

Supplementary Information Appendix

The Gypsy moth genome provides insights into flight capability and virus-host interactions

Jing Zhang^a, Qian Cong^a, Emily A. Rex^b, Winnie Hallwachs^c, Daniel H. Janzen^c, Nick V. Grishin^{a,d,e} and Don B. Gammon^b

^aDepartment of Biophysics, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA

^bDepartment of Microbiology, University of Texas Southwestern Medical Center, Dallas, USA, TX 75390, USA

^cDepartment of Biology, University of Pennsylvania, Philadelphia, PA 19104, USA

^dDepartment of Biochemistry, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA

^eHoward Hughes Medical Institute, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA

To whom correspondence may be addressed:

Daniel H. Janzen
djanzen@sas.upenn.edu

Nick V. Grishin
grishin@chop.swmed.edu

Don B. Gammon
don.gammon@utsouthwestern.edu

This PDF file includes:

Supplementary text
References for SI reference citations

Other supplementary materials for this manuscript include the following:

Datasets S1 to S7

S1. Additional Methods.

Genome annotation

We used both RNA-seq- and homology-based methods to annotate protein-coding regions. For RNA-seq-based annotation, we used two pipelines: TopHat followed by cufflinks and Trinity followed by PASA were used as described (1). We also obtained eight sets of homology-based annotations by aligning proteins from *Drosophila* and seven other Lepidopteran genomes using exonerate as described (1). In addition, alignment of proteins from insects in the entire UniRef90 database (2) by genblastG (3) was used to obtain another gene prediction set. The *de novo* gene prediction software packages AUGUSTUS v2.6.1 (4), GlimmerHMM v3.0.1 (5) and SNAP v2006-07-28 (6) were also used for gene predictions as described (1). Finally, all sets of predictions were supplied to EvidenceModeller vr20120625 (7) to generate the final gene models as described (1).

We predicted the function of proteins by transferring annotations and GO-terms from the closest BLAST hits (E -value $< 10^{-5}$) in both the UniProt database (8) and Flybase (9). Finally, we performed InterproScan v5-440 (10) to identify conserved protein domains and functional motifs, predict coiled-coil domains, transmembrane and signal peptide motifs, to detect 3D structure templates, and to assign proteins to protein families and map them to metabolic pathways as described (1).

Transposons and repeats regions were identified by two methods, coverage-based annotation and RepeatModeler v1.0.11 as described (1). The regions that are covered three times more than average are identified as putative repetitive regions. Both repeats identified by coverage and RepeatModeler were submitted to the CENSOR server (11) to assign them to the repeat classification hierarchy as described (1). The species-specific repeat library and all repeats classified in RepBase v18.12 were used to mask repeats in the genome by RepeatMasker v4.0.5 as described (1).

Sex chromosomes

We mapped genomic sequence reads from male and female contigs to identify sex-linked contigs. Reads with MAPQ scores ≥ 40 were used to calculate coverage. To remove the transposons' effects on coverage calculation, we calculated the coverage only for gene regions instead of whole contigs. The median of the coverage was used to normalize the coverage of each gene. The distribution of normalized gene coverage ratios (male:female ratios, M:F ratios) was manually checked to empirically determine the thresholds for Z-linked (M:F ratio > 1.5) contigs. In addition, candidate Z-linked protein-coding genes were also identified through finding orthologs of Z-linked proteins of *B. mori* in *L. dispar* by reciprocal best hit blast method, considering the high degree of gene synteny between Lepidoptera species. Proteins identified by both methods were considered to be Z-linked in *L. dispar*.

Identification of orthologous proteins and gene expansion

We identified the orthologous groups from 18 Lepidopteran genomes using OrthoMCL v2.0.9 (12). If two OrthoMCL-defined orthologous groups overlapped in the *Drosophila* proteins that they mapped to, we merged them into one family. The total number and total length of proteins in a family were used to identify expanded gene families in *L. dispar*. If the total number and length of proteins from *L. dispar* in a family were >1.5x the average number and length across other Lepidopteran species, we considered this protein family to have undergone expansion in *L. dispar*. The most interesting gene expansions discussed in the paper were further investigated to include all relevant proteins using reciprocal BLAST results and function annotations. Proteins encoded by the genome but missed in the protein sets were predicted with the help of genblastG. Protein sequence from each family were aligned with MAFFT. Evolutionary trees were built with RAxML v8.2.6 (model: PROTGAMMAGTP) (13) and visualized in FigTree as described (1).

Population analysis of wild-caught specimens

We sequenced 26 specimens (Dataset S1) from five populations (continental Asia, Japan, Europe, Iran, and USA). The reads were trimmed by Trimmomatic v0.36 (14) and then aligned to the *L. dispar* genome using Burrows-Wheeler Aligner v0.7.13-r1126 with default settings (15). Alignments were then sorted by Picard v4 and PCR duplicates were removed (16).

Identification and analysis of diverged proteins

We used two criteria to identify diverged proteins between AGM and EGM that may be important for their phenotypic differences. First, we estimated the fixation index for both AGM and EGM using the following formula: $F_{ST} = (\pi_{between} - \pi_{within}) / \pi_{between}$, where $\pi_{between}$ is the average divergence between subspecies, and π_{within} is the average divergence within subspecies. Second, we detected all the positions that are conserved (sharing a common amino acid in over 80% of sequences) within but different between subspecies, and evaluated the statistical enrichment of such positions in each protein. The enrichment is quantified using a binomial test (p = rate of divergent positions in the alignment, m = the number of divergent positions in a protein, n = the total number of aligned positions in a protein). We identified diverged proteins with high fixation indices in both subspecies and enriched in divergent positions. We chose a cutoff of 0.1 for the fixation index and a p-value cutoff of <0.05 for enrichment of divergent positions, resulting in a set of diverged proteins. A similar strategy was used for identifying diverged proteins between European and North American EGM populations. Using the diverged set of proteins between EGM and AGM, we identified enriched GO terms ($p < 0.01$) associated with them using binomial tests (m = the number of diverged proteins that were associated with this GO term, N = number of diverged proteins, p = the probability for this GO term to be associated with any gene). Representative GO terms that are significantly enriched from this analysis were graphed in Figure 4C using REVIGO (17).

Cell and virus culture

LD652 cells were maintained in a 1:1 mixture of Ex-Cell 420 (Sigma) and Graces insect medium (Invitrogen) as described (18, 19). VSV infections were performed using VSV-LUC, which encodes a firefly luciferase reporter gene (18, 19). VSV-LUC stocks were amplified in baby hamster kidney (BHK) cells and titrated on African Green Monkey kidney cells (BSC-40) cells as described (18, 19). VACV infections were performed with VACV-FL-GFP, which encodes a firefly luciferase-green fluorescent protein (GFP) fusion protein (18). VACV-FL-GFP was amplified in BSC-40 cells and also titrated on these cells (18). Virus infections in LD652 cells were carried out for 2 h in Sf-900 II serum free media (Invitrogen) at 27°C (19). After 2 h, inocula were replaced with normal growth medium. AmEPV infections were performed with strain vAm Δ sph/gfp, which encodes GFP in the *spheroidin* gene locus (20). AmEPV was titrated on LD652 cells using a fluorescence-based plaque assay with an EVOS-FL fluorescence microscope (Invitrogen). The following multiplicities of infection (MOIs) were used: AmEPV (10), VACV-FL-GFP (50), VSV (10), which result in infection of 100% of cells (18, 19, 21).

Transcriptomic analyses of host response to infection

Total RNA was extracted from three independent mock- or virus-infected LD652 cell cultures 24 hpi using RNeasy isolation kits (Qiagen). Total mRNA was further isolated using NEBNext Poly(A) mRNA Magnetic Isolation. Libraries were constructed using NEBNext Ultra Directional RNA library prep kit for Illumina according to the manufacturer's protocol. Libraries were sequenced on an Illumina HiSeq2500 instrument. DEGs with adjusted p-values (q-values) <0.05 were considered significant and Protein products of DEGs were annotated based on their closest homolog from UniProt databases.

Additional Dataset S1 (separate file)

Specimens used in study and their heterozygosity.

Additional Dataset S2 (separate file)

L. dispar proteins that map to the *B. mori* Z chromosome.

Additional Dataset S3 (separate file)

Human orthologs of *L. dispar* proteins.

Additional Dataset S4 (separate file)

Significantly diverged proteins between European and North American *L. dispar dispar*.

Additional Dataset S5 (separate file)

Significantly diverged proteins between *L. dispar dispar* (EGM) and *L. dispar asiatica* (AGM).

Additional Dataset S6 (separate file)

Differentially expressed *L. dispar* genes 24 hpi.

Additional Dataset S7 (separate file)

KEGG Pathway Analyses of *L. dispar* DEGs after virus infection.

References

1. Cong Q, Li W, Borek D, Otwinowski Z, & Grishin NV (2018) The Bear Giant-Skipper genome suggests genetic adaptations to living inside yucca roots. *Mol. Genet. Genomics*.
2. Suzek BE, Huang H, McGarvey P, Mazumder R, & Wu CH (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* 23(10):1282-1288.
3. She R, *et al.* (2011) genBlastG: using BLAST searches to build homologous gene models. *Bioinformatics* 27(15):2141-2143.
4. Stanke M, Schoffmann O, Morgenstern B, & Waack S (2006) Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* 7:62.
5. Majoros WH, Pertea M, & Salzberg SL (2004) TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* 20(16):2878-2879.
6. Korf I (2004) Gene finding in novel genomes. *BMC Bioinformatics* 5:59.
7. Haas BJ, *et al.* (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* 9(1):R7.
8. UniProt C (2014) Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* 42(Database issue):D191-198.
9. St Pierre SE, Ponting L, Stefancsik R, McQuilton P, & FlyBase C (2014) FlyBase 102--advanced approaches to interrogating FlyBase. *Nucleic Acids Res.* 42(Database issue):D780-788.
10. Jones P, *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30(9):1236-1240.
11. Jurka J, Klonowski P, Dagman V, & Pelton P (1996) CENSOR--a program for identification and elimination of repetitive elements from DNA sequences. *Comput. Chem.* 20(1):119-121.
12. Li L, Stoeckert CJ, Jr., & Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13(9):2178-2189.
13. Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312-1313.
14. Bolger AM, Lohse M, & Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15):2114-2120.
15. Li H & Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26(5):589-595.
16. Li H, *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078-2079.
17. Supek F, Bosnjak M, Skunca N, & Smuc T (2011) REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* 6(7):e21800.

18. Rex EA, Seo, D., and Gammon, D.B (2018) Arbovirus infections as screening tools for identifying viral immunomodulatory proteins and host antiviral factors. *JoVE* In press.
19. Gammon DB, *et al.* (2014) A single vertebrate DNA virus protein disarms invertebrate immunity to RNA virus infection. *Elife* 3.
20. Li Q, Liston P, & Moyer RW (2005) Functional analysis of the inhibitor of apoptosis (iap) gene carried by the entomopoxvirus of *Amsacta moorei*. *J. Virol.* 79(4):2335-2345.
21. Li Y, Yuan S, & Moyer RW (1998) The non-permissive infection of insect (gypsy moth) LD-652 cells by Vaccinia virus. *Virology* 248(1):74-82.