Sequencing, assembly and annotation of the whole-insect genome of *Lymantria dispar dispar*, the European gypsy moth

Michael E. Sparks1,2,*, François Olivier Hebert,2 J. Spencer Johnston,3 Richard C. Hamelin,2,4 Michel Cusson,2,5 Roger C. Levesque6 and Dawn E. Gundersen-Rindal1

1USDA-ARS Invasive Insect Biocontrol and Behavior Laboratory, Beltsville, MD 20705, USA,
2Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval, Quebec City, QC, G1V 0A6, Canada,
3Department of Entomology, Texas A&M University, College Station, TX 77843, USA,
4Department of Forest and Conservation Sciences, Faculty of Forestry, University of British Columbia, Vancouver, BC V6T 1Z4, Canada, and
5Laurentian Forestry Centre, Canadian Forest Service, Natural Resources Canada, Quebec City, QC G1V 4C7, Canada

*Corresponding author: michael.sparks2@usda.gov

Abstract

The European gypsy moth, *Lymantria dispar dispar* (LDD), is an invasive insect and a threat to urban trees, forests and forest-related industries in North America. For use as a comparator with a previously published genome based on the LD652 pupal ovary-derived cell line, as well as whole-insect genome sequences obtained from the Asian gypsy moth subspecies *L. dispar asiatica* and *L. dispar japonica*, the whole-insect LDD genome was sequenced, assembled and annotated. The resulting assembly was 998 Mb in size, with a contig N50 of 662 Kb and a GC content of 38.8%. Long interspersed nuclear elements constitute 25.4% of the whole-insect genome, and a total of 11,901 genes predicted by automated gene finding encoded proteins exhibiting homology with reference sequences in the NCBI NR and/or UniProtKB databases at the most stringent similarity cutoff level (i.e., the gold tier). These results will be especially useful in developing a better understanding of the biology and population genetics of *L. dispar* and the genetic features underlying Lepidoptera in general.

Keywords: whole-genome sequencing; PacBio long-read assembly; Illumina short-read polishing; automated gene finding; Lepidoptera; European gypsy moth; gypsy moth genomics

Introduction

Feeding on over 300 species of woody plants, the European gypsy moth, *Lymantria dispar dispar* (LDD), poses significant economic and ecological threats to urban trees, forests and forest-related industries in North America, particularly with respect to forest-related industries (Leonard 1981). Having a representative genome sequence for this forest pest is important to facilitate various aspects of biosurveillance, including discovery of traits associated with invasiveness and mapping of population resequencing data to determine sources and pathways of spread (Roe et al. 2019; Hamelin and Roe 2020). An Illumina-based genome assembly of biomaterials prepared from the LDD-derived in vitro cell line, LD652 (Goodwin et al. 1978), was recently published (Zhang et al. 2019), however, this cell culture may not accurately portray the comprehensive genomic and/or genetic complements of the intact LDD insect (Sparks et al. 2013). Whole-insect genomes of two subspecies of Asian gypsy moths, *L. dispar asiatica* (LDA) and *L. dispar japonica* (LDJ), were also recently sequenced using PacBio technology (Hebert et al. 2019). This study reports the sequencing of LDD at the whole-insect level using PacBio and Illumina technologies, done specifically to enable nuclear genome-informed population genetics and comparative genomics analyses among these subspecies, thereby augmenting earlier mitogenome-based research findings (Djoumad et al. 2017). These resources will contribute to an improved understanding of lepidopteran genetics in general (Triant et al. 2018).

Materials and methods

The European gypsy moth genome size was estimated using previously described methods (Johnston et al. 2019) applied to LDD New Jersey Standard Strain (NJSS) moth heads obtained from the USDA-APHIS Otis Laboratory (Buzzards Bay, MA, USA): four individuals were processed apiece for each of the male and female sexes. Genomic DNA was prepared from a single female pupa, which had been reared from an NJSS egg mass—this biomaterial was used for sequencing a total of 96 SMRT-Cells by the University of Washington PacBio Sequencing Services group (Seattle, WA, USA). Subread data were extracted in fastq format from bas.h5 archives (without imposition of quality filters) using Pacific Biosciences’ pbhtools (https://github.com/PacificBiosciences/pbhtools). Read correction, trimming and assembly were performed using version 1.9 of the CANU assembler (Koren et al. 2017). Subread sets were compiled into BAM-formatted archives using the bax2bam utility from Pacific Biosciences’ SMRT Analysis software suite (https://www.
pacbio.com/support/software-downloads) with the following pulse feature options enabled: DeletionQV, DeletionTag, InsertionQV, IPD, MergeQV, SubstitutionQV, PulseWidth and SubstitutionTag. BAM-formatted subreads were aligned to assembled genomic templates using the BLASR long read aligner via the pbalign wrapper (https://github.com/PacificBiosciences/blasr; https://github.com/PacificBiosciences/pbalign), the results of which were subsequently used as inputs to the Arrow genome polishing algorithm as implemented in the Pacific Biosciences GenomicConsensus package (https://github.com/PacificBiosciences/GenomicConsensus). Three rounds of Arrow-based assembly polishing were performed.

An additional library was prepared from a single non-sexed, 3rd-instar larval insect reared from an NJSS egg mass. A set of 359,599,922 Illumina PE100 short read pairs was sequenced from this library by the Georgia Genomics Facility (Athens, GA, USA) and aligned to the Arrow-polished template with bowtie2 (Langmead and Salzberg 2012), the results of which were used as input to the Filion polishing program operating in its diploid mode (Walker et al. 2014). The fully polished assembly was then processed with Purge Harplotigs (Roach et al. 2018) to minimize residual heterogeneity; empirically selected read depth cutoffs of $L = 19$, $mid = 58$, and $high = 107$ were used. The genome was then submitted to the NCBI WGS division, which filtered for potential vector contaminants and mitochondrial sequences. The NCBI-vetted assembly was compared with the endopterygota_or9 database using BUSCO (Simão et al. 2015) to assess genome quality in terms of known single-copy orthologs. Repetitive content was identified using RepeatModeler (Smit and Hubley 2020) — which wraps the LTRharvest (Ellinghaus et al. 2008), RepeatScout (Price et al. 2005) and RECON (Bao and Eddy 2002) pipelines — and masked using RepeatMasker (Smit et al. 2019).

Gene annotation was conducted using a modified version of the insect-specific pipeline described in Hebert et al. (2019) (see Figure 1). In addition to extrinsic hints provided to Augustus (Stanke et al. 2006) by Exonerate (Slater and Birney 2005) and PASA (Haas et al. 2003), proteins predicted by GenomeThreader (Gremme et al. 2005), used to align Trinity-based RNA-Seq assemblies (Haas et al. 2013) to genomic templates and compute consensus gene structures from the results, were also used. In particular, GenomeThreader’s splice site models were retrained using BSSM4GQS (Sparks and Brendel 2005) with Bombyx mori data (obtained from NCBI’s Annotation Release 102; https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Bombyx_mori/102/) and its predictions were post-processed with MetaWAm (https://github.com/scentiant/MetaWAm; Sparks and Brendel 2008) to identify translation initiation sites. If a GenomeThreader-derived consensus gene structure was ideal (i.e., canonical translation start and stop codons were present and, for multi-exon models, introns uniformly exhibited GT donors and AG acceptors), it was provided to Augustus as an immutable anchor; otherwise, it was supplied as a hint. Both of the alternatives-from-evidence and alternatives-from-sampling Augustus flags were set to false. RNA-Seq data used to prepare Trinity assemblies were obtained from Sparks et al. (2013) and other unpublished studies conducted by MES and DEGR.

Protein sequences predicted by the overall pipeline were constrained to 50 or more amino acids (aa) in length. These were then compared with the NCBI NR and UniProtKB’s Swiss-Prot and TrEMBL protein databases (UniProt Consortium 2019) using the DIAMOND aligner in its BLASTp-like mode (Buchfink et al. 2015). Gene models were sorted into gold, silver and bronze tiers based on their encoded products’ similarity to exemplar proteins: gold-tier entries are at least 150 aa in length and exhibit a single high-scoring segment pair with a reference protein (also of at least 150 aa in length) such that 75% or more of aligned residues are positively similar, and the hit length-to-subject sequence length ratio is 90% or greater. Silver-tier entries are a minimum of 100 amino acid residues long with a hit spanning at least 75% of a reference protein sequence’s length (also at least 100 aa in length), and bronze-tier entries are similar, but with the hit covering 50% or more of the reference protein’s length.

The annotation pipeline was also used to independently analyze the LDA and LDJ genomes (available from https://osf.io/umz2v/files/ as “lda genome v0.3.arow-polished.fasta” and “ldj genome v0.3.arow-polished.fasta”, respectively), as well as the two LD652 cell line assemblies described in Zhang et al. (2019): one of these LD652 assemblies incorporates information from Novogene-prepared Hi-C libraries (“LD652_Lymantria_dispar_FINAL.fasta”, obtained from http://prodata.swmed.edu/download/pub/LD652/) and the other (available from GenBank under assembly accession GCA_004115105.1), does not. Predicted protein sequences from each of the gold, silver and bronze tiers were combined in an assembly-specific manner, and these were then globally pooled and clustered with CD-HIT (Fu et al. 2012) using a sequence identity threshold of 0.975. Clusters having exactly one representative sequence from each of the five assemblies were identified, providing a set of putative single-copy gene models universally present in these L. dispar genomes.

Data availability
All data used in the preparation of the final LDD assembly are available at the NCBI under BioProject accession number PRJNA694036: PacBio subreads and Illumina short-read data are available at the SRA database (accession numbers SRR13505170-SRR13505265, and SRR13518687-SRR13518688, respectively). The Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession JAEVLZ00000000. The version described in this paper is JAEVLZ010000000. In addition, the whole genome assembly and automated protein-coding gene finding results (see below) are available at the Open Science Framework (OSF) repository: https://doi.org/10.17605/OSF.IO/JX2VS. Lastly, interested parties are welcome to contact the corresponding author to request relevant data and software.

Results and discussion
The fully processed, NCBI-vetted genome assembly of the whole-insect European gypsy moth (i.e., LDD) exhibited a size of 998 Mb, a contig N50 of 662 Kb and a GC content of 38.8% (see Table 1). The assembly size is 1.015 times larger than the 983 Mb mean estimate obtained from a sample of four females using flow cytometry [female: 983.2 $\pm$ 6.7 Mb ($n = 4$) and male: 993.3 $\pm$ 6.2 Mb ($n = 4$)]. Of a total of 2,442 known BUSCOs in the Endopterygota, 95.5% were complete (93.7% single-copy and 1.8% duplicated), 2.7% were fragmented and 1.8% were missing. Long interspersed nuclear elements (LINEs) constitute 25.40% of the LDD genome (see Table 2). This amount is only slightly less than the 25.74% and 26.86% fractions of LINEs observed by the revised annotation pipeline as applied to LDA and LDJ, respectively, although it is notably greater than the 18.96% and 17.34% amounts observed for the LD652 assemblies with and without Hi-C library integration, respectively (see Table 3). Similarly, although LDD, LDA and LDJ exhibit commensurate levels of total repetitive DNA (68.74%, 67.45% and 67.09%, respectively), the LD652 assemblies comprised lesser amounts of such content (with Hi-C = 59.76%, sans Hi-C = 59.66%).

The protein-coding gene annotation pipeline identified 75,018 unique LDD genes, of which 11,901 sorted to the gold tier, 17,588 to the silver, and 26,773 to bronze (the remaining 18,746 “non-podium”
Figure 1  *L. dispar dispar* genome assembly and annotation pipeline. The automated pipeline used to assemble the European gypsy moth genome, as well as annotate repetitive content and nuclear protein coding genes, is a modified version of the methods used for annotating the Asian gypsy moth genomes, *L. dispar asiatica* and *L. dispar japonica* (Hebert et al. 2019). (The *Bombyx mori* image is copyright Freepik 2018 and reproduced here with permission.)
genes failed to satisfy the criteria for any of these levels, thus representing either false positive gene calls or novel sequences not yet present in reference databases). Reannotation of the LDA and LDJ genomes with this updated annotation system yielded commensurate counts of genes encoding proteins corroborated by records in the NCBI and UniProtKB databases (i.e., gold, silver and bronze tier models): LDA = 9,581 + 16,081 + 27,583 = 53,245 "podium-level" genes, LDJ = 57,331 and LDD = 56,272 (see Table 3). These amounts are roughly double what was observed for the LD652 assemblies: 24,345 unique protein-coding gene loci were identified in the Hi-C library-based assembly and 26,504 in the other version. Clustering of pooled, podium-level sequences identified a total of 3,265 putative single-copy proteins shared across all five of these L. dispar genomes.

Given that the LD652 cell line was derived from ovarian tissue over four decades ago (Goodwin et al. 1978) and has undergone long-term storage, multiple serial passages and exposure to unknown selective pressures in the interim (Lynn 2000, 2006), the possibility exists that loss of chromosomal-level genetic material may partially explain the discrepancies in genome size, repetitive content and gene content observed between LD652 and LDD (Table 3). More likely, however, is that the short-read technology used to sequence LD652 (Zhang et al. 2019) was comparably less effective at characterizing such a highly repetitive genome as that of L. dispar than were the longer PacBio reads used in this study (and in Hebert et al. 2019). What is more, the lower levels of BUSCOs detected and fewer protein-coding genes predicted in

| Table 1 | Genome assembly statistics following initial Canu assembly, each of three rounds of Arrow polishing, a round of Pilon short-read polishing, primary contig identification using the Purge Haplotigs pipeline, and purging of mitochondrial sequences and putative vector contaminants by software filters at the NCBI WGS division |
|---|---|---|---|---|---|---|
| Unpolished assembly (Canu) | Arrow polished (round 1) | Arrow polished (round 2) | Arrow polished (round 3) | Pilon polished | Purge Haplotigs processed | NCBI filters (final version) |
| N50 (a) | 355,683 | 355,730 | 355,728 | 355,719 | 354,801 | 661,876 |
| N90 (a) | 1,028 | 1,028 | 1,028 | 1,029 | 1,029 | 1,039 |
| Total size | 1,280,392,172 | 1,281,096,304 | 1,281,192,033 | 1,281,236,029 | 1,279,160,664 | 998,776,906 |
| Min contig | 1,028 | 1,028 | 1,028 | 1,029 | 1,029 | 1,039 |
| Max contig | 8,311,499 | 8,312,296 | 8,312,357 | 8,312,335 | 8,267,660 | 8,267,660 |
| Contig count | 16,115 | 16,115 | 16,115 | 16,115 | 16,115 | 4,625 |
| SMARTS cells | 96 |
| Bases in subreads (b) | 20,008,385 |
| Genome size estimate | 983,200,000 |
| Sequencing coverage | 115.11 |

aN50 was calculated using the definition provided in Miller et al. (2010), with N90 being similarly defined.

bSubreads determined using bax2bam (from smrtlink_7.0.0.63985) w/options "--subread" and "--pulsefeatures=DeletionQV,DeletionTag,InsertionQV,IPD,MergeQV,SubstitutionQV,PulseWidth,SubstitutionTag".

| Table 2 | Repetitive content of the L. dispar dispar genome assembly as identified using RepeatModeler/RepeatMasker |
|---|---|---|
| Number of elements | Length occupied (bp) | Percentage of sequence (%) |
| Bases masked | 686,312,151 | 68.74 |
| Retroelements | 1,688,175 | 336,176,424 | 33.67 |
| SINES | 150,440 | 12,344,986 | 1.24 |
| Penelope | 214,022 | 28,041,373 | 2.81 |
| LINEs | 1,210,601 | 253,631,498 | 25.40 |
| CRE/SLACS | 1,123 | 414,475 | 0.04 |
| L2/CRI/Rex | 120,505 | 29,326,164 | 2.94 |
| R1/LOA/Jockey | 289,120 | 100,744,241 | 10.09 |
| R2/R4/NeSL | 3,328 | 1,806,751 | 0.18 |
| RTE/Boy-B | 392,519 | 58,464,284 | 5.86 |
| L1/CIN4 | 0 | 0 | 0.00 |
| LTR elements: | 327,134 | 70,200,891 | 7.03 |
| BEL/Pao | 97,919 | 24,858,437 | 2.49 |
| Ty1/Copia | 20,415 | 6,797,796 | 0.68 |
| Gypsy/DIRS1 | 114,861 | 23,908,124 | 2.39 |
| Retroviral | 455 | 56,012 | 0.01 |
| DNA transposons | 410,996 | 75,109,994 | 7.52 |
| hobc-Activator | 39,043 | 10,065,906 | 1.01 |
| Tcl-1-IS630-Pogo | 324,022 | 56,591,826 | 5.67 |
| En-Spm | 0 | 0 | 0.00 |
| MuDR-ISIS05 | 0 | 0 | 0.00 |
| PiggyBac | 1,258 | 410,331 | 0.04 |
| Tourist/Harbinger | 6,160 | 810,421 | 0.08 |
| Other (Mirage, P-element, Transib) | 4,091 | 523,746 | 0.05 |
| Rolling-circles | 746,035 | 123,018,327 | 12.32 |
| Unclassified | 1,043,151 | 151,324,752 | 15.16 |
| Total interspersed repeats | 562,611,170 | 56.35 |
| Small RNA | 4,846 | 382,666 | 0.04 |
| Satellites | 192 | 130,399 | 0.01 |
| Simple repeats | 3,585 | 169,589 | 0.02 |
Table 3 A comparison of descriptive statistics assessing assembly quality, and repetitive element annotation and gene finding results

|          | LDD       | LDA       | LDJ       | LD652 (Hi-C) | LD652 |
|----------|-----------|-----------|-----------|--------------|-------|
| Size (bp)| 998,430,749 | 921,904,509 | 1,000,530,546 | 788,637,865 | 864,963,299 |
| Contig count | 4,622 | 8,189 | 11,303 | 134,187 | 135,446 |
| N50       | 661,876 | 212,365 | 137,117 | 5,068,179 | 249,594 |
| GC (%)    | 91,592 | 40,935 | 37,304 | 2,453 | 3,108 |
| LINE elements | 1,210,601 | 1,073,465 | 1,439,736 | 902,373 | 879,467 |
| BUSCO (%) | 25.40 | 25.74 | 38.51 | 35.18 | 35.25 |
| bp masked | 686,312,151 | 621,829,597 | 671,280,650 | 471,271,105 | 516,072,603 |
| bp masked (%) | 68.74 | 67.45 | 67.90 | 59.76 | 59.66 |
| Gold tier | 11,901 | 9,581 | 8,834 | 5,341 | 5,896 |
| Silver tier | 17,598 | 16,081 | 16,445 | 5,289 | 5,819 |
| Bronze tier | 26,773 | 27,583 | 32,052 | 13,715 | 14,789 |
| Non-podium | 18,746 | 21,614 | 26,86 | 18.96 | 17.34 |
| BUSCO (%) | 98.2 | 96.4 | 96.7 | 97.7 | 93.9 |

Assembly sources and methods used are as described in the Materials and Methods section. Reported gold-, silver- and bronze-tier, as well as non-podium level, values correspond to the number of unique protein-coding loci identified at the respective quality level by the gene annotation pipeline. Reported BUSCO values correspond to the fraction of complete (i.e., single-copy and duplicated) and fragmented BUSCOs detected.

LD652 genomes suggest that the longer PacBio reads were also more effective at recovering coding (and hence, typically less repetitive) regions of the LDD, LDA and LDJ genomes in a highly contiguous manner. Resequencing of LD652 using longer-read technologies will help with resolving this ambiguity.

In L. dispar, females constitute the heterogametic sex (Sahara et al. 2012), which in part motivated the decision to sequence a single female LDD pupa using long reads—representative sequence from both sex chromosomes would thereby be captured. Among Asian gypsy moths, females are also flight capable (Keena et al. 2008; Srivastava et al. 2021), whereas European gypsy moth females lack this capacity for flight. (Gypsy moth males are flight capable.) This results in a comparatively slower rate of host range expansion in North America for European gypsy moth populations, and a corresponding increase in concern for preventing the establishment of Asian gypsy moths. The genomic sequence from both sex chromosomes would thereby be captured.

**Conflicts of interest**
None declared.

**Literature cited**

Bao Z, Eddy SR. 2002. Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res. 12: 1269–1276.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 12:59–60.

Djoumad A, Nisole A, Zahiri R, Freschi L, Picq S, et al. 2017. Comparative analysis of mitochondrial genomes of geographic variants of the gypsy moth, Lymantria dispar, reveals a previously undescribed genotypic entity. Sci Rep. 7:14245.

Ellinghaus D, Kurtz S, Willhoeft U. 2008. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics. 9:18.

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 28: 3150–3152.

Goodwin RH, Tompkins GJ, McCawley P. 1978. Gypsy moth cell lines divergent in viral susceptibility. I. Culture and identification. In Vitro. 14:485–494.

Gremme G, Brendel V, Sparks ME, Kurtz S. 2005. Engineering a software tool for gene structure prediction in higher organisms. Inf Softw Technol. 47:965–978.

Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, et al. 2003. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 31: 5654–5666.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood FD, et al. 2013. De novo transcript sequence reconstruction from RNA-Seq reference generation and analysis with Trinity. Nat Protoc. 8: 1494–1512.

Hamelin RC, Roe AD. 2020. Genomic biosurveillance of forest invasive alien enemies: a story written in code. Evol Appl. 13:95–115.

Hebert FO, Freschi L, Blackburn G, Béliveau C, Dewar K, et al. 2019. Expansion of LINEs and species-specific DNA repeats drives genome expansion in Asian gypsy moths. Sci Rep. 9:16413.

Johnston JS, Bernardini A, Hjelmen CE. 2019. Genome size estimation and quantitative cytogenetics in insects. In: S Brown, M Pfrender, editors. Insect Genomics, Methods in Molecular Biology. New York, NY: Humana Press, p. 15–26.
