O. Box 65, FI-00014 University of Helsinki, Finland*
[2] *Department of Biology, Lund University, Sölvegatan 37, SE-223 62 Lund, Sweden*
[3] *Institute of Biotechnology, P.O. Box 56, FI-00014 University of Helsinki, Finland (*corresponding author's e-mail: mikko.frilander@helsinki.fi)*

Received 2 Dec. 2016, final version received 9 Feb. 2017, accepted 9 Feb. 2017

Ahola, V., Wahlberg, N. & Frilander, M. J. 2017: Butterfly genomics: insights from the genome of *Melitaea cinxia*. — *Ann. Zool. Fennici* 54: 275–291.

The first lepidopteran genome (*Bombyx mori*) was published in 2004. Ten years later the genome of *Melitaea cinxia* came out as the third butterfly genome published, and the first eukaryotic genome sequenced in Finland. Owing to Ilkka Hanski, the *M. cinxia* system in the Åland Islands has become a famous model for metapopulation biology. More than 20 years of research on this system provides a strong ecological basis upon which a genetic framework could be built. Genetic knowledge is an essential addition for understanding eco-evolutionary dynamics and the genetic basis of variability in life history traits. Here we review the process of the *M. cinxia* genome project, its implications for lepidopteran genome evolution, and describe how the genome has been used for gene expression studies to identify genetic consequences of habitat fragmentation. Finally, we introduce some future possibilities and challenges for genomic research in *M. cinxia* and other Lepidoptera.

Enter genomics

The past four decades have been a time of unprecedented progress in biological sciences, to which Ilkka Hanski made fundamental contributions. His insights into the causes and consequences of population dynamics at the ecological level paved the way for a more intimate look at how the dynamics and the evolutionary history of a species are reflected at the genomic level. In the last two decades, this intimacy advanced from allele frequency studies of a single genetic marker (Orsini *et al*. 2009, Hanski 2011) to a full-blown genomic scale investigation to identify the specific genes involved in dispersal and colonization (Wheat *et al*. 2011, Somervuo *et al*. 2014, Kvist *et al*. 2015). The genomic revolution in insects started with the publication of the *Drosophila melanogaster* genome in 2000 (Adams *et al*. 2000), which itself was a prelude to the human genome project. This huge project demonstrated that it is technically possible to delve into the genomes of complex eukaryotes and discover the genetic mechanisms behind various life history traits. About four years later, the genomic revolution reached Lepidoptera, with a draft silkworm (*Bombyx mori*) genome being published by two competing groups (Mita *et al*. 2004, Xia *et al*. 2004). This and the rapid development of sequencing technologies set off the race to sequence the first butterfly genome, and the main finalist contenders appeared to be Hanski's *M. cinxia*, as well as *Heliconius melpomene* and *Bicyclus anynana*, each of which had been extensively used as model systems to study different aspects of biology.

Through the subsequent years, researchers followed the progress of the three competing projects at various conferences and were all surprised when the first butterfly genome to be published was the monarch butterfly (*Danaus plexippus*) (Zhan *et al*. 2011). The *Heliconius melpomene* genome followed soon after (The Heliconius Genome Consortium 2012), and the *M. cinxia* genome two years later (Ahola *et al*. 2014). At present, there are 25 published Lepidoptera genomes of which 18 represent different butterfly species (NCBI database accessed November 2016). This genomic revolution has set the stage ready for investigations on genome-environment interactions on ecologically well-studied species.

Ilkka Hanski's dream of sequencing the genome of his favorite species was realized in time for him to be able to actively start integrating the population biology and genomics of his favorite species. Here we review the process of obtaining the whole genome sequence and highlight the published results as well as discuss future possibilities and challenges.

Why was the *Melitaea cinxia* genome important?

The Glanville fritillary butterfly (*Melitaea cinxia*) model system in the Åland Islands in Finland was originally developed to study the effects of habitat fragmentation (Ehrlich & Hanski 2004). The system represents a classic metapopulation consisting of ca. 4000 habitat patches of dry meadows that can be divided into 131 networks (Ojanen *et al*. 2013). Movement within networks is common, but long migration distances make such movements rare between the networks. The annual large-scale survey of the whole metapopulation has formed the backbone of these studies, which, coupled with experiments in common garden conditions and outdoor cages, have enabled a large and diverse set of studies on the ecology, evolution and genetics of this species. Ilkka Hanski and colleagues have made substantial contributions to the development of metapopulation theory using the system (Hanski 1999, Hanski & Ovaskainen 2000), but also for other ecological investigations at the landscape scale (Hanski & Saccheri 2006, Hanski 2011).

The development of next generation sequencing (NGS) has enabled individual research groups studying the ecology of natural populations to expand their scope to investigate the genetic and molecular basis of individual fitness traits at the scale of the whole genome. Thus the technological development has changed the entire research field from predominantly hypothesis-driven approaches with a small number of candidate genes to more hypothesis-free global approaches that can help to understand population-level phenomena such as local adaptation or population dynamics at the genomic level. For genetic research relying on genetic markers, NGS has been a tremendous help, enabling efficient and large-scale marker discovery. The benefits of NGS are particularly true in Lepidoptera research, for which it has been difficult to develop microsatellite markers (Van't Hof *et al*. 2007) — the probable reason why only a few studies using genetic markers existed for Lepidoptera prior to the genomic revolution. At the more technical front, the access to whole genome or transcriptome sequences has significantly accelerated development and implementation of the custom genotyping arrays that in turn allow simultaneous screening of a large number of different genetic markers for a large number of individuals. Therefore, the possibilities offered by the new technologies were exciting, promising a rational selection of (a large number of) genetic markers in contrast to the semi-arbitrary process of candidate gene selection.

The motivation for genomic studies of the Glanville fritillary arose not only from the aims to find association between genotypic and phenotypic variation in general, but from the special interest to understand the genetic basis of female reproductive success and dispersal related traits, both of which are essential for the species to persist in fragmented landscapes. It was important to define the genetic and molecular level impact of variation in dispersal traits but also the role of natural selection for population and evolutionary dynamics in the study system. Additionally, the genomic resources enable building a pedigree across the entire Åland Islands study system over several generations. Such a pedigree could be used for obtaining empirical assessments of dispersal and gene flow, but most importantly, for eco-evolutionary dynamics modeling with special interest in population growth and inbreeding depression.

More than 20 years of research on the study system has provided plenty of material and examples of quantitative and qualitative variables that appeared to have a genetic basis, and can be linked to the landscape level. For instance, on the Åland Islands, female Glanville fritillary use two plant species, *Plantago lanceolata* and *Veronica spicata*, during oviposition as hosts for larvae (Nieminen *et al*. 2004). Earlier studies have suggested that female oviposition preference on one of the host plant species is genetically determined (Kuussaari *et al*. 2000, Hanski & Singer 2001). Oviposition preference is an important fitness trait for herbivores with gregarious larval groups, and female host choice can be crucial for survival of offspring. Larval and pupal development traits are highly heritable (Saastamoinen 2008, Klemme & Hanski 2009, Kvist *et al*. 2013, de Jong *et al*. 2014). Late larval stages development is important for building body mass in the adult stage, and is related to reproductive success of females (Saastamoinen 2007), whereas larval development time can increase mating success of males (Saastamoinen *et al*. 2013).

Similarly, individual mobility and flight metabolic traits (both components of dispersal ability) are suggested to be highly heritable (Saastamoinen 2008, Klemme & Hanski 2009, Mattila & Hanski 2014), but both the genetic architecture and genes responsible for variation in those traits have been mostly unknown, and only rarely studied in wild populations. Only a few candidate genes have been identified to be associated with the above mentioned life history and dispersal traits in *M. cinxia* (Orsini *et al*. 2009, de Jong *et al*. 2014, Ahola *et al*. 2015, Wong *et al*. 2016). The most important candidate is the gene phosphoglucose isomerase (*Pgi*) discussed elsewhere in this issue (Niitepõld & Saastamoinen 2017).

The trials and tribulations of building the *Melitaea cinxia* genome

The idea for sequencing the whole genome of *M. cinxia* was initially born around 2006, before the first NGS platforms became available, but was at that time regarded as a “dream project”. The justification for the dream was that it would open the door to many genomics tools that were available for laboratory model organisms but which required full genome sequence. The dream started to become attainable in 2006 when the first NGS technology, 454 pyrosequencer (launched in 2005), landed at the Institute of Biotechnology (BI), University of Helsinki. Soon after publishing the *M. cinxia* transcriptome sequence using 454 sequencing technology (Vera *et al*. 2008), the decision for sequencing the whole genome became relevant. In retrospect it is now clear that none of the participants involved in the decision and initial planning had a clear idea of the full complexity of the project, which is fortunate as otherwise this project would probably have never been initiated.

The original plan was to use the 454 FLX Titanium sequencer to provide long (~400 bp) reads with lower 6–8X coverage, and supplement the data with short (~50 bp) SOLiD ver. 3 reads that produced 20 times higher throughput than the 454 sequencer. From the *B. mori* genome repeat annotation (Osanai-Futahashi *et al*. 2008) and the estimated genome size of 320 Mb for *M. cinxia,* we anticipated that the *M. cinxia* genome may have relatively high transposable element (TE) content. This information was critical for the final assembly of the genome as high repeat content can significantly impede genome assembly, particularly where TE composition is unknown, as was the case with *M. cinxia*. To overcome this obstacle, our plan was to use a hybrid assembly strategy: assemble long 454 reads to contiguous sequences (contigs), and build the longer chromosomal scaffolds by linking the contigs with SOLiD mate pair (MP) reads and the already existing transcriptome information. The outcome of this strategy was supposed to be a typical first-draft genome in which the contigs containing unique genome information were linked and ordered using MP and paired end (PE) reads of different insert lengths, effectively jumping over the unassembled TEs of different (and often unknown) lengths. Use of different technologies was also thought to correct biases caused by different sequencing technologies. The known challenge was that at the time the project started there was no assembly method that would be able to accommodate both the 454 and SOLiD data. Consequently, one of the aims was also to produce new bioinformatics methods and tools for the genome projects. This turned out to be one of the strengths of the project.

The project was launched in early 2009 by establishing the genome steering group that remained essentially the same during the entire project. The core steering group consisted of roughly ten people originating from the Metapopulation Research Group (MRG) and Institute of Biotechnology (BI) (DNA Sequencing and Genomics Laboratory, RNA splicing group, and Bioinformatics group) with several members from the Department of Computer Science joining soon after. This started the five-year cooperation and integration of scientists from various backgrounds and modalities.

One of the immediate questions was the selection of an individual or individuals to be sequenced. The best option would have been to obtain DNA from a highly inbred (> 6 generations) individual or individuals. Unfortunately, with the nine-month generation time of *M. cinxia* this was not an option. Therefore, DNA for 454 sequencing was extracted from a single 7th (last) instar male larvae, which lacks the female W chromosome known to be composed almost entirely of repetitive DNA. To obtain large amounts of high-quality DNA was however not trivial because none of the standard DNA extraction methods worked particularly well with *M. cinxia*. In most cases, DNA preparations contained a pigment from the black larvae that interfered with DNA quality/quantity measurements and — as we learned later the hard way — also decreased the stability of DNA preparations. Nevertheless, the first sequences with a read length of 360 bp started to accumulate, and the iterative process of assembly and sequencing was initiated.

The first contig assembly was finished in 2010 but, somewhat disappointingly, the contigs were very short, most probably due to TEs and heterozygous indels of different lengths that interfered in the assembly process. This lead to several attempts for obtaining longer contigs, which turned out to be mostly useless — contigs were and remained short (N50 = 2015 bp). In principle, increased amounts of data should have led to a steady increase in the contig length, but with *M. cinxia* this never happened. Contig length did not markedly increase with increasing amount of sequence data. Nevertheless, the first contig assemblies were sufficient for the launch of the genotyping branch of the project, and were used together with the transcriptome data in the design of the first custom genotyping array for *M. cinxia* (Ahola *et al*. 2015, Wong *et al*. 2016).

Within the steering group there was a strong belief or at least hope that scaffolding with SOLiD MP and Illumina PE sequencing data would at least partially compensate for the short contig length. Therefore, concurrent with contig assembly, read pair sequence data from several short libraries (< 5 kb insert sizes) were produced and employed in scaffolding using a step-wise strategy in which libraries were added to the assembly in ascending order of insert size. This approach, however, resulted in only slightly improved assembly, leaving the number or scaffolds high and N50 relatively low (N50 < 50 kb). In other words, the resulting genome assembly was too fragmented to be publishable or useful for supporting genomic research. It was obvious that longer insert lengths were needed for bridging the longer TEs.

This started the most challenging part of the whole project. The project was already behind the initial schedule and the pressure for finishing the assembly and moving to the next annotation phase was high. Libraries with longer insert lengths were attempted, but it appeared that the true insert length was always much less than what was expected. New libraries were tried and DNA was sent to other laboratories — with the same outcome. The whole project was stuck in the scaffolding stage and there was a real possibility that the project would never be finished. The pressure on everyone involved was hard — people responsible for the stage were pushed to the edge — the problem simply had to be solved.

At the darkest hour, the resolution came from an observation by the BI team that was responsible for DNA isolation of a large number of individuals for the purpose of population level genotyping. They noticed that DNA that was supposed to be good was in fact highly fragmented after a long period of storage. Heating during library preparation stages further exacerbated this problem. The most likely culprit was the dark pigment present in every DNA preparation that most likely caused chemical fragmentation of the DNA. Thus, there was no possibility of building long MP libraries using the available DNA. This observation was the turning point of the whole project.

With the problem finally identified, it was possible to consider work-arounds. Fortunately, two members of the steering group had long-enough experience in molecular biology to remember the CsCl purification method that originates from the 1980s, which is pretty much the Stone Age in this fast-evolving field. A DNA isolation protocol for *Drosophila* (Smoller *et al.* 1991) was quickly modified in ways that prevented the pigments from binding to the DNA. The resulting genomic DNA was pure, beautiful and of high molecular weight. The genome project was saved, people were smiling, and sequencing machines were running again!

The only small drawback was that several full-sib families of 100 individuals had to be used for DNA extraction thus bringing more variability to scaffolding. New data was rapidly used to build several possible assemblies. Here the delay that was so nerve wracking for the sequencing and assembly teams was in fact a blessing for the computational biologists. They had more time for method development, and when the new data was available, the new scaffolding method (Salmela *et al.* 2011) could be tuned up for the data.

One of the original aims in the genome project was to generate a linkage map as an independent resource to support the genome assembly. This turned out to be an excellent decision. The linkage map project was running concurrently with the sequencing and assembly project. The aim was to use RAD-sequencing for F1 families, and use the resulting data to build a high-throughput linkage map. It however turned out that there was no linkage map software that could efficiently use tens or hundreds of thousands markers and take into account the achiasmatic meiosis of female Lepidoptera. RAD-seq library preparation turned out to be challenging, and succeed only after development of a new protocol (Rastas *et al.* 2013). Therefore, it took a long time to obtain good quality data but luckily again, that delay enabled the development of the missing linkage map method and software (Rastas *et al.* 2013).

Linkage map with 40 000 SNPs turned out to be crucial when choosing between different scaffolding versions of the genome assembly. Scaffolding produced several different versions and naturally our initial decision was to choose the one with longest scaffolds. It turned out, however, that the longest assembly with N50 = 250 kb contained unacceptable high level (75%) of chimeric scaffolds. In other words, this particular assembly was too greedy, trading accuracy for increased (but false) scaffold length. This led us to choose a more humble but more correct version with 8262 scaffolds (N50 = 119 kb) of which only 8% were chimeric. Generally, we valued the quality assessment of the assembly as one of the paramount stages of the project, which was not so common at that time. We used altogether eight validation methods to assess correctness and completeness of the assembly (Ahola *et al.* 2014). The value of the linkage map was further emphasized during the final assembly stages when this data was used together with MP and long PacBio sequencing reads to generate superscaffolds that raised the statistics of the assembly to the same level as the other sequenced Lepidoptera.

The scaffolding statistics were somewhat lower but still comparable to the other published Lepidoptera genomes (The International Silkworm Genome Consortium 2008, Zhan *et al.* 2011, The Heliconius Genome Consortium 2012, You *et al.* 2013, Zhan & Reppert 2013), which enabled the project to be moved to the next stage. Gene models were built in collaboration with the European Bioinformatics Institute (EBI), as the genome was already from the beginning meant to be published in the *Ensembl Metazoa* database (http://metazoa.ensembl.org). *Heliconius melpomene* and *D. plexippus* genomes were published already in 2012, which together with the *B. mori* genome published earlier, provided an essential assistance for the genome annotation and gene model construction. The project quickly sprinted from gene model construction to functional annotation and to manual annotation. Gene models were built in Maker (Cantarel *et al.* 2008) that did not fully benefit from the available transcriptome data, therefore further iteration rounds would have been needed to improve the gene models. These iteration steps were cut to a minimum and the gene models were frozen because of the tight schedule, and the acute need for fixed gene models to analyze of RNA-seq data for gene expression studies.

After functional annotations and ortholog analyses were completed, the project was moved to the manual annotation stage. Up to this point, the project was in the hands of a small group of people from the University of Helsinki, but the manual annotation opened the project to about twenty other collaborators around the world. At this stage the individual gene models were examined in detail, often by a specialist in the field, by linking the gene to the orthologous genes in other species and providing a protein name. Manual annotation was easy for short and conserved genes but challenging for others. It attempted to address many issues such as fragmented and incomplete gene models arising from short contigs and inefficient use of transcriptome data, or presence of frame shifts or other mistakes due to errors in sequence data.

Finally, the genome assembly and annotation stages were finished and the genome was to be published as soon as possible in a highly ranked journal. This was supposed to be the fast phase of the project, but it turned out to be a second bottleneck. The focus on producing the genome itself meant that the research group had not given much thought to the actual scientific story required for a high-profile paper. Some five years earlier the genome itself would have been sufficient to propel the publication to one of the high-ranking journals, but as whole genome assemblies became more common it was now necessary to present an accompanying lead story. Most people expected that the *M. cinxia* genome paper would be about the ecology of the species with special focus to evolutionary consequences of habitat fragmentation. Two large-scale RNA-seq studies were performed to support the story of the genome paper, and there were multiple attempts at finding the appropriate storyline for more than a year. During the summer of 2013 it however became clear that we knew too little about the genomic and molecular consequences of habitat fragmentation in *M. cinxia* — it was too early and optimistic to think that we could make an attractive genome paper from that topic. We decided that the two RNA-seq studies would be published separately (Somervuo *et al.* 2014, Kvist *et al.* 2015), which left us again with an empty table — the project was once again completely halted. As with earlier delays, this was also useful as it allowed us to write the very sizable supplementary document (125 pages, 37 figures and 34 tables) that would accompany the actual paper.

Ilkka Hanski's strong opinion that was shared by many in the steering group was that we had to find something special from the *M. cinxia* genome itself, and use that for writing the genome paper. We did not want to write a "traditional" genome paper describing different gene families because lepidopteran genomes in general have very similar gene content (Ahola *et al.* 2014); it therefore seemed that there was no material for a high-profile genome paper.

The first genome meeting after the summer holidays in autumn 2013 was the second major turning point for the project. Early in the morning one of the bioinformaticians introduced a figure comparing the orthologous genes between *B. mori* and *M. cinxia* at the genome scale (Fig. 1a). This immediately generated a stir among the participants. The genome group was astonished to see the extremely high conservation of gene order in all of the *M. cinxia* 31 chromosomes between the two very divergent species. Furthermore, as *B. mori* has fewer chromosomes than *M. cinxia*, it has "fusion chromosomes" that continued the same homology trend with very sharp sites of chromosomal fusions (Fig. 1b). Further comparisons to other species reinforced the observation. The group was now

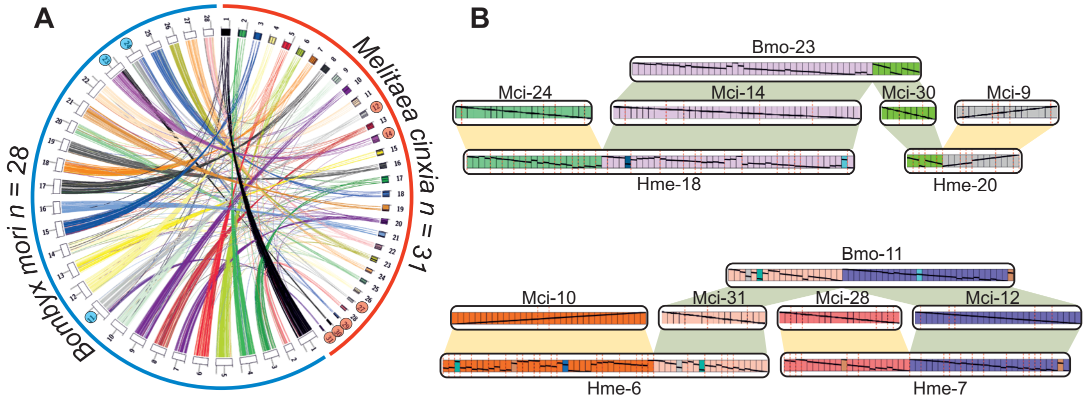

Fig. 1. Chromosome-level synteny in Lepidoptera. (**a**) A circos map showing one-to-one gene orthologues (4485) connecting *M. cinxia* and *B. mori* chromosomes. The *B. mori* chromosomes that are formed by a fusion of two ancestral chromosomes are shaded in blue and the corresponding orthologous chromosomes in *M. cinxia* with red. (**b**) Two detailed examples where the same *M. cinxia* chromosomes involved in fusion events in both *B. mori* and *H. melpomene*, but with non-orthologous fusion partners. Each box represents one superscaffold in *M. cinxia* and a scaffold in *H. melpomene*. Modified from the Ahola *et al.* (2014). For exhaustive description of the figure elements, see the original publication.

very excited and there was nothing that was able to prevent further analyses — not even the knowledge that the phenomenon was reported many times before (Yasukochi *et al*. 2006, Beldade *et al*. 2009, Yasukochi *et al*. 2009, The Heliconius Genome Consortium 2012, Sahara *et al*. 2013, Van't Hof *et al*. 2013), but mostly with low resolution data — and the fact that none of the members had any experience in Lepidoptera chromosome biology. Soon it was decided that ancient chromosome number among Lepidoptera, chromosome conservation and dynamics were going to be the leading story of the genome paper. Here the initial visualization of complex data played an important role in choosing the topic, but was also crucial in the very end of the process and appreciated later by the readers. The detailed analyses led to the unexpected conclusion that the chromosomal fusion dynamics in Lepidoptera is unlikely to be a random process (Ahola *et al*. 2014).

The publication process culminated in early summer 2014 when MRG was holding an annual meeting, with the scientific advisory board members of the Academy of Finland Center of Excellence present. Ilkka Hanski received the first (and only) reviewer comments from the journal in the early evening and it was immediately clear that the paper would be later formally accepted by the journal after a few modest clarifications asked by the reviewers. It was time to toast the successful completion of the project!

Melitaea cinxia genome sequence was the first eukaryotic genome totally finished and published in Finland. Although the group members had experience in sequencing and assembling microbial and fungal genomes, performing automatic and manual annotation as well as developing assembly and annotation methods, the *M. cinxia* genome was the first large genome project for all of us. In addition to the genome and the genome paper, it also yielded new scaffolding (Salmela *et al*. 2011), read error correction (Salmela 2010, Salmela & Schröder 2011), functional annotation (Koskinen *et al*. 2015), orthology prediction (Ta *et al*. 2011, Koskinen & Holm 2012), linkage mapping (Rastas *et al*. 2013), and RNA-seq (Kvist *et al*. 2015) and RAD-seq (Rastas *et al*. 2013) library preparation methods. The *M. cinxia* genome project was also a pioneering project in the sense that it gave many groups in Finland the confidence to initiate a similar investigation as exemplified by the ongoing silver birch (J. Salojärvi unpubl. data) and Saimaa ringed seal genome projects, and also helped those groups by identifying critical stages of the project.

Insights from the *Melitaea cinxia* genome

Dynamics of chromosome evolution in Lepidoptera

One of the unexpected outcomes arising from the *M. cinxia* genome project was that it was able to address a general question related to chromosome evolution in Lepidoptera. In fact, such work was started at the University of Helsinki already a long time ago. Professor and academician Esko Suomalainen (1910-1995) was instrumental in setting up the study of karyotypes in Lepidoptera and other insects (Suomalainen 1969), so it is only fitting that a Finnish genome project on a butterfly has brought us intriguing insights into chromosome evolution in this taxon.

Unlike in mammals and many other taxa the chromosomes in Lepidoptera are holocentric (Suomalainen 1966, Murakami & Imai 1974, Wolf 1996), meaning that the entire chromosome acts as a centromere during mitosis. Lepidoptera are also known for their large variation in chromosome numbers, even within species (Brown *et al*. 2004, Kandul *et al*. 2007, Saura *et al*. 2013). Variation in chromosome numbers has been suggested to be involved in species diversification (Mayr 1963), although a good understanding of the mechanisms behind such a process is lacking. Holocentric chromosomes in principle allow for individuals with different numbers of chromosomes to reproduce, if synteny is preserved across chromosomes. Additionally, holocentric chromosomes should be, due to their lack of localized centromere, more amenable to internal rearrangements because such rearrangements are not expected to lead to acrocentric chromosomes and chromosome loss during meiosis.

Contrary to these theoretical expectations, earlier studies in *B. mori* using linkage maps (Yasukochi *et al*. 2006, Beldade *et al*. 2009, Baxter *et al*. 2011), FISH mapping (Yasukochi *et al*. 2009, Yoshido *et al*. 2011), and BAC sequencing (d'Alencon *et al*. 2010) suggested that the gene order on chromosomes in Lepidoptera is highly conserved. However, it required *M. cinxia* genome based on both sequence and linkage data, together with the genomes of *H. melpomene* and *B. mori* to pinpoint the conservative nature of synteny and gene order on lepidopteran chromosomes at high resolution (Ahola *et al*. 2014). The *n* = 31 chromosome number of *M. cinxia* represents the ancient karyotype number of the Lepidoptera, and therefore it offers an excellent comparison point against the other genomes.

Our comparisons between *M. cinxia* and *H. melpomene* and *B. mori* genomes provided both expected and unexpected results. As expected from the earlier work the gene order in chromosomes was indeed conserved between the three species (Fig. 1a). Given the large evolutionary distance between the three species ~117 My (Wahlberg *et al*. 2013), and additional inclusion of *Plutella xylostella* with *n* = 31 into the analyses, it was possible to conclude that this is a general lepidopteran trait. Our work added a more detailed scale to the comparison by confirming that this is not just a macro-scale event but is reproduced at the gene level.

The unexpected results came when we asked how the dissimilar chromosome number (*n* = 31 for *M. cinxia*, *n* = 21 for *H. melpomene*, *n* = 28 for *B. mori*) between the three species affected the comparison. The lepidopteran karyotypes evolve predominantly through chromosomal fusions and fissions, of which the fusions are more common (White 1973). As the ancestral chromosome number is 31, this means that most species have 31 or fewer chromosomes. The exciting result came from detailed investigations of the *H. melpomene* and *B. mori* "fusion chromosomes", which have been formed by joining of two ancestral chromosomes. Our work revealed that fusion chromosomes are formed by direct joining of the chromosome ends, which seem to prevent any further intrachromosomal rearrangements across the fusion boundary (Fig. 1). The other important result was that the fusion process does not seem to be random, but prefers short chromosomes, or to be more precise, chromosomes that are short in *M. cinxia* genome. This has led to a situation where the orthologs of the same *M. cinxia* chromosomes participate in chromosomal fusion events in different species despite their long divergence time of ~117 My (Wahlberg *et al*. 2013) (Fig. 1b). Furthermore, we showed that the propensity for fusions is associated with TE content of the chromosome, and speculated that such elements may be involved either in forming telomeres (chromosome ends) in Lepidoptera or at least are associated with telomeres. Together our results suggest that the lepidopteran karyotypes evolve in a significantly different way than in most other metazoan species.

The foray into karyotype evolution was an unexpected but pleasant conclusion of the genome project. Furthermore, it is gratifying to note that our main conclusions relating to fusion chromosomes have already been independently confirmed for both *H. melpomene* (Davey *et al.* 2016) and *B. mori* (Yasukochi *et al.* 2016). As a practical outcome, our conclusions on karyotype evolution have significantly simplified any future lepidopteran genome project because the evolutionarily conserved gene content for each chromosome can now be assumed during the genome assembly phase. At the same time, it also poses the yet unanswered question of whether such conserved synteny is in any way reflected in the ecology and evolution of lepidopteran species.

Ecological and evolutionary consequences of habitat fragmentation

Ecological and evolutionary consequences of habitat fragmentation in *M. cinxia* have been well-described but their genetic basis is mostly unknown. However, several studies have linked physiological characteristics to habitat fragmentation, suggesting that these characteristics may have a genetic basis. Specifically, in the Åland Islands study system on average 100 new habitat patches are colonized every year. Physiological studies have shown that females from newly established populations have higher flight metabolic rate (FMR) than females from old local populations (Haag *et al.* 2005, Wheat *et al.* 2011). More generally, higher FMR predicts higher dispersal capacity (Niitepõld *et al.* 2009), which, together with a field study (Ovaskainen *et al.* 2008) suggests that new population individuals have higher dispersal rate. When comparing the average FMR in four *M. cinxia* populations across the Baltic Sea: Åland Islands (AL, Finland), Saaremaa (SA, Estonia), Uppland and Öland (UP and OL, Sweden), individuals from fragmented populations (AL, UP) have ca. 30% higher FMR than those from continuous populations (SA, OL) (Duplouy *et al.* 2013).

Additionally, Duplouy *et al.* (2013) suggested that habitat fragmentation can also affect other life history traits. Post-diapause larvae from fragmented populations have higher body mass than those from continuous landscapes, and their development time is shorter. There are also slight morphological differences related to flight: thorax size is greater and wing load is smaller with the individuals living in fragmented landscapes. Higher FMR in fragmented landscapes is also reflected in behavioral differences: males in fragmented landscapes are more mobile and possibly for that reason have higher mating success than males from continuous landscapes.

These empirical results show that habitat fragmentation selects for genotypes and individuals that are more dispersive, further suggesting that there must be genetic raw material to fuel this selection process. Consistently, genome and transcriptome sequencing a large number of individuals has revealed that genetic variation within the Åland Islands is relatively high. The estimates of single nucleotide (SNP) and indel polymorphism vary from 15 to 36 SNPs/kb but are likely to be underestimated (Ahola *et al.* 2014; first author's unpubl. data). Thus the genome sequence includes a lot of material for natural selection to act on. This variation can in principle affect phenotypic characteristics either through allelic variation where different alleles code for proteins that have slightly different biological properties. Alternatively, variation can influence biological properties by changing expression levels of genes involved in phenotypic characteristics.

Earlier studies have investigated allelic variation within a single gene, *Pgi* functioning in the glycolysis pathway at the very center of the cellular energy metabolism network. These studies suggested that SNP variation within the *Pgi* coding region explains a surprisingly large amount of the variation affecting FMR (Niitepõld *et al.* 2009). Similarly, comparison between old and new population individuals have shown allele frequency differences in the *Pgi* locus (Haag *et al.* 2005), and subsequent studies have shown that FMR is correlated with gene expression level of *Pgi* (Kvist *et al.* 2015).

These studies were, however, concentrated on *Pgi* only. While the whole transcript of the gene was known, the majority of the results relied on a single SNP (or two allozymes), and hence provided a rather restricted view on genetic and molecular mechanisms leading to population and landscape level differences. The transcriptome and genome sequencing projects enabled studies beyond *Pgi*. The transcriptome sequence (Vera *et al.* 2008) enabled development of a custom microarray, and the genome sequencing (Ahola *et al.* 2014) facilitated RNA-seq studies based on mapping of reads to the genome, both of which in turn enabled global genome- and/or transcriptome-wide analyses.

To date four transcriptome-wide gene expression studies have been completed for *M. cinxia* (Wheat *et al.* 2011, Kvist *et al.* 2013, 2015, Somervuo *et al.* 2014). Together they have uncovered extensive gene expression differences among populations, different environmental conditions, sexes and families. Of these, the family-level differences suggest that the underlying genetic variation acts at least partially through gene expression regulation, while the population level differences suggest complex interactions between the genome and environment/landscape and physiology. At the same time these studies also provide evidence for adaptive responses in the gene expression programs. While the sex-biased gene expression is most prevalent in gonads, it can be detected in the other tissues similarly as in other systems (Perry *et al.* 2014, Grath & Parsch 2016). For example, 37% of *M. cinxia* genes that show gene expression differences after flight exhibit sex-biased gene expression that likely relates to sexual dimorphism, and physiological and behavioral differences between the sexes. These expression differences result in part from the incomplete dosage compensation of the sex chromosome (66% in Kvist *et al.* 2015), where the second copy of the Z chromosome is only partially silenced in males resulting in male-biased gene expression patterns. This interesting observation is an example of an interplay between gene expression and important morphological, physiology and behavioral characteristics between females and males, which may affect sexual selection and antagonism in this species.

Because flight is the key biological factor in butterfly dispersal it is not surprising that three of the four gene expression studies have concentrated directly or indirectly on the flight-induced gene expression changes with *M. cinxia*. Kvist *et al.* (2015) concentrated directly on flight: Which genes and gene groups are differentially expressed in thorax (i.e. flight muscle) due to an induced flight treatment which attempts to simulate a long distance migration that is typical for the females establishing new populations? The results showed that flight has long-lasting effects on roughly 1500 genes (~9% of all *M. cinxia* genes; Kvist *et al.* 2015). Such a large number of genes suggested a general regulatory change. Consistently, many transcription factors were indeed activated immediately after the flight treatment, suggesting that they represent upstream regulators that induce or suppress the expression of larger groups of target genes or pathways. As expected, these relate to biochemical pathways involved in energy metabolism, but genes involved in hypoxia response were also observed. Both pathways were downregulated after induced flight, which is expected because 15 minutes of continuous flight is both energy and oxygen consuming for *M. cinxia* (Suarez *et al.* 2000), and the downregulation of genes may protect butterflies from damaging effects of flight. Interestingly, genes related to immune system and cellular biosynthetic processes were upregulated as a response to flight, which suggests that flight induces general protective mechanism also seen with energy metabolism and hypoxia responsive genes.

Can we detect population level differences suggesting that selection may favor certain gene expression patterns in a subset of populations? The answer is a tentative yes. A gene expression study comparing adult butterflies from two fragmented (AL and UP) and two continuous (SA and OL) populations (Somervuo *et al.* 2014) reported that genes that were up-regulated in flight were also up-regulated in fragmented populations when compared with continuous populations (Somervuo *et al.* 2014, Kvist *et al.* 2015). These genes included several antimicrobial peptides (AMPs), which in *Drosophila* are known to improve tolerance to oxidant stress (Zhao *et al*. 2011). Similar flight-induced gene expression differences have been reported between new and old populations in the Åland system. Wheat *et al*. (2011) reported differences in gene expression profiles and physiological characteristics between individuals from new versus old populations, such that the new population individuals showed higher FMR and higher expression of genes responsible for maintenance of muscle proteins in thorax. Somervuo *et al*. (2014) on the other hand showed low put positive correlation: the genes that were higher expressed in fragmented than continuous populations were also somewhat more higher expressed in new than old populations.

Together, several studies have shown that flight does indeed have an effect on gene expression and the consequences of some of these effects can also be observed at population or landscape level, but the signal is often weak or missing. Why? One should always bear in mind that selection takes into account the whole life-history of the organism and operates at the whole metapopulation level. Therefore, alleles or gene expression patterns that are optimal for flight may not be best for larval development or optimal for continuous landscape or in old established population. Furthermore, gene expression is a very plastic trait varying between time points, samples, tissues and even cells, and it is not always repeatable (Grath & Parsch 2016). Therefore population-level expression studies with high-level hypotheses have low power to produce gene or sequence-level results. For this reason the potential new candidate genes arising from these analyses is low. Next we will highlight a few genes and pathways that have potentially a role in flight capacity or other life history trait differences between populations.

Hexmerins are general storage proteins found in insect hemolymph, and used as reservoirs of amino acids particularly for synthesis of adult proteins during pupation (Telfer & Kunkel 1991), but also for egg production (Pan & Telfer 2001). Gene expression levels of hexamerins are found to be positively correlated with post-diapause larval development time but the expression differences also appear between families, again showing positive correlation with the fast developing families with higher hexamerin expression levels (Kvist *et al*. 2013). Hexamerin genes form a large gene family with overlapping functions. They are co-expressed, and thus most probably co-regulated by the juvenile and ecdysteroid hormones. The results suggest that variation in larval development is likely to be due to underlying genetic variation of hormonal regulators. At the population level the average larval development time is higher in fragmented than in continuous populations (Duplouy *et al*. 2013). Habitat fragmentation may thus select for faster developing individuals. The selection has possibly occurred in the juvenile or ecdysteroid hormone pathway genes, but causal genes under selection and the precise genetic variation or molecular processes remain to be determined.

Another important gene is a hypoxia response gene Hypoxia-inducible factor 1α (*Hif-1α*), a transcription factor and a main regulator of the hypoxia pathway (Kaelin & Ratcliffe 2008). Hypoxia response is a stress reaction triggered by low oxygen level such as long-lasting flight. *Hif-1α* is mainly regulated at the protein level and provides an immediate response to hypoxia. In Kvist *et al*. (2015) *Hif-1α* transcript and other hypoxia-responsive genes were downregulated in butterflies after 20 h of induced flight treatment when the butterflies were recovering from the exercise, and the energy metabolism pathways were downregulated. Somervuo *et al*. (2014) reported that individuals from fragmented populations showed in general lower basal expression levels of *Hif-1α* than the individuals from continuous populations. In contrast, individuals from fragmented landscape have higher FMR (Duplouy *et al*. 2013), and hence higher flight performance, suggesting that they may be less sensitive to hypoxia. This gene is an example of selection at the landscape level where habitat fragmentation favors more dispersive individuals that is then reflected at the genetic or molecular level. Consistently, one SNP in *Hif-1α* is associated with significant expression differences between genotypes, and the genotype associated with lower gene expression appears to be more frequent in the fragmented populations (Somervuo *et al*. 2014).

The third example is succinate-dehydrogenase-d (*Sdhd*) gene. Wheat *et al*. (2011) identified indel variation at 3´ untranslated region (3´UTR) of the transcript that was associated with higher expression of energy metabolism genes, and thereby to higher flight endurance. Interestingly, the new population individuals show higher flight endurance, higher indel abundance and higher expression of *Sdhd* as compared with individuals from the old populations. Later studies have found that *Sdhd* in fact influences hypoxia regulation through the *Hif-1α* protein, and thereby directly impacts the flight performance (Marden *et al*. 2013). Additionally, the *Sdhd* indel region includes a predicted microRNA binding site (Wheat *et al*. 2011). As microRNAs are well-characterized negative regulators of translation, variation in this regulatory event may provide a mechanistic explanation for the differences in flight performance, and hence could at least partly explain differences among more and less dispersive populations.

Together these few studies of gene expression variation in experimental conditions and among individuals derived from populations representing either newly colonized *vs*. old populations, or from continuous vs fragmented landscapes are starting to build a picture of how the genomic information is influenced by the population structure and eventually the landscape. Particularly these studies demonstrate that habitat fragmentation and more generally metapopulation dynamics leads to selection pressure that acts on an intricately connected set of genes and pathways.

Future possibilities and challenges

Lepidopteran genomics is still a new research field. Although draft reference genomes are presently available for ca. 25 species, they do not equally cover clades of Lepidoptera. *Melitaea cinxia* represents a non-tropical Nymphalidae, of which the evolutionary distance to the nearest species sequenced is ~70 My. Importantly, all the genomic resources of *M. cinxia* and most other Lepidoptera species have been made publicly available. This welcomes other research groups to use *M. cinxia* as a study species but also significantly benefits more general comparative studies between other Lepidoptera or between insect species or beyond. Additionally, the presently sequenced Lepidoptera species, including *M. cinxia*, will be serving as comparison points and benchmarks for the larger ongoing genome sequencing initiatives, such as the i5K, aiming to sequence the genomes of 5000 arthropod species (i5K Consortium 2013).

Despite of the availability of reference genomes, many basic properties of genome organisation and function are still poorly understood among the Lepidoptera. The sequence and function of the female-specific W sex chromosome is poorly explored — mostly because of a high TE content (which makes sequence assembly difficult if not nearly impossible) and the lack of protein coding genes. Epigenetic modifications and transcription factor binding sites are mostly unknown for Lepidoptera, as well as the extent, role and significance, especially evolutionary/population biology significance, of the mRNA processing pathways such as pre-mRNA splicing. Noncoding RNA genes have drawn significant attention in mammalian biology in the recent years, but they also remain virtually uncharacterized in the lepidopteran genomes. TEs comprise at least 25%–35% of lepidopteran genomes (Osanai-Futahashi *et al*. 2008, The Heliconius Genome Consortium 2012). Recent studies have provided intriguing examples on the evolutionary role of TEs (Nadeau *et al*. 2016, Van't Hof *et al*. 2016), but at a larger scale the role of TEs in genome and adaptive evolution is mostly unexplored. In the near future, PacBio and other NGS technologies producing long reads will be the key to open up new avenues to these types of research questions that could not be addressed with short read NGS.

One of Ilkka Hanski's specific research topics was to show that allele frequency of one SNP pgi:c.331A>C in the *Pgi* gene is controlled by eco-evolutionary dynamics in the heterogeneous Åland study system (Zheng *et al*. 2009, Hanski & Mononen 2011, Hanski 2011, Hanski *et al*. 2017). The studies of Hanski and colleagues demonstrated that allelic variation in the *Pgi* gene explains large proportion of variation in metapopulation size (Hanski *et al*. 2017), and can thus be regarded as an example of evolutionary consequences of habitat loss and fragmentation

(Hanski & Mononen 2011). These studies were based on a single SNP, but should these classical studies now be extended to the genomic scale? Whole genome sequence and variation information is now available for *M. cinxia*, and enable studies for detecting candidate genes or linked loci responsible for variability in dispersal and reproductive success related traits. The use of those in the eco-evolutionary modelling with the Åland study system would continue Ilkka Hanski's legacy, and broaden the knowledge of evolutionary consequences of habitat fragmentation.

Future challenges include the development of *M. cinxia* lines that would be more sustainable under laboratory conditions. As this would preferably require skipping of the obligate winter diapause, it would be necessary to first define the molecular basis of diapause to develop lines. Furthermore, identifying genes and genetic loci associated with other life history traits, such as those affecting development, would be important for future studies of *M. cinxia*, especially when aiming at understanding the role of phenotypic plasticity (Saastamoinen *et al.* 2013, Ahola *et al.* 2015), microbiota (Ruokolainen *et al.* 2016) and microhabitat to the life history traits.

Currently, an emerging tool in Lepidoptera genomics is to use genome editing, i.e. the ability to make specific changes to the genome of an organism, to demonstrate the effect of potential causal variant to phenotype. Genome editing has already successfully been applied to *B. mori*, *Papilio xuthus* and *D. plexippus* (Wei *et al.* 2014, Li *et al.* 2015, Markert *et al.* 2016) and there are no technical reasons for it to not work with *M. cinxia*. Thus in the future it will be possible to test ecologically relevant candidate genes in their native hosts in a similar fashion as already done with the laboratory model organisms. This makes it possible to move from exclusively correlation-based studies to investigations of causal relationships that will provide deeper understanding of underlying mechanisms of ecologically important traits.

Acknowledgements

We thank all those who contributed to the *Melitaea cinxia* genome project. We are particularly grateful to Panu Somervuo, Leena Salmela, Patrik Koskinen, Pasi Rastas, Niko Välimäki, Alan Schulman, Jaakko Tanskanen, Esko Ukkonen, Veli Mäkinen and Liisa Holm for their role in the development of bioinformatics methods and participation in data analysis and interpretation of the results; Rainer Lehtonen, Petri Auvinen and Ilkka Hanski for their role in result interpretation and the overall design and management of the project; Jouni Kvist and Lars Paulin for management and execution of the laboratory work. Additionally, we thank Panu Somervuo, Petri Auvinen and Lars Paulin for comments for this manuscript, and Leena Salmela and Pasi Rastas for providing additional statistics. This research was supported by grants from the Academy of Finland (project 284601 (MJF) and 283108 (VA)).

References

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., George, R. A., Lewis, S. E., Richards, S., Ashburner, M., Henderson, S. N., Sutton, G. G., Wortman, J. R., Yandell, M. D., Zhang, Q., Chen, L. X., Brandon, R. C., Rogers, Y. H. C., Blazej, R. G., Champe, M., Pfeiffer, B. D., Wan, K. H., Doyle, C., Baxter, E. G., Helt, G., Nelson, C. R., Gabor, G. L., Miklos, Abril, J. F., Agbayani, A., An, H. J., Andrews-Pfannkoch, C., Baldwin, D., Ballew, R. M., Basu, A., Baxendale, J., Bayraktaroglu, L., Beasley, E. M., Beeson, K. Y., Benos, P. V., Berman, B. P., Bhandari, D., Bolshakov, S., Borkova, D., Botchan, M. R., Bouck, J., Brokstein, P., Brottier, P., Burtis, K. C., Busam, D. A., Butler, H., Cadieu, E., Center, A., Chandra, I., Cherry, J. M., Cawley, S., Dahlke, C., Davenport, L. B., Davies, P., Pablos, B. d., Delcher, A., Deng, Z., Mays, A. D., Dew, I., Dietz, S. M., Dodson, K., Doup, L. E., Downes, M., Dugan-Rocha, S., Dunkov, B. C., Dunn, P., Durbin, K. J., Evangelista, C. C., Ferraz, C., Ferriera, S., Fleischmann, W., Fosler, C., Gabrielian, A. E., Garg, N. S., Gelbart, W. M., Glasser, K., Glodek, A., Gong, F., Gorrell, J. H., Gu, Z., Guan, P., Harris, M., Harris, N. L., Harvey, D., Heiman, T. J., Hernandez, J. R., Houck, J., Hostin, D., Houston, K. A., Howland, T. J., Wei, M. H., Ibegwam, C., Jalali, M., Kalush, F., Karpen, G. H., Ke, Z., Kennison, J. A., Ketchum, K. A., Kimmel, B. E., Kodira, C. D., Kraft, C., Kravitz, S., Kulp, D., Lai, Z., Lasko, P., Lei, Y., Levitsky, A. A., Li, J., Li, Z., Liang, Y., Lin, X., Liu, X., Mattei, B., McIntosh, T. C., McLeod, M. P., McPherson, D., Merkulov, G., Milshina, N. V., Mobarry, C., Morris, J., Moshrefi, A., Mount, S. M., Moy, M., Murphy, B., Murphy, L., Muzny, D. M., Nelson, D. L., Nelson, D. R., Nelson, K. A., Nixon, K., Nusskern, D. R., Pacleb, J. M., Palazzolo, M., Pittman, G. S., Pan, S., Pollard, J., Puri, V., Reese, M. G., Reinert, K., Remington, K., Saunders, R. D. C., Scheeler, F., Shen, H., Shue, B. C., Sidén-Kiamos, I., Simpson, M., Skupski, M. P., Smith, T., Spier, E., Spradling, A. C., Stapleton, M., Strong, R., Sun, E., Svirskas, R., Tector, C., Turner, R., Venter, E., Wang, A. H., Wang, X., Wang,

Z. Y., Wassarman, D. A., Weinstock, G. M., Weissenbach, J., Williams, S. M., Woodage, T., Worley, K. C., Wu, D., Yang, S., Yao, Q. A., Ye, J., Yeh, R. F., Zaveri, J. S., Zhan, M., Zhang, G., Zhao, Q., Zheng, L., Zheng, X. H., Zhong, F. N., Zhong, W., Zhou, X., Zhu, S., Zhu, X., Smith, H. O., Gibbs, R. A., Myers, E. W., Rubin, G. M. & Venter, J. C. 2000: The genome sequence of *Drosophila melanogaster*. — *Science* 287: 2185–2195.

Ahola, V., Koskinen, P., Wong, S. C., Kvist, J., Paulin, L., Auvinen, P., Saastamoinen, M., Frilander, M. J., Lehtonen, R. & Hanski, I. 2015: Temperature- and sex-related effects of serine protease alleles on larval development in the Glanville fritillary butterfly. — *Journal of Evolutionary Biology* 28: 2224–2235.

Ahola, V., Lehtonen, R., Somervuo, P., Salmela, L., Koskinen, P., Rastas, P., Välimäki, N., Paulin, L., Kvist, J., Wahlberg, N., Tanskanen, J., Hornett, E. A., Ferguson, L. C., Luo, S., Cao, Z., de Jong, M. A., Duplouy, A., Smolander, O.-P., Vogel, H., McCoy, R. C., Qian, K., Chong, W. S., Zhang, Q., Ahmad, F., Haukka, J. K., Joshi, A., Salojärvi, J., Wheat, C. W., Grosse-Wilde, E., Hughes, D., Katainen, R., Pitkänen, E., Ylinen, J., Waterhouse, R. M., Turunen, M., Vähärautio, A., Ojanen, S. P., Schulman, A. H., Taipale, M., Lawson, D., Ukkonen, E., Mäkinen, V., Goldsmith, M. R., Holm, L., Auvinen, P., Frilander, M. J. & Hanski, I. 2014: The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera. — *Nature Communications* 5, 4737, doi:10.1038/ncomms5737.

Baxter, S. W., Davey, J. W., Johnston, J. S., Shelton, A. M., Heckel, D. G., Jiggins, C. D. & Blaxter, M. L. 2011: Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism. — *PLoS ONE* 6(4): e19315, doi:10.1371/journal.pone.0019315.

Beldade, P., Saenko, S. V., Pul, N. & Long, A. D. 2009: A gene-based linkage map for *Bicyclus anynana* butterflies allows for a comprehensive analysis of synteny with the lepidopteran reference genome. — *PLoS Genetics* 5(2): e1000366, doi:10.1371/journal.pgen.1000366.

Brown, K. S. Jr., Von Schoultz, B. & Suomalainen, E. 2004: Chromosome evolution in Neotropical Danainae and Ithomiinae (Lepidoptera). — *Hereditas* 141: 216–236.

Cantarel, B. L., Korf, I., Robb, S. M., Parra, G., Ross, E., Moore, B., Holt, C., Sanchez Alvarado, A. & Yandell, M. 2008: MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. — *Genome Research* 18: 188–196.

d'Alencon, E., Sezutsu, H., Legeai, F., Permal, E., Bernard-Samain, S., Gimenez, S., Gagneur, C., Cousserans, F., Shimomura, M., Brun-Barale, A., Flutre, T., Couloux, A., East, P., Gordon, K., Mita, K., Quesneville, H., Fournier, P. & Feyereisen, R. 2010: Extensive synteny conservation of holocentric chromosomes in Lepidoptera despite high rates of local genome rearrangements. — *Proceedings of the National Academy of Sciences of the United States of America* 107: 7680–7685.

Davey, J. W., Chouteau, M., Barker, S. L., Maroja, L., Baxter, S. W., Simpson, F., Merrill, R. M., Joron, M., Mallet, J., Dasmahapatra, K. K. & Jiggins, C. D. 2016: Major improvements to the *Heliconius melpomene* genome assembly used to confirm 10 chromosome fusion events in 6 million years of butterfly evolution. — *G3: Genes|Genomes|Genetics* 6: 695–708.

de Jong, M. A., Wong, S. C., Lehtonen, R. & Hanski, I. 2014: Cytochrome P450 gene *CYP337* and heritability of fitness traits in the Glanville fritillary butterfly. — *Molecular Ecology* 23: 1994–2005.

Duplouy, A., Ikonen, S. & Hanski, I. 2013: Life history of the Glanville fritillary butterfly in fragmented versus continuous landscapes. — *Ecology and Evolution* 3: 5141–5156.

Ehrlich, P. & Hanski, I. 2004: *On the wings of checkerspots: A model system for population biology*. — Oxford University Press, New York.

Grath, S. & Parsch, J. 2016: Sex-biased gene expression. — *Annual Review of Genetics* 50: 29–44.

Haag, C. R., Saastamoinen, M., Marden, J. H. & Hanski, I. 2005: A candidate locus for variation in dispersal rate in a butterfly metapopulation. — *Proceedings of the Royal Society B* 272: 2449–2456.

Hanski, I. 1999: *Metapopulation ecology*. — Oxford University Press, New York.

Hanski, I. A. 2011: Eco-evolutionary spatial dynamics in the Glanville fritillary butterfly. — *Proceedings of the National Academy of Sciences of the United States of America* 108: 14397–14404.

Hanski, I. & Mononen, T. 2011: Eco-evolutionary dynamics of dispersal in spatially heterogeneous environments. — *Ecology Letters* 14: 1025–1034.

Hanski, I. & Ovaskainen, O. 2000: The metapopulation capacity of a fragmented landscape. — *Nature* 404: 755–758.

Hanski, I. & Singer, M. C. 2001: Extinction–colonization dynamics and host-plant choice in butterfly metapopulations. — *The American Naturalist* 158: 341–353.

Hanski, I. & Saccheri, I. 2006: Molecular-level variation affects population growth in a butterfly metapopulation. — *PLoS Biology* 4(5): e129, doi:10.1371/journal.pbio.0040129.

Hanski, I., Schulz, T., Wong, S. C., Ahola, V., Ruokolainen, A. & Ojanen, S. P. 2017: Ecological and genetic basis of metapopulation persistence of the Glanville fritillary butterfly in fragmented landscapes. — *Nature Communications* 8, 14504, doi:10.1038/ncomms14504.

i5K Consortium 2013: The i5K Initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. — *Journal of Heredity* 104: 595–600.

Kaelin, W. G. Jr. & Ratcliffe, P. J. 2008: Oxygen sensing by metazoans: the central role of the HIF hydroxylase pathway. — *Molecular Cell* 30: 393–402.

Kandul, N. P., Lukhtanov, V. A. & Pierce, N. E. 2007: Karyotypic diversity and speciation in Agrodiaetus butterflies. — *Evolution* 61: 546–559.

Klemme, I. & Hanski, I. 2009: Heritability of and strong single gene (*Pgi*) effects on life-history traits in the Glanville fritillary butterfly. — *Journal of Evolutionary Biology* 22: 1944–1953.

Koskinen, J. P. & Holm, L. 2012: SANS: high-throughput retrieval of protein sequences allowing 50% mismatches. — *Bioinformatics* 28: 438–443.

Koskinen, P., Törönen, P., Nokso-Koivisto, J. & Holm, L. 2015: PANNZER — High-throughput functional annotation of uncharacterized proteins in an error-prone environment. — *Bioinformatics* 31: 1544–1552.

Kuussaari, M., Singer, M. & Hanski, I. 2000: Local specialization and landscape-level influence on host use in an herbivorous insect. — *Ecology* 81: 2177–2187.

Kvist, J., Wheat, C. W., Kallioniemi, E., Saastamoinen, M., Hanski, I. & Frilander, M. J. 2013: Temperature treatments during larval development reveal extensive heritable and plastic variation in gene expression and life history traits. — *Molecular Ecology* 22: 602–619.

Kvist, J., Mattila, A. L. K., Somervuo, P., Ahola, V., Koskinen, P., Paulin, L., Salmela, L., Fountain, T., Rastas, P., Ruokolainen, A., Taipale, M., Holm, L., Auvinen, P., Lehtonen, R., Frilander, M. J. & Hanski, I. 2015: Flight-induced changes in gene expression in the Glanville fritillary butterfly. — *Molecular Ecology* 24: 4886–4900.

Li, X., Fan, D., Zhang, W., Liu, G., Zhang, L., Zhao, L., Fang, X., Chen, L., Dong, Y., Chen, Y., Ding, Y., Zhao, R., Feng, M., Zhu, Y., Feng, Y., Jiang, X., Zhu, D., Xiang, H., Feng, X., Li, S., Wang, J., Zhang, G., Kronforst, M. R. & Wang, W. 2015: Outbred genome sequencing and CRISPR/Cas9 gene editing in butterflies. — *Nature Communications* 6: 8212.

Marden, J. H., Fescemyer, H. W., Schilder, R. J., Doerfler, W. R., Vera, J. C. & Wheat, C. W. 2013: Genetic variation in hif signaling underlies quantitative variation in physiological and life-history traits within lowland butterfly populations. — *Evolution* 67: 1105–1115.

Markert, M. J., Zhang, Y., Enuameh, M. S., Reppert, S. M., Wolfe, S. A. & Merlin, C. 2016: Genomic access to monarch migration using TALEN and CRISPR/Cas9-mediated targeted mutagenesis. — *G3: Genes|Genomes|Genetics* 6: 905–915.

Mattila, A. L. K. & Hanski, I. 2014: Heritability of flight and resting metabolic rates in the Glanville fritillary butterfly. — *Journal of Evolutionary Biology* 27: 1733–1743.

Mayr, E. 1963: *Animal species and evolution*. — Harvard University Press, Cambridge, MA.

Mita, K., Kasahara, M., Sasaki, S., Nagayasu, Y., Yamada, T., Kanamori, H., Namiki, N., Kitagawa, M., Yamashita, H., Yasukochi, Y., Kadono-Okuda, K., Yamamoto, K., Ajimura, M., Ravikumar, G., Shimomura, M., Nagamura, Y., Shin, I. T., Abe, H., Shimada, T., Morishita, S. & Sasaki, T. 2004: The genome sequence of silkworm, *Bombyx mori*. — *DNA Research* 11: 27–35.

Murakami, A. & Imai, H. T. 1974: Cytological evidence for holocentric chromosomes of the silkworms, *Bombyx mori* and *B. mandarina*, (Bombycidae, Lepidoptera). — *Chromosoma* 47: 167–178.

Nadeau, N. J., Pardo-Diaz, C., Whibley, A., Supple, M. A., Saenko, S. V., Wallbank, R. W. R., Wu, G. C., Maroja, L., Ferguson, L., Hanly, J. J., Hines, H., Salazar, C., Merrill, R. M., Dowling, A. J., ffrench-Constant, R. H., Llaurens, V., Joron, M., McMillan, W. O. & Jiggins, C. D. 2016: The gene cortex controls mimicry and crypsis in butterflies and moths. — *Nature* 534: 106–110.

Nieminen, M., Siljander, M. & Hanski, I. 2004: Structure and dynamics of *Melitaea cinxia* metapopulations. — In: Ehrlich, P. R. & Hanski, I. (eds.), *On the wings of checkerspots: A model system for population biology*: 63. Oxford University Press, New York.

Niitepõld, K. & Saastamoinen, M. 2017: A candidate gene in an ecological model species: Phosphoglucose isomerase (*Pgi*) in the Glanville fritillary butterfly (*Melitaea cinxia*). — *Annales Zoologici Fennici* 54: 259–273.

Niitepõld, K., Smith, A. D., Osborne, J. L., Reynolds, D. R., Carreck, N. L., Martin, A. P., Marden, J. H., Ovaskainen, O. & Hanski, I. 2009: Flight metabolic rate and *Pgi* genotype influence butterfly dispersal rate in the field. — *Ecology* 90: 2223–2232.

Ojanen, S. P., Nieminen, M., Meyke, E., Pöyry, J. & Hanski, I. 2013: Long-term metapopulation study of the Glanville fritillary butterfly (*Melitaea cinxia*): survey methods, data management, and long-term population trends. — *Ecology and Evolution* 3: 3713–3737.

Orsini, L., Wheat, C. W., Haag, C. R., Kvist, J., Frilander, M. J. & Hanski, I. 2009: Fitness differences associated with *Pgi* SNP genotypes in the Glanville fritillary butterfly (*Melitaea cinxia*). — *Journal of Evolutionary Biology* 22: 367–375.

Osanai-Futahashi, M., Suetsugu, Y., Mita, K. & Fujiwara, H. 2008: Genome-wide screening and characterization of transposable elements and their distribution analysis in the silkworm, *Bombyx mori*. — *Insect biochemistry and molecular biology* 38: 1046–1057.

Ovaskainen, O., Smith, A. D., Osborne, J. L., Reynolds, D. R., Carreck, N. L., Martin, A. P., Niitepõld, K. & Hanski, I. 2008: Tracking butterfly movements with harmonic radar reveals an effect of population age on movement distance. — *Proceedings of the National Academy of Sciences* 105: 19090–19095.

Pan, M. L. & Telfer, W. H. 2001: Storage hexamer utilization in two lepidopterans: differences correlated with the timing of egg formation. — *Journal of Insect Science* 1: 2.

Perry, J. C., Harrison, P. W. & Mank, J. E. 2014: The ontogeny and evolution of sex-biased gene expression in *Drosophila melanogaster*. — *Molecular Biology and Evolution* 31: 1206–1219.

Rastas, P., Paulin, L., Hanski, I., Lehtonen, R. & Auvinen, P. 2013: Lep-MAP: fast and accurate linkage map construction for large SNP datasets. — *Bioinformatics* 29: 3128–3134.

Ruokolainen, L., Ikonen, S., Makkonen, H. & Hanski, I. 2016: Larval growth rate is associated with the composition of the gut microbiota in the Glanville fritillary butterfly. — *Oecologia* 181: 895–903.

Saastamoinen, M. 2007: Life-history, genotypic, and environmental correlates of clutch size in the Glanville fritillary butterfly. — *Ecological Entomology* 32: 235–242.

Saastamoinen, M. 2008: Heritability of dispersal rate and other life history traits in the Glanville fritillary butterfly. — *Heredity* 100: 39–46.

Saastamoinen, M., Ikonen, S., Wong, S. C., Lehtonen, R. &

Hanski, I. 2013: Plastic larval development in a butterfly has complex environmental and genetic causes and consequences for population dynamics. — *Journal of Animal Ecology* 82: 529–539.

Sahara, K., Yoshido, A., Shibata, F., Fujikawa-Kojima, N., Okabe, T., Tanaka-Okuyama, M. & Yasukochi, Y. 2013: FISH identification of *Helicoverpa armigera* and *Mamestra brassicae* chromosomes by BAC and fosmid probes. — *Insect Biochemistry and Molecular Biology* 43: 644–653.

Salmela, L. 2010: Correction of sequencing errors in a mixed set of reads. — *Bioinformatics* 26: 1284–1290.

Salmela, L. & Schröder, J. 2011: Correcting errors in short reads by multiple alignments. — *Bioinformatics* 27: 1455–1461.

Salmela, L., Mäkinen, V., Välimäki, N., Ylinen, J. & Ukkonen, E. 2011: Fast scaffolding with small independent mixed integer programs. — *Bioinformatics* 27: 3259–3265.

Saura, A., Von Schoultz, B., Saura, A. O. & Brown, K. S. Jr. 2013: Chromosome evolution in Neotropical butterflies. — *Hereditas* 150: 26–37.

Smoller, D. A., Petrov, D. & Hartl, D. L. 1991: Characterization of bacteriophage P1 library containing inserts of *Drosophila* DNA of 75–100 kilobase pairs. — *Chromosoma* 100: 487–494.

Somervuo, P., Kvist, J., Ikonen, S., Auvinen, P., Paulin, L., Koskinen, P., Holm, L., Taipale, M., Duplouy, A., Ruokolainen, A., Saarnio, S., Sirén, J., Kohonen, J., Corander, J., Frilander, M. J., Ahola, V. & Hanski, I. 2014: Transcriptome analysis reveals signature of adaptation to landscape fragmentation. — *PLoS ONE* 9(7): e101467, doi:10.1371/journal.pone.0101467.

Suarez, R. K. 2000: Energy metabolism during insect flight: Biochemical design and physiological performance. — *Physiological and Biochemical Zoology* 73: 765–771.

Suomalainen, E. 1966: Achiasmatische Oogenese bei Trichopteren. — *Chromosoma* 18: 201–207.

Suomalainen, E. 1969: Chromosome evolution in the Lepidoptera. — *Chromosomes Today* 2: 132–138.

Ta, H. X., Koskinen, P. & Holm, L. 2011: A novel method for assigning functional linkages to proteins using enhanced phylogenetic trees. — *Bioinformatics* 27: 700–706.

Telfer, W. H. & Kunkel, J. G. 1991: The function and evolution of insect storage hexamers. — *Annual Review of Entomology* 36: 205–228.

The Heliconius Genome Consortium 2012: Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. — *Nature* 487: 94–98.

The International Silkworm Genome Consortium 2008: The genome of a lepidopteran model insect, the silkworm *Bombyx mori*. — *Insect Biochemistry and Molecular Biology* 38: 1036–1045.

Wahlberg, N., Wheat, C. W. & Pena, C. 2013: Timing and patterns in the taxonomic diversification of Lepidoptera (butterflies and moths). — *PLoS ONE* 8(11): e80875, doi:10.1371/journal.pone.0080875.

Van't Hof, A. E., Brakefield, P. M., Saccheri, I. J. & Zwaan, B. J. 2007: Evolutionary dynamics of multilocus microsatellite arrangements in the genome of the butterfly *Bicyclus anynana*, with implications for other Lepidoptera. — *Heredity* 98: 320–328.

Van't Hof, A. E., Nguyen, P., Dalikova, M., Edmonds, N., Marec, F. & Saccheri, I. J. 2013: Linkage map of the peppered moth, *Biston betularia* (Lepidoptera, Geometridae): a model of industrial melanism. — *Heredity* 110: 283–295.

Van't Hof, A. E., Campagne, P., Rigden, D. J., Yung, C. J., Lingley, J., Quail, M. A., Hall, N., Darby, A. C. & Saccheri, I. J. 2016: The industrial melanism mutation in British peppered moths is a transposable element. — *Nature* 534: 102–105.

Wei, W., Xin, H., Roy, B., Dai, J., Miao, Y. & Gao, G. 2014: Heritable genome editing with CRISPR/Cas9 in the silkworm, *Bombyx mori*. — *PLoS ONE* 9(7): e101210, doi:10.1371/journal.pone.0101210.

Vera, J. C., Wheat, C. W., Fescemyer, H. W., Frilander, M. J., Crawford, D. L., Hanski, I. & Marden, J. H. 2008: Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. — *Molecular Ecology* 17: 1636–1647.

Wheat, C. W., Fescemyer, H. W., Kvist, J., Tas, E. V. A., Vera, J. C., Frilander, M. J., Hanski, I. & Marden, J. H. 2011: Functional genomics of life history variation in a butterfly metapopulation. — *Molecular Ecology* 20: 1813–1828.

White, M. J. D. 1973: *Animal cytology and evolution*, 3rd ed. — Cambridge Univ. Press, Cambridge.

Wolf, K. W. 1996: The structure of condensed chromosomes in mitosis and meiosis of insects. — *International Journal of Insect Morphology & Embryology* 25: 37–62.

Wong, S. C., Oksanen, A., Mattila, A. L. K., Lehtonen, R., Niitepõld, K. & Hanski, I. 2016: Effects of ambient and preceding temperatures and metabolic genes on flight metabolism in the Glanville fritillary butterfly. — *Journal of Insect Physiology* 85: 23–31.

Xia, Q., Zhou, Z., Lu, C., Cheng, D., Dai, F., Li, B., Zhao, P., Zha, X., Cheng, T., Chai, C., Pan, G., Xu, J., Liu, C., Lin, Y., Qian, J., Hou, Y., Wu, Z., Li, G., Pan, M., Li, C., Shen, Y., Lan, X., Yuan, L., Li, T., Xu, H., Yang, G., Wan, Y., Zhu, Y., Yu, M., Shen, W., Wu, D., Xiang, Z., Yu, J., Wang, J., Li, R., Shi, J., Li, H., Li, G., Su, J., Wang, X., Li, G., Zhang, Z., Wu, Q., Li, J., Zhang, Q., Wei, N., Xu, J., Sun, H., Dong, L., Liu, D., Zhao, S., Zhao, X., Meng, Q., Lan, F., Huang, X., Li, Y., Fang, L., Li, C., Li, D., Sun, Y., Zhang, Z., Yang, Z., Huang, Y., Xi, Y., Qi, Q., He, D., Huang, H., Zhang, X., Wang, Z., Li, W., Cao, Y., Yu, Y., Yu, H., Li, J., Ye, J., Chen, H., Zhou, Y., Liu, B., Wang, J., Ye, J., Ji, H., Li, S., Ni, P., Zhang, J., Zhang, Y., Zheng, H., Mao, B., Wang, W., Ye, C., Li, S., Wang, J., Wong, G. K. & Yang, H. 2004: A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). — *Science* 306: 1937–1940.

Yasukochi, Y., Ashakumary, L. A., Baba, K., Yoshido, A. & Sahara, K. 2006: A second-generation integrated map of the silkworm reveals synteny and conserved gene order between lepidopteran insects. — *Genetics* 173: 1319–1328.

Yasukochi, Y., Ohno, M., Shibata, F., Jouraku, A., Nakano, R., Ishikawa, Y. & Sahara, K. 2016: A FISH-based chromosome map for the European corn borer yields insights into ancient chromosomal fusions in the silkworm. — *Heredity* 116: 75–83.

Yasukochi, Y., Tanaka-Okuyama, M., Shibata, F., Yoshido, A., Marec, F., Wu, C. C., Zhang, H. B., Goldsmith, M. R. & Sahara, K. 2009: Extensive conserved synteny of genes between the karyotypes of *Manduca sexta* and *Bombyx mori* revealed by BAC-FISH mapping. — *PLoS ONE* 4(10): e7465, doi:10.1371/journal.pone.0007465.

Yoshido, A., Yasukochi, Y. & Sahara, K. 2011: *Samia cynthia* versus *Bombyx mori*: comparative gene mapping between a species with a low-number karyotype and the model species of Lepidoptera. — *Insect Biochemistry and Molecular Biology* 41: 370–377.

You, M. S., Yue, Z., He, W. Y., Yang, X. H., Yang, G., Xie, M., Zhan, D. L., Baxter, S. W., Vasseur, L., Gurr, G. M., Douglas, C. J., Bai, J. L., Wang, P., Cui, K., Huang, S. G., Li, X. C., Zhou, Q., Wu, Z. Y., Chen, Q. L., Liu, C. H., Wang, B., Li, X. J., Xu, X. F., Lu, C. X., Hu, M., Davey, J. W., Smith, S. M., Chen, M. S., Xia, X. F., Tang, W. Q., Ke, F. S., Zheng, D. D., Hu, Y. L., Song, F. Q., You, Y. C., Ma, X. L., Peng, L., Zheng, Y. K., Liang, Y., Chen, Y. Q., Yu, L. Y., Zhang, Y. N., Liu, Y. Y., Li, G. Q., Fang, L., Li, J. X., Zhou, X., Luo, Y. D., Gou, C. Y., Wang, J. Y., Wang, J., Yang, H. M. & Wang, J. 2013: A heterozygous moth genome provides insights into herbivory and detoxification. — *Nature Genetics* 45: 220–225.

Zhan, S. & Reppert, S. M. 2013: MonarchBase: the monarch butterfly genome database. — *Nucleic Acids Research* 41: 758–763.

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

Zhao, H. W., Zhou, D. & Haddad, G. G. 2011: Antimicrobial peptides increase tolerance to oxidant stress in *Drosophila melanogaster*. — *Journal of Biological Chemistry* 286: 6211–6218.

Zheng, C., Ovaskainen, O. & Hanski, I. 2009: Modelling single nucleotide effects in phosphoglucose isomerase on dispersal in the Glanville fritillary butterfly: coupling of ecological and evolutionary dynamics. — *Philosophical Transactions of the Royal Society B* 364: 1519–1532.