DATA NOTE

A high-coverage draft genome of the mycalesine butterfly Bicyclus anynana

Reuben W. Nowell1,∗, Ben Elsworth1, Vicencio Oostra2, Bas J. Zwaan3, Christopher W. Wheat4, Marjo Saastamoinen5, Ilik J. Saccheri6, Arjen E. van’t Hof6, Bethany R. Wasik7, Heidi Connahs8, Muhammad L. Aslam8, Sujai Kumar1, Richard J. Challis1, Antónia Monteiro7,8,9, Paul M. Brakefield10 and Mark Blaxter1,∗

1Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3FL, UK, 2Department of Genetics, Evolution and Environment, University College London, London, UK, 3Laboratory of Genetics, Wageningen University, Wageningen, the Netherlands, 4Department of Zoology, Stockholm University, Stockholm, Sweden, 5Metapopulation Research Centre, Department of Biosciences, University of Helsinki, Helsinki, Finland, 6Institute of Integrative Biology, University of Liverpool, Liverpool L69 7ZB, UK, 7Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT 06511, USA, 8Department of Biological Sciences, National University of Singapore, Singapore 117543, 9Yale-NUS College, Singapore 138609 and 10Department of Zoology, University of Cambridge, Cambridge, CB2 3EJ, UK

∗Correspondence address. Reuben W. Nowell, Institute of Evolutionary Biology, University of Edinburgh, Ashworth Laboratories, Kings Buildings, Charlotte Auerbach Road, Edinburgh EH9 3FL, UK; E-mail: reubennowell@gmail.com; Mark Blaxter, Institute of Evolutionary Biology, University of Edinburgh, Ashworth Laboratories, Kings Buildings, Charlotte Auerbach Road, Edinburgh EH9 3FL, UK; E-mail: mark.blaxter@ed.ac.uk

Abstract

The mycalesine butterfly Bicyclus anynana, the “Squinting bush brown,” is a model organism in the study of lepidopteran ecology, development, and evolution. Here, we present a draft genome sequence for B. anynana to serve as a genomics resource for current and future studies of this important model species. Seven libraries with insert sizes ranging from 350 bp to 20 kb were constructed using DNA from an inbred female and sequenced using both Illumina and PacBio technology; 128 Gb of raw Illumina data was filtered to 124 Gb and assembled to a final size of 475 Mb (∼×260 assembly coverage). Contigs were scaffolded using mate-pair, transcriptome, and PacBio data into 10 800 sequences with an N50 of 638 kb (longest scaffold 5 Mb). The genome is comprised of 26% repetitive elements and encodes a total of 22 642 predicted protein-coding genes. Recovery of a BUSCO set of core metazoan genes was almost complete (98%). Overall, these metrics compare well with other recently published lepidopteran genomes. We report a high-quality draft genome sequence for Bicyclus anynana. The genome assembly and annotated gene models are available at LepBase (http://ensembl.lepbase.org/index.html).

Keywords: bicyclus anynana; squinting bush brown; nymphalidae, nymphalid; satyrid; lepidopteran genome
Data Description

The squinting bush brown butterfly, Bicyclus anynana, is a member of the remarkably speciose nymphalid subtribe Mycalesina, which is distributed across the Old World tropics (Fig. 1). B. anynana is an important model organism for the study of lepidopteran ecology, development, speciation, behaviour, and evolution [1–6]. B. anynana are found primarily in woodland habitats across East Africa (from southern Sudan in the north to Swaziland in the south), and adults are typically observed flying close to the ground, where they feed on fallen fruit [1]. Strikingly, B. anynana exhibits seasonal polyphenism, a form of phenotypic plasticity whereby individuals that develop during the wet season differ in behaviour, appearance, and life history to those that develop during the dry season [7–9]. Wet season butterflies are smaller, have shorter lifespans, are more active, and show larger and more conspicuous eyespots on their wings in comparison to dry season individuals. The genetic basis of this plasticity and its impacts on various other life history and developmental characteristics are ongoing research questions to which the availability of a B. anynana reference genome will contribute [10–12].

Sampling and sequencing

Genomic DNA was extracted from a B. anynana female that had been inbred via 7 generations of brother-sister matings. The captive laboratory stock population from which these individuals originated was established in 1988 from 80 wild-caught individuals and has been maintained at large effective population sizes to minimise the loss of genetic diversity [1]. Two short-insert libraries with insert sizes of 350 and 550 bp were constructed using Illumina TruSeq Nano reagents and sequenced (125 base, paired-end) on an Illumina HiSeq2500 at Edinburgh Genomics (Edinburgh, UK). DNA from a sister to this focal animal was used to construct four long-insert (mate-pair) libraries with insert sizes of 3 and 5 kb (2 of each) at the Centre for Genomic Research, University of Liverpool (Liverpool, UK); libraries of both insert-sizes were then sequenced on an Illumina HiSeq2500 and an Illumina MiSeq at Edinburgh Genomics (Table 1). DNA from a female descendant of the same inbred line was used to construct 2 long read libraries with insert sizes of 10 and 20 kb, sequenced on the PacBio platform at the Genome Institute of Singapore at ~×10 coverage using 16 P6 SMRT cells. All raw data have been deposited in the Short Read Archive under the accessions given in Table 1.

A total of 128.2 Gb of raw Illumina data was filtered for low-quality bases and adapter contamination using Skewer v. 0.2.2 [13], and both raw and trimmed reads were inspected using FastQC v. 0.11.4 [14]. Only 4 Gb of data (3.1%) was discarded, indicating the high quality of the raw data. Kmer frequency distributions were estimated using the “kmercountexact” program from the BBMap v. 36.02 package [15] and showed 2 major coverage peaks at ~×105 and ~×210 (Fig. 2).

<table>
<thead>
<tr>
<th>Library type</th>
<th>Platform</th>
<th>Read length</th>
<th>Insert size (expected)</th>
<th>Number of reads (raw)</th>
<th>Number of reads (trimmed)</th>
<th>Number of bases (trimmed)</th>
<th>SRA run accessions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short insert</td>
<td>Illumina</td>
<td>125 bp paired-end</td>
<td>350 bp</td>
<td>271 808 057 pairs</td>
<td>267 241 712 (98.3%)</td>
<td>66 334 099 834 (97.6%)</td>
<td>ERR1102671-2, ERR1102675-6</td>
</tr>
<tr>
<td>Short insert</td>
<td>Illumina</td>
<td>125 bp paired-end</td>
<td>550 bp</td>
<td>241 050 065 pairs</td>
<td>234 269 871 (97.2%)</td>
<td>57 913 474 128 (96.1%)</td>
<td>ERR1102673-4, ERR1102677-8</td>
</tr>
<tr>
<td>Mate pair</td>
<td>Illumina</td>
<td>100 bp paired-end</td>
<td>3 kb</td>
<td>77 105 680 pairs</td>
<td>31 848 200 (41.3%)</td>
<td>5 758 856 502 (37.3%)</td>
<td>ERR1750945</td>
</tr>
<tr>
<td>Mate pair</td>
<td>Illumina</td>
<td>100 bp paired-end</td>
<td>3 kb</td>
<td>5 641 764 pairs</td>
<td>2 170 610 (38.5%)</td>
<td>397 993 018 (35.3%)</td>
<td>ERR754051</td>
</tr>
<tr>
<td>Mate pair</td>
<td>Illumina</td>
<td>100 bp paired-end</td>
<td>5 kb</td>
<td>77 614 870 pairs</td>
<td>45 676 725 (58.9%)</td>
<td>8 203 769 131 (52.8%)</td>
<td>ERR1750946</td>
</tr>
<tr>
<td>Long read</td>
<td>PacBio P6</td>
<td>0.80–50 kb</td>
<td>10 kb</td>
<td>1 388 796</td>
<td>1 199 064 (86.3%)</td>
<td>4 086 394 966</td>
<td>ERR1797559-74</td>
</tr>
</tbody>
</table>
peak (×105) represents the proportion of the genome that is heterozygous and has an approximate span of 87.7 Mb (18.4% of the genome; calculated as one-half of the area under the ×105 curve, from ×50 to ×150). The expected proportion of heterozygous sites given 7 brother-sister (full-sib) matings is 0.75^7 = 13.3%, or 63.5 Mb. Thus, the greater than expected heterozygosity is likely to be due primarily to selection against highly inbred individuals during the course of the inbreeding regime [16].

Contaminant filtering and assembly

Short-insert libraries were screened for the presence of contaminant reads using Taxon-Annnotated GC-Coverage (TAGC) plots, or “blobplots” [17]. An initial draft assembly was constructed using the CLC assembler (CLCBio, Copenhagen) and compared to the NCBI nucleotide database (nt) using Megablast v. 2.3.0+ [18], and against the UniRef90 protein database using Diamond v. 0.7.10 [19]. Read coverage for each contig was calculated by mapping both libraries to the CLC assembly using CLC mapper (CLCBio, Copenhagen), and blobplots were generated using Blobtools v. 0.9.19.4 [20] using the “bestsumorder” rule for taxonomic annotation of contigs (Fig. 3). Contigs that showed a substantially different coverage relative to that of the main cluster of contigs and/or good hits to sequences annotated as non-Arthropoda were classed as putative contaminants. A total of 237 394 pairs of reads (~59 Mb) that were classed as either “mapped/mapped” or “mapped/unmapped” to a putative contaminant were subsequently discarded from further analysis.

Filtered libraries were reassembled using the heterozygous-aware assembler Platanus v. 1.2.4 [21], with default parameters. Contigs were further scaffolded with the mate pair libraries using SSPACE v. 3.0 [22] and with 35 747 assembled B. anynana transcripts [23] using a combination of L_RNA_scaffold [24] and SCUBAT v. 2 [25]. A final round of scaffolding was performed with PacBio long reads (fastq files error-corrected using the RS_Preassembler.2 protocol) using SSPACE-LongRead v. 1.1 [26]. Finally, gaps between scaffolds were filled using GapFiller v. 1.10 [27] and PBJelly v. 15.8.24 [28].

Our final assembly (v. 1.2) comprised 10 800 scaffolds spanning a total of 475.4 Mb, with a scaffold N50 of 638 kb (Table 2). The genome-wide proportion of G+C was 36.5%, while the number of underdetermined bases (Ns) was 5.8 Mb (~1.2% of the total span). We determined assembly completeness by mapping both genomic and transcriptomic reads from B. anynana (SRA whole genome sequencing accessions ERR1102671-8 and transcriptome accessions ERR1022636-7, ERR1022640-1, and ERR1022644-5, downloaded October 2016) to the genome using BWA mem v. 0.7.12 [29] and STAR v. 020201 [30], respectively. Over 99% of reads from the 2 short-insert libraries mapped to the assembly, suggesting that the vast majority of the genome represented by these data has been assembled. In addition, 94.9% of RNA-Seq reads mapped to the assembly, suggesting that the majority of transcribed genes are present. Gene-level completeness was assessed using CEGMA v. 2.5 [31] and BUSCO v. 2.0 [32]. The proportion of CEGMA genes “completely” recovered (n = 248) was 81%, increasing to 97% when partially recovered genes were included. The recovery of BUSCO genes specific to the metazoan (n = 978) was higher, at 98% for complete genes, increasing to 99% when partial genes were included. An almost complete set (99.2%) of BUSCO genes specific to the Arthropoda (n = 1066) was also recovered. In addition, CEGMA indicated a duplication rate of 1.1 while BUSCO estimated only ~2% of genes were present in multiple copies. The high complete CEGMA/BUSCO scores suggest that a good assembly has captured the majority of core metazoan/Arthropod genes in full length and that the fragmentation of genes across multiple scaffolds is low. In addition, the low duplication rates suggest that most genes are present in single copy, and thus that the genome does not include significant duplicated segments representing alternative haplotypes.

Annotation

Prior to gene prediction, we masked the B. anynana assembly for repetitive elements to minimise the number of
spurious open-reading frames due to low-complexity repeat regions or transposable elements. Repetitive motifs in the B. anynana assembly were modelled ab initio using RepeatModeler v. 1.0.5 (http://www.repeatmasker.org/RepeatModeler.html). Repeats occurring within genuine coding regions were excluded by querying the proteins from a previous B. anynana assembly (v. 0.1) versus the RepeatModeler database using BLAST, removing any sequences showing a match at the E-value $\leq 1e^{-10}$ threshold. The filtered RepeatModeler database was combined with known repeats from the Lepidoptera using RepBase v. 20.05 and input to RepeatMasker v. 4.0.5 to mask the assembly. Overall, approximately one-quarter of the assembly (122.6 Mb) was masked from gene prediction (Table 3).

Gene finding was performed following a 2-pass approach [35]. Initial gene models were constructed with MAKER v. 2.31 using HMMs derived from SNAP v. 4.3 [37] in conjunction with a recently published B. anynana transcriptome as evidence. MAKER gene models were then passed to AUGUSTUS v. 3.0.3 for refinement, resulting in an initial set of 26,722 predicted protein-coding genes. A set of basic filters was applied to remove likely spurious gene models (Table 4), resulting in the deletion of 4080 gene models. Protein sequences from the filtered 22,642 genes were annotated using BLAST searches versus UniRef90 and the NCBI non-redundant protein database (nr), and domains/motifs were described using InterProScan5 [40]. Summary statistics for the 22,642 predicted gene models are given in Table 5.

Comparison to other lepidopteran genomes

To ascertain the relative quality of the B. anynana v. 1.2 assembly, we compared our results to 9 other published lepidopteran genomes available on LepBase (http://lepbase.org/) [41]: Bombyx mori ASM15162 v. 1 [42], Danaus plexippus v. 3 [43], Heliconius melpomene Hmel2 [44,45], Lermaea accius v. 1.1 [46], Melitaea cinxia MelCx1.0 [47], Papilio glaucus v. 1.1 [48], Papilio polytes Pp01.0 [49], Papilio xuthus Pp_xu.1.0 [49], and Plutella xylostella DBM_Pf_v1.1 [50]. The B. anynana v. 1.2 assembly was of high quality compared to other published genomes, with the majority of the genome represented in a relatively small number of
scaffolds despite being only marginally smaller than the largest lepidopteran genome, B. mori (Fig. 4a). Interestingly, B. anynana v. 1.2 encodes the highest number of proteins of the 10 species compared (Fig. 4b). Despite measures to eliminate potentially spurious ORFs caused by annotation error or by duplication, B. anynana encodes ~3250 more genes than the diamondback moth P. xylostella, and ~10 400 more than the swallowtail P. polytes. It is tempting to attribute the apparently high number of genes to the developmental plasticity and alternative seasonal forms with divergent morphologies and life histories in B. anynana. However, it remains to be determined whether the number of genes predicted in B. anynana is a function of its larger genome size or unusual life history characteristics, or if further curation of the v. 1.2 gene models will reduce the number of inferred genes.

Concluding remarks

We present a high-coverage, high-quality draft assembly and annotation of the mycalesine butterfly B. anynana. The assembly will be a core resource for ongoing analyses of population genomics, discovery of cis-regulatory elements of wing patterning and other genes, functional genetics and functional ecology of complex gene families, and the evolution of novel and plastic lifecycle strategies in lepidopterans and other arthropods.

Abbreviations


Acknowledgements

We thank Edinburgh Genomics and the Genome Institute of Singapore for genome sequencing, initial QC, and data delivery. We also thank 2 reviewers for helpful comments on a previous version of this manuscript. Funding for the Bicyclus anynana genome project was provided by the ERC Advanced Grant number 250325 (EMARES) to P.M.B. and by the South East Asian Biodiversity Genomics Center (NUS grants R-154-000-648-646 and R-154-000-648-733) to A.M. Funding for LepBase was provided by BBSRC grant number BB/K020161.

Availability of supporting data

All raw sequence data have been deposited in the Short Read Archive (SRA) and are available for download using the accession numbers provided in Table 1. The B. anynana v. 1.2 assembly, as well as final predicted gene models and protein annotations, are publicly available for viewing and download via LepBase [41], an Ensembl [51] genome database for the Lepidoptera (http://ensembl.lepbase.org/index.html). Data supporting the manuscript, including annotations as well as BUSCO and CEGMA results, are also available via the GigaScience database, GigaDB [52]. A previous B. anynana assembly (nBa.0.1) is also available on LepBase.

Competing interests

The authors declare that they have no competing interests.

Author contributions

P.M.B. and M.B. designed the study; A.M. and B.R.W. collected samples and produced the inbred line; A.E.V.H., I.J.S., and H.C. extracted DNA samples; R.W.N., B.E., and M.B. worked on the genome assembly and annotation; V.O., B.J.Z., C.W., and M.S. contributed transcriptome data; A.M., H.C., and M.L.A. contributed PacBio data; S.K. and R.J.C. uploaded the assembly to LepBase. R.W.N., V.O., A.M., P.M.B., and M.B. wrote the manuscript. All authors read and approved the final version of the manuscript.
Figure 4: Assembly and gene prediction comparison among 10 lepidopteran genomes. (a) Cumulative assembly curves showing the relationship between the number of scaffolds (x-axis) and the cumulative span of each assembly (y-axis), coloured by species. Higher-quality assemblies are represented by an almost-vertical line (e.g., H. melpomene Hmel2 assembly in black), indicating that a relatively small number of scaffolds is required to reach the final genome span; conversely, a long tail indicates that the assembly includes a large number of smaller scaffolds. The curve for B. anynana (brown and bold) suggests a good assembly for this species, with the majority of the assembly comprised of relatively few scaffolds. (b) B. anynana v. 1.2 encodes the greatest number of genes of the 10 genomes and is particularly different from B. mori, which is of equivalent length. Species names/colours are as follows: “bicyclus” (brown), B. anynana; “bombyx” (blue), B. mori; “danaus” (light green), D. plexippus; “heliconius” (black), H. melpomene; “lerema” (dark green), L. accius; “melitaea” (orange), M. cinxia; “glaucus” (red), P. glaucus; “polytes” (pink), P. polytes; “xuthus” (violet), P. xuthus; “plutella” (grey), P. xylostella.

References


