FEATURED ARTICLE

GENOME REPORT

G3 Genes | Genomes | Genetics

De Novo Genome Assembly of the Meadow Brown Butterfly, *Maniola jurtina*

Kumar Saurabh Singh,*,[1] David J. Hosken,* Nina Wedell,* Richard ffrench-Constant,* Chris Bass,* Simon Baxter,[†] Konrad Paszkiewicz,[‡] and Manmohan D Sharma*,[1]

*College of Life and Environmental Sciences, University of Exeter, Penryn, UK, [†]School of Biological Sciences, University of Adelaide, Australia, and [‡]College of Life and Environmental Sciences, University of Exeter, UK

ORCID IDs: 0000-0001-8352-5897 (K.S.S.); 0000-0002-9957-3153 (M.D.S.)

ABSTRACT Meadow brown butterflies (*Maniola jurtina*) on the Isles of Scilly represent an ideal model in which to dissect the links between genotype, phenotype and long-term patterns of selection in the wild - a largely unfulfilled but fundamental aim of modern biology. To meet this aim, a clear description of genotype is required. Here we present the draft genome sequence of *M. jurtina* to serve as a founding genetic resource for this species. Seven libraries were constructed using pooled DNA from five wild caught spotted females and sequenced using Illumina, PacBio RSII and MinION technology. A novel hybrid assembly approach was employed to generate a final assembly with an N50 of 214 kb (longest scaffold 2.9 Mb). The sequence assembly described here predicts a gene count of 36,294 and includes variants and gene duplicates from five genotypes. Core BUSCO (Benchmarking Universal Single-Copy Orthologs) gene sets of Arthropoda and Insecta recovered 90.5% and 88.7% complete and single-copy genes respectively. Comparisons with 17 other Lepidopteran species placed 86.5% of the assembled genes in orthogroups. Our results provide the first high-quality draft genome and annotation of the butterfly *M. jurtina*.

KEYWORDS
Genome assembly
Lepidoptera
comparative genomics
Maniola jurtina
meadow brown

The meadow brown butterfly (*Maniola jurtina*, NCBI:txid191418) is a member of the nymphalid subtribe Satyrini. It is an important model organism for the study of lepidopteran ecology and evolution and has been extensively studied by ecological geneticists for many years (Dowdeswell *et al.* 1949; Dowdeswell 1961; Ford 1965). Found across the Palearctic realm it primarily habituates in grasslands, woodland rides, field-margins and can even be found in overgrown gardens.

The species displays marked sexual dimorphism. Females are more colorful than males and have large upper-wing eyespots (Figure 1). It also exhibits considerable quantitative variation in the sub-marginal spot pattern of its wings (Brakefield and van Noordwijk 1985) and therefore represents an ideal model in which to dissect the links between genotype, phenotype and long-term patterns of selection in the wild (Baxter *et al.* 2017) - a largely unfulfilled but fundamental aim of modern biology. This draft genome and corresponding annotations will offer a core resource for ongoing work in lepidopterans and other arthropods of ecological importance.

doi: https://doi.org/10.1534/g3.120.401071
Manuscript received September 6, 2019; accepted for publication March 4, 2020; published Early Online March 11, 2020.

Supplemental material available at figshare: https://doi.org/10.25387/g3.11594187.
[1]Corresponding authors: College of Life and Environmental Sciences, University of Exeter, Penryn, TR10 9FE, UK. E-mail: m.d.sharma@exeter.ac.uk; k.saurabh-singh@exeter.ac.uk

MATERIALS AND METHODS

Sampling and sequencing

Adult meadow brown (*Maniola jurtina*) butterflies were collected from multiple fields (Isles of Scilly, Cornwall) in June 2012, anesthetized by refrigeration for 2 hr and then killed by subsequent freezing. High molecular weight genomic DNA was isolated from whole body (pooled, excluding wings) of five individual females using the genomic-tip 100/G kit (Qiagen, Hilden, Germany) supplemented with RNase A (Qiagen, Hilden, Germany) and Proteinase K (New England Biolabs, Hitchin, UK) treatment, as per the manufacturer's instructions. DNA quantity and quality were subsequently assessed using a NanoDrop-2000 (Thermo Scientific, Loughborough, UK) and a Qubit 2.0 fluorometer (Life Technologies). Molecular integrity was confirmed using pulse-field gel electrophoresis.

Illumina data (100bp paired-end) was generated using standard Illumina protocols for a 250-500 bp PE library and multiple mate-pair

Figure 1 Female *Maniola jurtina* (picture credit: Richard ffrench-Constant).

libraries ranging between 180 to 7k bp (Table S1). 20 kb PacBio libraries were generated and size-selected following the manufacturers recommended protocols and sequenced on 18 SMRT cells of the RSII instrument. Finally, long reads (longest read 300Kb) were obtained using the Oxford Nanopore Technologies MinION platform (R7.4) (Table S2). Illumina, PacBio and MinION library preparation and sequencing were performed by the Exeter Sequencing Service, University of Exeter.

Genome assembly

The genome characteristics of *M. jurtina* were estimated using a k-mer based approach implemented in GenomeScope (Vurture *et al.* 2017). Short-read Illumina reads were quality filtered and subjected to 19-mer frequency distribution analysis using Jellyfish -v2.2.0 (Marçais and Kingsford 2011).

Genome assembly was performed by adopting a novel hybrid approach (Figure 2). Paired-end Illumina reads were trimmed and filtered for quality values using Trim Galore -v0.4.2 (Krueger 2016) and assembled using Spades -v3.9.1 (Bankevich *et al.* 2012). Long reads obtained from MinION were mixed with PacBio reads and assembled using Canu -v1.3.0 (Koren *et al.* 2017). The short-read assembly was further assembled along with long-reads using DBG2OLC -v20160205 (Ye *et al.* 2016). Canu and DBG2OLC assemblies were later merged using QuickMerge -v0.3.0 (Chakraborty *et al.* 2016) and redundancy reduction, scaffolding and gap closing were carried using Redundans -v0.14c (Pryszcz and Gabaldon 2016). The draft assembly was polished using arrow (part of the genomicconsensus package in PacBio tools) -v2.3.2 (Pacific Biosciences of California), which exclusively mapped long PacBio reads against the draft assembly using the BLASR pipeline (Chaisson and Tesler 2012). The draft assembly was also polished with the Illumina short-reads using Pilon -v1.23.0 (Walker *et al.* 2014).

Evaluation of the completeness of the genome assembly

The completeness of the draft genome was assessed by mapping raw short and long reads against the assembly. BUSCO (Benchmarking universal single-copy orthologs) -v3.0.2 (Simão *et al.* 2015) and CEGMA (Core Eukaryotic genes mapping approach) -v2.5.0 (Parra *et al.* 2007) were used to check genomic completeness of the assembly. In the case of BUSCO, Arthropoda and Insecta gene sets were compared against the assembly. We also assessed the completeness of this assembly by aligning complete genomes of *M. jurtina* genome against *H. melpomene* and *B. anynana* (a close relative) using Mummer -v3.1.0 (Kurtz *et al.* 2004).

Genome annotation

Before predicting gene models, the genome of *M. jurtina* was masked for repetitive elements using RepeatMasker -v4.0.7 (Smit 2013–2015). RepeatModeler -v1.0.11 (Smit 2008–2015) was used to model the repeat motifs and transposable elements. Repeats originating from coding regions were removed by performing a BLAST search against the *B. anynana* proteins. Sequence with hits at E-value $> 1e^{-10}$ were filtered out. The RepBase -*v*24.05 library was then merged with the repeats predicted by RepeatModeler and used to mask the *M. jurtina* genome. Protein coding genes were predicted using GeneMark-ES

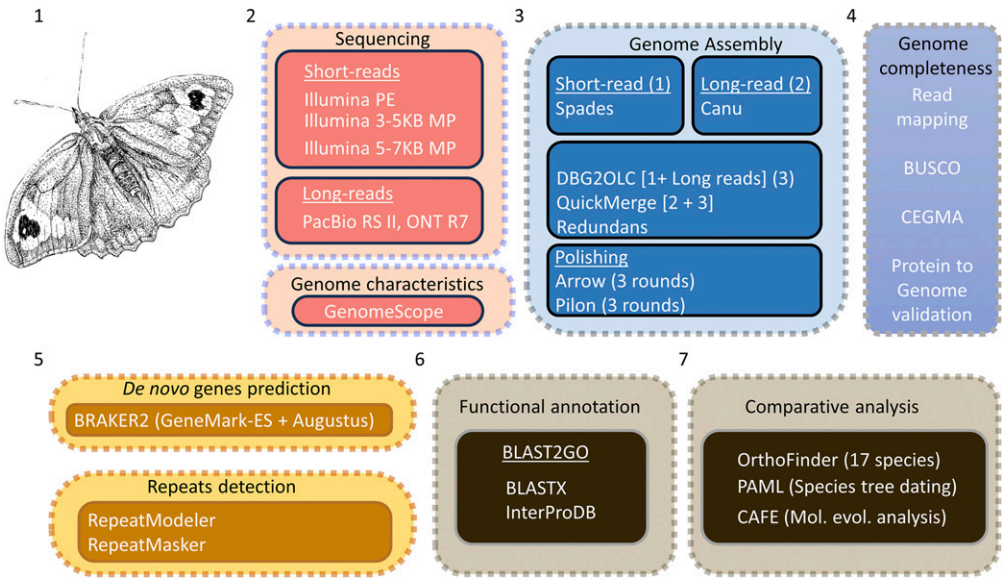

Figure 2 Schematic overview of the workflow used for sequencing, genome size estimation, assembly and annotation of the *M. jurtina* genome. 1. An artist's impression of a female *M. jurtina* (samples collected from multiple fields and processed for DNA extraction); 2. Multiple sequencing approaches adopted along with genome characterization using genome scope; 3. Genome assembly using a hybrid approach; 4. Genome completeness assessment; 5. *De novo* genes prediction and repeat detection; 6. Functional annotation; 7. Comparative analysis. Note that transcriptome data (orange segment) were obtained from publicly available sources at NCBI and only used for genome annotation.

Table 1 *jurtina* genome properties

Properties	Genome
# scaffolds (> 1000 bp)	10,860
Total length (>= 1000 bp)	618,415,580
Largest scaffold	2,944,739
Total length	618,415,580
GC (%)	36.90
N50	214,423
N75	78,459
L50	658
L75	1,875
# N's per 100 kbp	8,864.86

-v4.3.8 (Lomsadze *et al.* 2005) and AUGUSTUS -v3.3.0 (Stanke and Morgenstern 2005) implemented in the BRAKER -v 2.1.2 (Hoff *et al.* 2016) pipeline using species-specific RNA-seq alignments as evidence. Publicly available *M. jurtina* RNA-seq datasets (SRR3724201, SRR3724266, SRR3724269, SRR3724271, SRR3724198, SRR3724196, SRR3724195, SRR3721773, SRR3721752, SRR3721684, SRR3721695) were downloaded from NCBI and mapped individually against the repeat masked genome using STAR -v2.7.1 (Dobin *et al.* 2013). The bam files from individual samples were then combined using custom scripts and then fed into BRAKER. Functional annotation of the *de-novo* predicted gene models was carried out using homology searches against the NCBI nr database and Interpro database using BLAST2GO -v5.2.5 (Gotz *et al.* 2008).

Comparison to other Lepidopteran species

To characterize orthology and investigate gene family evolution across Lepidoptera, the final annotation set of *M. jurtina* was compared to 17 other genomes including a dipteran (*Drosophila melanogaster*), and a trichopteran (*Limnephilus lunatus*) as outgroups. The proteomes of *Amyelois transitella* v1.0, *B. anynana* v1.2, *Bombyx mori* v1.0, *Calycopis cecrops* v1.1, *Chilo suppressalis* v1.0, *Danaus plexippus* v3.0, *Heliconius melpomene* v2.0, *Junonia coenia* v1.0, *Limnephilus lunatus* v1.0, *Melitaea cinxia*, *Operophtera brumata* v1.0, *Papilio polytes* v1.0, *Phoebis sennae* v1.1, *Plodia interpunctella* v1.0, *Plutella xylostella* v1.0 were downloaded from Lepbase. OrthoFinder -v1.1.8 (Emms and Kelly 2018 *preprint*) was used to define orthologous groups (gene families) of genes between these peptide sets.

Phylogenetic tree construction and divergence time estimation

Phylogenetic analysis was performed using 39 single-copy orthologous genes, conserved among 17 species, using OrthoFinder. Additionally, OrthoFinder generated a species tree where *D. melanogaster* was used as the outgroup. The species tree was rooted using the STRIDE -v1.0.0 (Emms and Kelly 2017) algorithm implemented in OrthoFinder. MCMCTREE, as implemented in PAML -v4.9e (Yang 2007), was then used to estimate the divergence times of *M. jurtina* with approximate likelihood calculation. For this, the substitution rate was estimated using *codeml* by applying root divergence age between the Diptera, Lepidoptera and Trichoptera as 350 MY (Kjer *et al.* 2015). This is a simple fossil calibration of 350 MY for the root. The estimated substitution rate was the per site substitution rate for the amino acid dataset and used to set priors for the mean substitution rate in Bayesian analysis. As a second step, the gradient (g) and Hessian (H) of branch lengths for all 17 species were also estimated. Finally, the tree file with fossil calibrations, the gradient vector and hessian matrices file and the concatenated genes alignment information were used in the approximate likelihood calculation. The parameter settings of MCMCTREE were as follows: clock = 2, model = 3, BDparas = 110, kappa_gamma = 6 2, alpha_gamma = 11, rgene_gamma = 9.09, and sigma2_gamma = 1 4.5. Finally, Gene family evolution across arthropods was investigated using CAFE -v3.0 (De Bie *et al.* 2006). Scripts used for the analysis of genomic data are available at: https://github.com/kumarsaurabh20/Maniola_jurtina_genome_sequencing

Analysis of spot pattern related genes

To test whether any genes involved in wing or spot -pattern formation across Lepidoptera were identifiable in the current *Maniola* assembly, we first performed a wide literature search on PubMed (https://www.ncbi.nlm.nih.gov/pubmed/) using the keywords *Lepidoptera*, *butterfly*, *wing*, *spot*, *pattern*, *gene* and then manually filtered through the results to generate a list of candidate genes (Table 1 and Table 3).

This includes a selection of regulators possibly responsible for pattern variation (*APC*, *Naked cuticle*), transcription factors linked with eyespot patterning (*Distal-less*, *Dll*, and *Engrailed*, *En*), along with other transcription regulators such as *Apterous* and *DP*. Additionally, we considered *poik* (HM00025), also known as *cortex*, *Optix*, *Doublesex*, *Hox*, *Vermilion* and *black* (pigment synthesis) along with the *Ecdysone receptor* (*EcR*) involved in wing pattern plasticity.

NCBI esearch and efetch tools were used to filter (*NOT partial NOT hypothetical NOT uncharacterized*), and query individual spot pattern proteins across Lepidoptera using both, full and abbreviated protein names where available (Table 3; total 1347 homologs) and then these proteins were queried against the *Maniola* genome using Exonerate -v2.2.0 (Slater and Birney 2005) protein2genome model with the following customised options *–refine region–score 900–percent 70 -S FALSE –softmasktarget TRUE –bestn 1–ryo \">%ti (%tab - %tae) coding (%tcb - %tce) cds_length (%tcl)\n%tcs\n"*.

Data availability

The raw sequencing data and genome assembly have been deposited at the NCBI SRA database under the BioProject PRJNA498046 and genome accession number VMKL00000000. Blast results, annotation and proteome associated with this manuscript are available at https://zenodo.org/record/3352197. Scripts used for the analysis of

Table 2 Different assembly versions, data, software used and summary statistics

Version	Data	Assembler	N50	#Sequences	Total length
1	Short-read PE	Spades	48,073	53,043	319,930,151
2	Long-reads (PacBio + MinION)	Canu	32,954	10,463	296,564,618
3	Version 1 + Long-read (PacBio + Minion)	DBG2OLC	60,269	46,361	317,966,984
4	Version 2 + 3	QuickMerge	92,579	30,457	762,970,634
5	Version 4 + PE + MP + PacBio + MinION	Redundans	213,669	10,863	616,464,047

genomic data are available at: https://github.com/kumarsaurabh20/Maniola_jurtina_genome_sequencing. Supplemental material available at figshare: https://doi.org/10.25387/g3.11594187.

RESULTS AND DISCUSSION

Genome assembly

Sequencing of short-read libraries, both paired-end and mate-pairs, produced 317.1 million read pairs with an average insert size of 524.8 bp. Analysis of the unimodal 19-mer histogram with a coverage peak at 17x suggested an expected genome size of 576 MB (see Materials and Methods). Note here, that although the genome size estimated via this method is strongly dependent on the sequencing read-depth, based on the genome size of the most closely related species *Bicyclus anynana* (475 Mb), this estimate does not seem inordinate. The estimated heterozygosity rate was in the range of 1.89–1.93% (Table S3) and the genome was comprised of approximately 76% repetitive elements that are likely to contain units of highly repetitive W chromosome as the samples used in this study were all female (Table S3). We next performed a *de novo* genome assembly using a hybrid approach (see Materials and Methods).

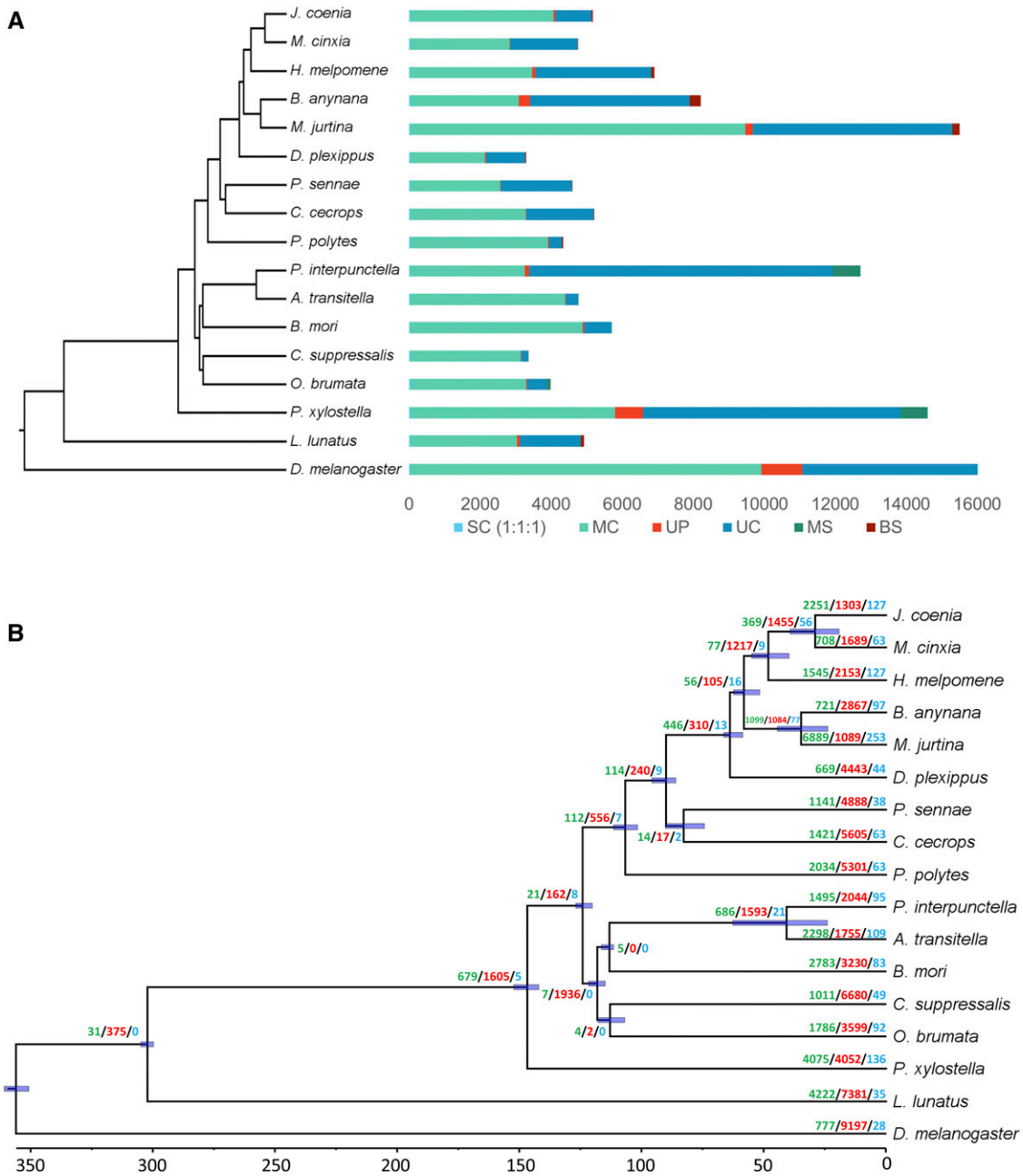

Figure 3 Evolutionary and comparative genomic analysis. (A) Ortholog analysis of *M. jurtina* with 16 other arthropod species. SC indicates common orthologs with the same number of copies in different species, MC indicates common orthologs with different copy numbers in different species, UP indicates species specific paralogs, UC indicates all genes which were not assigned to a gene family, MS indicates moths specific genes and BS indicates butterfly specific genes. (B) Species phylogenetic tree and gene family evolution. Numbers on the node indicate counts of the gene families that are expanding (green), contracting (red) and rapidly evolving (blue).

Spades assembly using multiple k-mer values produced 53,043 scaffolds having a total length of 319.9 Mb and N50 of 48.073 Kb. Long-read library sequencing produced 18.08 Gb (total 2398917 reads greater than 1000 bp) of data giving 21.7x overall sequencing coverage (Table S2). Canu assembled 10,463 contigs with N50 of 32.9 Kb. To further improve genome contiguity, we used DBG2OLC which is based on a hybrid approach of using both long- and short-reads. This assembly resulted in 46,361 short-read polished contigs with N50 of 60.26 kb which is an improvement of 12Kb over Spades assembly. In view of the recent developments in the hybrid assemblers, we further explore combining DBG2OLC assembly with long-read only Canu assembly using Quickmerge, an approach known to achieve high genomic contiguity with modest long- and short-read sequencing coverage. Merging of two assemblies with Quickmerge produced 30,457 contigs with a further improved N50 of 92.57 Kb. The assembly size, however, in Quickmerge step (762.9 Mb) surpassed the expected genome size of 576 Mb. To remove the alternate haplotypes from the assembly and reduce the inflated genome size, we added a redundancy removal step by using Redundans. This step improved the N50 by removing haplotypes and reducing the total assembly size. The final genome assembly comprised 618 Mb with 36.9 GC% and N50 of 214Kb (Table 2). Detailed assembly properties are given in Table 1 and Table S4.

Evaluation of the completeness of the genome assembly

To evaluate the completeness of the genome assembly, we first mapped raw short and long reads against it. The percentage of aligned reads ranged from 94 to 95% using paired-end and mate-paired short reads. Then we assessed the gene completeness using BUSCO and CEGMA. About 90.5% and 88.7% total BUSCO genes were identified in the Arthropoda and Insecta sets respectively. Additionally, 91% CEGMA genes, both complete and partials, were successfully found in the assembly (Table S5 and S6). The number of matches found between *M. jurtina* and *B. anynana*, after whole genome alignment, were significantly more as compared to *H. melpomene*. The genome size of *H. melpomene* (~250 MB) is smaller than *B. anynana* (~475 MB). Therefore many *M. jurtina* genomic sequences ended up with no hits.

Genome annotation

Annotation of the *M. jurtina* genome was carried out using the BRAKER pipeline. 11 publicly available datasets (See Material and Methods) were downloaded from NCBI totalling 116.4 million single-end transcriptomic reads. To predict genes, the reads were aligned against the *M. jurtina* assembly. BRAKER pipeline resulted in 38,101 genes after removing low quality genes with fewer than 50 amino-acid and/or exhibiting premature termination. In the final gene set, mean gene length, mean CDS length, mean intron length and exon number per gene were 4,144 bp, 976 bp, 921 bp and 5 respectively (Table S7). Approximately 34,263 out of 38,101 genes (90%) of the predicted genes could be assigned functional annotation based on BLAST searches against the non-redundant protein database of NCBI and InterPro.

Comparison to other Lepidopteran species

For comparative genomics analysis, we analyzed the orthologous gene relationships among several species (see Materials and Methods and Table S8). The combined gene count of these species was 349,442 of which 86.5% were assigned to 15,064 orthogroups. 50% of all genes were in orthogroups with 23 or more genes and were contained in the largest 4439 orthogroups. There were 2915 orthogroups with all species present and 39 of these consisted entirely of single-copy genes. A total of 216 gene families were specific to *M. jurtina* compared to 627 and 1716 in butterfly and moths respectively (Figure 3A).

Phylogenetic tree construction and divergence time estimation

The phylogenetic analysis showed that *M. jurtina* is more closely related to *B. anynana* than to *H. melpomene* or *M. cinxia*. The divergence time between *M. jurtina* and *B. anynana* was estimated to be around 34 MYA and that between *M. jurtina* and *H. melpomene* is estimated as 57 MYA (Figure 3B and see Table S9 for divergence time calibrations). Whole genome alignments, using Mummer -v3.1.0 (Kurtz *et al.* 2004) between *M. jurtina* – *B. anynana* and *M. jurtina* - *H. melpomene* were also performed to confirm this relatedness (Figure 4). In the dated phylogeny, the most species rich family Nymphalidae has remained stable and diverged from Papilionidae around 90 MY ago. This age is also supported by previously published butterfly phylogenies (Wahlberg *et al.* 2013; Espeland *et al.* 2018).

Analysis of gene family evolution

CAFE models the evolution of gene family size across a species phylogeny under a ML birth–death model of gene gain and loss and simultaneously reconstructs ML ancestral gene family sizes for all

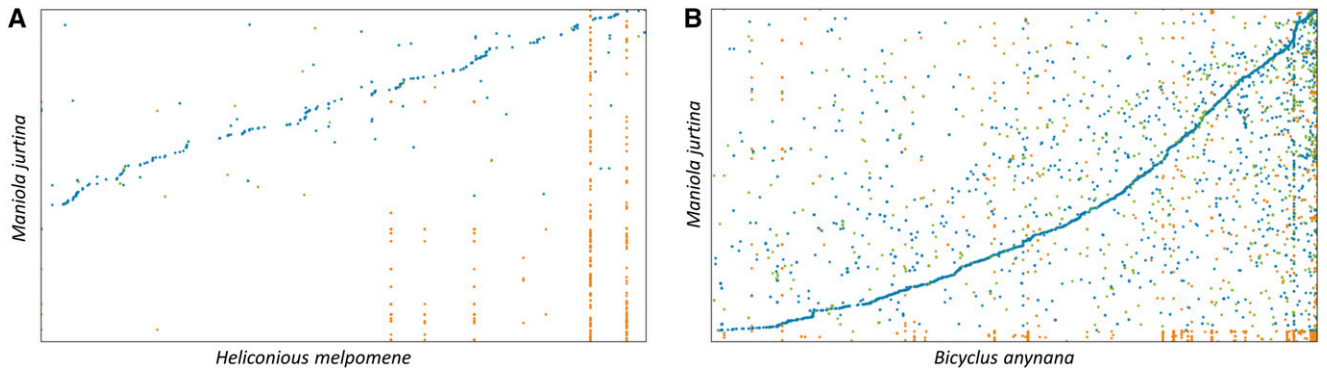

Figure 4 Genome comparisons *Comparison of* the *Maniola jurtina genome with Heliconious melpomene and Bicyclus anynana*. The dot plots were generated using Mummer. The plots show relatedness of *M. jurtina* with (A) *H. melpomene* and (B) *B. anynana*. Both of these genomes were taken as references (x-axis) and queried using *M. jurtina* (y-axis) genome. In both plots, blue, green and orange colored dots represent the unique forward, unique reverse and repetitive alignments respectively. Plot B shows more consistent and contiguous alignments than plot A. The dot plots were generated using https://dnanexus.github.io/dot/.

Table 3 Candidate wing spot patterning genes obtained from a literature search are listed in column 1. Column 2 has the number of annotated orthologs across Lepidoptera in our NCBI protein database search using the full gene name as listed and alternate names (comma separated). Column 3 presents the number of proteins per gene that matched in our Exonerate workflow

	Candidate gene *(alt. name in brackets) [Reference]*	NCBI Protein Matches within Lepidoptera *(n)*	protein2genome matches against the *Maniola* genome *(n)*
1	*Ap (Beldade et al. 2005)*	1	0
2	*APC (Saenko et al. 2010)*	18	4
3	*Apterous / apterousA (apA) (Beldade et al. 2005; Prakash and Monteiro 2018)*	25	4
4	*Black (Walker and Monteiro 2013)*	5	0
5	*C2 domain-containing protein 5 (C2CD5) (Rivera-Colón et al. 2018 preprint)*	32	7
6	*Calcium-activated potassium channel slowpoke (slo) (Özsu and Monteiro 2017; Rivera-Colón et al. 2018 preprint)*	205	191
7	*Cortex (Nadeau et al. 2016; Callier 2018)*	26	0
8	*Decapentaplegic (dpp) (Monteiro et al. 2013; Connahs et al. 2017 preprint)*	145	86
9	*Distal-less (Dll) (Koch et al. 2003; Reed and Serfas 2004; Monteiro et al. 2013)*	47	1
10	*Doublesex (dsx) (Kunte et al. 2014; Nishikawa et al. 2015)*	204	101
11	*DP transcription factor (dp) (Beldade et al. 2005)*	4	2
12	*Ecdysone receptor (ecr) (Koch et al. 1996; Koch et al. 2003)*	101	33
13	*engrailed (en) (Brunetti et al. 2001)*	19	0
14	*Geranylgeranyl pyrophosphate synthase (GGPS1) (Rivera-Colón et al. 2018 preprint)*	26	10
15	*Hox (Hombría 2011)*	39	6
16	*Invected (Brunetti et al. 2001)*	15	0
17	*Naked cuticle (Saenko et al. 2010)*	16	3
18	*Neutral Ceramidase (Cdase) (Özsu and Monteiro 2017; Rivera-Colón et al. 2018 preprint)*	32	22
19	*Notch (N) (Reed and Serfas 2004)*	117	72
20	*numb (Rivera-Colón et al. 2018 preprint)*	40	0
21	*Optix (Reed et al. 2011; Callier 2018)*	5	3
22	*Phosphoinositide 3-kinase adapter protein 1 (PIK3AP1) (Rivera-Colón et al. 2018 preprint)*	58	0
23	*poikilomousa / poik (HM00025) (Nadeau et al. 2015 preprint)*	2	0
24	*spaetzle (spz) (Özsu and Monteiro 2017; Rivera-Colón et al. 2018 preprint)*	65	2
25	*spalt (Sal) (Brunetti et al. 2001)*	1	1
26	*Transient receptor potential channel pyrexia (pyx) (Özsu and Monteiro 2017; Rivera-Colón et al. 2018 preprint)*	74	23
27	*Vermilion (Beldade et al. 2005)*	2	0
28	*wingless (wg) (Monteiro et al. 2006)*	0	—
29	*WntA (Callier 2018)*	1	1
30	*Zinc finger CCCH domain-containing protein 10 (ZC3H10) (Rivera-Colón et al. 2018 preprint)*	23	1

internal nodes, allowing the detection of expanded gene families within lineages. We ran CAFE on our matrix of gene family sizes generated by OrthoFinder and modeled their evolution along the dated species tree. Genes involved in binding, metabolism and transport of natural or synthetic allelochemicals are particularly found to be rapidly evolving in *M. jurtina* (Figure 3B).

Analysis of spot pattern related genes

Dowdeswell, Fisher and Ford first studied the island-specific wing-spot patterns in *M. jurtina* on the isles of Scilly (Dowdeswell *et al.* 1949), and this work was continued for more than 20 years (reviewed in (Ford 1965)). Their major findings, which became a cornerstone of ecological genetics, have been re-visited and largely re-confirmed with contemporary data (Baxter *et al.* 2017). Patterns of wing-spot polymorphism have remained unchanged on some islands over 60 years and there is some evidence of genetic differentiation across the Scillies (Baxter *et al.* 2017). Nonetheless, much remains to be done to better understand the underlying genetics of spot pattern variation in this species.

Butterfly wing patterns have long been suggested to be polygenic (Beldade and Brakefield 2002) and recent evidence from *B. anynana* (very closely related to *M. jurtina*) has confirmed this to be the case and strongly suggested that 10-11 different genomic regions may be involved in eye-spot number variation (Rivera-Colón *et al.* 2018 *preprint*) and see (Monteiro and Prudic 2010).

Protein to genome matches were found for 20 out of the 30 candidate genes (Table 3). We further cross checked this by creating a blast database of the 1347 homolog spot pattern related proteins from Lepidoptera and then searching the homologs within the *M. jurtina* proteome for matches. This resulted in over 1500 matches (see Table S10).

Specific experiments now need to be undertaken to further test candidate genes and their possible roles in wing-spot polymorphism, and to revisit other findings from Ford and co-workers (reviewed in (Ford 1965)) in the iconic Scillies study system.

Concluding remarks

Here we present a high-quality draft assembly and annotation of the butterfly *M. jurtina*. The assembly, along with the cross-species comparisons and elements of key spot-pattern genes will offer a core genomic resource for ongoing work in lepidopterans and other arthropods of ecological importance.

ACKNOWLEDGMENTS

The authors would like to acknowledge the use of University of Exeter's Advanced Computing Resources (Athena and Carson) and

Maisy Inston (University of Exeter) for the graphical illustration of a *M. jurtina* sample. KSS and CB received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement n646625). DJH was funded by NERC (NE/G005303/1) and the Leverhulme Trust (RF-2015-001), and NW by the Royal Society (Wolfson award). Part funding for this project was also received via internal grant funding within the University of Exeter. There are no competing interests to declare.

LITERATURE CITED

Bankevich, A., S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin *et al.*, 2012 SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J. Comput. Biol. 19: 455–477. https://doi.org/10.1089/cmb.2012.0021

Baxter, S. W., J. I. Hoffman, T. Tregenza, N. Wedell, and D. J. Hosken, 2017 EB Ford revisited: assessing the long-term stability of wing-spot patterns and population genetic structure of the meadow brown butterfly on the Isles of Scilly. Heredity 118: 322–329. https://doi.org/10.1038/hdy.2016.94

Beldade, P., and P. M. Brakefield, 2002 The genetics and evo-devo of butterfly wing patterns. Nat. Rev. Genet. 3: 442–452. https://doi.org/10.1038/nrg818

Beldade, P., P. M. Brakefield, and A. D. Long, 2005 Generating phenotypic variation: prospects from "evo-devo" research on *Bicyclus anynana* wing patterns. Evol. Dev. 7: 101–107. https://doi.org/10.1111/j.1525-142X.2005.05011.x

Brakefield, P. M., and A. J. van Noordwijk, 1985 The Genetics of Spot Pattern Characters in the Meadow Brown Butterfly *Maniola jurtina* (Lepidoptera, Satyrinae). Heredity 54: 275–284. https://doi.org/10.1038/hdy.1985.37

Brunetti, C. R., J. E. Selegue, A. Monteiro, V. French, P. M. Brakefield *et al.*, 2001 The generation and diversification of butterfly eyespot color patterns. Curr. Biol. 11: 1578–1585. https://doi.org/10.1016/S0960-9822(01)00502-4

Callier, V., 2018 How the butterfly got its spots (and why it matters). Proc. Natl. Acad. Sci. USA 115: 1397–1399. https://doi.org/10.1073/pnas.1722410115

Chaisson, M. J., and G. Tesler, 2012 Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics 13: 238. https://doi.org/10.1186/1471-2105-13-238

Chakraborty, M., J. G. Baldwin-Brown, A. D. Long, and J. J. Emerson, 2016 Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. Nucleic Acids Res. 44: e147.

Connahs, H., S. Tlili, J. van Creij, T. Y. Loo, T. Banerjee *et al.*, 2017 Disrupting different Distal-less exons leads to ectopic and missing eyespots accurately modeled by reaction-diffusion mechanisms. bioRxiv. (Preprint posted September 5, 2017). https://doi.org/10.1101/183491

De Bie, T., N. Cristianini, J. P. Demuth, and M. W. Hahn, 2006 CAFE: a computational tool for the study of gene family evolution. Bioinformatics 22: 1269–1271. https://doi.org/10.1093/bioinformatics/btl097

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski *et al.*, 2013 STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29: 15–21. https://doi.org/10.1093/bioinformatics/bts635

Dowdeswell, W. H., 1961 Experimental studies on natural selection in the butterfly, *Maniola jurtina*. Heredity 16: 39–52. https://doi.org/10.1038/hdy.1961.3

Dowdeswell, W. H., R. W. Fisher, and E. B. Ford, 1949 The quantitative study of populations in the Lepidoptera; *Maniola jurtina L*. Heredity 3: 67–84. https://doi.org/10.1038/hdy.1949.3

Emms, D. M., and S. Kelly, 2017 STRIDE: Species Tree Root Inference from Gene Duplication Events. Mol. Biol. Evol. 34: 3267–3278. https://doi.org/10.1093/molbev/msx259

Emms, D. M., and S. Kelly, 2018 OrthoFinder2: fast and accurate phylogenomic orthology analysis from gene sequences. bioRxiv. (Preprint posted November 8, 2018). https://doi.org/10.1101/466201

Espeland, M., J. Breinholt, K. R. Willmott, A. D. Warren, R. Vila *et al.*, 2018 A Comprehensive and Dated Phylogenomic Analysis of Butterflies. Curr. Biol. 28: 770–778.e5. https://doi.org/10.1016/j.cub.2018.01.061

Ford, E. B., 1965 *Ecological genetics*, Methuen Ltd., London.

Gotz, S., J. M. Garcia-Gomez, J. Terol, T. D. Williams, S. H. Nagaraj *et al.*, 2008 High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 36: 3420–3435. https://doi.org/10.1093/nar/gkn176

Hoff, K. J., S. Lange, A. Lomsadze, M. Borodovsky, and M. Stanke, 2016 BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. Bioinformatics 32: 767–769. https://doi.org/10.1093/bioinformatics/btv661

Hombría, J. C., 2011 Butterfly eyespot serial homology: enter the Hox genes. BMC Biol. 9: 26. https://doi.org/10.1186/1741-7007-9-26

Kjer, K. M., J. L. Ware, J. Rust, T. Wappler, R. Lanfear *et al.*, 2015 Response to Comment on "Phylogenomics resolves the timing and pattern of insect evolution". Science 349: 487. https://doi.org/10.1126/science.aaa7136

Koch, P. B., P. M. Brakefield, and F. Kesbeke, 1996 Ecdysteroids control eyespot size and wing color pattern in the polyphenic butterfly *Bicyclus anynana* (Lepidoptera: Satyridae). J. Insect Physiol. 42: 223–230. https://doi.org/10.1016/0022-1910(95)00103-4

Koch, P. B., R. Merk, R. Reinhardt, and P. Weber, 2003 Localization of ecdysone receptor protein during colour pattern formation in wings of the butterfly *Precis coenia* (Lepidoptera: Nymphalidae) and co-expression with Distal-less protein. Dev. Genes Evol. 212: 571–584. https://doi.org/10.1007/s00427-002-0277-5

Koren, S., B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman *et al.*, 2017 Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 27: 722–736. https://doi.org/10.1101/gr.215087.116

Krueger, F., 2016 TrimGalore; https://github.com/FelixKrueger/TrimGalore. The Babraham Institute.

Kunte, K., W. Zhang, A. Tenger-Trolander, D. H. Palmer, A. Martin *et al.*, 2014 doublesex is a mimicry supergene. Nature 507: 229–232. https://doi.org/10.1038/nature13112

Kurtz, S., A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway *et al.*, 2004 Versatile and open software for comparing large genomes. Genome Biol. 5: R12. https://doi.org/10.1186/gb-2004-5-2-r12

Lomsadze, A., V. Ter-Hovhannisyan, Y. O. Chernoff, and M. Borodovsky, 2005 Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. 33: 6494–6506. https://doi.org/10.1093/nar/gki937

Marçais, G., and C. Kingsford, 2011 A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27: 764–770. https://doi.org/10.1093/bioinformatics/btr011

Monteiro, A., B. Chen, D.M. Ramos, J.C. Oliver, X. Tong *et al.*, 2013 Distal-less regulates eyespot patterns and melanization in Bicyclus butterflies. J. Exp. Zool. (Mol. Dev. Evol.) 320: 321–331.

Monteiro, A., G. Glaser, S. Stockslager, N. Glansdorp, and D. Ramos, 2006 Comparative insights into questions of lepidopteran wing pattern homology. BMC Dev. Biol. 6: 52.

Monteiro, A., and K. M. Prudic, 2010 Multiple approaches to study color pattern evolution in butterflies. Trends Evol. Biol. 2: e2. https://doi.org/10.4081/eb.2010.e2

Nadeau, N. J., C. Pardo-Diaz, A. Whibley, M. Supple, R. Wallbank *et al.*, 2015 The origins of a novel butterfly wing patterning gene from within a family of conserved cell cycle regulators. bioRxiv. doi: 10.1101/016006 (Preprint posted March 24, 2015).

Nadeau, N. J., C. Pardo-Diaz, A. Whibley, M. A. Supple, S. V. Saenko *et al.*, 2016 The gene cortex controls mimicry and crypsis in butterflies and moths. Nature 534: 106–110. https://doi.org/10.1038/nature17961

Nishikawa, H., T. Iijima, R. Kajitani, J. Yamaguchi, T. Ando *et al.*, 2015 A genetic mechanism for female-limited Batesian mimicry in Papilio butterfly. Nat. Genet. 47: 405–409. https://doi.org/10.1038/ng.3241

Özsu, N., and A. Monteiro, 2017 Wound healing, calcium signaling, and other novel pathways are associated with the formation of

butterfly eyespots. BMC Genet. 18: 788. https://doi.org/10.1186/s12864-017-4175-7

Pacific Biosciences of California, I., GenimicConsensus; PacBio tools - https://github.com/PacificBiosciences/GenomicConsensus.

Parra, G., K. Bradnam, and I. Korf, 2007 CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23: 1061–1067. https://doi.org/10.1093/bioinformatics/btm071

Prakash, A., and A. Monteiro, 2018 apterous A specifies dorsal wing patterns and sexual traits in butterflies. Proc. Biol. Sci. 285. https://doi.org/10.1098/rspb.2017.2685

Pryszcz, L.P., and T. Gabaldon, 2016 Redundans: an assembly pipeline for highly heterozygous genomes. Nucleic Acids Res. 2016 Jul 8: e113. https://doi.org/10.1093/nar/gkw294

Reed, R. D., R. Papa, A. Martin, H. M. Hines, B. A. Counterman *et al.*, 2011 optix drives the repeated convergent evolution of butterfly wing pattern mimicry. Science 333: 1137–1141. https://doi.org/10.1126/science.1208227

Reed, R. D., and M. S. Serfas, 2004 Butterfly wing pattern evolution is associated with changes in a Notch/Distal-less temporal pattern formation process. Curr. Biol. 14: 1159–1166. https://doi.org/10.1016/j.cub.2004.06.046

Rivera-Colón, A. G., E. L. Westerman, S. M. Van Belleghem, A. Monteiro, and R. Papa, 2018 The genetic basis of hindwing eyespot number variation in *Bicyclus anynana* butterflies. bioRxiv. (Preprint posted December 21, 2018) https://doi.org/10.1101/504506

Saenko, S. V., P. M. Brakefield, and P. Beldade, 2010 Single locus affects embryonic segment polarity and multiple aspects of an adult evolutionary novelty. BMC Biol. 8: 111. https://doi.org/10.1186/1741-7007-8-111

Simão, F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov, 2015 BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31: 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Slater, G. S., and E. Birney, 2005 Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6: 31. https://doi.org/10.1186/1471-2105-6-31

Smit, A., R. Hubley, and P. Green, 2013–2015 RepeatMasker Open-4.0. http://www.repeatmasker.org.

Smit, A. F. A., and R. Hubley, 2008–2015 RepeatModeler Open-1.0. http://www.repeatmasker.org.

Stanke, M., and B. Morgenstern, 2005 AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res. 33: W465–W467. https://doi.org/10.1093/nar/gki458

Vurture, G. W., F. J. Sedlazeck, M. Nattestad, C. J. Underwood, H. Fang *et al.*, 2017 GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics 33: 2202–2204. https://doi.org/10.1093/bioinformatics/btx153

Wahlberg, N., C. W. Wheat, and C. Pena, 2013 Timing and Patterns in the Taxonomic Diversification of Lepidoptera (Butterflies and Moths). PLoS One 8: e80875. https://doi.org/10.1371/journal.pone.0080875

Walker, B. J., T. Abeel, T. Shea, M. Priest, A. Abouelliel *et al.*, 2014 Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9: e112963. https://doi.org/10.1371/journal.pone.0112963

Walker, J. F., and A. Monteiro, 2013 Determining the putative source of a morphogen underlying black spot development in *Pieris rapae* butterflies. Integr. Comp. Biol. 53: E389.

Yang, Z., 2007 PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24: 1586–1591. https://doi.org/10.1093/molbev/msm088

Ye, C., C. M. Hill, S. Wu, J. Ruan, and Z. Ma, 2016 DBG2OLC: Efficient Assembly of Large Genomes Using Long Erroneous Reads of the Third Generation Sequencing Technologies. Sci. Rep. 6: 31900. https://doi.org/10.1038/srep31900

Communicating editor: S. Celniker