De Novo Genome Assembly and Annotation of an Andean Caddisfly, *Atopsyche davidsoni* Sykora, 1991, a Model for Genome Research of High-Elevation Adaptations

Blanca Ríos-Touma[1,*], Ralph W. Holzenthal[2], Ernesto Rázuri-Gonzales[2,3], Jacqueline Heckenhauer[3,4], Steffen U. Pauls[3,4,5], Caroline G. Storer[6], and Paul B. Frandsen [ID][4,7,*]

[1]Facultad de Ingenierías y Ciencias Aplicadas, Ingeniería Ambiental, Grupo de Investigación en Biodiversidad, Medio Ambiente y Salud (BIOMAS), Universidad de las Américas, Quito, Ecuador

[2]Department of Entomology, University of Minnesota, St. Paul, Minnesota, USA

[3]Department of Terrestrial Zoology, Entomology III, Senckenberg Research Institute and Natural History Museum Frankfurt, Germany

[4]LOEWE Centre for Translational Biodiversity Genomics (LOEWE-TBG), Frankfurt, Germany

[5]Institute of Insect Biotechnology, Justus-Liebig University, Gießen, Germany

[6]McGuire Center for Lepidoptera and Biodiversity, Florida Museum of Natural History, University of Florida, Gainesville, Florida, USA

[7]Department of Plant and Wildlife Sciences, Brigham Young University, Provo, Utah, USA

*Corresponding authors: E-mails: blanca.rios@udla.edu.ec; paul_frandsen@byu.edu.

Abstract

We sequence, assemble, and annotate the genome of *Atopsyche davidsoni* Sykora, 1991, the first whole-genome assembly for the caddisfly family Hydrobiosidae. This free-living and predatory caddisfly inhabits streams in the high-elevation Andes and is separated by more than 200 Myr of evolutionary history from the most closely related caddisfly species with genome assemblies available. We demonstrate the promise of PacBio HiFi reads by assembling the most contiguous caddisfly genome assembly to date with a contig N50 of 14 Mb, which is more than $6\times$ more contiguous than the current most contiguous assembly for a caddisfly (*Hydropsyche tenuis*). We recover 98.8% of insect BUSCO genes indicating a high level of gene completeness. We also provide a genome annotation of 12,232 annotated proteins. This new genome assembly provides an important new resource for studying genomic adaptation of aquatic insects to harsh, high-altitude environments.

Key words: caddisfly genomics, Trichoptera, Insecta, PacBio HiFi, high altitude, extreme environments.

Significance

We provide the first de novo genome assembly for the caddisfly genus *Atopsyche* in the family Hydrobiosidae, a lineage separated by more than 200 Myr of evolution from the nearest sequenced families. The sampled species occupies high-elevation streams in the tropical Andes that are characterized by very cold water and intense ultraviolet radiation, among other harsh environmental conditions. This is the first caddisfly genome assembled from PacBio HiFi sequencing reads and represents the highest quality caddisfly genome to date, both in terms of contiguity (contig N50) and gene completeness (BUSCO). This new genome assembly will provide an opportunity for examining the genomic basis of adaptation to such extreme environments.

Introduction

In the tropical Andes of South America, the páramo represents a unique ecosystem lying between the tree line, approximately 3,000 m asl (above sea level), and the higher elevation snow line (fig. 1A and supplementary video, Supplementary Material online). Although relatively nonseasonal in terms of day length, these generally humid ecosystems can experience daily fluctuations in air temperature from below 0 °C to 25 °C (Sklenář et al. 2010; Duchicela et al. 2021). Streams originating in the páramo comprise the headwaters of some of Earth's largest rivers (Jacobsen 2008; Encalada et al. 2019). The uplift and active volcanism of the Andes over the past 50 Myr have produced complex geomorphology and a diversity of soil types (Jacobsen 2008; Encalada et al. 2019). As a result, high-altitude Andean aquatic ecosystems can be extremely hydrologically and physico-chemically variable and range from more stable spring-fed systems to dynamic glacial streams.

Although air temperature can vary widely in the páramo, high-elevation equatorial Andean streams are characterized by low water temperature throughout the year, varying from approximately 0–15 °C (Jacobsen 2008; Jacobsen et al. 2014; Finn et al. 2016). Jacobsen (2004) hypothesized that consistently low water temperature was a primary reason aquatic insect richness decreased along an elevational gradient toward high-altitude Andean streams, at least among insect families. In addition to cold water, aquatic insect growth and development are constrained by the decline in O_2 solubility with decreasing atmospheric pressure at high altitudes (Jacobsen et al. 2003; Jacobsen 2020). Finally, high ultraviolet radiation and the strong drying effects of wind in these generally exposed ecosystems can be relevant factors affecting biological processes and diversity, especially for flying adult insects (Buytaert et al. 2011). Combined, the geological history and harsh abiotic conditions of the high-elevation Andean páramo have contributed to conditions well-suited to specialists, resulting in species endemism of up to 60% in the highlands of the tropical Andes (Buytaert et al. 2011).

Among Andean caddisflies, the primarily Neotropical genus *Atopsyche* includes many species apparently adapted and restricted to high Andean rivers and streams. For example, in Ecuador, the 26 recorded species show strong elevational stratification, occurring between 330 and 3,810 m asl, with many species restricted to high Andean habitats (Ríos-Touma et al. 2017). These narrow geographic distributions and elevational ranges seem to be the case for many of the 139 species known in the genus, except for some more broadly distributed, low elevation species (Gomes and Calor 2019). As such, the species in the genus are ideal for studying biogeography, speciation, physiological adaptations to broad ecological conditions, and, especially, adaptations to extreme high Andean aquatic habitats. In addition, the genus belongs to one of only two families of caddisflies that do not construct portable larval cases or fixed retreats with silken capture nets (Wiggins 2004). Also of note, the "free-living" larvae, as known so far, are solely predaceous and have modified chelate forelegs to aid in the capture of invertebrate prey (fig. 1B).

Aquatic insect species are underrepresented in available insect genome assemblies (Hotaling et al. 2020). One way to close that gap is to sequence more aquatic insects, especially those that might be key to understanding the evolution of certain adaptations. Here, we provide the first whole-genome assembly and annotation for *Atopsyche davidsoni* Sykora, 1991, a member of the caddisfly family Hydrobiosidae, representing 200 Myr of evolution from *Agrypnia vestita* and *Hesperophylax magnus* the most closely related caddisfly species with genome assemblies available (Thomas et al. 2020; Olsen et al. 2021) (fig. 1C). Further, *A. davidsoni* inhabits high Andean páramo streams, and our specimen was collected near the base of Cotopaxi volcano, constituting the highest recorded elevation for *Atopsyche* in Ecuador (3,810 m asl). Cotopaxi has experienced more than 50 eruptions in the last 200 years, with the most recent aggressive eruptions occurring between 1877 and 1880 (Hall and Mothes 2008), and with significant eruptions of ash as recently as 2015. Stream water chemistry and dynamics are influenced by the volcanic material from past eruptions. We hope this new, high-quality whole-genome assembly will provide an important resource for studying the genomic basis of adaptation to extreme aquatic habitats.

Results and Discussion

Sequencing and Genome Size Estimation

High-molecular weight DNA was extracted from a single larval specimen and generated a band above 50 kb when visualized with pulse-field gel electrophoresis. After running the pbccs read consensus tool, we recovered 18.6 Gb of sequence in HiFi reads (those reads with a quality score greater than Q20), corresponding to ~50× read coverage. The estimated genome size generated in GenomeScope 2.0 (Ranallo-Benavidez et al. 2020) was 341.6 Mb with 90.1% unique sequence (supplementary fig. 1, Supplementary Material online).

Genome Assembly and QC

The genome assembly of *A. davidsoni* was highly contiguous with a total assembly length of 370.8 Mb contained in 80 contigs with a GC content of 34%. The contig L50 was 10 and the contig N50 was >14 Mb, making this assembly the most contiguous caddisfly genome assembly produced to date (table 1). In addition, BUSCO (Manni et al. 2021) analysis using the Insecta and Endopterygota gene sets recovered 98.8% and 97.5% of expected single copy orthologs within the assembly, respectively, providing support for a high level of gene completeness. Blobtools (Laetsch and Blaxter 2017)
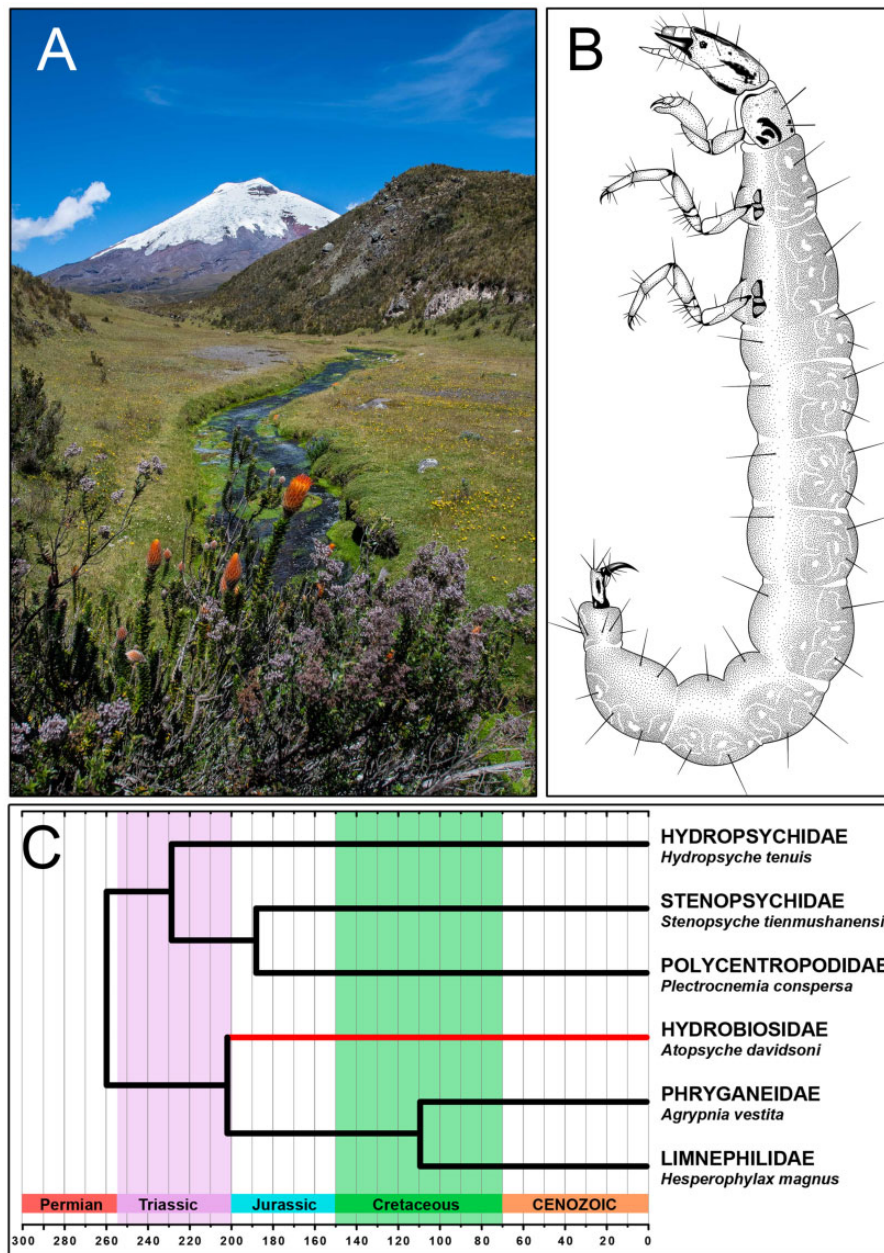
Fig. 1.—(*A*) The Ecuadorian páramo, the high-elevation environment where Atopsyche was collected; Volcan Cotopaxi in the background. Photo credit: Xavier Amigo. (*B*) *Atopsyche davidsoni*, larva. (*C*) Phylogenetic tree, redrawn with dates from Thomas et al. (2020), showing the evolutionary divergence among caddisfly families with long-read-based genome assemblies available.

analysis assigned 99.9% of the genome to the Phylum Arthropoda, indicating low levels of contamination.

Annotation

Following RepeatModeler and RepeatMasker analysis, 17.6% of the genome was classified as repetitive, a higher amount of repetitive sequence than estimated with GenomeScope 2.0 (9.1%). This may partially explain why the genome size estimate from GenomeScope 2.0 (341.7 Mb) was lower than the recovered assembly length (370.8 Mb). Of the repeats that were classified, DNA transposons were the most abundant, comprising 4.59% of the genome. Notably, LINE elements only comprised 1.85% of the genome, a feature more similar to that of retreat-making caddisflies than to the more closely related tube-case making caddisflies (Olsen et al. 2021). Genome size variation in caddisflies has recently been linked to the

Table 1

Comparison of Currently Published Long-Read-Based Genome Assemblies of Caddisflies

Species	Accession	Suborder	Sequencing Platform Coverage	Assembly Length (bp)	Contig N50 (kb)	BUSCOs Present (%)[a,b]	% Repetitive Elements[c]
Atopsyche sp. (present study)	SUB10108826	Integripalpia	PacBio HiFi (50.19×)	370,818,532	14,095.1	98.8	17.67
Agrypnia vestita (Olsen et al. 2021)	JADDOH000000000	Integripalpia	PacBio + Illumina (17.86× + 87.96×)	1,353,059,149	111.8	94.4	71.41
Hesperophylax magnus (Olsen et al. 2021)	JADDOG000000000	Integripalpia	Nanopore + Illumina (26.38× + 49.30×)	1,233,588,871	768.2	95.9	74.36
Hydropsyche tenuis (Heckenhauer et al. 2019)	GCA_009617725.1	Annulipalpia	Nanopore + Illumina (16.5× + 167.6×)	229,663,394	2,190.1	98.4	22.65
Plectrocnemia conspersa (Heckenhauer et al. 2019)	GCA_009617715.1	Annulipalpia	Nanopore + Illumina (17.1× + 82.9×)	396,695,105	869.0	98.7	42.9
Stenopsyche tienmushanensis (Luo et al. 2018)	GCA_008973525.1	Annulipalpia	PacBio + Illumina (153× + 150×)	451,494,475	1,296.9	98.1	45.67

Note.—The genome assembly for *Atopsyche davidsoni* is currently the highest quality both in terms of contiguity and gene completeness (BUSCO scores).
[a]Present = complete + fragmented.
[b]$N_{insecta}$ = 1,367.
[c]Values for taxa not included this study were taken from Olsen et al. (2021).

expansion of repetitive elements (Heckenhauer et al. 2021; Olsen et al. 2021). The relatively lower proportion of repetitive sequence and the relatively smaller genome size of *A. davidsoni* reiterates the findings that the bulk of repetitive element expansion occurred in lineages of tube case-making caddisflies (e.g., *Agrypnia* and *Hesperophylax*, table 1). Of 12,232 predicted proteins, 10,370 were verified by BLAST with the NCBI nonredundant protein database, 6,543 were assigned to GO terms, and 4,216 were functionally annotated with Blast2GO (Götz et al. 2008). Metabolic and cellular processes were, by far, the top functional categories assigned to annotated genes.

Conclusion

Here, we provide the first genome assembly for the caddisfly family Hydrobiosidae and the first caddisfly genome assembled from PacBio HiFi reads. The genome assembly is the most contiguous generated for a caddisfly to date, demonstrating the promise of high-quality, long reads (table 1). In addition, the species sampled occupies high-elevation streams in the harsh environment of the Andean páramo. We hope that these genomic resources will serve as a baseline for future studies on the genomic adaptation to such environments, while also providing valuable data for a species that occupies a habitat imperiled by global climate change.

Materials and Methods

Specimen Collection and Preparation

We collected a larval specimen with an aquatic kick net from a tributary to the Pita River, a perennial stream flowing in the páramo (Esmeraldas River basin), at 3,818 m asl in Cotopaxi National Park in Ecuador. The specimen was placed into 100% EtOH and stored in a −20 °C freezer upon return from the field. Prior to DNA extraction, we dissected and removed the gut contents, froze the specimen in liquid N_2, and macerated the tissue. We extracted high-molecular weight DNA from the specimen using a Zymo Quick-DNA HMW MagBead Kit (Zymo Research DS1710-03). To ensure recovery of suitable high-molecular weight DNA, we visualized the extracted molecules using pulse-field gel electrophoresis.

Library Prep and Sequencing

We prepared a PacBio HiFi library with the SMRTbell Express Template Prep Kit 2.0 (PacBio 100-938-900) using the PacBio low input protocol (DNA sheared to 15 kb) for HiFi sequencing followed by AMPure bead cleanup. The genomic library was sequenced on a single 30-h movie 8M SMRT cells in CCS mode on the PacBio Sequel II system at the BYU DNA Sequencing Center.

HiFi Read Generation, Assembly, and Quality Assessment

We generated HiFi reads from the raw subreads using the pbccs tool implemented in the pbbioconda package (Pacific Biosciences 2021). We counted k-mers of length 21 from the HiFi reads with KMC v.3.1.1 (Kokot et al. 2017) using the following command *kmc -k21 -t80 -m1064 -ci1 -cs10000 @files.txt reads tmp/*, and created a histogram of the k-mer frequencies with the command *kmc_tools transform reads histogram reads.histo -cx10000*. We then ran GenomeScope 2.0 (Ranallo-Benavidez et al. 2020) with the exported k-mer count histogram with the online web tool (http://qb.cshl.edu/genomescope/genomescope2.0/, last accessed September 1, 2021) using the following parameters: *k-mer length=21* and *max kmer coverage=10,000*.

We then assembled the HiFi reads into contigs using Hifiasm v0.13-r307 (Cheng et al. 2021) with the "-l 2" option enabled for aggressive duplicate purging. To assess whether the genome contained expected single copy orthologs, we used BUSCO v.4.1.4 (Manni et al. 2021) with both the Insecta odb10 and Endopterygota odb10 core ortholog sets (Kriventseva et al. 2019) with the –long option enabled for species specific gene model training (supplementary note 3, Supplementary Material online).

We screened the final genome assembly for potential contamination with taxon-annotated GC-coverage plots using BlobTools v1.0 (Laetsch and Blaxter 2017). We first mapped all raw HiFi reads against the final genome assembly using minimap2 (Li 2018) with the *-ax asm20* option and then sorted the resulting bam file with *samtools sort*. We assigned taxonomy with MegaBLAST (Zhang et al. 2000) using the following parameters: *-task megablast and -e-value 1e-25*. We used the blobtools module *map2cov* to calculate coverage and created the BlobDB using *blobtools create*. We used blobtools *plot* to visualize the DB (supplementary fig. 2, Supplementary Material online). We calculated assembly statistics with assembly_stats v.0.1.4 (Trizna 2020). Assembly statistics and BUSCO completeness for comparison to previously published genomes can be found in table 1 and supplementary table 1, Supplementary Material online.

Annotation

We annotated the *A. davidsoni* genome assembly using MAKER v3.01.03 (Campbell et al. 2014). To do this, we first masked and annotated repeats using RepeatModeler2 (Flynn et al. 2020) and RepeatMasker (Smit et al. 2013–2015) as described by Heckenhauer et al. (2019). We then predicted genes with the homology–based gene prediction tool GeMoMa v1.6.4 (Keilwagen et al. 2016, 2018) using the annotated proteins from the previously published genome of *Rhyacophila brunnea* (Heckenhauer et al. 2021) as reference organism with the following command: *GeMoMa -Xmx50G*

GeMoMaPipeline threads=80 outdir=annotation_out GeMoMa.Score=ReAlign AnnotationFinalizer.r=NO o=true t=./atopsyche.masked.fasta i=rhy a=rhybru.gff g=rhybru.fasta. We then ran the MAKER pipeline to generate additional ab initio gene predictions (Campbell et al. 2014). We used the proteins predicted from GeMoMa for protein homology evidence with the augustus-generated gene prediction models from BUSCO. For EST evidence, we used the 1KITE transcriptome (1kite.org), 131012_I246_FCC2J5BACXX_L6_RINSinlTCBRABPEI-56.tsa.fas.

We added functional annotations to the predicted genes by first querying all predicted proteins against the ncbi-blast 2.9.0+ nonredundant protein databases using BlastP with an e-value cutoff of 10^{-4} and –max_target_seqs set to 10. We then used the command line version of Blast2GO v.1.4.4 (Götz et al. 2008) to assign GO terms and functional annotations to the predicted proteins (supplementary figs. 3–5, Supplementary Material online).

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Author Contributions

Conceptualization: B.R.-T., R.W.H., S.U.P., and P.B.F. Methodology: B.R.-T., J.H., and P.B.F. Validation: E.R.-G., J.H., and P.B.F. Formal analysis: B.R.-T., E.R.-G., R.W.H., J.H., C.G.S., and P.B.F. Investigation: P.B.F. Resources: B.R.-T., R.W.H., and P.B.F. Data curation: E.R.-G., R.W.H., C.G.S., and P.B.F. Writing-original draft preparation: B.R.-T., R.W.H., E.R.-G., J.H., and P.B.F. Writing-review and editing: all authors. Visualization: R.W.H. Supervision: B.R.-T. and P.B.F. Project administration: B.R.-T., R.W.H., and P.B.F. Funding acquisition: B.R.-T., R.W.H., S.U.P., and P.B.F.

Data Availability

This project has been deposited at NCBI under BioProject: PRJNA741212, BioSample: SAMN20492399, assembly accession: JAILWY000000000, and the Sequence Read Archive (SRA): SRR15654800. The annotation and predicted proteins have been deposited in FigShare and at the following doi: https://doi.org/10.6084/m9.figshare.16836871.v1. Supplementary video 1, which shows the sampling locality can be found in FigShare at the following doi: https://doi.org/10.6084/m9.figshare.16837963.v1.

Literature Cited

Buytaert W, Cuesta-Camacho F, Tobón C. 2011. Potential impacts of climate change on the environmental services of humid tropical alpine regions. Global Ecol Biogeogr. 20(1):19–33.

Campbell MS, Holt C, Moore B, Yandell M. 2014. Genome annotation and curation using MAKER and MAKER-P. Curr Protoc Bioinformatics. 48(1):4.11.1–4.11.39.

Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. 18(2):170–175.

Duchicela SA, et al. 2021. Microclimatic warming leads to a decrease in species and growth form diversity: insights from a tropical alpine grassland. Front Ecol Evol. 9:673655.

Encalada AC, et al. 2019. A global perspective on tropical Montane rivers. Science 365(6458):1124–1129.

Finn DS, Encalada AC, Hampel H. 2016. Genetic isolation among mountains but not between stream types in a tropical high-altitude mayfly. Freshw Biol. 61(5):702–714.

Flynn JM, et al. 2020. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A. 117(17):9451–9457.

Gomes V, Calor AR. 2019. *Atopsyche* Banks (Trichoptera, Hydrobiosidae): new species, redescription, and new records. Zootaxa 4567(3):zootaxa.4567.3.8.

Götz S, et al. 2008. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 36(10):3420–3435.

Hall M, Mothes P. 2008. The rhyolitic–andesitic eruptive history of Cotopaxi volcano, Ecuador. Bull Volcanol. 70(6):675–702.

Heckenhauer J, et al. 2019. Annotated draft genomes of two Caddisfly species *Plectrocnemia conspersa* CURTIS and *Hydropsyche tenuis* NAVAS (Insecta: Trichoptera). Genome Biol Evol. 11(12):3445–3451.

Heckenhauer J, et al. 2021. Genome size evolution in the diverse insect order Trichoptera. bioRxiv: 2021. 2005.2010.443368. doi: 10.1101/2021.05.10.443368.

Hotaling S, Kelley JL, Frandsen PB. 2020. Aquatic insects are dramatically underrepresented in genomic research. Insects 11(9):601.

Jacobsen D. 2004. Contrasting patterns in local and zonal family richness of stream invertebrates along an Andean altitudinal gradient. Freshwater Biol. 49(10):1293–1305.

Jacobsen D. 2008. Tropical high-altitude streams. In: Dudgeon D, editor. Tropical stream ecology. Amsterdam, Netherlands: Elsevier. p. 219–256.

Jacobsen D. 2020. The dilemma of altitudinal shifts: caught between high temperature and low oxygen. Front Ecol Environ. 18(4):211–218.

Jacobsen D, et al. 2014. Temporal variability in discharge and benthic macroinvertebrate assemblages in a tropical glacier-fed stream. Freshwater Sci. 33(1):32–45.

Jacobsen D, Rostgaard S, Vasconez JJ. 2003. Are macroinvertebrates in high altitude streams affected by oxygen deficiency? Freshwater Biol. 48(11):2025–2032.

Keilwagen J, et al. 2016. Using intron position conservation for homology-based gene prediction. Nucleic Acids Res. 44(9):e89.

Keilwagen J, Hartung F, Paulini M, Twardziok SO, Grau J. 2018. Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. BMC Bioinformatics 19(1):189.

Kokot M, Długosz M, Deorowicz S. 2017. KMC 3: counting and manipulating k-mer statistics. Bioinformatics 33(17):2759–2761.

Kriventseva EV, et al. 2019. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic Acids Res. 47(D1):D807–D811.

Laetsch DR, Blaxter ML. 2017. BlobTools: interrogation of genome assemblies. F1000Research 6:1287.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34(18):3094–3100.

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol. 38(10):4647–4654.

Olsen LK, et al. 2021. Draft genome assemblies and annotations of *Agrypnia vestita* Walker, and *Hesperophylax magnus* Banks reveal substantial repetitive element expansion in tube case-making caddisflies (Insecta: Trichoptera). Genome Biol Evol. 13(3):evab013.

Pacific Biosciences. 2021. pbbioconda. [cited 2021 Jun]. Available from: https://github.com/PacificBiosciences/pbbioconda.

Ranallo-Benavidez TR, Jaron KS, Schatz MC. 2020. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat Commun. 11(1):1432.

Ríos-Touma B, Holzenthal RW, Huisman J, Thomson R, Rázuri-Gonzales E. 2017. Diversity and distribution of the caddisflies (Insecta: Trichoptera) of Ecuador. PeerJ 5:e2851.

Sklenář P, Kučerová A, Macek P, Macková J. 2010. Does plant height determine the freezing resistance in the páramo plants? Austral Ecol. 35(8):929–934.

Smit A, Hubley R, Green P. 2013–2015. RepeatMasker Open-4.0. [cited 2021 Jun]. Available from: http://www.repeatmasker.org.

Thomas JA, Frandsen PB, Prendini E, Zhou X, Holzenthal RW. 2020. A multigene phylogeny and timeline for Trichoptera (Insecta). Syst Entomol. 45(3):670–686.

Trizna M. 2020. assembly_stats 0.1.4. doi:10.5281/zenodo.3968775.

Wiggins GB. 2004. Caddisflies: the underwater architects. Toronto, Canada: University of Toronto Press.

Zhang Z, Schwartz S, Wagner L, Miller W. 2000. A greedy algorithm for aligning DNA sequences. J Comput Biol. 7(1–2):203–214.