Genomics reveals the origins of ancient specimens

Qian Cong[2], Jinhui Shen[2], Jing Zhang[2], Wenlin Li[2], Lisa N. Kinch[1,2], John V. Calhoun[3,4], Andrew D. Warren[3] and Nick V. Grishin[1,2]

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

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

[3]McGuire Center for Lepidoptera and Biodiversity, Florida Museum of Natural History, University of Florida, 3215 Hull Rd., UF Cultural Plaza, PO Box 112710, Gainesville, Florida, 32611-2710 USA

[4]Mailing address: 977 Wicks Drive, Palm Harbor, FL USA 34684-4656

Contact: grishin@chop.swmed.edu

Centuries of zoological studies amassed billions of specimens in collections worldwide. Genomics of these specimens promises to rejuvenate biodiversity research. The obstacles stem from DNA degradation with specimen age. Overcoming this challenge, we set out to resolve a series of long-standing controversies involving a group of butterflies. We deduced geographical origins of several ancient specimens of uncertain provenance that are at the heart of these debates. Here, genomics tackles one of the greatest problems in zoology: countless old, poorly documented specimens that serve as irreplaceable embodiments of species concepts. The ability to figure out where they were collected will resolve many on-going disputes. More broadly, we show the utility of genomics applied to ancient museum specimens to delineate the boundaries of species and populations, and to hypothesize about genotypic determinants of phenotypic traits.

A study of an animal starts from its name, which is governed by strict rules[1]. A specimen termed the name-bearing type represents a species[1]. The lectotype specimen[1] of *Homo sapiens* is Carl Linnaeus, the father of taxonomy[2]. Populations conspecific with the type carry its name. If two types are conspecific, the earlier name is used. Because most animals were named over a century ago, their types are old and lack details about their collection localities. Historically, phenotypic similarity to the type and its geographic locality determined conspecificity. For cryptic species and uncertain type localities, the phenotypic approach is problematic, leading to heated debates[3-7]. We devised a strategy to obtain genomic sequences of old type specimens and determine to which present-day populations they correspond. To solve a daunting biological problem, we applied this strategy to the skipper butterfly *Hesperia comma* and its relatives. It is *the* most important skipper, because the whole family of butterflies (Hesperiidae, the skippers) is typified by the genus *Hesperia*, and the type of that genus is *H. comma*, as designated by Carl Linnaeus himself. Several mysteries surround these butterflies. (1) Like the Linnaean type of *comma*, the type of its American counterpart, *Hesperia colorado*, collected by noted naturalist Theodore Mead[3,4], lacks a locality label. Where was this type collected? The answer will seal the fate of names later given to populations of this species. (2) Are American *comma*-like butterflies the same species as *comma* in Europe? (3) Unusually for butterflies, *Hesperia* inhabits a wide range of elevations, from lowlands to alpine zone at over 3500 m. What are the genetic determinants of this elevational plasticity? Genomic data provide clear answers to all these questions.

***Hesperia colorado* lectotype mapped to Lake County in Colorado.** The biggest puzzle and controversy is the origin of the *H. colorado* lectotype specimen. It was collected by Mead in 1871 in Colorado, but where exactly? This question is critical because the answer determines what names apply to *Hesperia* populations. Some[5,6] suggested that the lectotype is from the subalpine zone, while others[3,4,7] argued that it is from the Arkansas River Basin. It was even proposed to be a hybrid of these two populations, and thus a poor choice for the name-bearing type[5,6]. Therefore, phenotypic traits without precise locality data are insufficient to attribute the lectotype to a population. Such situations are common in zoology, and offering a general solution, we ventured to place the lectotype on the map using genomics.

First, we sequenced and assembled a reference genome of *Hesperia colorado* from a specimen collected and preserved for that purpose. The genomic sequence was about 650 million base pairs, larger than most butterflies. Second, we overcame the challenge of sequencing a 150-year-old small butterfly pinned in a collection without damaging such an invaluable specimen. This resulted in ~25% complete nuclear and entire mitochondrial genomes of the *colorado* lectotype. Third, we obtained a whole genome shotgun of 85 specimens across Colorado (Table S1). The ages of these specimens varied from 150 years (from Mead's same expedition) to recently collected. These sequences were mapped onto the reference genome and analyzed by a combination of population genetics tools. The result was unambiguous: the lectotype of *colorado* was geographically placed in an area about 15 km in diameter in Lake County, Colorado, and belongs to the Arkansas River Basin population (Fig. 1, specimen #45).

Using PCA, t-SNE, STRUCTURE and TREEMIX analyses, we determined that *H. colorado* is represented by 4 major populations in Colorado (blue, green, red and yellow in Fig. 1). These populations correspond to subspecies. The alpine subspecies *sublima* is surrounded by three others along the three major river basins: the Platte (*ochracea*), the Arkansas (*colorado*) and, the largest, Colorado (assigned to *idaho*) (Fig. 1b). The subspecies intergrade at the boundaries of their ranges, forming hybrids (Fig. 1e). Hybridization patterns are best revealed by the PCA analysis, where the hybrids line up between the centers of populations (along dotted lines in Fig. 1a). The transition between the alpine (red) and Arkansas River Basin (blue) populations is in Lake county (cyan numbers in Fig. 1a), where three subspecies meet and hybridize. The lectotype of *colorado* (#45 Fig. 1a) is surrounded exclusively by the specimens from Lake Co., implying where it was collected. Moreover, the mitogenome of the lectotype was a 100% match to a single specimen (Fig. 2c), and that specimen was collected near Twin Lakes in Lake Co. Another of Mead's specimens (#49, a paralectotype, Fig. 1a), collected during the same 1871 expedition, maps to the same area. STRUCTURE (Fig. 1e) reveals their genomic composition. Lake County specimens are characterized by having some *sublima* (red) and *idaho* (green) components, making them attributable to the narrow geographic zone of transition between these populations. Nevertheless, as t-SNE shows (Fig. 1c), the lectotype clusters within the Arkansas River Basin population (called *oroplata*; #44 is its holotype), and STRUCTURE reveals that only a minor fraction of its genome is of hybrid origin. Thus the name for this population is *colorado*, superseding *oroplata*, a younger name[8].

As a control of our methods applied to ancient specimens, we used three others (two collected by Mead) with less questionable localities[3,4]. These specimens (#6, 19 & 20, Fig. 1) mapped accordingly with their presumed localities. Specimen #6, from near Leadville, is indeed placed with the Lake Co. specimens, but closer to *sublima* as suggested by its genetic makeup. The other two specimens were attributed to *ochracea* in agreement with their associated data. Placing these controls builds the confidence in inferring localities of specimens #45 & 49 by genomics. Moreover, historical evidence suggests that the lectotype was indeed collected in Lake County, as the specimen bears a label "7-13" (July 13). According to Mead's

journal, he was at Twin Lakes on that date. We think that our success in deducing the localities is due to *Hesperia* being non-migratory, local butterflies that form well-diverged populations.

Another type specimen with a questionable collection locality, *H. colorado manitoba*, trees next to a specimen from Lac la Hache in British Columbia (Fig. 2a-c), in accordance with its label data, thereby validating *manitoba*'s type locality[9]. Topology of the nuclear genomic trees follows the geographic distribution of these populations: those that are close on the map tend to cluster in the trees. For example, all specimens from Colorado, including the *colorado* lectotype, form a clade, which also includes the nearby populations from eastern Arizona. However, the mitogenome tree (Fig. 2c) is incongruent with the nuclear genome trees, revealing a history of mitochondria different from that of nuclear genomes, a phenomenon commonly observed in closely related populations and species[10]. Nevertheless, all three trees support the distinction between Old World *comma* (also found in Alaska) and the North American species *H. colorado*, and place both *manitoba* and *colorado* type specimens in agreement with their collection localities[3,9], contrary to speculations by Scott et al[5,6].

Hesperia colorado is a species distinct from Hesperia comma. Many American species have similar-looking counterparts in Europe and Asia, posing a question about whether we should regard them as the same species. Phylogenetic trees of three different genomic regions consistently reveal a deep split between American *Hesperia colorado* and the Old World *Hesperia comma* (Fig. 2), strongly suggesting that they are not conspecific. The genome-wide Fixation index (Fst)[11] between *comma* and *colorado* is about 0.5, corresponding to a divergence between species. Despite its broad distribution, populations of *comma* from Europe, Africa and Asia are much closer to each other than any of them to Nearctic *colorado*. A surprise was a specimen from northern Alaska, which is true *comma*, most similar to specimens from northeastern Russia across the Bering Strait (Fig. 2ab). Therefore, *comma* is indeed present in America, but all other *comma*-like specimens we sequenced from the US and Canada were *colorado*. Functional analysis (Table S2) of proteins that differ the most between the two species reveals the prevalence of nuclear-encoded mitochondrial respiratory chain components, which is consistent with the profound divergence (about 3%) in the mitogenomes. Furthermore, the two species show divergence in circadian rhythm, DNA, and histone methylation, which may have a direct effect on the reproductive barrier between *comma* and *colorado*.

Adaptations to high elevation in Hesperia colorado sublima. Unlike other *Hesperia colorado* populations, *sublima* is confined to high elevations. Comparing *sublima* with others, we identified 162 unique amino acid replacements enriched in 80 proteins and positively selected in 7 (Table S3). These proteins function in metabolism, response to light, and development of the respiratory system and muscles (Fig. 3ab). This diversity, more than in human high-elevation populations dated up to 45,000 ya[12], is likely due to the longer time of adaptation (~100,000 ya) and one to two year generation span in *Hesperia*. Human adaptations are largely restricted to variations in hypoxia-inducible factors, which upregulate glycogen and ATP synthesis under hypoxic conditions. Similarly, in *sublima*, variations map to regulators of insulin pathway and glycogen synthase (GS), in addition to metabolic enzymes, electron carriers and heme-binding proteins. These variations may reprogram *sublima* metabolism to ensure sufficient energy production with reduced oxygen. One GS variation is in the helix interacting with glycogenin (Fig. 3c). The interaction between the GS and glycogenin is crucial[13], and the *sublima*-specific variation affecting this interaction may lead to more efficient glycogen production.

Insects lack the dedicated oxygen-carrying blood cells of vertebrates, and a network of tracheal tubes directly delivers oxygen throughout the body. Similar to humans growing more capillaries in muscle with exercise at high elevation[14], *sublima*-specific variations in factors governing the development of this network may promote elaboration of tracheal tubes. Other adaptations in *sublima* lack apparent human analogies. We observe variations in UV photoreceptors and circadian regulators, changes likely related to stronger UV radiation and colder climate at higher altitudes. Most notably, 4 out of 7 proteins positively selected in *sublima* are related to flight ability, including the motor neuron axon guidance factor *trol* with a variation near its calcium binding site: in *sublima*, hydrophobic Leucine replaces hydrophilic Glutamine of lower elevation populations (Fig. 3d). Strong fliers, including swallowtails and monarch butterflies, all have hydrophobic residues in this position, while in weak fliers, such as the cabbage white, this residue is hydrophilic. These changes should make *sublima* a better flier adapted to stronger wind on mountaintops.

Genetic basis for paler appearance of *Hesperia colorado ochracea.* Out of all populations in Colorado, *ochracea* is perhaps most recognizable due to its paler wings and poorly defined white spots. Surprisingly, we found that *ochracea*-specific polymorphisms are largely confined to a short 200kb region on the Z chromosome (Fig. 4). This unusually high density of SNPs suggests introgression from some population outside of Colorado rather than gradual evolution by point mutations. ABBA-BABA tests[15] with other populations reveal that this region was introgressed (P-value < 6e-66) from *Hesperia colorado assiniboia*, a northern butterfly of the Central Plains. Because the wing patterns of *ochracea* resemble *assiniboia* (Fig. 4a-f), it is likely that this introgressed region is inducing *assiniboia*-like wing patterns in *ochracea.* Proteins encoded by this region include Shank3, a regulator of Wnt signaling pathway modulating internalization of the Wnt receptor Fz2[16]. The Wnt pathway has been implicated in wing patterning[17-19], and Wnt receptor Fz2 is expressed in developing wings[20], suggesting a role of Shank3 in wing pattern formation.

Materials and methods. Specimens used in this project were either collected in the field (and stored in RNAlater or EtOH), or borrowed from collections listed in the Acknowledgements. The collection dates of specimens ranged from 1871 to 2016. See Table S1 for complete specimen data. A piece of thoracic tissue from fresh specimens, and either the abdomen or a leg from pinned museum specimens were used for DNA extraction and genomic library preparation according to our protocols developed previously[21-24]. We used mate-pair libraries to assemble a reference genome of *Hesperia colorado* from a single wild-collected specimen, also accompanied by RNAseq for gene annotation. Paired-end libraries were constructed for all other specimens, sequenced for 150 bp from both ends targeting 5 to 10X coverage and mapped to the reference genome. Z chromosome scaffolds were found as those containing Z chromosome proteins in *Heliconius*[25]. The following computational tools were used: PCA[26], t-SNE[27], TREEMIX[28], STRUCTURE[29], and IQ-TREE[30]. Details of adapting these methods to work with data from ancient specimens are described in the Supplement.

References Cited

1   ICZN. *International Code of Zoological Nomenclature. Fourth edition.* The Code Online (ICZN) edn, (The International Trust for Zoological Nomenclature, 1999).

2   Spamer, E. E. Know thyself: responsible science and the lectotype of *Homo sapiens* Linnaeus, 1758. *Proceedings of the Academy of Natural Sciences of Philadelphia* **149**, 109-114 (1999).

3   Calhoun, J. V. An updated itinerary of Theodore L. Mead in Colorado in 1871, with type locality clarifications and a lectotype designation for *Melitaea eurytion* Mead (Nymphalidae). *Journal of the Lepidopterists' Society* **69**, 1-38 (2015).

4   Calhoun, J. V. Assessing specimen provenance through the writings of Theodore L. Mead, with notes on his specimens of *Hesperia colorado* (Hesperiidae). *News of the Lepidopterists' Society* **57**, 173, 176–181 (2015).

5   Scott, J. A. Mead's butterflies in Colorado, 1871. *Papilio (New Series)* **24**, 1-69 (2016).

6   Scott, J. A. *et al.* Case 3709 – Conservation of the usage of names of western North American Hesperia *comma*-group subspecies (Lepidoptera, Hesperiidae) through designation of neotypes for *Pamphila colorado* Scudder, 1874, *Pamphila manitoba* Scudder, 1874, *Pamphila cabelus* Edwards, 1881, *Pamphila harpalus* Edwards, 1881, *Pamphila idaho* Edwards, 1883, *Pamphila oregonia* Edwards, 1883, and *Hesperia colorado ochracea* Lindsey, 1941. *Bulletin of Zoological Nomenclature* **75**, 88–121 (2018).

7   Warren, A. D. & Calhoun, J. V. A new name for the alpine populations of *Hesperia colorado* (Scudder) (Hesperiidae, Hesperiinae) in Colorado. *Journal of the Lepidopterists' Society* **69**, 39-49 (2015).

8   Scott, J. A. New Papilionoidea and Hesperioidea from North America. *Papilio (New Series)* **1**, 1-12 (1981).

9   Calhoun, J. V. Butterflies collected by George R. Crotch in N America in 1873, with notes on the identity of *Pamphila manitoba* and a type locality clarification for *Argynnis rhodope*. *News of the Lepidopterists' Society* **57**, 135-143 (2015).

10   Cong, Q. *et al.* When COI barcodes deceive: complete genomes reveal introgression in hairstreaks. *Proceedings. Biological sciences* **284**, doi:10.1098/rspb.2016.1735 (2017).

11   Holsinger, K. E. & Weir, B. S. Genetics in geographically structured populations: defining, estimating and interpreting F(ST). *Nat Rev Genet* **10**, 639-650, doi:10.1038/nrg2611 (2009).

12   Ossendorf, G. *et al.* Middle Stone Age foragers resided in high elevations of the glaciated Bale Mountains, Ethiopia. *Science* **365**, 583-587, doi:10.1126/science.aaw8942 (2019).

13   Zeqiraj, E. *et al.* Structural basis for the recruitment of glycogen synthase by glycogenin. *Proceedings of the National Academy of Sciences of the United States of America* **111**, E2831-2840, doi:10.1073/pnas.1402926111 (2014).

14   Mathieu-Costello, O. Muscle adaptation to altitude: tissue capillarity and capacity for aerobic metabolism. *High Alt Med Biol* **2**, 413-425, doi:10.1089/15270290152608598 (2001).

15   Martin, S. H., Davey, J. W. & Jiggins, C. D. Evaluating the use of ABBA-BABA statistics to locate introgressed loci. *Mol Biol Evol* **32**, 244-257, doi:10.1093/molbev/msu269 (2015).

16   Harris, K. P., Akbergenova, Y., Cho, R. W., Baas-Thomas, M. S. & Littleton, J. T. Shank Modulates Postsynaptic Wnt Signaling to Regulate Synaptic Development. *J Neurosci* **36**, 5820-5832, doi:10.1523/JNEUROSCI.4279-15.2016 (2016).

17   Mazo-Vargas, A. *et al.* Macroevolutionary shifts of WntA function potentiate butterfly wing-pattern diversity. *Proceedings of the National Academy of Sciences of the United States of America* **114**, 10701-10706, doi:10.1073/pnas.1708149114 (2017).

18   Martin, A. & Reed, R. D. Wnt signaling underlies evolution and development of the butterfly wing pattern symmetry systems. *Dev Biol* **395**, 367-378, doi:10.1016/j.ydbio.2014.08.031 (2014).

19   Martin, A. *et al.* Diversification of complex butterfly wing patterns by repeated regulatory evolution of a Wnt ligand. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 12632-12637, doi:10.1073/pnas.1204800109 (2012).

20   Hanly, J. J., Wallbank, R. W. R., McMillan, W. O. & Jiggins, C. D. Conservation and flexibility in the gene regulatory landscape of heliconiine butterfly wings. *Evodevo* **10**, 15, doi:10.1186/s13227-019-0127-4 (2019).

21   Cong, Q., Borek, D., Otwinowski, Z. & Grishin, N. V. Tiger Swallowtail Genome Reveals Mechanisms for Speciation and Caterpillar Chemical Defense. *Cell reports*, doi:10.1016/j.celrep.2015.01.026 (2015).

22   Li, W. *et al.* Genomes of skipper butterflies reveal extensive convergence of wing patterns. *Proceedings of the National Academy of Sciences of the United States of America* **116**, 6232–6237, doi:doi.org/10.1073/pnas.1821304116 (2019).

23   Zhang, J., Cong, Q., Shen, J., Brockmann, E. & Grishin, N. V. Genomes reveal drastic and recurrent phenotypic divergence in firetip skipper butterflies (Hesperiidae: Pyrrhopyginae). *Proceedings. Biological sciences* **286**, 20190609, doi:10.1098/rspb.2019.0609 (2019).

24   Zhang, J. *et al.* Gypsy moth genome provides insights into flight capability and virus-host interactions. *Proceedings of the National Academy of Sciences of the United States of America*, doi:10.1073/pnas.1818283116 (2019).

25   Davey, J. W. *et al.* Major Improvements to the Heliconius melpomene Genome Assembly Used to Confirm 10 Chromosome Fusion Events in 6 Million Years of Butterfly Evolution. *G3 (Bethesda)* **6**, 695-708, doi:10.1534/g3.115.023655 (2016).

26   Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet* **2**, e190, doi:10.1371/journal.pgen.0020190 (2006).

27   van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9**, 2579-2605 (2008).

28   Pickrell, J. K. & Pritchard, J. K. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet* **8**, e1002967, doi:10.1371/journal.pgen.1002967 (2012).

29   Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959 (2000).

30   Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32**, 268-274, doi:10.1093/molbev/msu300 (2015).

Figures
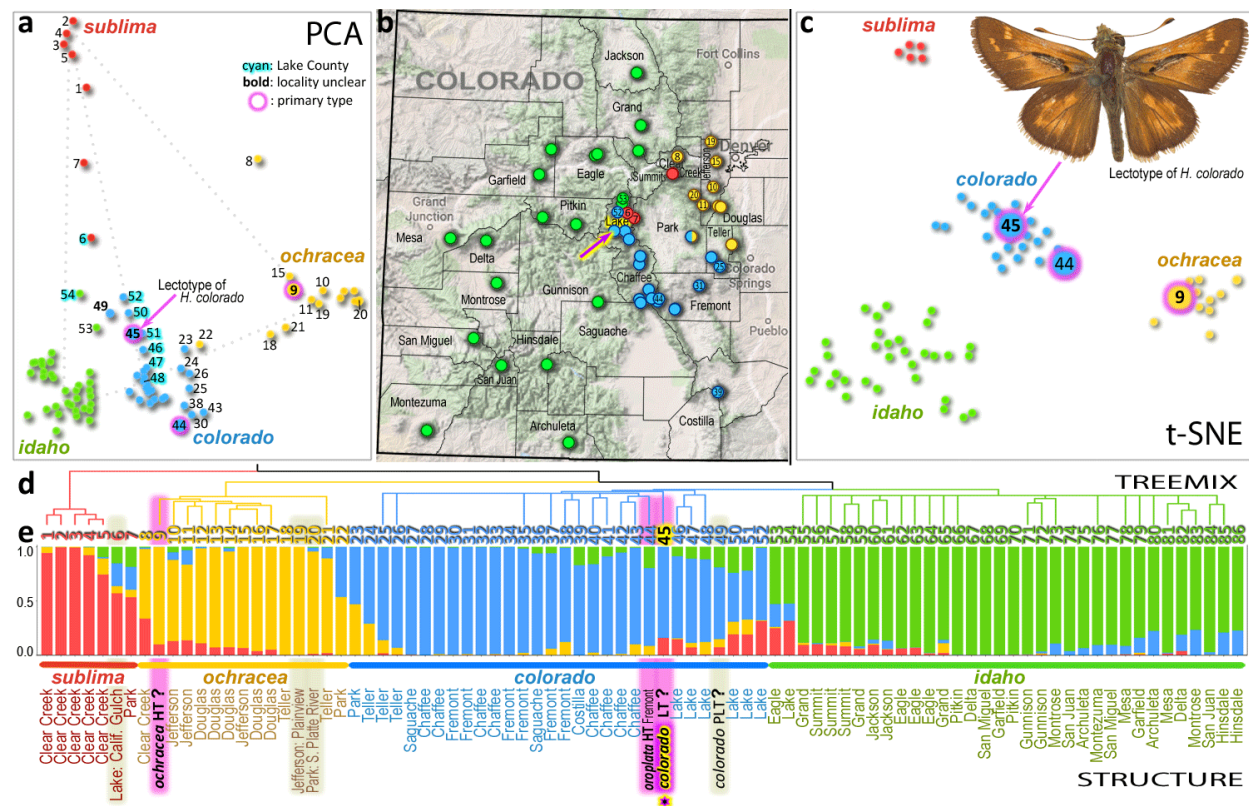
Fig. 1. *Hesperia colorado* **type specimen traced to Lake County in Colorado.** Different methods consistently partition specimens into four populations: *sublima* (red), *ochracea* (orange), *colorado* (blue), and *idaho* (green). **(a)** Principal Component Analysis (PCA) of covariance between the SNPs in samples from Colorado using Eigensoft[26]. The 1st and 2nd components are shown. **(b)** Truncated map of Colorado showing specimen localities. Arrow points to the locality in Lake County we deduced for the *colorado* lectotype. **(c)** t-SNE (parameter: perplexity = 10) reduced the first 10 PCA dimensions to two, revealing populations as clusters. **(d)** TREEMIX results showing the clustering and evolutionary history of specimens. **(e)** Population structure inferred by STRUCTURE, showing the proportion of each population's features in each specimen. Names of taxa (*colorado*, *oroplata*, *ochracea*) are given for type specimens (HT holotype, LT lectotype, PLT paralectotype), and primary types are highlighted in magenta; other ancient specimens are highlighted in gray; county names are given for other specimens (see other data in Table S1, referred to by the numbers from 1 to 86 given as leaves in the TREEMIX tree); a question mark after the name of a type indicates a questionable collection locality that we identify here using genomic comparisons.
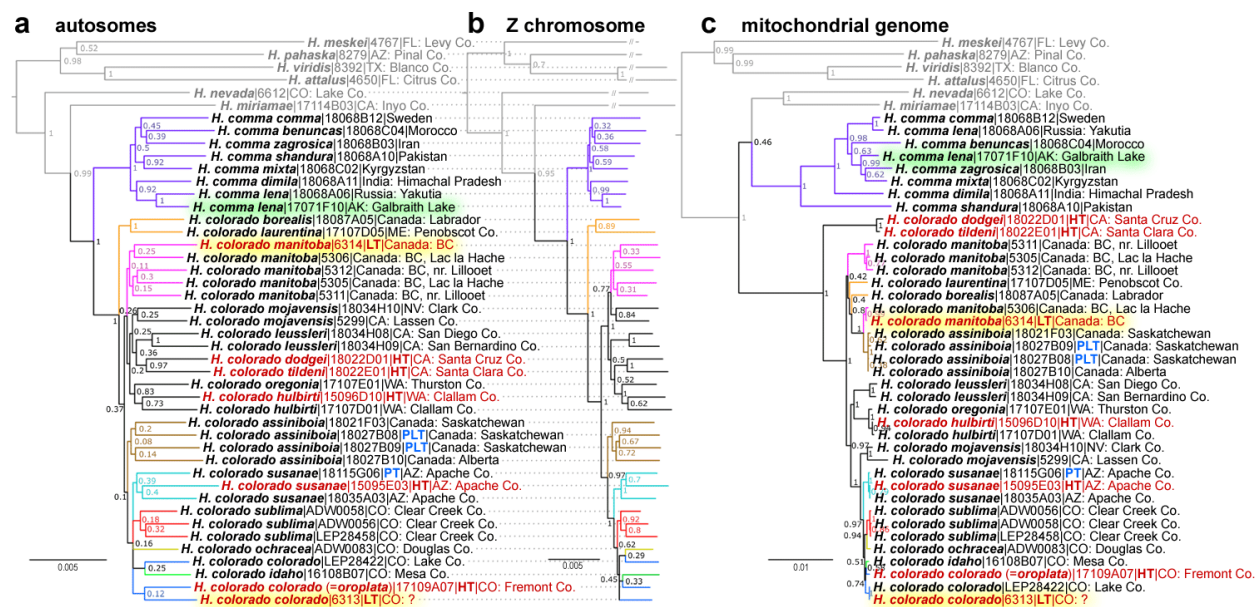
Fig. 2. Phylogenetic analysis suggests that *Hesperia colorado* is a species distinct from *Hesperia comma*. The trees are based on concatenated genomic alignments from **(a)** autosomes, **(b)** Z chromosome, and **(c)** mitogenome. Name, voucher number and general locality are given for each specimen, see Table S1 for additional data. *Hesperia* species outside the *comma* group are shown in gray. The trees are rooted with *Pompeius verna* (NVG-18014H01, not shown). Names of primary type specimens are shown in red. The two types with questionable localities are highlighted in yellow and an American *H. comma* specimen from Alaska is highlighted in green.
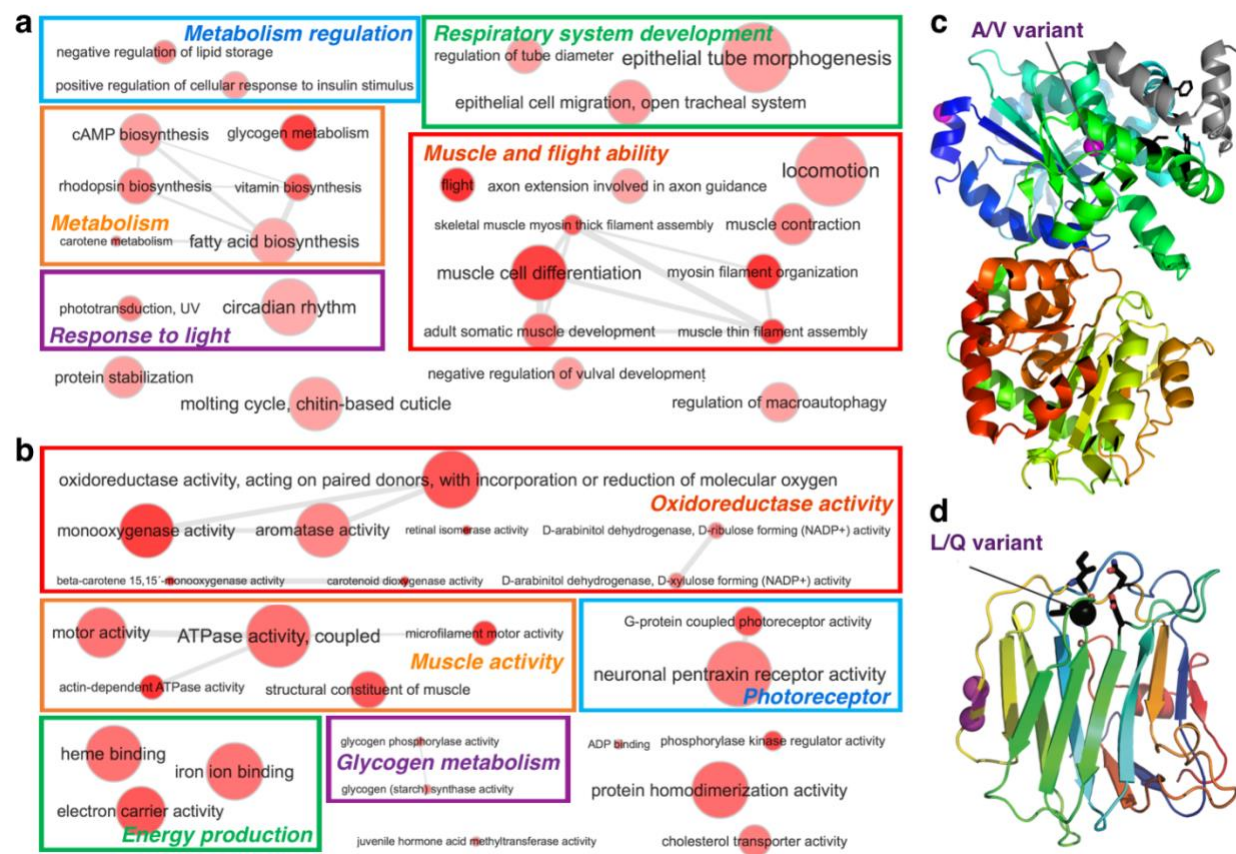
Fig. 3. Molecular adaptation to high elevation in *Hesperia colorado sublima*. (a-b) Functions of proteins related to high-elevation adaptation revealed by Gene Ontology (GO) terms in the categories of **(a)** biological process and **(b)** molecular function. The size of a circle correlates with the number of proteins in the genome associated with that GO term and its color indicates the significance (darker color corresponds to lower P-value) of a GO term enrichment among proteins showing high-elevation adaptation. **(c)** 3D structure model of glycogen synthase (template PDB: 4QLB. A variant in *sublima* (from A to V) is near the functional sites (black sticks) mediating its interaction with glycogenin. **(d)** 3D structure model (template PDB:3POY) of motor neuron guidance factor *trol* with a variant (black sphere, L in *sublima*, and Q in low-elevation populations) near the Ca2+ binding site. Other variants are shown as magenta spheres.
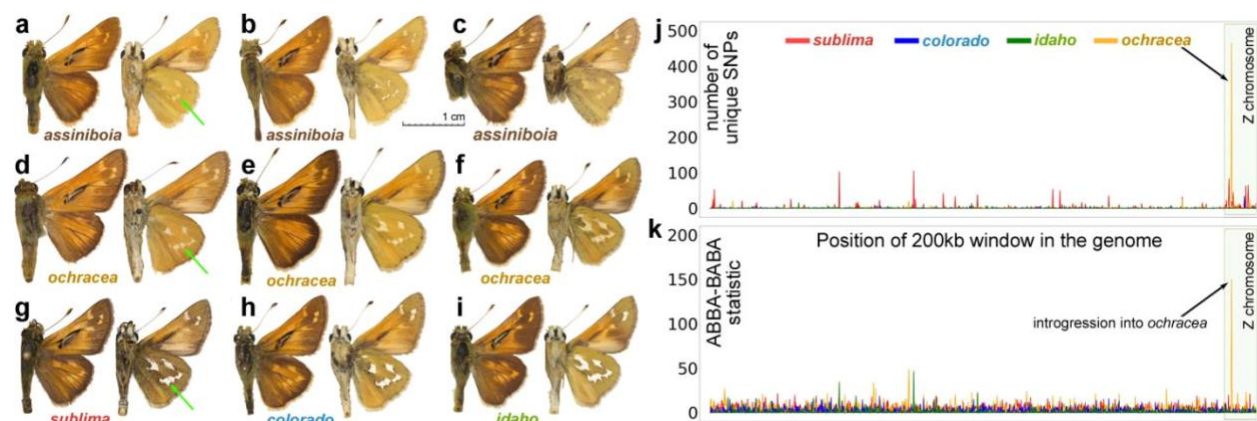
Fig. 4. Similarity in wing pattern between *Hesperia colorado ochracea* and *H. c. assiniboia* likely caused by the introgression of a 200kb Z-linked genomic region. Wing pattern variation in **(a-c)** *assiniboia* and **(d-f)** *ochracea* specimens; **(g-i)** typical specimens of other *Hesperia* populations in Colorado. Green arrows denote pale spots on the hindwing differing between populations. Voucher numbers for (a-i) are NVG-18027B08, NVG-18027B09, NVG-18027B10, NVG-15111B01, NVG-5533, NVG-16108A05, NVG-5532, NVG-16108C02, and NVG-16108B04, respectively; see Table S1 for additional data. **(j)** Number of unique SNPs in different populations in 200kb windows throughout the genome. **(k)** Introgression from *assiniboia* to 4 *Hesperia* populations in Colorado identified using the ABBA-BABA test. The y-axis shows the difference between the number of positions with the pattern ABBA and the number of positions with the pattern BABA in a 200kb genomic window (negatives omitted). The x-axis on both plots is the position of the window in the genome (concatenated scaffolds). Z chromosome scaffolds are placed last (highlighted pale olive). Counts are colored by taxon: *sublima* (red), *colorado* (blue), *idaho* (green), and *ochracea* (orange).