9-5-2014

The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera

Virpi Ahola

Marian R. Goldsmith
University of Rhode Island, mk101@etal.uri.edu

et al

Citation/Publisher Attribution

Ahola, Virpi et al. "The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera." *Nature Communications*. September 5, 2014.
Available at: http://dx.doi.org/10.1038/ncomms5737

The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera

NATURE COMMUNICATIONS | ARTICLE OPEN

The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera

Virpi Ahola, Rainer Lehtonen, Panu Somervuo, Leena Salmela, Patrik Koskinen, Pasi Rastas, Niko Välimäki, Lars Paulin, Jouni Kvist, Niklas Wahlberg, Jaakko Tanskanen, Emily A. Hornett, Laura C. Ferguson, Shiqi Luo, Zijuan Cao, Maaike A. de Jong, Anne Duplouy, Olli-Pekka Smolander, Heiko Vogel, Rajiv C. McCoy *et al.*

Nature Communications **5**, Article number: 4737 doi:10.1038/ncomms5737
Received 11 April 2014 Accepted 17 July 2014 Published 05 September 2014

Abstract

Print

Previous studies have reported that chromosome synteny in Lepidoptera has been well conserved, yet the number of haploid chromosomes varies widely from 5 to 223. Here we report the genome (393 Mb) of the Glanville fritillary butterfly (*Melitaea cinxia*; Nymphalidae), a widely recognized model species in metapopulation biology and eco-evolutionary research, which has the putative ancestral karyotype of *n*=31. Using a phylogenetic analyses of Nymphalidae and of other Lepidoptera, combined with orthologue-level comparisons of chromosomes, we conclude that the ancestral lepidopteran karyotype has been *n*=31 for at least 140 My. We show that fusion chromosomes have retained the ancestral chromosome segments and very few rearrangements have occurred across the fusion sites. The same, shortest ancestral chromosomes have independently participated in fusion events in species with smaller karyotypes. The short chromosomes have higher rearrangement rate than long ones. These characteristics highlight distinctive features of the evolutionary dynamics of butterflies and moths.

Subject terms: Biological sciences Evolution Genetics

Introduction

Butterflies and moths (Lepidoptera) have a large number of short holocentric chromosomes[1, 2, 3] with substantial variation in chromosome number[4, 5, 6] (*n*=5–223). However, the most common chromosome numbers are *n*=29–31 (refs 7, 8), and the distribution is markedly skewed with only a few species having *n*>31. The ancestral chromosome number has been inferred to be 31 (refs 9, 10), but until recently this has been difficult to confirm due to lack of comprehensive phylogenies. In spite of much variation in the number of chromosomes, the amount of DNA is approximately the same in different species, suggesting that species with fewer chromosomes have, on average, longer chromosomes[7]. Lepidopteran karyotypes are thought to have evolved via fusion and fission events[7]. Due to the holocentric chromosome structure with dispersed kinetochore activity, such events have been expected to be less deleterious than in monocentric chromosomes[11, 12]. Conversely, holocentricity may restrict gene flow[13] through meiotic[14] and recombination suppression mechanisms[15].

Detailed sequence-level studies of structural variation in lepidopteran chromosomes became feasible with the publication of the whole-genome sequence of *Bombyx mori* (*n*=28)[16]. Paradoxically, in spite of their holocentric chromosome structure, genomic comparison of *Heliconius melpomene* (*n*=21) with *B. mori* suggested a highly conserved chromosomal gene content and an ancestral chromosome number of *n*=31 (ref. 10), supporting previous low-resolution comparisons in other lepidopteran species[17, 18, 19, 20, 21, 22, 23, 24]. Linkage maps suggest that the karyotypes of *B. mori* and *H. melpomene* have evolved via several fusion events. Potential fusion chromosomes have been identified in both species[10, 21, 22, 25], but confirmation has not been possible without a whole-genome sequence and a high-density linkage map for a species with the putative ancestral karyotype.

Here we describe the genome of the Glanville fritillary butterfly (*Melitaea cinxia*), the first butterfly species with *n*=31 and for which both the genome and a high-density linkage map[26] are now available. Our analysis of chromosome fusions in *B. mori* and *H. melpomene* suggests unexpected features shaping karyotype evolution in Lepidoptera.

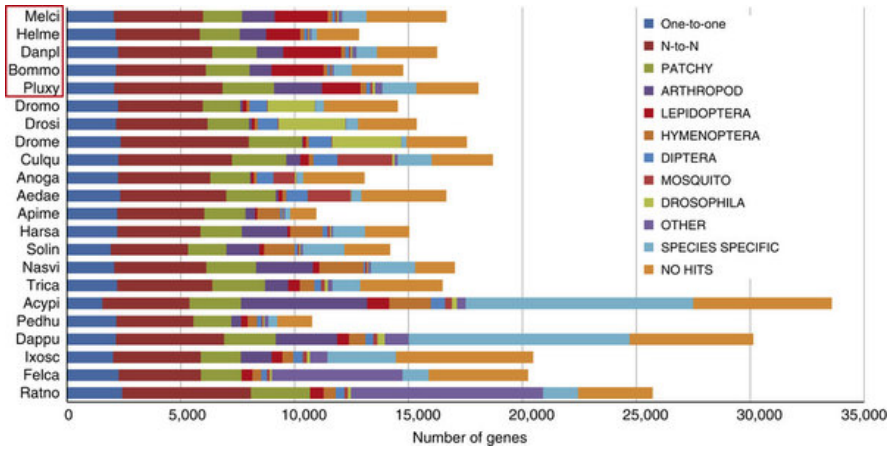
Results

Sequencing of the *M. cinxia* genome was based on samples from a polymorphic natural population (Supplementary Note 1), as the yield of DNA from a single individual was insufficient and no inbred lines were available (Supplementary Note 2). The initial 393 Mb assembly of the genome was performed using 454 and Illumina paired-end (PE) reads from a single male and Illumina PE reads from a pool of 10 full-sibs (Supplementary Notes 2 and 3). The contigs were then scaffolded[27] with PE and mate-pair (MP) library data from several full-sib families and an unrelated individual, which were sequenced with SOLiD, Illumina and 454 platforms (Supplementary Figs 1–3; Supplementary Table 1). The final assembly of the nuclear genome comprises 49,851 contigs (N50=13 kb) and 8,262 scaffolds (N50=119 kb), with an overall coverage of 95 × (Supplementary Figs 4–6; Supplementary Tables 1 and 3). A linkage map based on 40,718 single nucleotide polymorphisms (SNPs)[26] assigned 3,507 scaffolds (318 Mb) to 31 linkage groups, matching the 31 chromosomes reported for this species in a cytogenetic study[28] (Supplementary Table 6; Supplementary Note 5). For subsequent superscaffolding, we applied an in-house method utilizing the linkage map, long MP data and PacBio reads. The resulting 1,453 superscaffolds (N50=331 kb) cover 72% of the genome (Supplementary Fig. 7; Supplementary Tables 4 and 5). The remaining 4,846 scaffolds covering 111 Mb lack consistent map information. The mitochondrial genome (15,171 bp) was assembled and annotated separately (Supplementary Figs 8 and 13; Supplementary Tables 23 and 24; Supplementary Notes 3 and 8).

The quality of the assembly was assessed using several approaches and data sets, including PE, MP and PacBio read data and independently assembled transcriptome data, which were mapped to genomic contigs and scaffolds (Supplementary Note 6). Estimates of genome completeness and correctness indicate that the assembly is of high quality (Supplementary Figs 9 and 10; Supplementary Tables 7–14). Most importantly, the scaffolds show high consistency with the linkage map (91.4% non-chimeric scaffolds). The quality of the superscaffolds was assessed by comparison with an independent high-density linkage map, which showed that only 2.4% of superscaffolds have short chimeric stretches, mostly at their ends.

We predicted 16,667 gene models, the vast majority of which (96%) were supported by transcriptome data (Supplementary Tables 2, 17 and 18; Supplementary Note 8). Clustering the protein sequences into orthologous groups shows that, consistent with previous reports[29], the sequenced lepidopteran genomes have very similar gene content despite 140 My[30] of independent evolution (Fig. 1, Supplementary Figs 17–19). Functional annotations were performed separately for the predicted gene models and the assembled transcripts, yielding protein descriptions and gene ontology (GO) classifications for 12,410, KEGG pathways for 3,685 and InterProScan hits for 8,529 gene models (Supplementary Figs 14–16). We identified noncoding RNA genes including miRNA precursors, ribosomal RNAs, transfer RNAs and spliceosomal small nuclear RNAs (Supplementary Figs 11 and 12, Supplementary Tables 19–22). Moreover, we carried out manual curation of gene models and descriptions for 558 genes (Supplementary Fig. 20; Supplementary Table 25; Supplementary Data 1), including the Hox gene cluster. We identified all canonical Hox genes and four copies of the special homeobox (*Shx*) genes, two *ShxA*, and one *ShxB* and *C* (Supplementary Figs 21 and 22; Supplementary Note 8). All the Hox genes follow the gene order and location described for other Lepidoptera, but the duplication of *ShxA* and lack of *ShxD* are distinct from other nymphalid butterflies [10, 31].

Figure 1: Number of proteins in different classes of orthologous groups.



The statistics for *M. cinxia* are very similar to those for other Lepidoptera, including 5,977 conserved core proteins, 7,177 taxonomic order-, family- or species-specific proteins and 3,513 proteins without detectable sequence similarity to others. There appears to be rapid turnover of gene content in the genomes, indicated by the large proportion of order- and family-specific groups, which represent either rapidly evolving genes or dispensable

ancient paralogues that have been deleted from most other lineages. Species codes are as used in SwissProt: Melci=*Melitaea cinxia*, Helme=*Heliconius melpomene*, Danpl=*Danaus plexippus*, Bommo=*Bombyx mori*, Pluxy=*Plutella xylostella*, Dromo=*Drosophila mojavensis*, Drosi=*Drosophila simulans*, Drome=*Drosophila melanogaster*, Culqu=*Culex quinquefasciatus*, Anoga=*Anopheles gambiae*, Aedae=*Aedes aegypti*, Apime=*Apis mellifera*, Harsa=*Harpegnathos saltator*, Solin=*Solenopsis invicta*, Nasvi=*Nasonia vitripennis*, Trica=*Tribolium castaneum*, Acypi=*Acyrthosiphon pisum*, Pedhu=*Pediculus humanus*, Dappu=*Daphnia pulex*, Ixosc=*Ixodes scapularis*, Felca=*Felis catus* and Ratno=*Rattus norvegicus*. The lepidopteran species are highlighted with a box. See Supplementary Note 8 for the definition of classes.

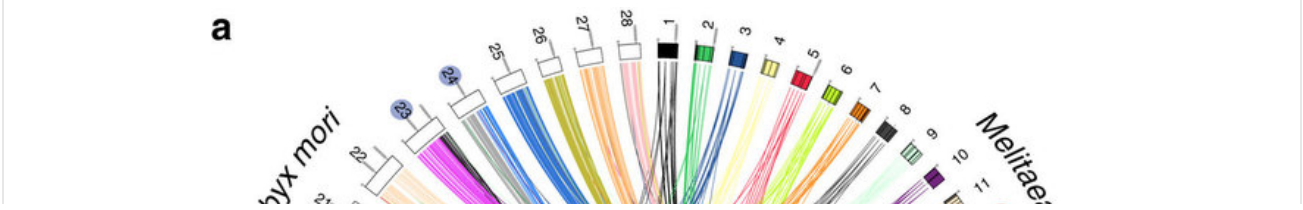
Genomic variation was characterized with three independent data sets (Supplementary Figs 23 and 24; Supplementary Tables 26 and 27; Supplementary Note 9). In a group of 10 full-sibs and an independent individual sequenced with Illumina, more than five million SNPs were identified corresponding to an average density of 13.2 SNPs per kb. The SNP density was 8.2 SNPs/kb in the coding exons, which is roughly half of the density in introns (15.3 SNPs per kb). Approximately half a million indels with an average density of 1.7 per kb were identified. Longer indel variants (>50 bp) were detected using the PacBio data comprising 2,165 deletions and 313 insertions. We have described elsewhere genetic variation in four regional metapopulations of *M. cinxia* using extensive RNA-seq data[32].

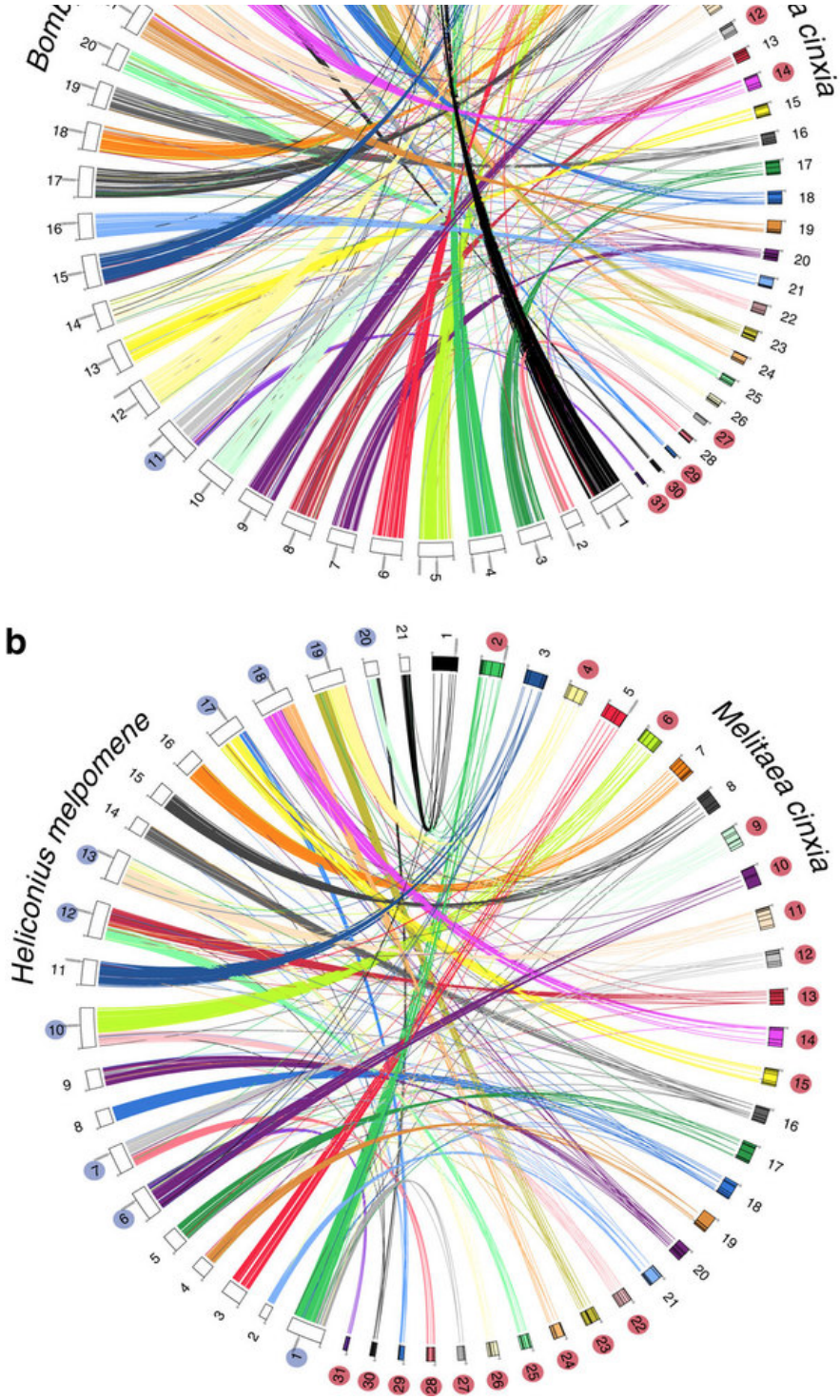
While the GC content varies among Lepidoptera (Supplementary Table 18), the average GC content of the *M. cinxia* genome (33%) is distributed remarkably uniformly across all the chromosomes, similarly to that found in *B. mori* (Supplementary Figs 26, 27, 30 and 31; Supplementary Note 10). The median gene density is 3 per 100 kb in both species (Supplementary Fig. 28). Uniform GC and gene content distributions across the chromosomes are characteristics of species with holocentric chromosomes[33, 34, 35, 36], contrasting with species that have monocentric chromosomes with localized centromeres, in which the genome is compartmentalized to GC-rich and GC-poor regions with higher and lower gene densities[37].

Repetitive elements comprise 28% of the assembled *M. cinxia* genome (Supplementary Tables 15 and 16; Supplementary Note 7). The proportion of repetitive elements fluctuates across the chromosomes from 7 to 42% within 100 kb sliding windows (Supplementary Figs 29–31), but it does not show a clear pattern. The distribution of repeats is strikingly different from that in human and mouse, which have a high repeat content in the pericentromeric and subtelomeric regions[38], but it also differs from holocentric nematodes, in which repeats are enriched in distal chromosome regions[34, 35].

With this study, a whole-genome sequence and a high-resolution linkage map are available for three lepidopteran species, *M. cinxia*, *B. mori*[16, 17] and *H. melpomene*[10, 39]. In interspecific chromosomal comparisons, 4,485 one-to-one orthologous genes with map information were identified between *M. cinxia* and *B. mori*, and 3,869 between *M. cinxia* and *H. melpomene*. The majority (96%) of these orthologues mapped to orthologous chromosomes among the three species (Fig. 2; Supplementary Tables 28 and 29; Supplementary Note 11). The remaining 4%, representing putative translocated genes, were relatively evenly distributed and comprise <6% of the genes on any chromosome (Supplementary Figs 33–35; Supplementary Note 12). Comparison of *M. cinxia* with other lepidopteran species carrying the putative ancestral chromosome number of *n*=31, namely, *Plutella xylostella*[40] and *Biston betularius*[22] (Figs 3a and 4), together with the results of previous studies[21, 22, 23], confirms that overall chromosome synteny has been strikingly well conserved in all 31 chromosomes among distantly related (>140 My)[30] Lepidoptera (Supplementary Tables 30–32; Supplementary Note 11). The phylogenetic range covers almost all Ditrysia and thus represents at least 95% of existing species (Fig. 3a). The distribution of karyotypes on a phylogeny of 312 species in the family Nymphalidae (Fig. 3b; Supplementary Fig. 32; Supplementary Note 11; Supplementary Data 2) further indicates that *n*=31 is unambiguously the ancestral karyotype in this family, although there are some subfamilies (for example, Danainae and Satyrinae) that show much variation in chromosome number even among closely related lineages. Our data argue against the suggestion[7] that repeated fusion and fission events followed by selection would have maintained the *n*=31 karyotype in Lepidoptera (see also Saura *et al.*[6]). Rather, the results indicate that high macrosynteny is a manifestation of the exceptional stability of the ancestral karyotype[18, 22, 23].

Figure 2: Chromosome mapping of *Melitaea cinxia* to *Bombyx mori* and *Heliconius melpomene*.

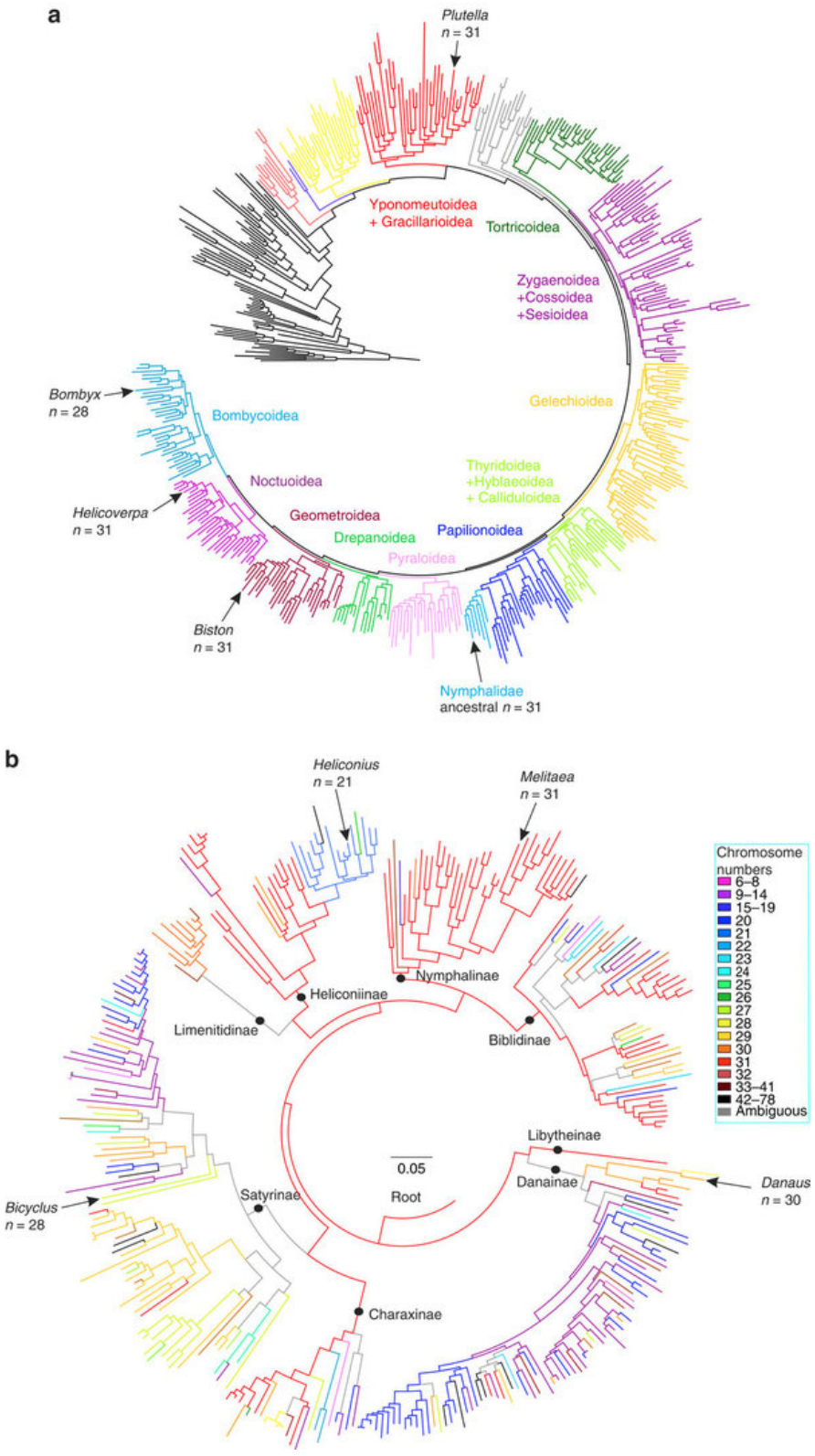




(**a**) One-to-one orthologues (4,485) connecting *M. cinxia* and *B. mori* chromosomes and (**b**) 3,869 one-to-one orthologues connecting *M. cinxia* and *H. melpomene* chromosomes. *M. cinxia* chromosomes are numbered according to chromosome length from the largest to the smallest. The links leading from *M. cinxia* chromosomes are pooled into bins that are ordered within chromosomes. Bands drawn in *M. cinxia* chromosomes represent bin borders. Chromosome 1 is the Z chromosome in *M. cinxia* and *B. mori*. The fusion chromosomes are shaded with blue and the orthologous chromosomes in *M. cinxia* with red.

Figure 3: Haploid chromosome numbers mapped onto the lepidopteran tree of life and a phylogenetic hypothesis for

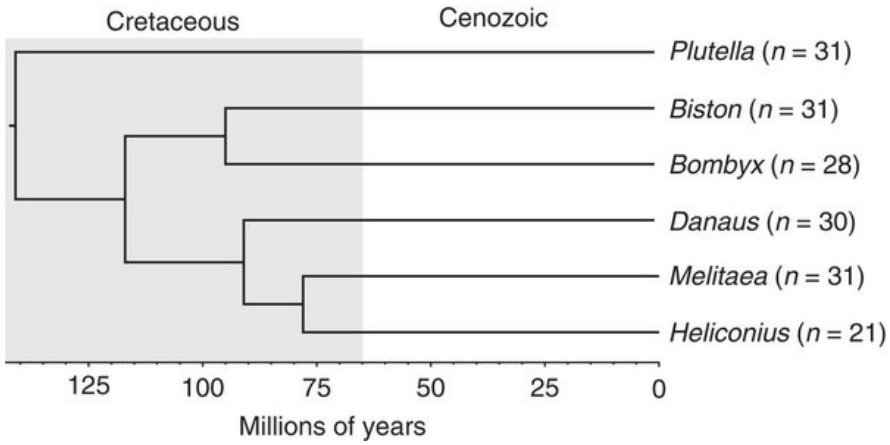
Nymphalidae.



(**a**) The lepidopteran tree of life showing the placement of focal species with their haploid chromosome number (*n*). The named species are those for which whole-genome sequence and linkage map are available (only linkage map for *Biston*). Major clades, often defined as superfamilies, are coloured. In the Papilionoidea clade (the butterflies), the family Nymphalidae is highlighted in light blue, with the putative ancestral chromosome

number of *n*=31 (for justification see panel **b**). The topology is taken from Regier *et al.*[69] (**b**) Haploid chromosome number mapped onto a phylogenetic hypothesis for Nymphalidae. The character state '31' is shown to be the most likely ancestral state for the family. The four species with whole-genome sequences are highlighted. Details of the source of phylogenetic hypothesis as well as chromosome numbers are found in the Supplementary Note 11.

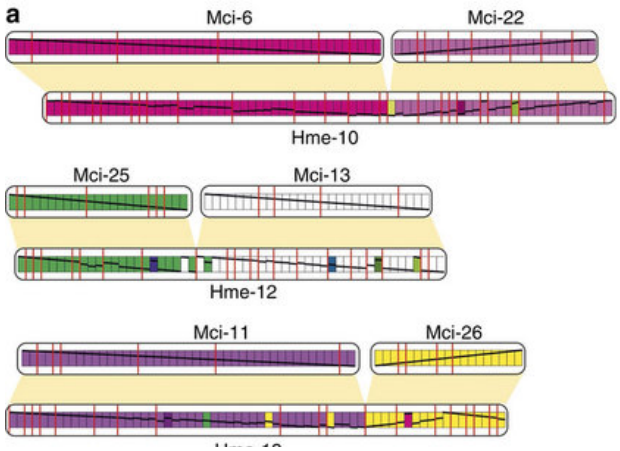
Figure 4: The chronogram of the lepidopteran species for which the whole-genome sequence or linkage map, or both, are available together with haploid chromosome numbers.

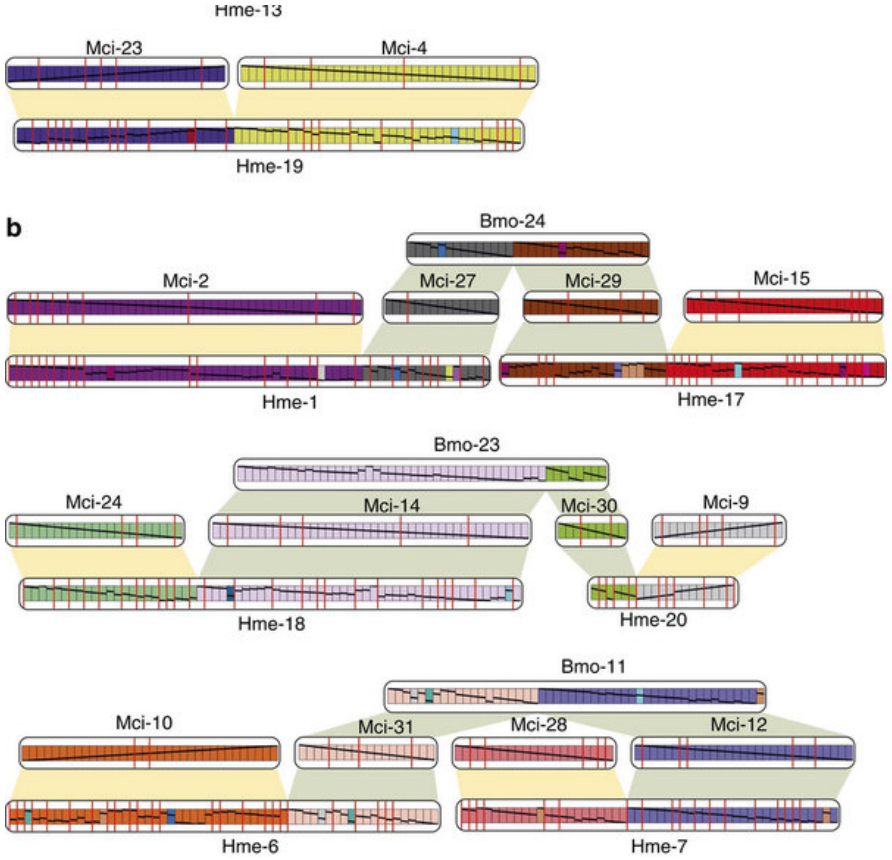


Times of divergence for *Danaus*, *Heliconius* and *Melitaea* are taken from Wahlberg *et al.*[70], the other times of divergence are from Wahlberg *et al.*[30].

The *M. cinxia* genome allows us to identify potential fusion and fission events that have shaped the *B. mori* (*n*=28) and *H. melpomene* (*n*=21) genomes from the ancestral karyotype. Our data confirm 3 fusion events in *B. mori* and 10 fusions in *H. melpomene*[10, 22] (Figs 2 and 5; Supplementary Figs 36 and 37; Supplementary Note 12). A prominent feature of the fusions in both species is the participation of the shortest orthologous *M. cinxia* chromosomes (chrs 29–31 and 22–31 in *B. mori* and *H. melpomene* fusions, respectively; Fig. 2). The bias towards the shortest chromosomes is highly significant, *P*=0.001. Reconstruction of the fusion chromosomes revealed that four of the *H. melpomene* fusions are lineage specific (Fig. 2; Fig. 5a). Surprisingly, the six ancestral chromosomes participating in the fusion events in *B. mori* are also involved in *H. melpomene* fusions, albeit with non-orthologous fusion partners (Fig. 5b), suggesting a preference for a subset of chromosomes to participate in fusion events in evolutionarily distant lineages. The probability of the same six chromosomes being involved in independent fusion events in the two species by chance is low, *P*=0.05. These results suggest that selection favours a subset of possible fusion events, possibly at the level of chromosome segregation or through the hypothetical sequence elements associated with the shortest chromosomes.

Figure 5: *Melitaea cinxia* chromosomes involved in fusion events in *Bombyx mori* and *Heliconius melpomene*.

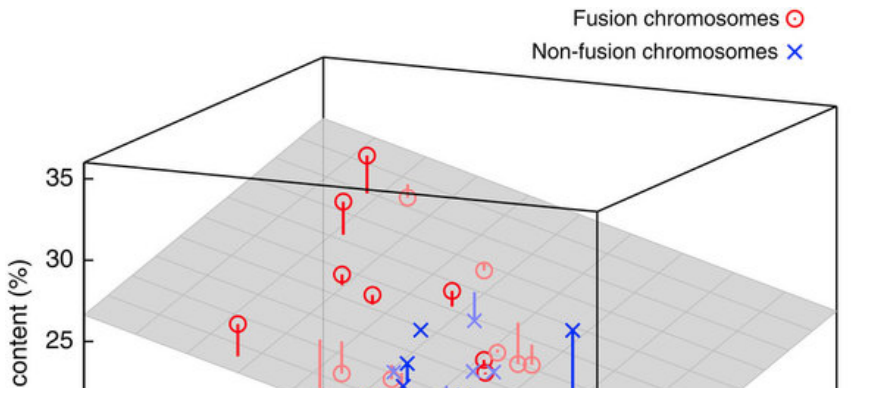


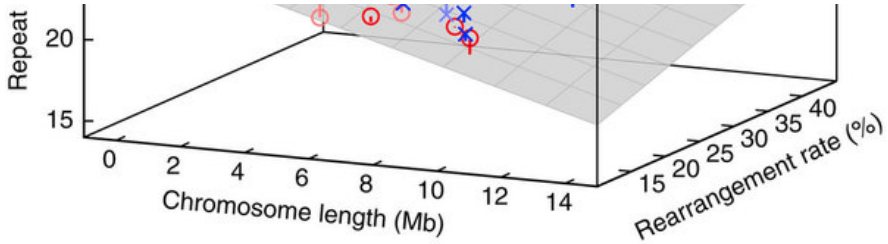


(**a**) Fusions unique to *H. melpomene*. (**b**) Fusions in which the same *M. cinxia* orthologous chromosomes have been fused in both *B. mori* and *H. melpomene*. Each box represents one superscaffold in *M. cinxia* and a scaffold in *H. melpomene*. The colour of each box is derived from the chromosomal origin of the orthologous segment in the *M. cinxia* genome (compare with Supplementary Fig. 36). Horizontal lines within the boxes show the corresponding loci in *M. cinxia* chromosomes, and red vertical lines show recombination sites (bin boundaries) in the linkage map.

A preference for short chromosomes in fusions may be related to a negative relationship between the rate of intrachromosomal rearrangement and chromosome length in *M. cinxia* (Fig. 6; r=−0.48, P=0.007), in which chromosome length is furthermore inversely related to the percentage of repetitive sequence (r=−0.73, P<2.8e−6). These relationships imply that longer chromosomes containing fewer repetitive elements are more stable and have fewer intrachromosomal rearrangements, whereas shorter chromosomes are more prone to elevated inter- and intrachromosomal rearrangement. This result is in agreement with the observation that 7 out of the 11 chromosomes with the strongest indication of introgression between four species of *Heliconius* butterflies [10] are fusion chromosomes.

Figure 6: Relationships among chromosome length, chromosome rearrangement rate and percentage of repetitive elements in the 31 chromosomes of *Melitaea cinxia*.





The rearrangement rate is described by the number of chromosomal breakpoints scaled by the number of orthologues. A plane minimizing squared error was fitted to the data and is shown in grey. Drop lines are drawn from each point to this plane.

The putative fusion regions in *B. mori* show 4% higher abundance of retrotransposons compared with the whole genome (Supplementary Tables 33 and 34; Supplementary Note 12). This may reflect the persistence of retrotransposons at ancient telomere sites, given the role of the LINE and PLE elements in telomere maintenance[41, 42], but also a possible role for retrotransposons in facilitating the fusion process[43]. It is noteworthy that even though the same six orthologous chromosomes have participated in the fusions in *B. mori* and *H. melpomene*, the orientation of the chromosomes is different between the two species in two of the three cases (Fig. 5b; Supplementary Fig. 36). Comparing the *B. mori* and *H. melpomene* fusion chromosomes with the orthologous *M. cinxia* non-fused chromosomes, we observe that once the fusion event took place, there was very little further rearrangement across the fusion site, with a sharp boundary separating the fused chromosomes (Fig. 5; Supplementary Figs 36 and 37). Thus, no orthologous genes have crossed the fusion boundary in *B. mori*, and only 13 such events were detected in *H. melpomene* over 10 chromosomes.

The reference genome of the Glanville fritillary, featuring the ancient karyotype with *n*=31, provides new insight into karyotype evolution in Lepidoptera. Comparisons between *M. cinxia*, *B. mori* and *H. melpomene* suggest that lepidopteran chromosomal fusion events favour the shortest chromosomes, which have high rates of chromosomal evolution and high frequency of transposable elements (TEs). The fusion chromosomes retain the ancestral chromosome segments and gene content with very little rearrangement across sharp fusion boundaries. Features such as conserved gene content in chromosomes and constrained fusions offer practical advantages for further lepidopteran genomic studies, by providing a guide for scaffolding and validation of newly sequenced genomes. These characteristics emphasize the distinctive features of evolutionary dynamics in holocentric chromosomes of butterflies and moths, which appear to evolve in a different manner than most other metazoan genomes.

Methods

Samples and sequencing

Initial sequencing of genomic DNA (Supplementary Note 2) was carried out using a single male larva collected from the Åland Islands, Finland. A male larva was used to avoid sequencing of the female W chromosome known to have high repetitive sequence content[44, 45]. In addition, a single female larva was used for sequencing of mitochondrial DNA (mtDNA). DNA was extracted using modified glass rod spooling methods. The male sample was used for the production of several 454 single-read fragment libraries, and 454 sequencing was carried out according to the manufacturer's instructions, and subsequently used for contig assembly.

In the sequencing of PE and MP libraries (Supplementary Note 2), we used a single male sample and several sets of 10–100 full-sibs to obtain a sufficient yield of high-molecular-weight DNA for library preparation (Supplementary Table 1). PE and MP data were produced using Illumina, SOLiD (ABI) and 454 (Roche) platforms and used for scaffolding (Supplementary Fig. 1). PE data consisted of Illumina libraries with insert size of 500 and 800 bp. The MP data included Illumina and SOLiD libraries with insert sizes of 1, 2, 3 and 5 kb, and 454 libraries with insert sizes of 8 and 16 kb (Supplementary Fig. 3; Supplementary Table 1). DNA for the short-insert (≤1 kb) libraries was extracted from thorax tissues using NucleoSpin Tissue Kits. For the long-insert (>1 kb) MP libraries and PacBio sequencing, DNA was extracted from pools of thorax or abdomen tissues using the CsCl purification method[46], which was modified to increase the yield, integrity and purity of DNA (Supplementary Fig. 2). The Illumina PE libraries were prepared according to Tuupanen *et al.*[47] but using PE adapters and larger size selection, and sequenced with an Illumina Genome Analyzer IIx (500 bp library) or a HiSeq 2000 (800 bp library) following standard PE-sequencing protocols. SOLiD and Illumina MP libraries were produced as described by the manufacturer (SOLiD MP library kit, Life technologies, CA, USA) with in-house modifications, and sequenced using SOLiD 5500XL and HiScan SQ. The 454 MP libraries were constructed by Roche 454 Life Sciences Sequencing Services (Branford, CT, USA) and sequenced with 454 FLX.

Libraries for PacBio sequencing were constructed following the manufacturer's protocols, and run on PacBioRS.

Transcriptome data from RNA-seq experiments were used in gene prediction, functional annotation and variation and linkage disequilibrium (LD) analyses (Supplementary Table 2; Supplementary Note 2). For gene prediction and functional annotation, we used pooled abdomen and mixed tissue samples consisting of head, thorax and larval tissues. For variation analyses, only thorax samples were used. RNA was extracted using the Trizol method (Life Technologies) followed by acid phenol–chloroform–isoamyl alcohol and chloroform extractions. RNA-sequencing libraries for the pooled samples were constructed using the Illumina TruSeq RNA Sample Preparation kit (A) and sequenced with Illumina HiSeq 2000 according to the manufacturer's instructions. For the variation analyses, two RNA-seq libraries were prepared for each individual using an in-house polyA-anchoring-based RNA-seq library protocol (Supplementary Note 2). These libraries were sequenced according to the manufacturer's instructions with Illumina HiSeq 2000 and HighScan SQ sequencers using the PE mode.

Genome assembly

Before the assembly, raw reads were filtered and trimmed as described in Supplementary Note 3. To correct sequencing errors and to eliminate additional variation from heterogenous DNA samples, we used two in-house error-correction methods, Coral[48] for 454 and Illumina reads, and HybridSHREC[49] for SOLiD colour-space reads.

Error-corrected 454 and Illumina PE reads were assembled using Newbler (Roche) and SOAPdenovo[50] (Supplementary Note 3). Contigs with a minimal length of 500 bp were used in scaffolding. For scaffolding, we used in-house MIP Scaffolder software[27], and required at least two read pairs for connecting a pair of contigs. Scaffolding was performed in seven stages in which the PE and MP libraries were added in ascending order of insert size. The most substantial increase in the N50 was observed with the 16 kb 454 MP library (Supplementary Fig. 4). After scaffolding, we used Illumina PE libraries to close the gaps between the contigs using SOAPdenovo GapCloser[50]. Only scaffolds longer than 1,500 bp were included into the final scaffold set. Ribosomal DNA and mtDNA were assembled separately using contigs, which were excluded from genome scaffolding due to their high abundance (Supplementary Note 3). In addition, assembly of 454 reads from a single female was used in the assembly of mtDNA.

To increase the continuity of the genome assembly, we constructed superscaffolds using an in-house-developed method that utilizes the linkage map, MP and PacBio data (Supplementary Note 3). First, MP reads and PacBio reads were aligned against existing scaffolds using BWA[51] and an in-house SANS aligner[52], respectively. After subsequent filtering, the linkage map was used as a guide to determine the most reliable path between the scaffolds to yield individual superscaffolds.

Linkage map

The linkage map for *M. cinxia* has been described by Rastas *et al.*[26] and in Supplementary Note 5. For the purpose of validating the superscaffolds, another linkage map was constructed by a similar procedure as in Rastas *et al.*[26] using 12,109 SNPs from an independent full-sib family with 19 offspring.

Genome validation

Correctness and consistency of the genomic assembly were assessed using eight approaches (Supplementary Table 7; Supplementary Note 6). (1) Correctness of the contigs was assessed by mapping PE and MP reads to the genome, and calculating the concordant mappings. (2) Correctness of the scaffolds was evaluated by re-scaffolding the contigs using PacBio reads and calculating the contig joins concordant with the scaffolds. (3) Consistency of the scaffolds was estimated by counting non-chimeric scaffolds based on the linkage map. (4) Completeness of the genome was evaluated by aligning assembled transcripts to scaffolds and calculating the proportion of aligned transcripts. (5) Completeness was further assessed by estimating the proportion of conserved core genes found in the genome and (6) the level of sequence synteny among other lepidopteran species. (7) Correctness and consistency of the superscaffolds were assessed by comparing gene order against *B. mori* within superscaffolds and (8) by estimating the proportion of non-chimeric scaffolds using an independent linkage map.

Prediction of repetitive elements

M. cinxia-specific TEs were predicted *de novo* as described in Supplementary Note 7. Long terminal repeat retrotransposons were searched using LTR_Finder[53]. The predicted TEs of *M. cinxia* were combined with the RebBase (v. 20120418) library of consensus TEs from different species[54]. The consensus sequences of TE families were collected as a *M. cinxia* repeat library and annotated using

Repbase18.05, RepeatPeps[54] and Dfam 1.1 (ref. 55) databases. RepeatMasker-open-4-0-2 (Smit, A.F.A., Hubley, R. and Green, P. RepeatMasker at http://repeatmasker.org) was used to estimate the distribution of TEs and other interspersed repeat elements in the genome.

Gene model prediction and functional annotation

Gene models were predicted for the repeat-masked genome using an evidence-based approach in MAKER[56], which combines *ab initio* modelling with RNA-seq and protein sequence evidence (Supplementary Note 8). *Ab initio* gene prediction was performed with SNAP[57]. Protein data consisted of all Arthropoda proteins from UniProtKB (UniProt release 2012_02) and whole proteomes of four species from Ensembl and two unpublished proteomes. As RNA-seq data, we used *de novo*-assembled transcripts[58] and TopHat/Cufflinks[59] mappings (Supplementary Note 4). Noncoding and mtDNA genes were predicted as described in Supplementary Note 8.

Functional descriptions, gene ontologies and enzyme commission numbers were predicted for the protein sequences translated from the gene models and from the assembled transcripts using an in-house PANNZER annotation pipeline[60] (Supplementary Fig. 14; Supplementary Note 8). Protein domains and other functional elements were detected and annotated using InterProScan[61]. Metabolic pathways and KEGG orthologues were predicted using the KAAS server[62]. Gene orthologies were predicted for the 5 lepidopteran species for which genome sequence information is available, 15 other arthropoda and 2 outgroups using an in-house EPT method[63] (Supplementary Fig. 17).

Variation analyses

SNPs and indels were detected from four data sets as described in Supplementary Note 9. The variation statistics described in the main paper are based on Illumina PE reads from a genomic pool of 10 full-sibs, which were also used in the genome assembly. The reads were mapped to the genome using BWA[51], and variants were detected using a GATK pipeline[64]. Long indels were detected using a PacBio genomic pool from 100 individuals (Supplementary Table 1). PacBio reads were mapped onto genomic scaffolds with BWA-SW[51] and indels exceeding 50 bp were detected. Linkage disequilibrium (r^2) was estimated from the Illumina RNA-seq data for the population in the Åland Islands (Finland) using an in-house script (Supplementary Fig. 25; Supplementary Note 9).

Phylogenetic analyses

The phylogenetic analyses were based on 312 species of the family Nymphalidae for which chromosome number and DNA sequences of 3–11 genes were available (Supplementary Note 11). DNA sequences were manually aligned, and a phylogenetic hypothesis was inferred in the maximum likelihood framework using RAxML[65]. The haploid chromosome numbers were mapped onto the tree using Mesquite (http://mesquiteproject.org).

Genome scans and synteny analyses

GC, gene and repeat contents were calculated within 100 kb sliding windows and 10 kb shifts for the superscaffolds of *M. cinxia* and the genome sequence of *B. mori*[16] (Supplementary Note 10). Since the superscaffolds were not ordered within bins in the current linkage map, the order and orientation of the superscaffolds within each bin were determined based on synteny to *B. mori*.

Chromosome mapping was carried out using orthologous genes between *M. cinxia* and *B. mori* and between *M. cinxia* and *H. melpomene* to define the level of gene conservation and translocations among chromosomes (Supplementary Notes 11 and 12). Furthermore, the mapped genes were used for the identification of fusion chromosomes in *B. mori* and *H. melpomene*. The same data set was used for calculating the number of breakpoints in the chromosomes, which were scaled by the number of one-to-one orthologues in the chromosomes, and used to measure the intrachromosomal rearrangement rate. Pairwise correlations between rearrangement rate, repeat content and chromosome lengths were calculated using the Pearson correlation coefficient. Chromosome mappings (Fig. 2) are illustrated using Circos[66].

Possible bias in the ancestral chromosomes that are involved in fusion events in *B. mori* and *H. melpomene* was measured as follows. First, the probability for the same six ancestral chromosomes to be involved in independent fusions in both species was calculated as

$$P=\frac{\binom{6}{6}\times\binom{31-6}{20-6}}{\binom{31}{20}}=0.053. \qquad (1)$$

We assumed that *B. mori* has 6 and *H. melpomene* 20 fusion chromosomes, and each ancestral chromosome fused only once. Second, we measured the bias towards small ancestral chromosomes being involved in these fusions. The ancestral (*M. cinxia*) chromosomes were ranked according to chromosome number, which reflects the length (*M. cinxia* chromosomes are numbered from the largest to the smallest). The median rank is 28 for *B. mori* and 18.5 for *H. melpomene*. The probabilities of obtaining at least as large medians by chance are 0.00092 and 0.14, respectively. The former is the probability of obtaining either chromosomes 28–31, or 27 and 29–31 from randomly chosen 6 chromosomes (out of 31), thus

$$P=\frac{\binom{27}{2}+\binom{26}{2}}{\binom{31}{6}}. \qquad (2)$$

The latter probability was computed by simulating random draws of 20 chromosomes (out of 31). These *P* values were combined by Fisher's method[67] to obtain the single *P* value of 0.0013.

The approximate fusion sites were detected by aligning the fusion chromosomes of *B. mori* (11, 23 and 24) against the orthologous chromosomes of *M. cinxia* (12+31, 14+30 and 27+29) using Mauve[68] (Supplementary Note 12). The content of TEs within the potential fusion regions was compared with the genome-wide content using RepeatMasker-open-4-0-2 with *B. mori*- and *H. melpomene*-specific repeat libraries[10, 16]. Chromosome fusions in Fig. 5 and Supplementary Fig. 36 are visualized using in-house scripts.

The annotated genome will be included in the EnsemblMetazoa http://metazoa.ensembl.org/Melitaea_cinxia/Info/Index. Further information, including superscaffolds, linkage map and annotations are available through our website at http://www.helsinki.fi/science/metapop/research/mcgenome.html.

Additional information

Accession codes: The genome sequence of the Glanville fritillary butterfly, *Melitaea cinxia,* has been deposited in DDBJ/EMBL/GenBank nucleotide core database under the accession code APLT00000000.

How to cite this article: Ahola, V. *et al.* The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera. *Nat. Commun.* 5:4737 doi: 10.1038/ncomms5737 (2014).

Accession codes

Referenced accessions

GenBank/EMBL/DDBJ

APLT00000000

References

1. Suomalainen, E. Achiasmatische oogenese bei Trichopteren. *Chromosoma* **18**, 201–207 (1966).
2. Murakami, A. & Imai, H. T. Cytological evidence for holocentric chromosomes of the silkworms, *Bombyx mori* and *B. mandarina*, (Bombycidae, Lepidoptera). *Chromosoma* **47**, 167–178 (1974).
3. Wolf, K. W. The structure of condensed chromosomes in mitosis and meiosis of insects. *Int. J. Insect Morphol.* **25**, 37–62 (1996).
4. Brown, K. S. Jr, Von Schoultz, B. & Suomalainen, E. Chromosome evolution in Neotropical Danainae and Ithomiinae (Lepidoptera). *Hereditas* **141**, 216–236 (2004).
5. Kandul, N. P., Lukhtanov, V. A. & Pierce, N. E. Karyotypic diversity and speciation in *Agrodiaetus* butterflies. *Evolution* **61**, 546–559 (2007).
6. Saura, A., Von Schoultz, B., Saura, A. O. & Brown, K. S. Jr Chromosome evolution in Neotropical butterflies. *Hereditas* **150**, 26–37 (2013).
7. White, M. J. D. *Animal Cytology and Evolution* 3rd edn Cambridge University Press (1973).

8. Robinson, R. *Lepidoptera Genetics* Pergamon Press (1971).

9. Suomalainen, E. Chromosome evolution in the Lepidoptera. *Chrom. Today* **2**, 132–138 (1969).

10. The Heliconius Genome Consortium. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* **487**, 94–98 (2012).

11. Maddox, P. S., Oegema, K., Desai, A. & Cheeseman, I. M. "Holo"er than thou: Chromosome segregation and kinetochore function in *C. elegans*. *Chromosome Res.* **12**, 641–653 (2004).

12. Marec, F., Sahara, K. & Traut, W. Rise and fall of the W chromosome in Lepidoptera. In:*Molecular Biology and Genetics of the Lepidoptera* eds Goldsmith M. R., Marec F. CRC Press Taylor & Francis Group LLC (2010).

13. Hipp, A. L., Rothrock, P. E., Whitkus, R. & Weber, J. A. Chromosomes tell half of the story: the correlation between karyotype rearrangements and genetic diversity in sedges, a group with holocentric chromosomes. *Mol. Ecol.* **19**, 3124–3138 (2010).

14. Baker, R. J. & Bickham, J. W. Speciation by monobrachial centric fusions. *Proc. Natl Acad. Sci. USA* **83**, 8245–8248 (1986).

15. Faria, R. & Navarro, A. Chromosomal speciation revisited: rearranging theory with pieces of evidence. *Trends Ecol. Evol.* **25**, 660–669 (2010).

16. The International Silkworm Genome Consortium. The genome of a lepidopteran model insect, the silkworm *Bombyx mori*. *Insect Biochem. Mol. Biol.* **38**, 1036–1045 (2008).

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

18. Yasukochi, Y. *et al.* Extensive conserved synteny of genes between the karyotypes of *Manducta sexta* and *Bombyx mori* revealed by BAC-FISH mapping. *PLoS ONE* **4**, e7465 (2009).

19. Beldade, P., Saenko, S. V., Pul, N. & Long, A. D. A gene-based linkage map for *Bicyclus anynana* butterflies allows for a comprehensive analysis of synteny with the lepidopteran reference genome. *PLoS Genet.* **5**, e1000366 (2009).

20. d'Alencon, E. *et al.* Extensive synteny conservation of holocentric chromosomes in Lepidoptera despite high rates of local genome rearrangements. *Proc. Natl Acad. Sci. USA* **107**, 7680–7685 (2010).

21. Baxter, S. W. *et al.* Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism. *PLoS ONE* **6**, e19315 (2011).

22. Van't Hof, A. E. *et al.* Linkage map of the peppered moth, *Biston betularia* (Lepidoptera, Geometridae): a model of industrial melanism. *Heredity* **110**, 283–295 (2013).

23. Sahara, K. *et al.* FISH identification of *Helicoverpa armigera* and *Mamestra brassicae* chromosomes by BAC and fosmid probes. *Insect Biochem. Mol. Biol.* **43**, 644–653 (2013).

24. Sichová, J., Nguyen, P., Dalíková, M. & Marec, F. Chromosomal evolution in Tortricid moths: conserved karyotypes with diverged features. *PLoS ONE* **8**, e64520 (2013).

25. Pringle, E. G. *et al.* Synteny and chromosome evolution in the Lepidoptera: evidence from mapping in *Heliconius melpomene*. *Genetics* **177**, 417–426 (2007).

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

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

28. Federley, H. Chromosomenzahlen finnländischer lepidopteren. I. Rhopalocera. *Hereditas* **24**, 397–464 (1938).

29. Suetsugu, Y. *et al.* Large scale full-length cDNA sequencing reveals a unique genomic landscape in a lepidopteran model insect, *Bombyx mori*. *G3 (Bethesda)* **3**, 1481–1492 (2013).

30. Wahlberg, N., Wheat, C. W. & Pena, C. Timing and patterns in the taxonomic diversification of Lepidoptera (butterflies and moths).

PLoS ONE **8**, e80875 (2013).

31. Chai, C. L. *et al.* A genomewide survey of homeobox genes and identification of novel structure of the hox cluster in the silkworm, *Bombyx mori*. *Insect Biochem. Mol. Biol.* **38**, 1111–1120 (2008).
32. Somervuo, P. *et al.* Transcriptome analysis reveals signature of adaptation to landscape fragmentation. *PLoS ONE* **9**, e101467 (2014).
33. Mandrioli, M. & Manicardi, G. C. Unlocking holocentric chromosomes: new perspectives from comparative and functional genomics? *Curr. Genomics* **13**, 343–349 (2012).
34. Stein, L. D. *et al.* The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. *PLoS Biol.* **1**, e45 (2003).
35. The C. elegans Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012–2018 (1998).
36. Grbic, M. *et al.* The genome of *Tetranychus urticae* reveals herbivorous pest adaptations. *Nature* **479**, 487–492 (2011).
37. Bernardi, G. *et al.* The mosaic genome of warm-blooded vertebrates. *Science* **228**, 953–958 (1985).
38. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
39. Martin, S. H. *et al.* Genome-wide evidence for speciation with gene flow in *Heliconius butterflies*. *Genome Res.* **23**, 1817–1828 (2013).
40. You, M. S. *et al.* A heterozygous moth genome provides insights into herbivory and detoxification. *Nat. Genet.* **45**, 220–225 (2013).
41. Fujiwara, H., Osanai, M., Matsumoto, T. & Kojima, K. K. Telomere-specific non-LTR retrotransposons and telomere maintenance in the silkworm, *Bombyx mori*. *Chromosome Res.* **13**, 455–467 (2005).
42. Gladyshev, E. A. & Arkhipova, I. R. Telomere-associated endonuclease-deficient penelope-like retroelements in diverse eukaryotes. *Proc. Natl Acad. Sci. USA* **104**, 9352–9357 (2007).
43. Tanaka, H. *et al.* Telomere fusions in early human breast carcinoma. *Proc. Natl Acad. Sci. USA* **109**, 14098–14103 (2012).
44. Sahara, K., Marec, F., Eickhoff, U. & Traut, W. Moth sex chromatin probed by comparative genomic hybridization (CGH). *Genome* **46**, 339–342 (2003).
45. Abe, H., Mita, K., Yasukochi, Y., Oshiki, T. & Shimada, T. Retrotransposable elements on the W chromosome of the silkworm, *Bombyx mori*. *Cytogenet. Genome Res.* **110**, 144–151 (2005).
46. Smoller, D. A., Petrov, D. & Hartl, D. L. Characterization of bacteriophage P1 library containing inserts of *Drosophila* DNA of 75-100 kilobase pairs. *Chromosoma* **100**, 487–494 (1991).
47. Tuupanen, S. *et al.* The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. *Nat. Genet.* **41**, 885–890 (2009).
48. Salmela, L. & Schröder, J. Correcting errors in short reads by multiple alignments. *Bioinformatics* **27**, 1455–1461 (2011).
49. Salmela, L. Correction of sequencing errors in a mixed set of reads. *Bioinformatics* **26**, 1284–1290 (2010).
50. Li, R. *et al. De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265–272 (2010).
51. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
52. Koskinen, J. P. & Holm, L. SANS: high-throughput retrieval of protein sequences allowing 50% mismatches. *Bioinformatics* **28**, i438–i443 (2012).
53. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).

54. Jurka, J. *et al.* Repbase update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).

55. Wheeler, T. J. *et al.* Dfam: a database of repetitive DNA based on profile hidden Markov models. *Nucleic Acids Res.* **41**, D70–D82 (2013).

56. Cantarel, B. L. *et al.* MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18**, 188–196 (2008).

57. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59–67 (2004).

58. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).

59. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).

60. Radivojac, P. *et al.* A large-scale evaluation of computational protein function prediction. *Nat. Methods* **10**, 221–227 (2013).

61. Zdobnov, E. M. & Apweiler, R. InterProScan--an integration platform for the signature-recognition methods in interpro. *Bioinformatics* **17**, 847–848 (2001).

62. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C. & Kanehisa, M. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* **35**, W182–W185 (2007).

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

64. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).

65. Stamatakis, A., Hoover, P. & Rougemont, J. A rapid bootstrap algorithm for the RAxML Web servers. *Syst. Biol.* **57**, 758–771 (2008).

66. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).

67. Mosteller, F. & Fisher, R. A. Questions and answers. *Am. Stat.* **2**, 30–31 (1948).

68. Darling, A. E., Mau, B. & Perna, N. T. ProgressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE* **5**, e11147 (2010).

69. Regier, J. C. *et al.* A large-scale, higher-level, molecular phylogenetic study of the insect order Lepidoptera (moths and butterflies). *PLoS ONE* **8**, e58568 (2013).

70. Wahlberg, N. *et al.* Nymphalid butterflies diversify following near demise at the Cretaceous/Tertiary boundary. *Proc. Biol. Sci.* **276**, 4295–4302 (2009).

Download references

Acknowledgements

We thank Tapio Heino, Frantisek Marec and Anssi Saura for comments on the manuscript; Eeva-Marja Turkki, Kirsi Lipponen, Matias Rantanen and Harri Kangas for constructing libraries and performing sequencing; Marja-Leena Peltonen for DNA isolation; and Annukka Ruokolainen, Toshka Nyman and Suvi Saarnio for sample management, DNA and RNA isolation and sequencing library preparation. Pia Välitalo and Alison Ollikainen are thanked for sample management; Jussi Nokso-Koivisto and Petri Törönen for help in bioinformatics analyses; Kimmo Mattila for help in using supercomputers; and the IT Center for Science Ltd (CSC), Espoo, Finland, the Institute for Molecular Medicine Finland (FIMM), Helsinki, Finland, and the European Bioinformatics Institute for supercomputing facilities. K.Q. was supported by the Viikki Doctoral Programme in Molecular Biosciences (VGSB) at the University of Helsinki, and R.M.W. by Marie Curie International Outgoing Fellowship PIOF-GA-2011-303312. This research was supported by grants from the Finnish National Research Council (141500, 250444 and 256453 to I.H.; 133132 to I.H., L.H., M.J.F., P.A. and R.L.; 118653 and 267591 to L.S.; and 135291 to M.J.F.) and the European Research Council (232826 to I.H.).

Author information

These authors contributed equally to this work
Virpi Ahola, Rainer Lehtonen & Panu Somervuo

Affiliations

Department of Biosciences, University of Helsinki, FI-00014 Helsinki, Finland
Virpi Ahola, Rainer Lehtonen, Panu Somervuo, Patrik Koskinen, Pasi Rastas, Maaike A. de Jong, Anne Duplouy, Wong Swee Chong, Jarkko Salojärvi, Sami P. Ojanen, Liisa Holm & Ilkka Hanski

Genome-Scale Biology Research Program, University of Helsinki, FI-00014 Helsinki, Finland
Rainer Lehtonen, Niko Välimäki, Riku Katainen, Esa Pitkänen, Mikko Turunen, Anna Vähärautio & Minna Taipale

Institute of Biomedicine, University of Helsinki, FI-00014 Helsinki, Finland
Rainer Lehtonen, Niko Välimäki, Riku Katainen & Esa Pitkänen

Center of Excellence in Cancer Genetics, University of Helsinki, FI-00014 Helsinki, Finland
Rainer Lehtonen

Institute of Biotechnology, University of Helsinki, FI-00014 Helsinki, Finland
Panu Somervuo, Patrik Koskinen, Lars Paulin, Jouni Kvist, Jaakko Tanskanen, Olli-Pekka Smolander, Kui Qian, Alan H. Schulman, Liisa Holm, Petri Auvinen & Mikko J. Frilander

Department of Computer Science & Helsinki Institute for Information Technology HIIT, University of Helsinki, FI-00014 Helsinki, Finland
Leena Salmela, Johannes Ylinen, Esko Ukkonen & Veli Mäkinen

Department of Biology, University of Turku, FI-20014 Turku, Finland
Niklas Wahlberg

Biotechnology and Food Research, MTT Agrifood Research Finland, FI-31600 Jokioinen, Finland
Jaakko Tanskanen & Alan H. Schulman

Department of Zoology, University of Cambridge, Cambridge CB2 3EJ, UK
Emily A. Hornett

Department of Biology, Pennsylvania State University, Pennsylvania 16802, USA
Emily A. Hornett

Department of Zoology, University of Oxford, Oxford OX1 3PS, UK
Laura C. Ferguson

College of Life Sciences, Peking University, Beijing 100871, P.R. China
Shiqi Luo & Zijuan Cao

School of Biological Sciences, University of Bristol, Bristol BS8 1UG, UK
Maaike A. de Jong

Department of Entomology, Max Planck Institute for Chemical Ecology, D-07745 Jena, Germany
Heiko Vogel

Department of Biology, Stanford University, Stanford, California 94305, USA
Rajiv C. McCoy

BioMediTech, University of Tampere, FI-33520 Tampere, Finland
Qin Zhang, Jani K. Haukka & Aruj Joshi

Department of Information Technology, University of Turku, FI-20014 Turku, Finland
Freed Ahmad

Department of Zoology, Stockholm University, SE-10691 Stockholm, Sweden
Christopher W. Wheat

Department of Evolutionary Neuroethology, Max Planck Institute for Chemical Ecology, D-07745 Jena, Germany
Ewald Grosse-Wilde

European Bioinformatics Institute, Hinxton CB10 1SD, UK
Daniel Hughes & Daniel Lawson

Baylor College of Medicine, Human Genome Sequencing Center, Houston, Texas 77030-3411, USA
Daniel Hughes

Department of Genetic Medicine and Development, University of Geneva Medical School & Swiss Institute of Bioinformatics, 1211 Geneva, Switzerland
Robert M. Waterhouse

Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
Robert M. Waterhouse

The Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA
Robert M. Waterhouse

Department of Pathology, University of Helsinki, FI-00014 Helsinki, Finland
Anna Vähärautio

Science for Life Laboratory, Department of Biosciences and Nutrition, Karolinska Institutet, SE-14183 Stockholm, Sweden
Anna Vähärautio & Minna Taipale

Department of Biological Sciences, University of Rhode Island, Kingston, Rhode Island 02881–0816, USA
Marian R. Goldsmith

These authors jointly supervised this work
Rainer Lehtonen, Liisa Holm, Petri Auvinen & Mikko J. Frilander

Contributions

R.L. coordinated the project. R.L., P.A., L.P. and M.J.F. designed the strategy for genome, transcriptome and RAD-tag sequencing and supervised the laboratory work. J.K. and M.J.F. prepared samples for genome sequencing. P.A. and L.P. developed in-house library construction methods. M. Turunen prepared Illumina PE libraries and A.V. full-length transcriptome libraries. P.A., L.P. and M. Taipale coordinated DNA and RNA sequencing. P.S. and L.S. processed sequence data. L.S., V.M., N.V., J.Y. and E.U. developed an in-house method for genome scaffolding. P.S. and L.S. assembled and scaffolded the genome. L.S. and E.A.H. assembled the mtDNA and L.S. the rDNA sequences. L.S. developed the superscaffolding method and implemented it on the genomic scaffolds. L.S., J.Y. and V.M. developed assembly validation software. L.S., P.R., P.S., N.V., P.K., J.Y. and V.M. contributed to assembly validation. P.R. developed the method for linkage map analyses and constructed linkage maps and LD analyses. V.A. and J.T. carried out *de novo* TE prediction. J.T. and A.H.S. annotated TE families, constructed the TE library and carried out repeat predictions. V.A. and D.H. predicted gene models. D.L. supervised gene prediction and functional annotation. P.K., M.J.F. and K.Q. predicted and annotated ncRNA genes. L.H. and P.K.

developed methods for functional annotation and orthologue prediction, and performed orthology analyses. P.K. performed functional annotation. R.M.W. performed OrthoDB orthology prediction. V.A. coordinated manual annotation and performed genome scans. V.A., S.L., Z.C., A.D., O.-P.S., M.A.d.J., H.V., R.C.M., L.C.F., E.A.H., W.S.C., J.K., P.S., P.R., Q.Z., L.H., F.A., J.K.H., A.J., J.S., C.W.W. and E.G.W. participated in manual annotation. E.A.H. annotated and performed analyses for mtDNA genes and L.C.F. for Hox cluster genes. V.A., P.K., N.V., L.S. and P.R. performed synteny analyses. V.A., M.J.F. and L.S. performed analyses of fusion chromosomes. N.W. carried out phylogenetic analyses. P.S., R.K. and E.P. detected SNP and indel variants. V.A. and P.S. performed variation analyses. P.S., P.K., V.A., L.S., W.S.C. and L.H. conducted various sequence analyses. V.A. coordinated writing of the Supplementary information. V.A., L.S., P.S., P.K., P.R., M.J.F., L.H., P.A., R.L., E.A.H., L.C.F., N.W., S.P.O., J.K., A.H.S., J.T., L.P., M. Taipale and K.Q. participated in writing the Supplementary information. V.A., M.J.F., I.H., P.A., R.L., L.H. and M.R.G. wrote the manuscript. All authors read and commented on the manuscript.

Competing financial interests

The authors declare no competing financial interests.

Corresponding author

Correspondence to: Ilkka Hanski

Supplementary information

PDF files

1. Supplementary Information (6,087 KB)
   Supplementary Figures 1-37, Supplementary Tables 1-34, Supplementary Notes 1-12 and Supplementary References

Excel files

1. Supplementary Data 1 (49 KB)
   List of manually annotated genes.

2. Supplementary Data 2 (45 KB)
   Thirty-five species for which haploid chromosome numbers and DNA sequence data were available. GenBank accession numbers are given for sequences used to infer the phylogenetic hypothesis. Outgroups were only used to root the tree and were not considered further for the analysis of chromosome number evolution.

Nature Communications ISSN (online) 2041-1723