Published in final edited form as:

Mol Genet Genomics. 2019 February ; 294(1): 211–226. doi:10.1007/s00438-018-1494-6.

The Bear Giant-Skipper genome suggests genetic adaptations to living inside yucca roots

Qian Cong², Wenlin Li², Dominika Borek², Zbyszek Otwinowski², and Nick V. Grishin^{1,2,#}

¹Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, Texas 75390-9050, USA.

²Department of Biophysics and Department of Biochemistry, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, Texas 75390-8816, USA.

Abstract

Giant-Skippers (Megathymini) are unusual thick-bodied, moth-like butterflies whose caterpillars feed inside Yucca roots and Agave leaves. Giant-Skippers are attributed to the subfamily Hesperini and they are endemic to southern and mostly desert regions of the North American continent. To shed light on the genotypic determinants of their unusual phenotypic traits, we sequenced and annotated a draft genome of the largest Giant-Skipper species, the Bear (*Megathymus ursus violae*). The Bear skipper genome is the least heterozygous among sequenced Lepidoptera genomes, possibly due to much smaller population size and extensive inbreeding. Their lower heterozygosity helped us to obtain a high-quality genome with an N50 of 4.2 Mbp. The ~430 Mb genome encodes about 14000 proteins. Phylogenetic analysis supports placement of Giant-Skippers with Grass-Skippers (Hesperini). We find that proteins involved in odorant and taste sensing as well as in oxidative reactions have diverged significantly in *Megathymus* as compared to *Lerema*, another Grass-Skipper. In addition, the Giant-Skipper has lost several odorant and gustatory receptors and possesses many fewer (1/3 ~ 1/2 of other skippers) anti-oxidative enzymes. Such differences may be related to the unusual life style of Giant-Skippers: they do not feed as adults, and their caterpillars feed inside Yuccas and Agaves, which provide a source of anti-oxidants such as polyphenols.

#Correspondence: Nick V. Grishin.

Qian Cong: qian.cong@utsouthwestern.edu

Wenlin Li: wenlin.li@utsouthwestern.edu

Dominika Borek: dominika@work.swmed.edu

Zbyszek Otwinowski: zbyszek@work.swmed.edu

Nick V. Grishin: grishin@chop.swmed.edu

AUTHORS' CONTRIBUTIONS

Q.C. and N.V.G. collected the specimens, Q.C. designed and carried out the experiments, performed the computational analyses and drafted the manuscript; W.L. performed the analysis of genome quality; D.B. and Z.O. designed and supervised experimental studies; N.V.G. directed the project and drafted the sections of the manuscript. All authors wrote the manuscript.

AVAILABILITY OF SUPPORTING DATA

See the Supplemental Information for the details of our protocols. Major scripts used in this projects and intermediate results are made available at <http://prodata.swmed.edu/LepDB/>.

COMPETING INTERESTS

The authors declare that they have no competing interests.

INTRODUCTION

Caterpillars of most butterflies feed on leaves from outside of the plants, while Giant-Skippers (family Hesperidae, tribe Megathymini) are highly unusual in their adaptation to living inside their food source as root borers or leaf miners (Freeman 1969; Roever 1975; Scott 1986; Zhang, Cong et al. 2017). The plants they inhabit are large, desert-dwelling succulents: yuccas with thick juicy roots and agaves with big fleshy leaves (Freeman 1969; Roever 1975; Scott 1986). Adaptation to internal feeding on large succulents has allowed Giant-Skippers to prosper in harsh, desert climates and become giants among their kin (Freeman 1969; Roever 1975; Scott 1986). Their large size, relatively small head (an adaptation for a caterpillar being able to turn around in a narrow burrow) and unique life history have led to an early classification of Giant-Skippers as a separate family (Freeman 1969; Roever 1975), which was later downgraded to a subfamily (Scott 1986). Previous studies based on several selected gene markers revealed a close phylogenetic connection between Giant-Skippers and other Grass-Skippers, confidently attributing them within the subfamily Hesperinae (Warren, Ogawa et al. 2008; Warren, Ogawa et al. 2009). Regardless of their taxonomic rank, Giant-Skippers are unique butterflies endemic to North America, and additional genomic studies promise to shed light on their evolution, unusual phenotypes and behavior.

Yucca-feeding Giant-Skippers are placed in the genus *Megathymus*. The largest representative of the genus is the Bear Giant-Skipper (*M. ursus*), frequently called the Ursine Giant-Skipper (Poling 1902; Stallings and Turner 1956; Freeman 1969; Wielgus, Wielgus et al. 1972; Roever 1975; Scott 1986) (Figs. 1–4). The Bear Giant-Skipper dwells in the Chihuahuan desert in southeastern Arizona, southern New Mexico, western Texas and north-central Mexico. It can be identified by its entirely white antennae. There are two main populations of *M. ursus* across its range. The nominotypical populations (*M. ursus ursus*) are distributed in the western part of the range (Poling 1902) and are characterized by yellow forewing bands separated by dark veins into spots with sharply defined edges. The skipper inhabits shaded mountain canyons and its caterpillars feed mostly on *Yucca schottii*. Adults fly once a year in the middle of summer, and the skipper spends most of the year as a caterpillar living inside its food plant. Eastern populations have orange continuous forewing bands with more diffuse edges (Fig. 1) and are classified as a different subspecies *M. ursus violae* (Stallings and Turner 1956). They prefer more open habitat on mesa-like mountains and slopes or shallow canyons (Fig. 3A–B), and caterpillars (Fig. 2C–J) typically feed on *Yucca torreyi* (Fig. 3C). Adults fly in late spring and early summer. In the northern parts of the range, both subspecies use *Yucca baccata* as the only available food plant.

Females of *M. ursus violae* lay eggs singly on leaves of young yucca plants. The egg is green when just laid, but turns beige with orange-brown blotchy spots in a couple of days (Fig. 2A). A caterpillar chews its way out through the apex of the egg, leaving a translucent egg shell behind (Fig. 2B). A pinkish caterpillar crawls to the tip of a leaf and webs a dome-like structure to cover itself. (Fig. 2C, D). It feeds on the leaf and deposits frass outside the silk dome (Fig. 2E). A later instar caterpillar bores into the center of the plant and burrows in towards the root. It maintains a “tent” or a “chimney”, a finger-shaped tube made of silk and frass particles (Fig. 3G) that protrudes from the center of the plant. The tent becomes larger

with time (Fig. 3C–F) and serves as a cocoon for pupation. A fully-grown caterpillar feeds inside the root on tissue and sap. It is grub-like and cream-colored (Fig. 2H, I) with black head (Fig. 2F) and anal plate (Fig. 2G). Prior to pupation the caterpillar secretes hydrophobic white powder through its glands on 10th and 11th sternites (Fig. 2J) into its tunnel and tent, in which it pupates (Fig. 2K). The silk covered in hydrophobic powder lines the pupation chamber (=cocoon) and protects it from water accumulation. The pupa has highly flexible abdominal segments and can move up and down the burrow by rotating the abdomen clockwise and counterclockwise (Scott 1986), which is unique to Giant-Skippers among butterflies. During cold nights with frost on the ground, pupae are deep in the root, and move up into the tent during hot days and “sun” themselves to accelerate their development with higher temperature. Eclosed adults crawl up from the tent and spread their wings hanging from a branch or a leaf (Fig. 4A). Males frequently “hilltop”, i.e., fly to the top of a hill or a mountain and perch on the ground and rocks (Fig. 4B), waiting for females.

Sequencing a high quality complete Giant-Skipper genome is the first and necessary step to understand the genomic determinants of their unique phenotypic traits. The genome can be compared to two other complete genomes of skippers we have sequenced previously: *Achalarus lyciades* from the subfamily Eudaminae and *Lerema accius* from the subfamily Hesperinae. Both of these skippers behave more like regular butterflies: their caterpillars feed on leaves and adults nectar on flowers. Comparative genomic analysis is expected to suggest hypotheses about phenotypic differences between the skipper species and shed light on their evolutionary history. The Bear Giant-Skipper is the largest and the most enigmatic representative of this group, so it was natural to select this species for genomic sequencing. We have chosen a specimen of *M. ursus violae* from west Texas for complete genome sequencing. Despite being the largest skipper, its genome is about the same size as that of a small Grass-feeding Skipper *Lerema accius* (Cong, Borek et al. 2015a). It has very low heterozygosity (0.1% vs. 1.5% in *L. accius*), which helped us obtain a high-quality assembly with a scaffold N50 of 4.2 Mb.

We compared the new genome with the genomes of Grass-Skipper *L. accius* and Spread-winged Skipper *A. lyciades* and revealed evidence for rapid divergence in odorant sensing mechanisms between *Megathymus* and *Lerema*, and a decrease in the number of anti-oxidative enzymes and odorant receptors in Giant-Skippers. Such observations may be related to the Giant-Skipper caterpillar living in food-plants that are particularly rich in antioxidants such as polyphenols (Cheeke, Piacente et al. 2006; Rizwan, Zubair et al. 2012) and that adults have a deteriorated proboscis and do not feed.

RESULTS

Assembly, annotation and quality of the Bear Giant-Skipper genome

We sequenced and assembled a 429 Mb genome draft of *Megathymus ursus violae* (*Mvi*), which is nearly the highest quality among currently available Lepidoptera genomes (International Silkworm Genome 2008; Duan, Li et al. 2010; Zhan, Merlin et al. 2011; Heliconius Genome 2012; You, Yue et al. 2013; Zhan and Reppert 2013; Ahola, Lehtonen et al. 2014; Tang, Yu et al. 2014; Cong, Borek et al. 2015b) according to LepBase (Challis, Kumar et al. 2016) (Fig. 5), coming second after *Heliconius erato* assemblies (Nadeau, Ruiz

et al. 2014; Challis, Kumar et al. 2016). The scaffold N50 of the *Mvi* genome assembly is 4.2 Mb, and the largest scaffold is 18.6 Mb, both better than the assembly of *Bombyx mori*, the classic model of Lepidoptera genomics (4.0 Mb and 16.2 Mb respectively, Fig. 5), but shorter than those of *Heliconius erato demophoon* (10.7 Mb and 23.9 Mb respectively). The *Mvi* draft assembly contains 1515 scaffolds (27 chromosomes (Freeman 1969)), compared to 43463 of *B. mori* (28 chromosomes (Pringle, Baxter et al. 2007)) and 196 of *H. erato* (21 chromosomes (Pringle, Baxter et al. 2007))). The genome assembly is similar to the best Lepidoptera genomes judged by the completeness of genes and proteins on the following benchmarks: Benchmarking Universal Single-Copy Orthologs (Waterhouse, Seppey et al. 2017) (BUSCO in Fig. 5), Core Eukaryotic Genes Mapping Approach (CEGMA) genes (Table S1) (Parra, Bradnam et al. 2007), cytoplasmic ribosomal proteins and independently assembled transcripts (Table 1). We deposited the Bear Giant-Skipper whole genome project at DDBJ/EMBL/GenBank under the accession PDGT000000000. The version described here is PDGT000000000. Furthermore, major results from the genome assembly, annotation and analysis can be downloaded from <http://prodata.swmed.edu/LepDB/>.

Next, we assembled the transcriptome of *Mvi* from a caterpillar collected at the same locality as the adult used for genomic sequencing. Using a combination of this transcriptome, available protein sequences from other Lepidoptera and *Drosophila melanogaster*, *Mvi* genomic sequence, *de novo* gene predictions, and repeats identification (Table S2), we predicted that the *Mvi* genome encodes 14024 proteins. We were able to annotate the functions of 10703 of these protein-coding genes (Table S3). The gene number in the *Mvi* genome is 10–20% smaller than that in the other two Skippers that we sequenced and annotated using the same pipeline (Table 1) (Cong, Borek et al. 2015a; Shen, Cong et al. 2017). Detailed analysis of protein families also revealed that the Giant-Skipper genome encodes fewer genes compared to other skippers (discussed below).

Phylogeny of Lepidoptera

We found orthologous proteins encoded in 14 Lepidoptera genomes (*Plutella xylostella*, *Bombyx mori*, *Manduca sexta*, *Megathymus ursus violae*, *Lerema accius*, *Achalarus lyciades*, *Danaus plexippus*, *Heliconius melpomene*, *Melitaea cinxia*, *Calycopis cecrops*, *Phoebis sennae*, *Pieris rapae*, *Papilio xuthus*, and *Pterourus glaucus*) and identified 5089 orthologous groups, from which 2117 groups have only a single gene in each of the species. Using RAXML, we constructed a phylogenetic tree from the concatenated alignment of these single-copy orthologs. In addition, we also performed coalescent-based species tree construction using ASTRAL based on individual gene trees (Mirarab, Reaz et al. 2014). The resulting species tree resembles the one obtained from a single concatenated alignment both in topology and the confidence of each node.

The tree places *Megathymus* as the sister to *Lerema* (Fig. 6), the sole member of the Hesperinae subfamily with a sequenced genome, and confirms the described close relationship between Giant- and other Grass-Skippers (Warren, Ogawa et al. 2008; Warren, Ogawa et al. 2009; Zhang, Cong et al. 2017). This placement was confident (100% bootstrap both on the entire dataset consisting of 355,045 positions and on 2% of positions selected from this dataset, see below). The lengths of the branches from *Lerema* and *Megathymus* to

their last common ancestor are comparable to the length of the branch from this ancestor to the last common ancestor of *Lerema*, *Megathymus* and *Achalarus*, indicating relatively close relationship between Giant-Skippers (*Megathymus*) and other Grass-Skippers (*Lerema*).

As in previously published genomic trees (Cong, Borek et al. 2015a; Shen, Cong et al. 2017), Papilionidae is placed as a sister to all other butterflies, including skippers (Hesperiidae) (Fig. 6). This topology contradicts the traditional, morphology-based phylogenetic view, but is supported by maximum-likelihood and Bayesian DNA-based trees published recently (Mutanen, Wahlberg et al. 2010; Heikkilä, Kaila et al. 2012; Kawahara and Breinholt 2014). When the concatenated alignment of all 2117 single-copy orthologs was used, all nodes received 100% bootstrap support. However, the nodes of trees constructed from very large sequence datasets could get 100% support for incorrect topology (Kubatko and Degnan 2007). Bootstrap measures the internal consistency of a phylogenetic signal in the alignment and not the correctness of the tree per se, and the tree may be biased by long-branch attraction and nucleotide composition. To detect nodes with the weakest support, we split the concatenated alignment of single-copy orthologs into 50 alignments (7101 positions in each alignment), thus reducing the amount of data used for the tree. The consensus tree built from these alignments gave the lowest support (88%) to the node referring to the relative position of swallowtails and skippers. Thus, the branching of skippers, swallowtails and other butterflies remains to be further investigated when better taxon sampling of complete genomes is achieved. Next, the position of *Calycopis* (family Lycaenidae) in the tree is less strongly supported (96%) than other nodes, which possibly results from its elevated evolutionary rate: *C. cecrops* forms a long branch compared to other species included in the tree (Fig. 6) (Cong, Shen et al. 2016; Cong, Shen et al. 2017a; Pellissier, Kostikova et al. 2017). However, this placement of Lycaenidae as a sister to Nymphalidae (represented by *Danaus*, *Heliconius* and *Melitaea*) in our tree agrees with morphology and other DNA evidence (Mutanen, Wahlberg et al. 2010; Heikkilä, Kaila et al. 2012; Cong, Shen et al. 2016; Cong, Shen et al. 2017c).

Genomic divergence between Giant-Skippers and other Grass-Skippers

As phylogenetic analysis indicates, Giant-Skippers (represented by *Megathymus ursus violae*) are closely related to other Grass-Skippers (represented by *Lerema accius*). However, they prominently differ in their morphology and life history. To study possible genomic determinants of these differences, we identified and analyzed orthologous proteins that are more divergent between *Megathymus* and *Lerema* than between *Lerema* and *Achalarus*. The logic of this criterion is as follows. In accord with the phylogenetic tree (Fig. 6), the majority (88.3%) of orthologs show higher sequence identity between *Lerema* and *Megathymus* than between *Lerema* and *Achalarus*. However, we hypothesized that Giant-Skippers experienced accelerated evolution in certain genomic regions resulting in rapid morphological and ecological differentiation: unique morphology such as larger body size with fat abdomen, caterpillar feeding inside roots of their food plants, lack of ability to feed as adults, among many others. Thus, the *Megathymus* genes that diverged more rapidly compared to *Lerema* and *Achalarus* would likely correlate with unique traits of Giant-Skippers, and we can use the divergence between *Lerema* and *Achalarus* as a reference point for each orthologous

group. For the most strongly divergent proteins in *Megathymus*, we expect that their sequence identity to *Lerema* may be lower than the identity between *Lerema* and *Achalarus*.

We found 991 such proteins and analyzed their predicted functions using GO terms (Fig. 7 and Table S4) (Ashburner, Ball et al. 2000). We found proteins involved in sensory organ precursor cell division (orange box in Fig. 7B), odorant and taste sensing (green box in Fig. 7A and magenta box in Fig. 7B), synthesis of fatty acids (red box in Fig. 7A), and oxidative reactions (blue box in Fig. 7A) are significantly (P -value < 0.01) over-represented in these 991 proteins.

Gene expansion events are rare in *Megathymus*

Gene expansions are common mechanisms of adaptation to evolve unique traits of certain organisms (Cong, Borek et al. 2015a; Cong, Borek et al. 2015b). We searched for gene expansions and found that such events are rare in *Megathymus* (Table S5) in comparison to other Lepidoptera species. For example, using the same criteria and dataset, we identified 50 families with gene expansions in *Megathymus* compared to 73 and 146 families for *Achalarus* and *Lerema*, respectively. The smaller number of expanded families is consistent with the overall smaller number of predicted genes in the *Megathymus* genome. Compared to *Achalarus* and *Lerema* genomes, the *Megathymus* genome encodes lower numbers of proteins from several families.

The *Megathymus* genome encodes fewer odorant receptors (Fig. 8A) compared to other Skippers (49, 56, and 58 copies for *Megathymus*, *Lerema* and *Achalarus*, respectively). In addition, *Megathymus* has fewer proteins to protect them from oxidative damage, such as peroxidases and catalases. *Achalarus* underwent gene expansion in peroxidases (Fig. 8B) while *Lerema* expanded in catalases (Fig. 8C). However, a similar gene expansion is absent in *Megathymus*. Finally, we see that a large expansion in chitinase-like proteins that we hypothesized may act as cellulases (Cong, Borek et al. 2015a; Shen, Cong et al. 2017) is still unique for *Lerema*, and is not present in either *Achalarus* or *Megathymus* (Fig. 8D). Unique expansion of putative cellulases may help Grass-Skippers to digest cellulose-rich but nutrient poor grasses, and such an adaptation is not needed for *Megathymus*, whose food plants are rich in carbohydrates other than cellulose.

DISCUSSION

Comparative genomics of butterflies suggests correlation between genotype and phenotype

With the advent of genomic sequencing, it became possible to obtain high quality, nearly complete genomic assemblies for essentially any organism (Palkopoulou, Mallick et al. 2015; Read, Petit et al. 2017; Li, Zhu et al. 2018). Butterflies are good subjects for genomic exploration due to their relatively small genomes (Talla, Suh et al. 2017), which are typically not more than a third of the size of the human genome. However, butterfly genomic assembly is challenging due to very high (up to 5%) heterozygosity (Cong, Shen et al. 2017b) compared to mammals (about 0.1%). Current assembly software has difficulty distinguishing between parental copies of the same genes and recently duplicated genes,

resulting in poor quality genomes (Tigano, Sackton et al. 2018). We were able to overcome these difficulties (Cong, Borek et al. 2015a; Cong, Borek et al. 2015b) and assemble good quality genomes for several butterfly species, three of which are skippers (Hesperiidae).

There are several reasons to obtain and study butterfly genomes. First, genomic sequence is the ultimate guide to evolution of these species. Phylogenetic analysis performed on a genomic scale is expected to produce the most accurate trees (Jarvis, Mirarab et al. 2014; Foley, Springer et al. 2016), and analyses of gene duplications and losses shed light on species evolution. Second, genome-scale comparative analysis reveals non-standard and poorly studied evolutionary phenomena such as incomplete lineage sorting and introgression (Thawornwattana, Dalquen et al. 2018). Third, and arguably most important, comparative genomics promises to elucidate genotypic determinants of phenotypic traits: wing patterns, ecological preferences, life histories and behavior (Kunte, Zhang et al. 2014; Janzen, Burns et al. 2017). Such comparative analyses projected on the biology of the species involved suggest hypotheses about molecular bases of phenotypes. These hypotheses would be more powerful when we obtain hundreds and thousands of genomic sequences. However, even with 3 sequenced skipper genomes today, we can make some interesting correlative observations about odorant, taste sensing and anti-oxidant proteins that could be linked to biological features of the skippers. It is important to understand that these are correlative hypotheses that need experimental exploration using genetic and genomic engineering to substantiate them fully.

Deteriorated odorant and gustatory sensing in *Megathymus*

Our study reveals signs of deteriorated odorant and gustatory sensing ability of *Megathymus*. On the one hand, we find that proteins involved in sensory organ precursor cell division, odorant, and taste sensing are significantly enriched among proteins that have been rapidly diverging in *Megathymus* compared to *Lerema* and *Achalarus*. On the other hand, the *Megathymus* genome encodes fewer odorant receptors compared to other Skippers. These observations are in agreement with the phenotypic and behavioral traits of *Megathymus*. First, their adults do not feed (Scott 1986) and therefore the odorant and gustatory receptors involved in adult feeding may have deteriorated. Second, many other butterflies are typically highly specific to their food plants and would rather starve to death than accept a different plant (Scott 1986). *Megathymus* caterpillars will accept many other roots instead of Yucca, such as Manioc (pers. observation) or even potatoes (Pettersson and Wielgus 1973; Minno 1994). Third, while other skippers may need more advanced odorant receptor system to aid their free-wandering caterpillars in selecting appropriate food plants that enable them to complete their development, *Megathymus* caterpillars mostly live inside its food plant throughout their life.

Food plant probably protect *Megathymus* from oxidative damage

We find that *Megathymus* has fewer proteins to protect them from oxidative damage. In addition, *Megathymus*' proteins involved in oxidative reactions tend to diverge more rapidly from *Lerema* than those from *Achalarus*. These observations may be related to the unique root-feeding life style of *Megathymus* caterpillar. The local oxygen and carbon dioxide concentrations inside the plant may differ from the outside, and the caterpillar may

submerge in the juice from the plants more frequently. These unique conditions may require enzymes involved in oxidative reactions to evolve fast and adapt to the environment. Meanwhile, their caterpillar's food plants, especially Yucca, are particularly rich in anti-oxidative chemicals such as polyphenols (Cheeke, Piacente et al. 2006). These chemicals may be able to protect the caterpillar against possible oxidative damage, relieving the evolutionary pressure to maintain many copies of anti-oxidative enzymes as in other skippers.

CONCLUSIONS

We obtained and comparatively analyzed a 430 Mb genomic sequence of the Bear Giant-Skipper (*Megathymus ursus violae*). We found that its genome is the least heterozygous among Lepidoptera with available genomic sequences and suggest that the reason may be its smaller population size and inbreeding. Due to lower heterozygosity, the assembly software, Platanus and Allpaths-LG, produced a high quality genome with N50 of 4.2 Mbp. About 14,000 proteins are encoded in the Bear Skipper genome. Comparison with the genomes of two other skippers we sequenced previously revealed that odorant and taste-sensing proteins and those involved in oxidative reactions diverged significantly in *Megathymus*. Furthermore, we find that Giant-Skipper genome has lost a number of proteins present in other butterflies, in particular, odorant and gustatory receptors and anti-oxidative enzymes. We connect these differences to unusual life histories of Giant-skippers whose adults do not feed (less need for odorant and gustatory receptors) and whose caterpillars feed and live inside anti-oxidant-rich Yuccas and Agaves (less need for anti-oxidative enzymes).

METHODS

Library preparation and sequencing

We took 2 legs and cut out a piece of muscle from a thorax of freshly eclosed *Megathymus ursus violae* male, voucher NVG-1504 (USA: TX, Pecos Co., Glass Mtns., ex larva found 4-May, pupated 7-May, eclosed 7-Jun-2013) and spread the specimen (Fig. 1). The specimen will be deposited in the National Museum of Natural History, Smithsonian Institution, Washington, DC, USA (USNM). Genomic DNA from specimen NVG-1504 was extracted using the ChargeSwitch gDNA mini tissue kit from ThermoFisher. We prepared 250 bp and 500 bp paired-end libraries with enzymes from NEBNext Modules following the Illumina TruSeq DNA sample preparation guide. We prepared 2 kbp, 6 kbp and 15 kbp mate pair libraries using a protocol similar to previously published Cre-Lox-based method (Van Nieuwerburgh, Thompson et al. 2012). For the 250 bp, 500 bp, 2 kbp, 6 kbp and 15 kbp libraries, we used approximately 250 ng, 250 ng, 2 µg, 3 µg and 5 µg of DNA, respectively. The amount of DNA from all the libraries was quantified with Qubit™ dsDNA HS Assay Kit from ThermoFisher, and we mixed 250 bp, 500 bp, 2 kbp, 6 kbp, 15 kbp libraries at relative molar concentration 40:20:8:4:3. The mixed library was sequenced for 150 bp at both ends on two lanes of Illumina HiSeq2500 at the UT Southwestern genomics core facility.

A caterpillar that we collected at the same locality as NVG-1504 on 30-Nov-2013 was used to extract RNA with the QIAGEN RNeasy Mini Kit. mRNA was further isolated using

NEBNext Poly(A) mRNA Magnetic Isolation Module and we prepared RNA-seq libraries using NEBNext Ultra Directional RNA Library Prep Kit for Illumina per manufacturer's protocol. We carried out paired-end sequencing of RNA-seq library for 150 bp on 1/8 (pooled with other samples from other projects) of an Illumina HiSeq2500 lane.

Genome and transcriptome assembly

We processed mate pair libraries using the Delox script (Van Nieuwerburgh, Thompson et al. 2012) to remove the loxP sequences and to separate true mate pair from paired-end reads. For all reads, including the ones for RNA-seq, we used mirabait (version: 3.4.0, parameters: -i -k 20) (Chevreux, Wetter et al. 1999) to remove contamination from the TruSeq adapters, an in-house script, quality_trim, to remove low quality portions (quality score < 20) at the ends of both reads. We used JELLYFISH (version: 1.1.2, k-mer length: 19) (Marcais and Kingsford 2011) to compute k-mer frequencies in all genomic DNA libraries, and QUAKE (version: 0.3, k-mer length: 19) (Kelley, Schatz et al. 2010) to correct sequencing errors. This data processing produced seven datasets used in genome assembly: 250 bp and 500 bp paired-end libraries, 2 kbp, 5 kbp, 10 kbp true mate pair libraries, a dataset consisting of all the paired-end reads from the mate pair libraries, and a single-end dataset containing all reads whose pairs were removed in the process. We used these libraries as input for the two *de novo* genome assemblers, Platanus (version: 1.2.1, default parameters) (Kajitani, Toshimoto et al. 2014) and Allpaths-LG (version r43762, default parameters) (Gnerre, Maccallum et al. 2011). The resulting assemblies were merged using Metassembler (Wences and Schatz 2015) (version: 1.5, mateAn parameters: -A 4000 -B 14000 for 10 kbp library, -A 2000 -B 8000 for 5 kbp library, -A 1000 -B 3000 for 2 kbp library; nucmer parameters: -c 20 -l 50)

The processed RNA-seq reads were assembled using three procedures: (1) *de novo* assembly by Trinity (Haas, Papanicolaou et al. 2013) (version: 20140413p1), (2) reference-based assembly by TopHat (Kim, Pertea et al. 2013) (v2.0.10, parameters: --read-edit-dist 5 --fusion-read-mismatches 3 --segment-mismatches 3 --read-mismatches 4 --read-gap-length 4 --read-realign-edit-dist 0 --mate-inner-dist 100 --mate-std-dev 50 --solexa1.3-quals --coverage-search --b2-sensitive --library-type frunstranded) and Cufflinks (Roberts, Pimentel et al. 2011) (v2.2.1), and (3) reference-guided assembly by Trinity (parameters: --normalize_reads --genome_guided_max_intron 100000). We used the Program to Assemble Spliced Alignment (PASA, version: r20130907, parameter: --ALIGNERS blat,gmap) (Haas, Salzberg et al. 2008) to integrate the results from all these three methods.

Repeats identification and gene annotation

We used two approaches to identify repeats in the genome: the RepeatModeler (Smit and Hubley 2008–2010) (version: 1.0.11) pipeline and in-house scripts to extract genomic segments with 3 times higher than expected coverage. We submitted these repeats to the CENSOR (Jurka, Klonowski et al. 1996) server (<http://www.girinst.org/censor/>, sequence resource: all) to classify them, and these repeats, in addition to the repeats in RepBase (Jurka, Kapitonov et al. 2005) (V18.12), were used to mask repeats in the *Mvi* genome by RepeatMasker (version: 4.0.5, parameter: -div 30) (Smit, Hubley et al. 1996–2010).

Transcript-based gene annotations were obtained from two pipelines: Trinity followed by PASA and TopHat followed by Cufflinks. Furthermore, four sets of homology-based annotations were obtained by aligning protein sets from *Drosophila melanogaster* (Misra, Crosby et al. 2002) and 3 Lepidoptera genomes (*Bombyx mori*, *Danaus plexippus* and *Heliconius melpomene*) to the *Megathymus ursus violae* genome using exonerate (version: 2.2.0, parameters: -model protein2genome -percent 30) (Slater and Birney 2005). We used insects proteins in the entire UniRef90 (Suzek, Huang et al. 2007) database to generate additional set of gene predictions by genblastG (version: 1.0.138, parameters: -g T -v 2 -c 0.5 -e 0.00001 -s 0) (She, Chu et al. 2011). 500 confident gene models were manually curated and selected by integrating the evidence from homologs and transcripts to train *de novo* gene predictors: AUGUSTUS (version 2.6.1) (Stanke, Schoffmann et al. 2006), SNAP (version 2006-07-28) (Korf 2004) and GlimmerHMM (version 3.0.1) (Majoros, Pertea et al. 2004). These predictors trained on our data, the self-trained (parameter: -max_nnn 1000) Genemark (version 2.3c) (Besemer and Borodovsky 2005) and a consensus-based pipeline Maker (version: 2.26) (Cantarel, Korf et al. 2008), were used to obtain another five sets of gene models. We supplied these homology-based and transcript-based and annotations to SNAP, AUGUSTUS and Maker to improve their performance. Overall, 11 sets of gene predictions were generated and integrated with EvidenceModeller (version: r20120625, weights: PROTEIN exonerate_Heliconius 4, PROTEIN exonerate_Danaus 8, PROTEIN exonerate_Bombyx 4, PROTEIN genBlastG_uniprot 4, TRANSCRIPT TOPHAT 10, TRANSCRIPT PASA 10, ABINITIO_PREDICTION maker 5, ABINITIO_PREDICTION augustus 4, ABINITIO_PREDICTION snap 3, ABINITIO_PREDICTION genemark 2, ABINITIO_PREDICTION GlimmerHMM 1; parameters: --search_long_introns 1, --re_search_intergenic 1) (Haas, Salzberg et al. 2008) to produce the final gene models.

Functions of *Megathymus* proteins were predicted by transferring annotations and GO-terms from the closest BLAST (version: 2.2.31+) (Altschul, Gish et al. 1990) hits (e-value < 0.00001) in the Swissprot (UniProt 2014) database and Flybase (St Pierre, Ponting et al. 2014). Finally, InterproScan (version: 5-44.0, parameters: -dp -goterms -pa) (Jones, Binns et al. 2014) was run to identify signal peptides, transmembrane helices, and coiled coils in the protein; and to assign proteins to protein families and map them to metabolic pathways, to find conserved protein domains and functional motifs, and to detect homologous 3D structures.

Orthologs identification and phylogenetic tree construction

OrthoMCL (version 2.0.9) (Li, Stoeckert et al. 2003) was used to identify the orthologous groups from 14 Lepidoptera genomes. Only a single gene from every species was present in 2117 orthologous groups. These groups were used for phylogenetic analysis. We used both global sequence aligner MAFFT (version: 7.299, parameters: --maxiterate 1000--genafpair) (Katoh and Standley 2013) and local sequence aligner BLASTP to build alignments for each single-copy orthologous group. From each individual alignment, we extracted the positions that were aligned consistently by both aligners and concatenated these positions to obtain an alignment of 355,045 positions.

We used RAXML (version: 8.2.6, model: PROTGAMMAAUTO) (Stamatakis 2014) to construct phylogenetic tree from this concatenated alignment and performed bootstrap resampling of the aligned positions to assign the confidence level of each node in the tree. Furthermore, in order to find the weakest nodes in the tree, we split the concatenated alignment into 50 alignments (about 7,101 positions in each alignment) to reduce the amount of data, and applied RAXML to each alignment. A 50% majority rule consensus tree was obtained and confidence level was assigned to each node in this tree as the percent of individual trees supporting this node. In addition, we selected the orthologous groups with at least 100 aligned positions (positions with any gaps are removed), and used these 1258 protein alignments to construct 1258 trees using RAXML (model: PROTGAMMAAUTO). These trees were summarized using ASTRAL (version 4.9.0) (Mirarab, Reaz et al. 2014).

Identification of fast-evolving protein in *Megathymus* compared to Grass-Skippers

We identified the fast-evolving genes in *Megathymus*, which may be related to its unique phenotypic traits compared to other Grass-Skippers. First, we started from the exons of *Lerema* proteins and searched for orthologous segments in the *Megathymus* and *Achalarus* genomes using TBLASTN (version: 2.2.31+). We ranked the confident hits (e-value < 0.01) and considered the top-ranking segment to be orthologous to the query if it covers more than 80% of its length and the sequence identity to the query at the protein level is above 40% and more than 10% higher than the second hit. Second, for each *Lerema* protein, we combined the orthologs of its exons to obtain an orthologous group, and the TBLASTN alignments were used for sequence comparison. We totally identified 9620 single-copy orthologous groups among the three skippers, and they were used for the following analysis.

Third, for each orthologous group, we evaluated the divergence between *Megathymus* and *Lerema* using an index of divergence that was defined by the following formula: $Index_{div} = Identity_{MvsL} / Identity_{AvsL}$, where $Identity_{MvsL}$ was the sequence identity between *Megathymus* and *Lerema*, and $Identity_{AvsL}$ represented the sequence identity between *Achalarus* and *Lerema*. Most proteins showed higher sequence identity between *Megathymus* and *Lerema* ($Index_{div} > 1$), as one would expect from the phylogeny of skippers. We identified orthologous groups with $Index_{div}$ below a certain cutoff (1, 0.98, 0.95, and 0.90), and considered them (namely, divergent groups) to be more likely related to the functional divergence between *Megathymus* and other Grass-Skippers. Fourth, the function of these divergent groups was analyzed using GO terms. We identified enriched GO terms associated with these proteins using binomial tests implemented in Python scipy package (https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.binom_test.html): N = number of divergent groups, m = the number of divergent groups associated with this GO term, p = the probability that this GO term is associated with any orthologous group. We used different cutoffs for $Index_{div}$: 1, 0.98, 0.95, and 0.90, and the significantly enriched GO-terms identified at each cutoff were combined.

Comparison of the loss and gain of genes among skippers

Starting from the orthologous groups from 14 Lepidoptera genomes identified by OrthoMCL (Li, Stoeckert et al. 2003), we classified Lepidoptera proteins into families by merging orthologous groups (defined by OrthoMCL) that had the same best hits in

Drosophila. We counted the number of proteins from each skipper species in each family. The families with well-defined functions and in which the number of proteins differed the most between *Megathymus* and any other skipper were carefully studied. For these families, we detected all relevant proteins from the genomes by reciprocal BLAST (-e 0.00001) searches and their function annotations. Proteins encoded in genomes and found in genomic assemblies but missed in the annotated protein sets were predicted using genblastG (-c 0.5 -e 0.00001). We used MAFFT to align protein sequences from each family. We constructed evolutionary trees using RAXML (model: PROTGAMMAAUTO) and visualized them in FigTree.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENT

We thank Lisa N. Kinch for suggestions and proofreading of the manuscript. We are grateful to Texas Parks and Wildlife Department (Natural Resources Program Director David H. Riskind) for the research permit #08–02Rev. Qian Cong was a Howard Hughes Medical Institute International Student Research fellow when these studies were performed. We thank Greg M. Lasley for the photograph of a live male shown in Fig 4b.

This work was funded in part by the National Institutes of Health (GM094575 to NVG) and the Welch Foundation (I-1505 to NVG). Authors declare that they have no conflict of interest. This article does not contain any studies with human participants performed by any of the authors. All applicable international, national, and institutional guidelines for the care and use of animals were followed.

REFERENCES

- Ahola V, Lehtonen R, et al. (2014) The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera. *Nature communications* 5: 4737. doi 10.1038/ncomms5737
- Altschul SF, Gish W, et al. (1990) Basic local alignment search tool. *Journal of molecular biology* 215(3): 403–410. doi 10.1016/S0022-2836(05)80360-2 [PubMed: 2231712]
- Ashburner M, Ball CA, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* 25(1): 25–29. doi 10.1038/75556 [PubMed: 10802651]
- Besemer J and Borodovsky M (2005) GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic acids research* 33(Web Server issue): W451–454 [PubMed: 15980510]
- Cantarel BL, Korf I, et al. (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome research* 18(1): 188–196 [PubMed: 18025269]
- Challis RJ, Kumar S, et al. (2016) Lepbase: the Lepidopteran genome database. *bioRxiv*. doi 10.1101/056994
- Cheeke PR, Piacente S, et al. (2006) Anti-inflammatory and anti-arthritis effects of *Yucca schidigera*: a review. *Journal of inflammation* 3: 6. doi 10.1186/1476-9255-3-6 [PubMed: 16571135]
- Chevreaux B, Wetter T, et al. (1999) Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. *Computer Science and Biology: Proceedings of the German Conference on Bioinformatics* 99: 45–56
- Cong Q, Borek D, et al. (2015a) Skipper genome sheds light on unique phenotypic traits and phylogeny. *BMC genomics* 16: 639. doi 10.1186/s12864-015-1846-0 [PubMed: 26311350]
- Cong Q, Borek D, et al. (2015b) Tiger Swallowtail Genome Reveals Mechanisms for Speciation and Caterpillar Chemical Defense. *Cell reports*. doi 10.1016/j.celrep.2015.01.026
- Cong Q, Shen J, et al. (2017a) When COI barcodes deceive: complete genomes reveal introgression in hairstreaks. *Proceedings Biological sciences* 284(1848). doi 10.1098/rspb.2016.1735

- Cong Q, Shen J, et al. (2016) Complete genomes of Hairstreak butterflies, their speciation, and nucleomitochondrial incongruence. *Sci Rep* 6: 24863. doi 10.1038/srep24863 [PubMed: 27120974]
- Cong Q, Shen J, et al. (2017b) The first complete genomes of Metalmarks and the classification of butterfly families. *Genomics* 109(5–6): 485–493. doi 10.1016/j.ygeno.2017.07.006 [PubMed: 28757157]
- Cong Q, Shen J, et al. (2017c) The first complete genomes of Metalmarks and the classification of butterfly families. *Genomics*. doi 10.1016/j.ygeno.2017.07.006
- Duan J, Li R, et al. (2010) SilkDB v2.0: a platform for silkworm (*Bombyx mori*) genome biology. *Nucleic acids research* 38(Database issue): D453–456 [PubMed: 19793867]
- Foley NM, Springer MS, et al. (2016) Mammal madness: is the mammal tree of life not yet resolved? *Philos Trans R Soc Lond B Biol Sci* 371(1699). doi 10.1098/rstb.2015.0140
- Freeman HA (1969) Systematic review of the Megathymidae. *J Lep Soc* 23(Suppl. 1): 1–59
- Gnerre S, Maccallum I, et al. (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences of the United States of America* 108(4): 1513–1518. doi 10.1073/pnas.1017351108 [PubMed: 21187386]
- Haas BJ, Papanicolaou A, et al. (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols* 8(8): 1494–1512 [PubMed: 23845962]
- Haas BJ, Salzberg SL, et al. (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome biology* 9(1): R7 [PubMed: 18190707]
- Heikkilä M, Kaila L, et al. (2012) Cretaceous origin and repeated tertiary diversification of the redefined butterflies. *Proceedings Biological sciences* 279(1731): 1093–1099. doi 10.1098/rspb.2011.1430 [PubMed: 21920981]
- Heliconius Genome C (2012) Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* 487(7405): 94–98 [PubMed: 22722851]
- International Silkworm Genome C (2008) The genome of a lepidopteran model insect, the silkworm *Bombyx mori*. *Insect biochemistry and molecular biology* 38(12): 1036–1045 [PubMed: 19121390]
- Janzen DH, Burns JM, et al. (2017) Nuclear genomes distinguish cryptic species suggested by their DNA barcodes and ecology. *Proceedings of the National Academy of Sciences of the United States of America* 114(31): 8313–8318. doi 10.1073/pnas.1621504114 [PubMed: 28716927]
- Jarvis ED, Mirarab S, et al. (2014) Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* 346(6215): 1320–1331. doi 10.1126/science.1253451 [PubMed: 25504713]
- Jones P, Binns D, et al. (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30(9): 1236–1240 [PubMed: 24451626]
- Jurka J, Kapitonov VV, et al. (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research* 110(1–4): 462–467 [PubMed: 16093699]
- Jurka J, Klonowski P, et al. (1996) CENSOR--a program for identification and elimination of repetitive elements from DNA sequences. *Computers & chemistry* 20(1): 119–121 [PubMed: 8867843]
- Kajitani R, Toshimoto K, et al. (2014) Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome research* 24(8): 1384–1395 [PubMed: 24755901]
- Katoh K and Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 30(4): 772–780. doi 10.1093/molbev/mst010 [PubMed: 23329690]
- Kawahara AY and Breinholt JW (2014) Phylogenomics provides strong evidence for relationships of butterflies and moths. *Proceedings Biological sciences* 281(1788): 20140970. doi 10.1098/rspb.2014.0970
- Kelley DR, Schatz MC, et al. (2010) Quake: quality-aware detection and correction of sequencing errors. *Genome biology* 11(11): R116 [PubMed: 21114842]
- Kim D, Pertea G, et al. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology* 14(4): R36 [PubMed: 23618408]

- Korf I (2004) Gene finding in novel genomes. *BMC bioinformatics* 5: 59 [PubMed: 15144565]
- Kubatko LS and Degnan JH (2007) Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst Biol* 56(1): 17–24. doi 10.1080/10635150601146041 [PubMed: 17366134]
- Kunte K, Zhang W, et al. (2014) doublesex is a mimicry supergene. *Nature* 507(7491): 229–232. doi 10.1038/nature13112 [PubMed: 24598547]
- Li L, Stoeckert CJ, Jr., et al. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research* 13(9): 2178–2189. doi 10.1101/gr.1224503 [PubMed: 12952885]
- Li S, Zhu S, et al. (2018) The genomic and functional landscapes of developmental plasticity in the American cockroach. *Nature communications* 9(1): 1008. doi 10.1038/s41467-018-03281-1
- Majoros WH, Pertea M, et al. (2004) TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* 20(16): 2878–2879 [PubMed: 15145805]
- Marcais G and Kingsford C (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27(6): 764–770 [PubMed: 21217122]
- Minno MC (1994) Immature Stages Of The Skipper Butterflies (Lepidoptera: Hesperidae) Of The United States; Biology, Morphology, And Descriptions. University of Florida
- Mirarab S, Reaz R, et al. (2014) ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* 30(17): i541–548. doi 10.1093/bioinformatics/btu462 [PubMed: 25161245]
- Misra S, Crosby MA, et al. (2002) Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review. *Genome biology* 3(12): RESEARCH0083
- Mutanen M, Wahlberg N, et al. (2010) Comprehensive gene and taxon coverage elucidates radiation patterns in moths and butterflies. *Proceedings Biological sciences* 277(1695): 2839–2848. doi 10.1098/rspb.2010.0392 [PubMed: 20444718]
- Nadeau NJ, Ruiz M, et al. (2014) Population genomics of parallel hybrid zones in the mimetic butterflies, *H. melpomene* and *H. erato*. *Genome research* 24(8): 1316–1333. doi 10.1101/gr.169292.113 [PubMed: 24823669]
- Palkopoulou E, Mallick S, et al. (2015) Complete genomes reveal signatures of demographic and genetic declines in the woolly mammoth. *Curr Biol* 25(10): 1395–1400. doi 10.1016/j.cub.2015.04.007 [PubMed: 25913407]
- Parra G, Bradnam K, et al. (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 23(9): 1061–1067 [PubMed: 17332020]
- Pellissier L, Kostikova A, et al. (2017) High Rate of Protein Coding Sequence Evolution and Species Diversification in the Lycaenids. *Frontiers in Ecology and Evolution*. doi 10.3389/fevo.2017.00090
- Pettersson MA and Wielgus RS (1973) Acceptance of artificial diet by *Megathymus streckeri*. (Skinner) (Megathymidae). *Journal of Research on the Lepidoptera* 12(4): 197–198
- Poling OC (1902) A new *Megathymus* from Arizona. *Entomological News* 13(4): 97–98
- Pringle EG, Baxter SW, et al. (2007) Synteny and chromosome evolution in the lepidoptera: evidence from mapping in *Heliconius melpomene*. *Genetics* 177(1): 417–426. doi 10.1534/genetics.107.073122 [PubMed: 17603110]
- Read TD, Petit RA, 3rd, et al. (2017) Draft sequencing and assembly of the genome of the world's largest fish, the whale shark: *Rhincodon typus* Smith 1828. *BMC genomics* 18(1): 532. doi 10.1186/s12864-017-3926-9 [PubMed: 28709399]
- Rizwan K, Zubair M, et al. (2012) Phytochemical and biological studies of *Agave attenuata*. *International journal of molecular sciences* 13(5): 6440–6451. doi 10.3390/ijms13056440 [PubMed: 22754375]
- Roberts A, Pimentel H, et al. (2011) Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* 27(17): 2325–2329 [PubMed: 21697122]
- Roever K (1975) Family Megathymidae In: Howe WH (ed) *The Butterflies of North America*. Doubleday and Co., Garden City, NY, pp. 411–422.
- Scott JA (1986) *The Butterflies of North America: A Natural History and Field Guide*. Stanford University Press, Stanford, CA.
- She R, Chu JS, et al. (2011) genBlastG: using BLAST searches to build homologous gene models. *Bioinformatics* 27(15): 2141–2143 [PubMed: 21653517]

- Shen J, Cong Q, et al. (2017) Complete genome of *Achalarus lyciades*, the first representative of the Eudaminae subfamily of Skippers. *Current Genomics* 18(4): 366–374 [PubMed: 29081692]
- Slater GS and Birney E (2005) Automated generation of heuristics for biological sequence comparison. *BMC bioinformatics* 6: 31 [PubMed: 15713233]
- Smit AFA and Hubley R (2008–2010) (<http://www.repeatmasker.org>) RepeatModeler Open-1.0.
- Smit AFA, Hubley R, et al. (1996–2010) (<http://www.repeatmasker.org>) RepeatMasker Open-3.0.
- St Pierre SE, Ponting L, et al. (2014) FlyBase 102--advanced approaches to interrogating FlyBase. *Nucleic acids research* 42(Database issue): D780–788 [PubMed: 24234449]
- Stallings DB and Turner JR (1956) Notes on *Megathymus ursus*, with description of a related new species. *The Lepidopterists' News* 10(1–2): 1–8
- Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9): 1312–1313. doi 10.1093/bioinformatics/btu033 [PubMed: 24451623]
- Stanke M, Schoffmann O, et al. (2006) Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC bioinformatics* 7: 62 [PubMed: 16469098]
- Suzek BE, Huang H, et al. (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* 23(10): 1282–1288 [PubMed: 17379688]
- Talla V, Suh A, et al. (2017) Rapid Increase in Genome Size as a Consequence of Transposable Element Hyperactivity in Wood-White (*Leptidea*) Butterflies. *Genome Biol Evol* 9(10): 2491–2505. doi 10.1093/gbe/evx163 [PubMed: 28981642]
- Tang W, Yu L, et al. (2014) DBM-DB: the diamondback moth genome database. *Database : the journal of biological databases and curation* 2014: bat087
- Thawornwattana Y, Dalquen D, et al. (2018) Coalescent analysis of phylogenomic data confidently resolves the species relationships in the *Anopheles gambiae* species complex. *Mol Biol Evol*. doi 10.1093/molbev/msy158
- Tigano A, Sackton TB, et al. (2018) Assembly and RNA-free annotation of highly heterozygous genomes: The case of the thick-billed murre (*Uria lomvia*). *Mol Ecol Resour* 18(1): 79–90. doi 10.1111/1755-0998.12712 [PubMed: 28815912]
- UniProt C (2014) Activities at the Universal Protein Resource (UniProt). *Nucleic acids research* 42(Database issue): D191–198 [PubMed: 24253303]
- Van Nieuwerburgh F, Thompson RC, et al. (2012) Illumina mate-paired DNA sequencing-library preparation using Cre-Lox recombination. *Nucleic acids research* 40(3): e24 [PubMed: 22127871]
- Warren AD, Ogawa JR, et al. (2008) Phylogenetic relationships of subfamilies and circumscription of tribes in the family Hesperidae (Lepidoptera : Hesperioidea). *Cladistics* 24(5): 642–676. doi 10.1111/j.1096-0031.2008.00218.x
- Warren AD, Ogawa JR, et al. (2009) Revised classification of the family Hesperidae (Lepidoptera: Hesperioidea) based on combined molecular and morphological data. *Syst Entomol* 34(3): 467–523
- Waterhouse RM, Seppey M, et al. (2017) BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol Biol Evol*. doi 10.1093/molbev/msx319
- Wences AH and Schatz MC (2015) Metassembler: merging and optimizing de novo genome assemblies. *Genome biology* 16: 207. doi 10.1186/s13059-015-0764-4 [PubMed: 26403281]
- Wielgus RS, Wielgus JR, et al. (1972) A new subspecies of *Megathymus ursus* Poling (Megathymidae) from Arizona with observations and notes on its distribution and life history *Bulletin of the Allyn Museum* 9: 1–11
- You M, Yue Z, et al. (2013) A heterozygous moth genome provides insights into herbivory and detoxification. *Nature genetics* 45(2): 220–225 [PubMed: 23313953]
- Zhan S, Merlin C, et al. (2011) The monarch butterfly genome yields insights into long-distance migration. *Cell* 147(5): 1171–1185 [PubMed: 22118469]
- Zhan S and Reppert SM (2013) MonarchBase: the monarch butterfly genome database. *Nucleic acids research* 41(Database issue): D758–763 [PubMed: 23143105]

Zhang J, Cong Q, et al. (2017) Mitogenomes of Giant-Skipper Butterflies reveal an ancient split between deep and shallow root feeders. F1000Res 6: 222. doi 10.12688/f1000research.10970.1 [PubMed: 28408977]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 1. Sequenced specimen of *Megathymus ursus violae*, male, USA: Texas, Pecos Co., Glass Mtns., ex larva found 4-May, pupated 7-May, eclosed 7-Jun-2013, voucher NVG-1504. Dorsal view is shown above and ventral view is below.

Figure 2. Immature stages of *M. ursus violae*.

USA: Texas: Pecos Co., Glass Mtns. **A.** egg and **B.** egg shell, 18-Jun-2003; **C–E.** first instar caterpillars: **C, E.** feeding on yucca leaves inside constructed silk domes, 25- and 17-Jun-2003; **D.** removed from the silk dome, 18-Jun-2003; **F–I.** late instar caterpillar inside yucca root, RNAseq was done on this caterpillar, 30-Nov-2013; **J.** prepupal caterpillar and **K.** pupa in powdered burrows, 4-May-2003.

Figure 3. Habitat and tents of *M. ursus violae* in Texas.

A, B. habitat: desert mountain slopes and canyons with yucca plants, 7-Mar-2009: **A.** El Paso Co., Franklin Mtns.; **B.** Jeff Davis Co., Davis Mtns. **C–G.** yucca plants and tents in Pecos Co., Glass Mtns.: **C.** overview of infested yucca plant; **D, E.** closeup of the plant with the tent; **F.** close-up of the tent containing a pupa, all 4-May-2013; **G.** close-up of a tent being constructed by an early instar caterpillar, 25-Jun-2003.

Figure 4. Adults of *M. ursus violae*.

A. freshly eclosed female (left), Texas: Pecos Co., Glass Mtns., 10-Jun-2003. **B.** a perched male, Texas: Jeff Davis Co., Davis Mtns., Jun-1999 photograph by Greg M. Lasley.

Figure 5. Quality of *Megathymus ursus violae* genome and other model Lepidoptera genomes. **A. *Danaus plexippus*; B. *Heliconius erato demophoon*; C. *Bombyx mori*; D. *Megathymus ursus violae*.** The plots were prepared using the assembly statistic plot (http://github.com/rjchallis/assembly_stats) developed by authors of Lepbase²² (<http://lepbase.org/>). This plot shows both the continuity and completeness of a genome assembly. Briefly, scaffolds are sorted in descending order by the length. Each scaffold is represented by a circular sector in the plot and the central angle of the sector shows the fraction of the genome that is assembled in this scaffold. The scales for genome fraction are marked on the outer circumference. The grey shade in each sector originates from the circumference and it shows the length of each scaffold. The scales for scaffold length are marked on the inner radius. This grey shade can be covered by the light orange, orange, and red shades showing the N90, N50 and longest scaffold length, respectively. Meanwhile, the cumulative number of scaffolds within a fraction of the genome is plotted in purple originating from the center. In

addition, the completeness of a genome is evaluated by the presence and status (complete, fragmented and duplicated) of Benchmarking Universal Single-Copy Orthologs (BUSCO) in the assembly and the results are shown in the small circular plots on the upper right corners. Fraction of complete, fragmented, duplicated BUSCO are shown in mid, light and dark green, respectively.

Figure 6. Phylogenetic tree of the Lepidoptera species with complete genomes available. Majority-rule consensus tree of the maximal likelihood trees constructed by RAXML on the concatenated alignment of universal single-copy orthologous proteins. Numbers by the nodes refer to bootstrap percentages. The numbers above are obtained from the complete alignment, and the numbers below are obtained from samples that contain 2% positions of the full alignment.

Figure 7. Enriched GO terms associated with proteins that are more divergent between *Lerema accius* and *Megathymus ursus violae* than between *Lerema accius* and *Acharus lyciades*.
A. Molecular function; **B.** Biological process. The size of dots indicates the number of *Drosophila* proteins that are associated with this term; the color of the dots indicates the level of significance, and darker color implies higher level of significance. Related GO terms are connected by grey lines. GO terms in the boxes are related to fatty acid synthase (red), odorant and taste sensing (green in **A** and purple in **B**), oxidoreductase activity (blue).

Figure 8. Comparison of gene expansion events in *Megathymus ursus violae* (red), *Acharalus lyciades* (blue) and *Lerema accius* (green).

a. Odorant Receptors, *Mvi* has fewer (49 compared to 56, and 58 for *Lac* and *Aly*, respectively) odorant receptors compared with the other two species; **b.** Peroxidase enzymes; **c.** Catalase enzymes; **d.** endochitinase-like proteins, unique expansion in *Lerema accius*.

Table 1.

Quality and composition of Lepidoptera genomes

Property	<i>Mvi</i>	<i>Aly</i>	<i>Pra</i>	<i>Cce</i>	<i>Lac</i>	<i>Pgl</i>	<i>Dpl</i>	<i>Hme</i>	<i>Mci</i>	<i>Bmo</i>	<i>Pxy</i>	<i>Mse</i>	<i>Pse</i>	<i>Pxu</i>
Genome size (Mb)	429	567	246	729	298	375	249	274	390	481	394	419	406	244
Genome size without gap (Mb)	427	536	243	689	290	361	242	270	361	432	387	400	347	238
Heterozygosity (%)	0.1	1.5	1.5	1.2	1.5	2.3	0.55	n.a.	n.a.	n.a.	~2	n.a.	1.2	n.a.
CEGMA (%)	99.6	99.6	99.6	100	99.3	99.6	99.6	98.2	98.9	99.6	98.7	99.8	99.3	99.6
CEGMA coverage by single scaffold (%)	87.4	87.1	88.7	85.3	86.6	86.9	87.4	86.5	79.2	86.8	84.1	86.4	87.4	88.8
Cytoplasmic Ribosomal Proteins (%)	98.9	98.9	98.9	98.9	98.9	98.9	98.9	94.6	94.6	98.9	93.5	98.9	98.9	97.8
<i>De novo</i> assembled transcripts (%)	99	98	99	97	98	98	96	n.a.	97	98	83	n.a.	97	n.a.
Repeat (%)	25.8	25	22.7	34	15.5	22	16.3	24.9	28	44.1	34	24.9	17.2	n.a.
Exon (%)	4.59	3.57	7.91	3.11	6.96	5.07	8.4	6.38	6.36	4.03	6.35	5.34	6.2	8.59
Intron (%)	30.9	28.4	33.3	24	31.6	25.6	28.1	25.4	30.7	15.9	30.7	38.3	25.5	45.5
Number of proteins (thousands)	14.1	15.9	13.2	16.5	17.4	15.7	15.1	12.8	16.7	14.3	18.1	15.6	16.5	13.1

Mvi: *Megathymus ursus violae*; *Aly*: *Achalarus lyciades*; *Pra*: *Pieris rapae*; *Cce*: *Calycopis cecrops*; *Lac*: *Lerema accius*; *Pgl*: *Pterourus glaucus*; *Dpl*: *Danaus plexippus*; *Hme*: *Heliconius melpomene*; *Mci*: *Melitaea cinxia*; *Bmo*: *Bombyx mori*; *Pxy*: *Plutella xylostella*; *Mse*: *Manduca sexta*; *Pse*: *Phoebis sennae*; *Pxu*: *Papilio xuthus*. Heterozygosity is calculated by us as the percent of heterozygous positions found by the Genome Analysis Toolkit (GATK) for *Aly*, *Cce*, *Lac*, *Pgl*, *Pra* and *Pse*, taken from literature for *Dpl*¹⁶; or estimated based on the histogram of K-mer frequencies for *Pxy*^{14,37}. CEGMA: Core Eukaryotic Genes Mapping Approach genes: these are essential genes and the presence of them in a genome is used to evaluate the quality of an assembly (Parra, Bradnam et al. 2007).