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.

NATURE COMMUNICATIONS | 5:4737 | DOI: 10.1038/ncomms5737 | www.nature.com/naturecommunications

© 2014 Macmillan Publishers Limited. All rights reserved.
**Butterflies and moths (Lepidoptera) have a large number of short holocentric chromosomes with substantial variation in chromosome number \((n = 5–223)\). 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\(^2\). Lepidopteran karyotypes are thought to have evolved via fusion and fission events\(^1\). 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 and recombination suppression mechanisms\(^{14}\).**

**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 reads.

**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 = *Nasoxa vitripennis*, Trica = *Tricola castaneum*, Acypi = *Acrhisopsis 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.
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 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\(^10,31\).

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 \textit{M. cinxia} using extensive RNA-seq data\(^32\).

While the GC content varies among Lepidoptera (Supplementary Table 18), the average GC content of the \textit{M. cinxia} genome (33\%) is distributed remarkably uniformly across all the chromosomes, similarly to that found in \textit{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\(^33–36\), 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 \textit{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 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, \textit{M. cinxia}, \textit{B. mori}\(^16,17\) and \textit{H. melpomene}\(^10,39\). In interspecific chromosomal comparisons, 4,485 one-to-one orthologues genes with map information were identified between \textit{M. cinxia} and \textit{B. mori}, and 3,869 between \textit{M. cinxia} and \textit{H. melpomene}. The majority (96\%) of these orthologues mapped to orthologous chromosomes among the three species (Fig. 2; Supplementary Tables 28 and 29; Supplementary

Figure 2 | Chromosome mapping of Melitaea cinxia to Bombyx mori and Heliconius melpomene. (a) One-to-one orthologues (4,485) connecting \textit{M. cinxia} and \textit{B. mori} chromosomes and (b) 3,869 one-to-one orthologues connecting \textit{M. cinxia} and \textit{H. melpomene} chromosomes. \textit{M. cinxia} chromosomes are numbered according to chromosome length from the largest to the smallest. The links leading from \textit{M. cinxia} chromosomes are pooled into bins that are ordered within chromosomes. Bands drawn in \textit{M. cinxia} chromosomes represent bin borders. Chromosome 1 is the Z chromosome in \textit{M. cinxia} and \textit{B. mori}. The fusion chromosomes are shaded with blue and the orthologous chromosomes in \textit{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.
Our data confirm 3 fusion events in *H. melpomene* (*H. melpomene* number of lepidopteran species carrying the putative ancestral chromosome *B. mori* et al. closely related lineages. Our data argue against the suggestion that show much variation in chromosome number even among unambiguously the ancestral karyotype in this family, although divergence are from Wahlberg previous studies 21–23, confirms that overall chromosome *M. cinxia* Supplementary Note 12). Comparison of the genes on any chromosome (Supplementary Figs 33–35; *Heliconius* and *Nymphalidae* (Fig. 3b; Supplementary Fig. 32; Supplementary Note 11). The remaining 4%, representing putative translocated chromosomes (chrs 29–31 and 22–31 in *H. melpomene* fusions, albeit with non-orthologous fusion partners (Fig. 5b), are also involved in *H. melpomene* fusion events in the ancestral karyotype18,22,23. The *M. cinxia* genome allows us to identify potential fusion events in *B. mori* (28) and *H. melpomene* (21) genomes from the ancestral karyotype. Our data confirm 3 fusion events in *B. mori* and 10 fusions in *H. melpomene*, 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 content15,32. 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\(^4\), which was modified to increase the yield, integrity and purity of DNA (Supplementary Fig. 2). The Illumina PE libraries were prepared according to Tuupanen \textit{et al.}\(^4\) 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.

**Figure 5 | Melitaea cinxia chromosomes involved in fusion events in Bombyx mori and Heliconius melpomene.** (a) Fusions unique to \textit{H. melpomene}. (b) Fusions in which the same \textit{M. cinxia} orthologous chromosomes have been fused in both \textit{B. mori} and \textit{H. melpomene}. Each box represents one superscaffold in \textit{M. cinxia} and a scaffold in \textit{H. melpomene}. The colour of each box is derived from the chromosomal origin of the orthologous segment in the \textit{M. cinxia} genome (compare with Supplementary Fig. 36). Horizontal lines within the boxes show the corresponding loci in \textit{M. cinxia} chromosomes, and red vertical lines show recombination sites (bin boundaries) in the linkage map.
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.

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, Coral46 for 454 and Illumina reads, and HybridSHRECT49 for SOLID colour-space reads.

Error correction. 16kb Illumina PE reads were assembled using Newbler (Roche) and SOAPdenovo50 (Supplementary Note 3). Contigs with a minimal length of 500 bp were used in scaffolding. For scaffolding, we used in-house MIP Scaffold software27, 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 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 reads to close the gaps between the contigs using SOAPdenovo GapCloser50. 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 BWA51 and an in-house SANS aligner52, 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 assessed 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 the contig joins concordant with the scaffolds. (8) Consistency of the scaffolds was evaluated by re-scaffolding the contigs using PacBio reads and calculating the proportion of non-chimeric scaffolds using an independent linkage map.

Prediction of repetitive elements. M. cinxia-specific TE families were predicted de novo as described in Supplementary Note 7. Long terminal repeat retrotransposons were searched using LTR_Finder53. The predicted TEs of M. cinxia were combined with the Rebase (v. 20120418) library of consensus TEs from different species24. The consensus sequences of TE families were collected as a M. cinxia repeat library and annotated using RepeatMasker v.18.05. RepeatPeps54 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 MAKER56, which combines ab initio modelling with RNA-seq and protein sequence evidence (Supplementary Note 8). Ab initio gene prediction was performed with SNAP57. Protein data consisted of all Arthropoda proteins from UniProtKB (UniProt release 2012_02) and all proteins of four species from Ensembl and two unpublished proteomes. As RNA-seq data, we used de novo-assembled transcripts58 and TopHat/Cufflinks59 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 predicted pathways by mapping the contigs to the genome using an in-house PANNZER annotation pipeline60 (Supplementary Fig. 14; Supplementary Note 8). Protein domains and other functional elements were detected and annotated using InterProScan61. Metabolic pathways and KEGG orthologues were predicted using the KAAS server26. Genetic orthologies were predicted for the 5 lepidopteran species for which genome sequences information is available, 15 other arthropoda and 2 outgroups using an in-house EPT method63 (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 BWA51, and variants were detected using a GATK pipeline64. Long indels were detected using a PacBio genomic pool from 100 individuals (Supplementary Table 1). PacBio reads were mapped onto genomic scaffolds with BWA-SW51 and indels exceeding 50 bp were detected. Linkage disequilibrium (r2) 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 RAxML65. 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 scaffolds of M. cinxia and the genome sequence of B. mori16 (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 chromosome orthologues (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 both the orthologous and the paralogous species, 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:

$$P = \frac{6 \times 31 - 6}{31 \times 20 - 6} = 0.053.$$

We assumed that B. mori has 6 and H. melpomene 2 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 their 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

\[ P = \frac{\binom{27}{2} + \binom{26}{2}}{\binom{31}{6}}. \]

(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 to obtain the single \( P \) value of 0.0013.

The approximate fusion sites were detected by aligning the fusion chromosomes of 26, 27, 28, 29, and 30 from randomly chosen 6 chromosomes (out of 31), thus 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

\[ P = \frac{\binom{27}{2} + \binom{26}{2}}{\binom{31}{6}}. \]

(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 to obtain the single \( P \) value of 0.0013.

The approximate fusion sites were detected by aligning the fusion chromosomes of 26, 27, 28, 29, and 30 from randomly chosen 6 chromosomes (out of 31), thus 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

\[ P = \frac{\binom{27}{2} + \binom{26}{2}}{\binom{31}{6}}. \]

(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 to obtain the single \( P \) value of 0.0013.

The approximate fusion sites were detected by aligning the fusion chromosomes of 26, 27, 28, 29, and 30 from randomly chosen 6 chromosomes (out of 31), thus 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

\[ P = \frac{\binom{27}{2} + \binom{26}{2}}{\binom{31}{6}}. \]

(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 to obtain the single \( P \) value of 0.0013.

The approximate fusion sites were detected by aligning the fusion chromosomes of 26, 27, 28, 29, and 30 from randomly chosen 6 chromosomes (out of 31), thus 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

\[ P = \frac{\binom{27}{2} + \binom{26}{2}}{\binom{31}{6}}. \]

(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 to obtain the single \( P \) value of 0.0013.
Author 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., 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 deep 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 orthology 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.F., J.S., C.W.C. 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.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., L.R., L.H. and M.R.G. wrote the manuscript. All authors read and commented on the manuscript.

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 APLT0000000.

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/ reprintsandpermissions/

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).

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/4.0/