# **The impact of sequencing depth and relatedness of the reference genome in population genomic studies: a case study with two caddisfly species (Trichoptera, Rhyacophilidae, *Himalopsyche*)**

**Impact of reference genome and sequencing depth**

Xi-Ling Deng1,2,3,\*, Paul B. Frandsen3,4,5, Rebecca B. Dikow5, Adrien Favre1,6, Deep Narayan Shah7, Ram Devi Tachamo Shah8,9, Julio V. Schneider1, Jacqueline Heckenhauer1,3,+,Steffen U. Pauls1,2,3,+

1Department of Terrestrial Zoology,Entomology III, Senckenberg Research Institute and Natural History Museum, Senckenberganlage 25, D‐60325 Frankfurt/Main, Germany

2Institute of Insect Biotechnology, Justus-Liebig-University Gießen, Heinrich-Buff-Ring 26, Heinrich-Buff-Ring 26, 35392 Gießen, Germany

3LOEWE Centre for Translational Biodiversity Genomics (LOEWE-TBG), Senckenberganlage 25, D‐60325 Frankfurt/Main, Germany

4Department of Plant & Wildlife Sciences, Brigham Young University, Provo, UT 84602, USA

5Data Science Lab, Office of the Chief Information Officer, Smithsonian Institution, Washington, DC 20560, USA

6Regional nature park of the Trient Valley, la Place 24, 1922 Salvan, Switzerland

7Central Department of Environmental Science, Tribhuvan University, Kirtipur, Nepal

8Aquatic Ecology Centre, School of Science, Kathmandu University, Dhulikhel, Nepal

9Department of Life Sciences, School of Science, Kathmandu University, Dhulikhel, Nepal

\*Corresponding author: [xiling.deng@senckenberg.de](mailto:xiling.deng@senckenberg.de) / xiling.deng@outlook.com

+equal contribution authors

# **Abstract**

Whole-genome sequencing for generating SNP data is increasingly used in population genetic studies. However, obtaining genomes for massive numbers of samples is still not within the budgets of many researchers. It is thus imperative to select an appropriate reference genome and sequencing coverage to ensure the accuracy of the results for a specific research question, while balancing cost and feasibility. To evaluate the effect of the choice of the reference genome and sequencing coverage on downstream analyses, we used five confamilial reference genomes of variable relatedness and three levels of sequencing coverage (3.5x, 7.5x and 12x) in a population genomic study on two caddisfly species: *Himalopsyche digitata* and *H. tibetana*. Using these 30 datasets (five reference genomes × three coverages × two target species), we estimated population genetic indices (inbreeding coefficient, nucleotide diversity, pairwise and genome-wide *F*ST) based on variants and population structure (PCA and admixture) based on genotype likelihood estimates. The results showed that both distantly related reference genomes and lower sequencing coverage lead to degradation of resolution. In addition, choosing a more closely related reference genome may significantly remedy the defects caused by low coverage. Therefore, we conclude that population genetic studies would benefit from closely related reference genomes, especially as the costs of obtaining a high-quality reference genome continue to decrease. However, to determine a cost-efficient strategy for a specific population genomic study, a trade-off between reference genome relatedness and sequencing depth can be considered.

# ***Key words*:** aquatic insects; *de novo* genomes; whole genome resequencing; reference genomes; sequencing coverage; population genomics

# **Introduction**

As high-throughput sequencing (HTS) technologies and bioinformatic tools are rapidly becoming more accurate and increasingly affordable, it is possible to generate whole-genome resequencing (WGR) data for almost any species. High quality WGR data provide a remarkable amount of information, including a vast number of loci as well as a large number of genetic variants, thus enabling powerful population genomics analyses (Goodwin et al., 2016). Today, low-coverage whole genome resequencing (lcWGR) is widely applied in population studies (Lou et al., 2021; Nielsen et al., 2011). When initiating a project on population genetics using lcWGR, there are two prerequisites: (i) the availability of a reference genome representing the focal species (Ellegren, 2014) and (ii) estimating the necessary sequencing coverage to support accurate results (Meisner & Albrechtsen, 2018).

The selection of reference genome and sequencing depth are therefore two important features in a population genetic study. However, with a fixed budget, it is important to find a balance between data quality and sequencing costs by compromising on the reference genome or sequencing depth. For every study, the reference genome needs to be carefully chosen to avoid bias in mapping and variant calling, considering the amount of sequence identity between the reference genome and the data resulting from resequencing (Nielsen et al., 2011). Ideally, such studies should include a high-quality species-specific reference genome, which is often not available for non-model organisms. Given the costs and time associated with generating a *de novo* reference genome, it can be more realistic to use an existing one, yet more distantly related, as is done most often in population genomic studies (Duchen & Salamin, 2021). Currently, several empirical studies have examined the impact of non-conspecific reference genomes in population genomics. For example, Gopalakrishnan et al. (2017) compared the use of dog and wolf genomes as reference genomes for either domestic dog or wolf populations. Their results showed that the selection of the reference genome only had a minor influence on downstream analyses, probably because of the close relatedness of these two taxa, which diverged approx. 30,000 years ago (Skoglund et al., 2015; Wang et al., 2013). In contrast, other studies have found that the use of distantly related reference genomes bias the results of resequencing analyses in bacteria (Valiente-Mullor et al., 2021), fungi (Garcia-Rubio et al., 2018), and mammals (Yang et al., 2019), but no studies are yet available for insects.

In comparison to choosing the reference genome based upon relatedness, sequencing depth needs more preliminary knowledge to determine. The sequencing depth needs to be tailored to each particular study based on many aspects, for instance genome size of the target species, the availability of funding, and of course the research question. Using low coverage sequencing in a population genetic study may result in data loss, thus causing statistical uncertainty during genotype and variant calling (Crawford & Lazzaro, 2012; Meisner & Albrechtsen, 2018; Nielsen et al., 2012). This statistical uncertainty is mainly due to the limited information provided by the low amount of reads, leading to poor discrimination between sequencing error and real variation, i.e. SNPs (Meisner & Albrechtsen, 2018). To improve the accuracy of estimates in a cost-limited population genetic study, especially using low coverage sequences, some researchers therefore demonstrated that employing a large sample size is efficient (Buerkle & Gompert, 2013; Fumagalli, 2013; Han et al., 2013; Sims et al., 2014). However, we argue that in the first place, an appropriate reference genome may improve the accuracy of the results even with a low sequencing depth and regardless of the sampling size. Therefore, comprehensively exploring the influence of reference choice and sequencing depth on downstream analyses is urgently needed, particularly for insects. In this study, we will evaluate the effects of the reference genome and depth of sequencing coverage on a genus of caddisflies, for which we already have significant molecular and taxonomic data (Deng et al., 2021).

*Himalopsyche* is a genus of caddisflies (Insecta: Trichoptera) that is primarily distributed in the mountainous areas of central and east Asia (Hjalmarsson et al., 2018). Their larvae live as free-roaming predators in cool fast-flowing rivers and streams, and are regarded as bioindicators of water quality (Hjalmarsson, 2019; Morse et al., 2019; Tsuruishi et al., 2006). There are currently 56 named species of *Himalopsyche* (Hjalmarsson, 2019), which are divided into five species groups: the *tibetana* group, the *lepcha* group, the *kuldschensis* group, the *phryganea* group, and the *navasi* group (Hjalmarsson et al., 2019). The species *H. digitata* and *H. tibetana* both belong to the *tibetana* group, which is distributed in the Himalayas. This area is characterized by a number of parallel north-south running river systems (such as the Karnali, the Narayani and the Koshi) with sharp elevational gradients. Currently, climate change is causing a number of cascading effects on river flow via rapidly receding glaciers greatly affecting aquatic biodiversity (Xu et al., 2009). It is therefore crucial to investigate the patterns of genetic diversity of aquatic insects in the region to understand the current and past ecological processes, in order to promote the conservation of freshwater biodiversity (Geist, 2011). Genome-wide analysis is an important conservation tool that can provide novel insights essential for e.g., identifying hotspots or reservoirs of genetic diversity, dispersal routes, ecological corridors and stepping stone habitats (Barbosa et al., 2018; Brandies et al., 2019; Hohenlohe et al., 2021; Jasper et al., 2019). However, the quality and quantity of aquatic insect genomes are still relatively low compared with terrestrial insects (Hotaling et al., 2020; Hotaling et al., 2021). Despite Trichoptera covering ~275 million years of evolution (Thomas et al., 2020) and comprising ~16,300 known species (Morse et al., 2019), only 29 Trichoptera genome assemblies (26 species) have been published to date (Heckenhauer et al., 2019; Heckenhauer et al., 2022; Luo et al., 2018; Olsen et al., 2021; Ríos-Touma et al. 2022). This represents less than 0.15% of all known species, which limits progress in genomics-based research of this ecologically relevant group.

The number of studies using population genomics is rapidly increasing. With it, the need to test the effect of reference genome's selection and sequencing coverage on the results. Indeed, such studies will allow to find a balance between data quality and sequencing costs by compromising on the reference genome or sequencing depth. Therefore, we conducted a case study using an empirical dataset to evaluate how reference genome selection, i.e., degree of relatedness, and sequencing depth affect downstream population genetic analyses of the species *H. digitata* and *H. tibetana*. In addition, we tried to reveal the correlation between the inferences from population genetics and drainage network, and provided insights for local biodiversity conservation of these enigmatic species.

# **Materials and methods**

## Study design

As illustrated in Fig. 1, first, we generated three new *de novo* whole genome assemblies for *H. tibetana* (*tibetana* group), *H. japonica* (*navasi* group)and *H. sp.* (*kuldschensis* group) sensu Hjalmarsson et al. 2019, in addition to two previously published genomes (i.e., *H. phryganea* (*phryganea* group) and *R. brunnea*; Heckenhauer et al., 2022). Since the specimens of *Himalopsyche sp* was collected as a larva that can not be identified to species level by morphologic diagnosis, and the CO1 sequence of this specimen differed slightly from all hitherto known sequences of named species, we can only classify this specimen as a species belonging to the *kuldschensis* group*.* These five genomes represent a gradient of genetic relatedness with respect to our target species, which were used as reference genomes. We used populations of two *Himalopsyche* species (*H. digitata* and *H. tibetana*), each species including four populations with each population containing six individuals except for one population from *H. digitata* and *H. tibetana* respectively which contained only five individuals. The reads of the populations were subsampled into three separate datasets with an average coverage of 12.5×, 7.5× and 3.5×, respectively. Reads from each dataset were mapped to the five different reference genomes separately. Afterward, variants were called from all datasets using two strategies: Genotype calling with GATK and direct estimation of the genotype likelihood using ANGSD. Variants identified with the first strategy were used to calculate the population genetic indices including inbreeding coefficient (*F*), nucleotide diversity (π) and pairwise fixation index (*F*ST); variants estimated from the second strategy were used in the principal component analysis (PCA) and admixture analysis. Finally, we compared population genetic indices and population structure with different references and depths of coverage.

## *De novo* genomes of three reference species

We used the genome assemblies of four species of *Himalopsyche* (*H. tibetana*, *H. sp.*, *H. phryganea*, *H. japonica*) and of one species from the closely-related genus *Rhyacophila (R. brunneae*, both family Rhyacophilidae) as reference genomes. The *Himalopsyche* speciesrepresent the four major taxonomic groups in the genus according to Hjalmarsson et al.(2018): the *tibetana* group (*H. tibetana*), the *kuldschensis* group (*H. sp.*), the *phryganea* group (*H. phryganea*), and the *navasi* group (*H. japonica).* Unfortunately, we could not obtain a sample of *H. lepcha* (the only species in the *lepcha* group). *Rhyacophila* is the most closely related genus to *Himalopsyche* (Thomas et al., 2020). The genomes of *H. phryganea* (JAGVSL000000000) and *Rhyacophila brunnea* (previous version of JAGYXB000000000, available at https://doi.org/10.6084/m9.figshare.c.6033011.v1) were previously sequenced and assembled (Heckenhauer et al., 2022). We generated new *de novo* assemblies of *H. tibetana* (collected from the Ê Ghunsa River, Nepal), *H. sp.* (*kuldschensis* group, collected from the Ê Ghunsa River, Nepal*),* and *H. japonica* (collected from the Nogami River, Kiso-machi, Nagano Prefecture, Japan). Tissue of abdomen and thorax segments were used for DNA extraction after removal of the intestinal tract. We extracted high molecular weight genomic DNA using a salting-out protocol adapted from Miller et al. (1988), as described in Heckenhauer et al. (2019). We quantified the DNA using a Qubit 4.0 fluorometer with the dsDNA Broad Range Kit (ThermoFisher Scientific, Waltham, USA) and checked its purity with a DS11 spectrophotometer (DeNovix, Wilmington, DE, USA). We used a low-cost sequencing strategy that has been shown to produce contiguous genome assemblies, i.e., employing a combination of short (Illumina) and long-read (Oxford Nanopore) technologies to sequence the three new reference genomes, as described in Supplementary Material 1 (section 1.1), Supplementary Material online.

We conducted a long-read assembly of the Oxford Nanopore Technology sequencing reads with wtdbg2 v2.4 (Ruan & Li, 2020), followed by mapping and polishing with long-reads with Minimap2 v14 (Li, 2018) and Racon v1.3.1 (Vaser et al., 2017). We then performed another round of long-read polishing with nanopolish 0.11.1 (Loman et al., 2015) by first indexing the signal-level data in the FAST5 files using nanopolish index, re-aligning the long-reads to the Racon-polished assembly with minimap2 and then sorting and indexing the bam file with samtools. We used nanopolish\_makerange.py to split our draft genome assembly into 50kb segments and generated a consensus for each segment in parallel with nanopolish variants (--consensus --min-candidate-frequency 0.1). We generated the polished genome in FASTA format using nanopolish vcf2fasta. Because noisy long reads can suffer from indel errors, even with polishing, we further polished the assembly with high-quality short-read data withPilon v1.22 (Walker et al., 2014). To do this, we mapped quality trimmed Illumina reads to the “nanopolished” assembly with bwa-mem and sorted the read alignments by leftmost coordinates using the sort options in SAMtools v1.9 (Li et al., 2009). Finally, we used Pilon v1.22 (option --fix indels) to polish the assembly. Following polishing, we used purge\_dups 1.2.3 (Roach et al., 2018) to purge haplotigs and overlaps in the assembly based on read depth.

For *H. japonica*, this pipeline did not meet the expected quality regarding contiguity and BUSCO completeness. Thus, we conducted a *de novo* hybrid assembly with the raw Illumina data together with the long reads using MaSuRCA v.3.1.1 (Zimin et al., 2013; Zimin et al., 2017). In the config file we specified the insert size and a standard deviation (10% of insert size) for the Illumina reads, as well as jellyfish hash size (estimated\_genome\_size\*~long read coverage). All other parameters were left as defaults. We used purge\_dups 1.2.3 to purge haplotigs and overlaps in the assembly based on read depth.

We calculated assembly statistics with QUAST v5.0.2 (Gurevich et al., 2013) and examined completeness with BUSCO v4.1.4 (Simão et al., 2015; Waterhouse et al., 2018) using the Endopterygota odb10 dataset with the options --long, -m = genome and -sp= fly. A summary of the assembly statistics and BUSCO completeness is given in Table 1. The final genome assemblies were screened for potential contaminations with taxon-annotated GC-coverage (TAGC) plots using BlobTools v1.0 (Laetsch & Blaxter, 2017). For this purpose, all preprocessed Illumina reads of the respective species were mapped against the final genome assemblies using BWA-MEM v0.7.17-r1188 (Li, 2013) and taxonomic assignment for BlobTools was done with blastn using -task megablast and -e-value 1e-25. Details are given in Supplementary Material 1 (section 1.2).

Genome size estimation and profiling was conducted from the short-read sequence data with GenomeScope 2.0 (Ranallo-Benavidez et al., 2020; Vurture et al., 2017) as described in Supplementary Material 1 (section 1.3).

### Annotation

We identified and classified repetitive elements *de novo* and generated a library of consensus sequences for each genome using RepeatModeler 2.0 (Flynn et al., 2020). We then annotated and masked repeats in each assembly with RepeatMasker 4.1.0 (Smit et al., 2015) using the custom repeat library for the species generated in the previous step. After masking repeats, genes were predicted using the homology-based gene prediction tool GeMoMa v1.6.4 (Keilwagen et al., 2019; Keilwagen et al., 2016) and the two previously published species (*H. phryganea* = RG 1 and *R. brunnea* = RG 2) as reference organisms as follows: GeMoMa -Xmx50G GeMoMaPipeline threads=$SLURM\_NPROCS outdir=<out\_dir> GeMoMa.Score=ReAlign AnnotationFinalizer.r=NO o=true t=<genome to be annotated> s=own i=<name of RG 1> a=<RG 1.gff> g=<RG 1 assembly.fasta> s=own i=i=<name of RG 2> a= a=<RG 2.gff> g=g=<RG 2 assembly.fasta>.

For functional annotation of predicted genes, we first split each amino acid FASTA file into multiple files with 50 sequences each using awk, and then used blastp to search against the ncbi-blast2.9.0+ nr database with an e-value cutoff of 10−4, -max\_target\_seqs set to 10 and -out format 6. The resulting xml files were merged using cat and then functionally annotated using the command-line version of Blast2GO (Götz et al., 2008).

### Species tree reconstruction

To determine the phylogenetic relatedness among the five species, we estimated a species tree using single-copy orthologs resulting from the BUSCO analyses of the five genome assemblies with an additional species, *Glossosoma conforme*, as an outgroup (Heckenhauer et al., 2022). For each single-copy ortholog, we generated an unaligned FASTA file with sequences from each species. We then aligned each ortholog with the MAFFT L-INS-i algorithm (Katoh & Standley, 2013). We selected the best-fit substitution model for each alignment using ModelFinder (option -m mfp, (Kalyaanamoorthy et al., 2017)) in IQtree v.2.0.6 (Minh et al., 2020) and estimated a maximum likelihood tree with 1000 ultrafast bootstrap replicates (Hoang et al., 2018) with the BNNI correction (options -bb 1000 -bnni). We then generated a multi-species coalescent species tree in ASTRAL-III (Zhang et al., 2018) using the best maximum likelihood tree from each ortholog as input. We visualized the trees using FigTree v.1.4.4 (<http://tree.bio.ed.ac.uk/software/figtree/>).

***Cactus alignment and Hal***

To further characterize the variation among the five genomes, we computed a whole genome alignment using Cactus v1.0.0 (Armstrong et al., 2020) with a star tree of the five genomes as input: ((Rhyacophila\_brunnea, Himalopsyche\_kuldschensis, Himalopsyche\_phryganea, Himalopsyche\_japonica, Himalopsyche\_tibetana)mr).

We used HALtools (Hickey et al., 2013) to obtain global alignment information with the halStats option. We then used “halSummarizeMutations” to generate mutation statistics, including the length of each mutation and the number of substitutions, transitions, transversions, gaps, insertions, deletions, inversions, duplications, and transpositions among genomes.

## Population genomic analyses

### Taxon sampling

We conducted whole genome re-sequencing on a total of 47 individuals from four *H. tibetana* populations and four *H. digitata* populations, with 6 individuals from each population except for pop 1 of *H. digitata*, which included 5 individuals. All 47 samples were collected as larvae in April 2018 and March 2019 in Nepal. All four *H. digitata* populations and two populations of *H. tibetana* were collected in the headwaters of the Gandaki basin with two additional populations of *H. tibetana* collected in the headwaters of Koshi basin (Fig. 6). Since it is presently not possible to identify *Himalopsyche* larvae to species-level solely based on morphological characters, we ensured correct identification of these samples based on two molecular markers: the mitochondrial COI and the nuclear CAD, using the methods outlined in Hjalmarsson et al. (2018).

All the samples were preserved in 95% ethanol and archived in the collections of the Senckenberg Research Institute and Natural History Museum (SMF). Specimen and voucher information is shown in Supplementary Material 2 and Fig. 6.

### DNA extraction, library preparation and sequencing

Genomic DNA was extracted following a modified salting-out protocol adapted from Miller et al. (1988), as described in Heckenhauer et al. (2019). In case of low purity DNA (A260/A230 purity ratio below 1.4 and DNA concentration higher than 100 μg/uL that was measured by DeNovix DS11 spectrophotometer), the template DNA was subject to an additional cleanup using magnetic beads as described in the Supplementary Material 1 (section 2.1). Afterward, all the samples were sent to Novogene Co., Ltd. (Hongkong, China) for DNA library preparation and sequencing. 150 bp paired-end reads were generated on an Illumina HiSeq 2000 platform.

### Quality control, data processing of population sequence data and variant calling

We assessed the quality of Illumina reads before and after each step using FastQC v0.11.8 (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) and MultiQC v1.7 (Ewels et al., 2016), details are described in Supplementary Material 1 (section 2.2). We trimmed overrepresented *k*-*mers* using autotrim.pl v0.6.1 (Waldvogel et al., 2018) with Trimmomatic v0.38 (Bolger et al., 2014), and we removed adapters with Cutadapt v2.23 (Martin, 2011). After quality filtering, the average sequence coverage was 18× and the minimal coverage was 12.5× for all the samples. To maintain even coverage of all the individuals and to simulate subsets with varying coverage levels, we chose 12.5×, 7.5× and 3.5× as our test coverage levels. We randomly subsampled the sequencing reads to a specified coverage using rasusa 0.3.0. (Hall, 2019) with a random seed of 1 (-s 1), the target coverage (--coverage 12.5, --coverage 7.5, --coverage 3.5) and the estimated genome size (--genome-size 550m) based on GenomeScope2 estimation (see below) (Ranallo-Benavidez et al., 2020; Vurture et al., 2017).

For population genomic analyses, FASTA files of the reference genomes were indexed with the function bowtie2-build of bowtie2 2.3.5 (Langmead & Salzberg, 2012). We then mapped the reads of the three different coverage datasets (12.5×, 7.5× and 3.5×) of the *H. digitata* and *H. tibetana* populations to each of the five different reference genomes separately using Bowtie2, which resulted in 30 datasets (2 species × 3 coverage × 5 reference genome, Fig. 1). After marking duplicate reads for each dataset using Picard v2.20.8 ([Picard Tools – By Broad Institute](http://broadinstitute.github.io/picard/)), we conducted variant calling for downstream population genetic analysis using two strategies. In the first strategy, we called the haplotype of each individual separately, by running GATK v4.1.7.0 ([GATK, broadinstitute.org](https://gatk.broadinstitute.org/hc/en-us)) HaplotypeCaller on each bam file. Then, we called genotypes with GATK GenotypeGVCFs across all resulting vcf files. Ultimately, we selected the variants and filtered missing sites using VCFtools v0.1.17 (Danecek et al., 2011). With the second strategy, we estimated genotype likelihoods with appropriate filtering (including reads/sites/alleles/depth filtering, non-missing individual, and SNP filtering) using ANGSD v0.931 (Korneliussen et al., 2014). For details and parameters used see Supplementary Material 1 (section 2.3 and 2.4).

### Genetic diversity and population structure analysis

The variants called by GATK with genotype calling were used to estimate π, individual *F*, and pairwise *F*ST (Weir & Cockerham, 1984). These are commonly used measurements for population genetics. *F* is used to directly quantify the alleles inherited from common ancestors in an individual’s lineage, π is an index for population-level genetic diversity quantification, and *F*ST provides a primary description of population genetic differentiation. In this study, π and *F*ST were calculated using VCFtools with a 50 kb window size. To better understand *F*ST within different datasets (species × reference genome × coverage), we applied pairwise *F*ST estimates (pairwise populations among the four populations) and global *F*ST estimates (including all the four populations).

We used the genotype likelihoods estimated by ANGSD to generated a PCA with [PCAngsd](http://www.popgen.dk/software/index.php/PCAngsd) (Meisner & Albrechtsen, 2018) and individual admixture proportions estimating using NgsAdmix (Skotte et al., 2013) after Linkage pruning with [ngsLD](https://github.com/fgvieira/ngsLD) (Fox et al., 2019). Further details about tools, parameters and commands were described in the Supplementary Material 1 (section 2.4). Most plots were generated in Rstudio v3.6.1 ([http://www.rstudio.org](http://www.rstudio.org/)) (scripts for plotting were included in the Supplementary Material 1 (section 3)).

# **Results**

## New genomic resources

We combined long- and short-read sequencing technologies to generate three new *de novo* genome assemblies for the genus *Himalopsyche* (Rhyacophilidae): *H. japonica*, *H. sp.* (*kuldschensis* group), and *H. tibetana*. For each species we obtained ~150–200× Illumina and ∼18–26× Oxford Nanopore sequence coverage. All three assemblies are of high-quality with respect to the number of contigs and contig N50 (Table 1). We identified >96% of the Endopterygota BUSCO gene set in the assemblies. BlobTools detected no contamination (Supplementary Material 1: figures 1–3). Re-mapping the Illumina reads back to the assemblies revealed more than 98% could be unambiguously placed (Supplementary Material 1: figures 1–3).

The estimated genome sizes resulting from the *k*-*mer* based estimation with Genomescope2 were 481 Mb (*H. japonica*), 495 Mb (*H. sp.* (*kuldschensis* group)) and 568 Mb (*H. tibetana*) (Supplementary Material 1: figures 4–6). Between 31% (*H. japonica*) and 44% (*H. tibetana*) of the genome assemblies were identified as repeats. A high percentage of the repeats were classified as interspersed repeats (approx. 28.5–40.3%). More than half of the interspersed repeats remain unclassified and therefore may be specific to Trichoptera. Details on repeat classes are given in Supplementary Material 1 (tables S1–3). The annotation of the genomes resulted in the prediction of 9,983 (*H. japonica*), 10,049 (*H. sp.* (*kuldschensis* group)) and 10,994 (*H. tibetana*) proteins. Most of the annotated proteins had functional Blast2GO annotations, were verified by BLAST or were mapped to GO terms. GO Distributions were similar to previously annotated caddisfly genomes, i.e., the major biological processes were cellular processes. Catalytic activity was the largest subcategory in molecular function and the cell membrane subcategories were the largest cellular component (Supplementary Material 1: figures 7–12).

The Cactus alignment of the five reference genomes showed that *Rhyacophila brunnea* had the longest sequence length and largest number of contigs (Supplementary Material 4). It also showed a higher level of mutations compared with the four *Himalopsyche* genomes, for instance gaps, insertions, inversions and duplications. The number of mutations among the four *Himalopsyche* genomes were at similar levels in the Cactus alignment over all.

## Phylogenetic relationships of the reference genomes – resolving the *Himalopsyche* backbone

We built a species tree for the five reference genomes to estimate their evolutionary history, and to verify the phylogenetic relationship between the reference genomes and the two population-level target species, *H. digitata* and *H. tibetana*. Phylogenetic relationships were strongly supported (Fig. 2). Our phylogenetic tree showed two pairs of sister species (*H. tibetana* + *H. sp.* (*kuldschensis* group); *H. phryganea* + *H. japonica*) with *R. brunnea* forming the sister clade of all four *Himalopsyche* species. According to Hjalmarsson et al. (2018; 2019), *H. digitata* and *H. tibetana* both belong to the *tibetana* group, *H. sp.* to the *kuldschensis* group, *H. phryganea* to the *phryganea* group and *H. japonica* to the *navasi* group. In contrast to our results Hjalmarsson et al. 2019 recovered the *navasi* group sister to all other Himalopsyche and a sister relationship between the *phryganea* and *kuldschensis* groups. Differences could be related to the different sampling strategies. Here we present much more genome-wide data per taxon, but reduced taxon sampling. For the purpose of assessing the impact of reference genome relatedness in SNP calling, *H. tibetana* and *H. digitata* (both belonging to *tibetana* group) are most closely related to *H. tibetana*, followed by *H. sp.* (*kuldschensis* group), *H. phryganea* and *H. japonica*, with *R. brunnea* being the most distantly related species.

## The impacts of sequencing depth and phylogenetic relatedness of the reference genome on population genetic studies

To better understand the impacts of sequencing depth and reference genome on population genomic analyses, we compared the number of quality-filtered variants, genetic diversity and population structure of the species *H. digitata* and *H. tibetana* among the datasets based on different reference genomes and varied sequencing coverage. The results revealed that, perhaps unsurprisingly, the most distantly related reference genome, or the lowest sequencing coverage resulted in the least accurate downstream analysis.

We observed that reference genome selection has a measurable impact on the numbers of variants (Fig. 3). With both strategies (variant calling using GATK and genotype likelihood estimation using Angsd), the number of variants sharply decreased when the reference genome was more distantly related, but remained similar with decreasing sequencing depth, especially when estimated by genotype likelihood. This was particularly striking for the *H. tibetana* populations when using ANGSD to call the variants, which dropped from millions to thousands when changing the reference genome from *H. tibetana* to the others (Fig. 3). Consequently, the massive loss of information resulting from choosing a distant reference genome is likely to lead to poor performance in the downstream analyses.

We observed that the choice of reference genome influenced population genetic values such as pairwise *F*ST (Supplementary material 3) and nucleotide diversity estimates (Supplementary Figure 14), the number of outliers in the inbreeding coefficient estimates (Supplementary Figure 13), as well as the resolution of the final outcome in genome-wide *F*ST estimates (Fig. 4), PCA and Admixture (Fig. 5). For example, in the case of pairwise *F*ST estimates, when changing the reference genome from *H. tibetana* to *R. brunnea* (with the same coverage, e.g., 12.5×), the *F*ST value between pop 1 and pop 3 of *H. tibetana* reduced from 0.17 to 0.06 (Supplementary material 3). The number of outliers (*F* < 0) in inbreeding coefficient estimates increased when using a reference genome that is distantly related to the target species, while the inbreeding coefficient values of individuals tended to fluctuate for both species, which might result in an ineffective or misleading conclusion (Supplementary Figure 13). The results of the PCA, admixture analysis and the genome-wide *F*ST clearly showed notable decreases in the resolution of the plots when selecting a more distantly related reference genome (Fig. 5). For instance, for populations of *H. digitata*, regardless of the coverage, the cluster of pop 1 was well defined when mapped to *H. tibetana* in the PCA plot and admixture analysis, but no distinguishable structure was discernible among all four populations when mapped to *R. brunnea* (Fig. 5a). In addition, using a conspecific reference genome significantly improved the accuracy of genome-wide *F*ST estimates, including both the global *F*ST value (the global *F*ST value of *H. tibetana* populations doubled when mapped to *H. tibetana*) and the distribution of *F*ST values from sliding windows. This effect was particularly pronounced with the abnormal value (highly qualified resolution, Fig. 5b).

We also observed extensive influence of sequencing coverage on the downstream analyses. In addition to decreasing numbers of variants, low sequence coverage also affected estimates of nucleotide diversity, the inbreeding coefficient, pairwise and global *F*ST, and population structure. In most cases, low coverage reduced the accuracy of these inferences, however, the influence proved to be negligible or limited in some treatments. For example, the population structure of *H. tibetana* was highly differentiated regardless of coverage when mapped to *H. tibetana*, and was less differentiated when coverage decreased while mapped to *H. phryganea*, but still sufficient to obtain a reliable result (Fig. 5b). Likewise, the *F* value of all individuals changed when the coverage decreased from 12.5× to 3.5×, but the differentiation among populations remained comparable (Supplementary Figure 13).

Furthermore, the results revealed that reference genome and coverage had variable impacts between the populations of *H. digitata* and *H. tibetana*. More specifically, the same treatment (same reference genome and coverage) might be sufficient to generate a reliable result for the populations of *H. tibetana*, but not for *H. digitata*. For instance, when mapped to *H. japonica*, the population structure of the *H. tibetana* populations was distinguishable regardless of coverage, but it was barely detectable for the *H. digitata* populations even with 12.5× coverage (Fig. 5). Considering the inherent genetic variation of the two species, it is not surprising that more distantly related reference genome or lower coverage were more tolerable for the populations with higher genetic diversity.

## Genetic diversity and population structure of *H. digitata* and *H. tibetana*

Each population of *H. digitata* had a similar level of nucleotide diversity, but less consistent *F* values: pop 4 was slightly higher (0.31), whereas the other three were similar (0.22 - 0.27, Fig. 6d ). Population 1, which is located in the main river of Gandaki basin in central western Nepal, formed a distinct cluster in the PCA analyses (Fig. 6a, b). Moreover, no admixture signal was detected in pop 1. As a population located in the middle of the basin, but more closely connected with pop3 and pop 4 by catchment, pop 2 formed a distinct cluster and adjoined pop3 and pop4 in the PCA analysis. Meanwhile, pop 2 was represented as genetic mixtures regardless of K values. Population 3 and 4 were mixed together in the PCA analyses, which was in line with the fact that they were located closely together in the most north-eastern tributaries of the Gandaki basin. In addition, pop 3 and pop 4 were homogeneous when K = 2, but mixed when K = 3 or 4. This population structure was further supported by the pairwise *F*ST estimates using the base-called variants (Fig. 6f).

Following the pattern of *H. digitata*, the populations of *H. tibetana* had a very similar level of both nucleotide diversity and inbreeding coefficient (Fig. 6e), and a more distinct population structure congruent with catchments (Fig. 6c, f). Moreover, compared with *H. digitata*, *H. tibetana* showed a greater population diversity among the four populations, including higher nucleotide diversity (~ 12-fold) and *F*ST (both pairwise and globalwise, Fig. 6f, Fig. 4), lower inbreeding coefficient (~ 0.5-fold), as well as a more distinct population structure (Fig. 6b, c). This is in accordance with the fact that the geographic locality and catchment connection of the *H. tibetana* populations were further apart from each other compared to the ones of the *H. digitata* populations.

# **Discussion**

## Reference genomes

In terms of BUSCO completeness and contiguity, the three *de novo* genomes provided in this study as references are of comparable quality than the other Trichoptera genomes previously published (Heckenhauer et al., 2019; Heckenhauer et al., 2022; Luo et al., 2018; Olsen et al., 2021; Rios-Touma et al. 2022). They contained 96% – 97% of an Endopterygota core gene collection indicating an almost complete coverage of known single copy orthologs in the assembly. The backmapping rate of Illumina reads to the assemblies ranged between 96 and 98% which also indicates the high quality of the assemblies.

Previous genomic studies have focused on sequencing a wide range of different families of Trichoptera and investigating variations across the order (Heckenhauer et al., 2022). These three new genome assemblies provide important new genomic resources for the scientific community, especially since Trichoptera and other aquatic insects are in general underrepresented in genomic research (Hotaling et al., 2020; Hotaling et al., 2021). Together with previously published ones (*H. phryganea* and *R. brunnea*), these newly available genomes adequately provide an initial perspective about the phylogenetic relationship of the four main taxonomic groups of *Himalopsyche*, as well as implying the genetic relatedness between the target species and the five different reference genomes respectively for this study.

## Impacts of reference genome and sequencing coverage on population genetic inferences

High-throughput sequencing of individual specimens is poised to become the state-of-the-art in population genetics studies. However, despite falling prices in sequencing this approach can be prohibitively costly for large sample sizes in species with large genomes. Many researchers may thus face the choice of either using an existing non-conspecific reference genome or lowering sequencing depth of individual samples. Previous studies have either demonstrated that the reference genome or the sequencing coverage have a direct influence on population genetic estimates (Garcia-Rubio et al., 2018; Gopalakrishnan et al., 2017; Valiente-Mullor et al., 2021; Yang et al., 2019). Studies assessing the joint effects of the reference genome and sequencing coverage are rare and hitherto lacking in insects. Here, we evaluated population diversity and structure based on a range of reference genomes and varied levels of low sequencing coverage in the first case study on insects. The results revealed that the choice of the reference genome and sequencing coverage both had an influence on estimating population genetic indices, including *F*, π and *F*ST, as well as the genetic structure inferred from genotype likelihoods. To some extent, the general trends are that a) the more closely related the reference genome is, the more stable the estimates of population genetic indices are, and b) the higher the sequencing depth, the better the resolution of the population structure analyses. However, the results also vary depending on the inherent genetic variation of the target species.

The choice of reference genome has a stronger influence on downstream analyses than resequencing coverage, which is consistent with previous studies (Garcia-Rubio et al., 2018; Günther & Nettelblad, 2019; Valiente-Mullor et al., 2021). This is mainly due to the massive loss of reads while mapping to a distantly related reference genome. For example, in *H. tibetana,* mapping rates decrease from ~98% when using a conspecific reference genome, to 10% with a reference genome of a species of another genus (*Rhyacophila brunnea*) (Supplementary material 5). This leads to a dramatic decline of variants (Fig. 3). Moreover, increasing mismatches may also occur due to alternative alleles, thus increasing the so-called “reference bias” (Günther & Nettelblad, 2019). Consequently, reference bias can impact variant calling by missing alternative alleles or by incorrectly calling heterozygous sites and therefore lead to an underestimation of variants, including rare/private variants (Günther & Nettelblad, 2019; Taub et al., 2010). Considering these two aspects, the effects related to the choice of reference genome may propagate to a certain degree to subsequent downstream analyses in a population genetic study, for example when investigating heterozygosity and genetic diversity, gene flow, as well as ancestry proportions (Brandt et al., 2015; Günther & Nettelblad, 2019). We have observed these in our results: the number of variants decreases logarithmically with genetic relatedness of the reference genome, especially for the genotype likelihood dataset called by ANGSD, which includes a strict filtering on base level, read level, sequencing depth and other levels during the genotype calling. Moreover, when the genetic relatedness between reference genomes and target species decreases, the population genetic indices including *F*, π and pairwise *F*ST tend to become less accurate especially at a low coverage (3.5×). In addition, estimates of population structure, including PCA and admixture, are strongly affected, resulting in unreliable, weakly supported findings.

We also show that sequencing depth must be considered when designing population genomic analyses. Unlike *de novo* genome assembly which demands high sequencing depth, highly accurate results can be achieved with lower sequencing depth in population genomics by combining information from a large number of individuals either during SNP calling or other processes (Buerkle & Gompert, 2013; Fumagalli, 2013; Han et al., 2013). Even though the accuracy of population genetic inferences can be improved by increasing the number of samples, the bias caused by low sequencing depth cannot be ignored, especially since it can often be difficult to obtain many samples for some populations of rare animals. As revealed by previous studies, low coverage may cause erroneous SNP calls, due to the errors introduced and amplified during PCR during library prep (Sims et al., 2014). Moreover, it may also produce ambiguous reads during mapping to the reference genome (Taub et al., 2010). Such biases may lead to incorrect conclusions in population genetics inferences, including population genetic differentiation, population structure and demography (Crawford & Lazzaro, 2012; Fumagalli, 2013; Han et al., 2013; Jiang et al., 2019; Korneliussen et al., 2013). Our results show that decreasing coverage massively reduced the number of informative sites, which may largely result from the variant calling step, especially when applying a series of filtering approaches, thus leading to a less accurate result in downstream analyses. However, sequencing coverage may have limited impact in some other cases: the results are not affected by coverage when the reference genome is either very closely or very distantly related to the target species. For example, when using *H. tibetana* as reference genome with *H. digitata* (the most closely related reference), population structures are distinguishable regardless of sequencing depth (Fig. 5a). However, when using *H. japonica* or *R. brunnea* as the reference genome (most distantly related), population structures are indistinguishable regardless of the sequencing depth. Therefore, increasing the sequencing depth may not improve the results in such cases where the reference genome is too distantly related.

To conclude, both reference genome and coverage have various degrees of influence on downstream analyses, whereas their respective impact is different for each target species. Our results imply that populations with a higher genetic diversity are less affected by the relatedness of the reference genome and the sequencing coverage in population structure analyses. In general, the results for *H. tibetana* appeared more robust despite the variation of reference genome and coverage in comparison to those obtained for *H. digitata*. This may result from (1) different pairwise relatedness between reference genome and target species, and (2) inherent population variation (or expected level of differentiation). Even though *H. tibetana* and *H. digitata* are very closely related and both belong to the *tibetana* group (Hjalmarsson et al., 2019), populations of *H. tibetana* show a higher level of population variation compared to those of *H. digitata*, as, for instance, shown by higher *F*ST value and nucleotide diversity of populations. This is likely caused by different distribution patterns with *H. tibetana* populations being more isolated at high elevations than *H. digitata.* A detailed investigation of this was not intended with this study and our sampling is insufficient and thus inconclusive regarding the underlying biological reasons for the differences in species-specific population structures.

Similar to the trade-off between sequencing depth and sample size demonstrated by previous studies (Buerkle & Gompert, 2013; Fumagalli, 2013), we reveal that the roles of reference genome and coverage in a study of population genetics could also be considered as a trade-off. We suggest that population genetic study using genomic data may benefit from applying a more closely related reference genome. Undoubtedly, the optimal option would be a conspecific reference genome even with a low coverage. Our results showed that a conspecific reference genome can significantly improve the accuracy and reliability of all kinds of analyses, in particular WGR-based SNP imputation which will be promising in other genomic investigations like genome-wide association studies. However, if resource limitations exist (in terms of funding, available biological material, time (i.e., wet- and dry lab efforts)), or the research is focused on populations with high inter-population variation, a trade-off between a more distant reference genome and a higher sequence coverage can be considered. In other words, in a population genetic study, the trade-off between reference genome and coverage is dictated by the focus of research. Although our case study is carried out on caddisflies and thus may not be universally applicable to other organisms, we believe that our results do provide a valuable example that enhances developing roadmaps involved in the choice of appropriate reference genome and sequencing depth in population genomic studies.

## Concordance between genetic patterns and biogeography of *H. digitata* and *H. tibetana*

Irrespective of coverage and reference genome, the results of the population genetic analyses for *H. tibetana* and *H. tibetana* are highly consistent with the geographic distribution of populations within the drainage and river networks. For example, in *H. tibetana*, the PCA plots show that pop1 and pop 2 form two distinct clusters respectively, while pop 3 and pop 4 cluster together and are separated from pop1 and pop 2. This is also consistent with their geographical location: pop 3 and pop 4 are both located in the Kanchenjunga region in far eastern Nepal; pop 2 is located in the most north-eastern tributaries of the Gandaki Basin, and pop 1 is located in the east of the Annapurna circuit, which is in the center of the Gandaki Basin. In the admixture plots, pop 1 and 2, as well as pop 3 and 4, share the same structure, respectively. These results are consistent with the basin structure of these populations (pop 1 and pop 2 in Gandaki basin; pop 3 and pop 4 in Koshi basin). Moreover, compared with *H. digitata* populations which are all located in one basin, *H. tibetana* populations show greater genetic diversity and clearer population structure. Due to the dependency of larvae on the aquatic habitat (De Moor & Ivanov, 2007), the limited dispersal capabilities of the adults (Griffith et al., 1998; Petersen et al., 2004) and the unique geographic feature of the Himalayan region, it is not surprising that the population genetics of the two target species show a notable correlation with drainages and river network, like in other Trichoptera studied in the region (Hoppeler et al. 2016) and elsewhere (Altermatt et al., 2013; Engelhardt et al., 2011). The resequencing approach, even with a low coverage of 12.5×, appears to be a suitable methodological avenue to study population genomics of insect populations when reference genomes of at least moderate relatedness are available.

# **Acknowledgments**

We are thankful to Raphael Coimbra providing valuable advice in population genetic analyses. We thank Koji Tojo and Robert Wisseman for providing specimens and numbers of Nepalese students for supporting our collection efforts in the field. We thank Subodh Sharma for obtaining the field work permit from Kathmandu University.

This research was funded by the Deutsche Forschungsgemeinschaft (DFG), grant number PA 1617/2‐1 and PA 1617/2‐2 awarded to Steffen Pauls, and AF 1117/1‐2 awarded to Adrien Favre. The sequencing of the reference genomes was funded through the LOEWE Center for Translational Biodiversity Genomics with a grant from the Hessen State Ministry of Higher Education, Research and the Arts (HMWK).

Analyses were conducted on the BYU Fulton Supercomputer, the TBG Deepfry clusters, and the Smithsonian High-Performance Cluster (SI/HPC), Smithsonian Institution (<https://doi.org/10.25572/SIHPC>).

[dataset] Jacqueline Heckenhauer et al.; 2022; Comparative genomics of Trichoptera Genome sequencing and assembly; NCBI BioProject ID: PRJNA558902; <https://doi.org/10.1093/gigascience/giac011>

# **Reference**

Altermatt, F., Seymour, M., & Martinez, N. (2013). River network properties shape α‐diversity and community similarity patterns of aquatic insect communities across major drainage basins. *Journal of Biogeography, 40*(12), 2249-2260.

Armstrong, J., Hickey, G., Diekhans, M., Fiddes, I. T., Novak, A. M., Deran, A., Fang, Q., Xie, D., Feng, S., & Stiller, J. (2020). Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature, 587*(7833), 246-251. <https://github.com/ComparativeGenomicsToolkit/cactus>

Barbosa, S., Mestre, F., White, T. A., Paupério, J., Alves, P. C., & Searle, J. B. (2018). Integrative approaches to guide conservation decisions: using genomics to define conservation units and functional corridors. *Molecular ecology, 27*(17), 3452-3465.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics, 30*(15), 2114-2120. <https://github.com/usadellab/Trimmomatic>

Brandies, P., Peel, E., Hogg, C. J., & Belov, K. (2019). The value of reference genomes in the conservation of threatened species. *Genes, 10*(11), 846.

Brandt, D. Y., Aguiar, V. R., Bitarello, B. D., Nunes, K., Goudet, J., & Meyer, D. (2015). Mapping bias overestimates reference allele frequencies at the HLA genes in the 1000 genomes project phase I data. *G3: Genes, Genomes, Genetics, 5*(5), 931-941.

Buerkle, C. A., & Gompert, Z. (2013). Population genomics based on low coverage sequencing: how low should we go? *Molecular ecology, 22*(11), 3028-3035.

Crawford, J. E., & Lazzaro, B. P. (2012). Assessing the accuracy and power of population genetic inference from low-pass next-generation sequencing data. *Frontiers in genetics, 3*, 66.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., & Sherry, S. T. (2011). The variant call format and VCFtools. *Bioinformatics, 27*(15), 2156-2158. <https://github.com/vcftools/vcftools>

De Moor, F., & Ivanov, V. (2007). Global diversity of caddisflies (Trichoptera: Insecta) in freshwater. In *Freshwater Animal Diversity Assessment* (pp. 393-407). Springer.

Deng, X.L., Favre, A., Lemmon, E.M., Lemmon, A.R. and Pauls, S.U. (2021). Gene Flow and Diversification in *Himalopsyche martynovi* Species Complex (Trichoptera: Rhyacophilidae) in the Hengduan Mountains. *Biology*, *10*(8), 816.

Duchen, P., & Salamin, N. (2021). A cautionary note on the use of genotype callers in Phylogenomics. *Systematic biology, 70*(4), 844-854.

Ellegren, H. (2014). Genome sequencing and population genomics in non-model organisms. *Trends in Ecology & Evolution, 29*(1), 51-63.

Engelhardt, C. H., Haase, P., & Pauls, S. U. (2011). From the Western Alps across Central Europe: Postglacial recolonisation of the tufa stream specialist Rhyacophila pubescens (Insecta, Trichoptera). *Frontiers in zoology, 8*(1), 1-14.

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics, 32*(19), 3047-3048. <https://github.com/ewels/MultiQC>

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences, 117*(17), 9451-9457. <https://github.com/Dfam-consortium/RepeatModeler>

Fox, E. A., Wright, A. E., Fumagalli, M., & Vieira, F. G. (2019). ngsLD: evaluating linkage disequilibrium using genotype likelihoods. *Bioinformatics, 35*(19), 3855-3856. <https://github.com/fgvieira/ngsLD>

Fumagalli, M. (2013). Assessing the effect of sequencing depth and sample size in population genetics inferences. *PloS one, 8*(11), e79667.

Garcia-Rubio, R., Monzon, S., Alcazar-Fuoli, L., Cuesta, I., & Mellado, E. (2018). Genome-wide comparative analysis of Aspergillus fumigatus strains: the reference genome as a matter of concern. *Genes, 9*(7), 363.

Geist, J. (2011). Integrative freshwater ecology and biodiversity conservation. *Ecological Indicators, 11*(6), 1507-1516.

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics, 17*(6), 333-351.

Gopalakrishnan, S., Castruita, J. A. S., Sinding, M.-H. S., Kuderna, L. F., Räikkönen, J., Petersen, B., Sicheritz-Ponten, T., Larson, G., Orlando, L., & Marques-Bonet, T. (2017). The wolf reference genome sequence (Canis lupus lupus) and its implications for Canis spp. population genomics. *BMC genomics, 18*(1), 1-11.

Götz, S., García-Gómez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., Robles, M., Talón, M., Dopazo, J., & Conesa, A. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic acids research, 36*(10), 3420-3435. <https://www.biobam.com/blast2go/>

Griffith, M. B., Barrows, E. M., & Perry, S. A. (1998). Lateral dispersal of adult aquatic insects (Plecoptera, Trichoptera) following emergence from headwater streams in forested Appalachian catchments. *Annals of the Entomological Society of America, 91*(2), 195-201.

Günther, T., & Nettelblad, C. (2019). The presence and impact of reference bias on population genomic studies of prehistoric human populations. *PLoS genetics, 15*(7), e1008302.

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics, 29*(8), 1072-1075. <https://github.com/ablab/quast>

Han, E., Sinsheimer, J. S., & Novembre, J. (2014). Characterizing bias in population genetic inferences from low-coverage sequencing data. *Molecular biology and evolution, 31*(3), 723-735.

Heckenhauer, J., Frandsen, P. B., Gupta, D. K., Paule, J., Prost, S., Schell, T., Schneider, J. V., Stewart, R. J., & Pauls, S. U. (2019). Annotated draft genomes of two caddisfly species Plectrocnemia conspersa CURTIS and Hydropsyche tenuis NAVAS (Insecta: Trichoptera). *Genome Biology and Evolution, 11*(12), 3445-3451.

Heckenhauer, J., Frandsen, P. B., Sproul, J. S., Li, Z., Paule, J., Larracuente, A. M., Maughan, P. J., Barker, M. S., Schneider, J. V., Stewart, R. J. & Pauls, S.U. (2022). Genome size evolution in the diverse insect order Trichoptera. *GigaScience, 11*.

Hickey, G., Paten, B., Earl, D., Zerbino, D., & Haussler, D. (2013). HAL: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics, 29*(10), 1341-1342. <https://github.com/ComparativeGenomicsToolkit/hal>

Hjalmarsson, A. E. (2019). Delimitation and description of three new species of Himalopsyche (Trichoptera: Rhyacophilidae) from the Hengduan Mountains, China. *Zootaxa, 4638*(3), 419–441-419–441.

Hjalmarsson, A. E., Graf, W., Jähnig, S. C., Vitecek, S., & Pauls, S. U. (2018). Molecular association and morphological characterisation of Himalopsyche larval types (Trichoptera, Rhyacophilidae). *ZooKeys*(773), 79.

Hjalmarsson, A. E., Graf, W., Vitecek, S., Jähnig, S. C., Cai, Q., Sharma, S., Tong, X., Li, F., Shah, D. N., & Shah, R. D. T. (2019). Molecular phylogeny of Himalopsyche (Trichoptera, Rhyacophilidae). *Systematic entomology, 44*(4), 973-984.

Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2018). UFBoot2: improving the ultrafast bootstrap approximation. *Molecular biology and evolution, 35*(2), 518-522.

Hohenlohe, P. A., Funk, W. C., & Rajora, O. P. (2021). Population genomics for wildlife conservation and management. *Molecular ecology, 30*(1), 62-82.

Hoppeler, F., Tachamo Shah, R. D., Shah, D. N., Jähnig, S. C., Tonkin, J. D., Sharma, S., & Pauls, S. U. (2016). Environmental and spatial characterisation of an unknown fauna using DNA sequencing–an example with Himalayan Hydropsychidae (Insecta: Trichoptera). *Freshwater Biology, 61*(11), 1905-1920.

Hotaling, S., Kelley, J. L., & Frandsen, P. B. (2020). Aquatic insects are dramatically underrepresented in genomic research. *Insects, 11*(9), 601.

Hotaling, S., Sproul, J. S., Heckenhauer, J., Powell, A., Larracuente, A. M., Pauls, S. U., Kelley, J. L., & Frandsen, P. B. (2021). Long-reads are revolutionizing 20 years of insect genome sequencing. *bioRxiv*.

Jasper, M., Schmidt, T. L., Ahmad, N. W., Sinkins, S. P., & Hoffmann, A. A. (2019). A genomic approach to inferring kinship reveals limited intergenerational dispersal in the yellow fever mosquito. *Molecular ecology resources, 19*(5), 1254-1264.

Jiang, Y., Jiang, Y., Wang, S., Zhang, Q., & Ding, X. (2019). Optimal sequencing depth design for whole genome re-sequencing in pigs. *BMC Bioinformatics, 20*(1), 1-12.

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K., Von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature methods, 14*(6), 587-589.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution, 30*(4), 772-780.

Keilwagen, J., Hartung, F., & Grau, J. (2019). GeMoMa: homology-based gene prediction utilizing intron position conservation and RNA-seq data. In *Gene Prediction* (pp. 161-177). Springer. <http://www.jstacs.de/index.php/GeMoMa>

Keilwagen, J., Wenk, M., Erickson, J. L., Schattat, M. H., Grau, J., & Hartung, F. (2016). Using intron position conservation for homology-based gene prediction. *Nucleic acids research, 44*(9), e89-e89. <http://www.jstacs.de/index.php/GeMoMa>

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics, 15*(1), 1-13. <https://github.com/ANGSD/angsd>

Korneliussen, T. S., Moltke, I., Albrechtsen, A., & Nielsen, R. (2013). Calculation of Tajima’s D and other neutrality test statistics from low depth next-generation sequencing data. *BMC Bioinformatics, 14*(1), 1-14.

Laetsch, D. R., & Blaxter, M. L. (2017). BlobTools: Interrogation of genome assemblies. *F1000Research, 6*(1287), 1287. <https://github.com/DRL/blobtools>

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods, 9*(4), 357-359. <http://bowtie-bio.sourceforge.net/bowtie2/index.shtml>

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics, 34*(18), 3094-3100. <https://github.com/lh3/minimap2>

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079. <https://doi.org/https://github.com/samtools/samtools>

Loman, N. J., Quick, J., & Simpson, J. T. (2015). A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature methods, 12*(8), 733-735. <https://github.com/jts/nanopolish>

Lou, R. N., Jacobs, A., Wilder, A. P., & Therkildsen, N. O. (2021). A beginner's guide to low‐coverage whole genome sequencing for population genomics. *Molecular ecology, 30*(23), 5966-5993.

Luo, S., Tang, M., Frandsen, P. B., Stewart, R. J., & Zhou, X. (2018). The genome of an underwater architect, the caddisfly Stenopsyche tienmushanensis Hwang (Insecta: Trichoptera). *Gigascience, 7*(12), giy143.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal, 17*(1), 10-12. <https://github.com/marcelm/cutadapt>

Meisner, J., & Albrechtsen, A. (2018). Inferring population structure and admixture proportions in low-depth NGS data. *Genetics, 210*(2), 719-731. <http://www.popgen.dk/software/index.php/PCAngsd>

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular biology and evolution, 37*(5), 1530-1534. <http://www.iqtree.org/>

Morse, J. C., Frandsen, P. B., Graf, W., & Thomas, J. A. (2019). Diversity and ecosystem services of Trichoptera. *Insects, 10*(5), 125.

Miller, S., Dykes, D., & Polesky, H. (1988). A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic acids res, 16*(3), 1215.

Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y., & Wang, J. (2012). SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. *PloS one, 7*(7), e37558.

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics, 12*(6), 443-451.

Olsen, L. K., Heckenhauer, J., Sproul, J. S., Dikow, R. B., Gonzalez, V. L., Kweskin, M. P., Taylor, A. M., Wilson, S. B., Stewart, R. J., & Zhou, X. (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 Biology and Evolution, 13*(3), evab013.

Petersen, I., Masters, Z., Hildrew, A., & Ormerod, S. J. (2004). Dispersal of adult aquatic insects in catchments of differing land use. *Journal of Applied Ecology, 41*(5), 934-950.

Ranallo-Benavidez, T. R., Jaron, K. S., & Schatz, M. C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications, 11*(1), 1-10. <https://github.com/tbenavi1/genomescope2.0>

Ríos-Touma, B., Holzenthal, R.W., Rázuri-Gonzales, E., Heckenhauer, J., Pauls, S.U., Storer, C.G. and Frandsen, P.B. (2022). De novo genome assembly and annotation of an Andean caddisfly, Atopsyche davidsoni Sykora, 1991, a model for genome research of high elevation adaptations. *Genome biology and evolution*, *14*(1), p.evab286.

Roach, M. J., Schmidt, S. A., & Borneman, A. R. (2018). Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics, 19*(1), 1-10. <https://doi.org/https://github.com/dfguan/purge_dups>

Ruan, J., & Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. *Nature methods, 17*(2), 155-158. <https://github.com/ruanjue/wtdbg2>

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics, 31*(19), 3210-3212. <https://busco-archive.ezlab.org/>

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., & Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics, 15*(2), 121-132.

Skoglund, P., Ersmark, E., Palkopoulou, E., & Dalén, L. (2015). Ancient wolf genome reveals an early divergence of domestic dog ancestors and admixture into high-latitude breeds. *Current Biology, 25*(11), 1515-1519.

Skotte, L., Korneliussen, T. S., & Albrechtsen, A. (2013). Estimating individual admixture proportions from next generation sequencing data. *Genetics, 195*(3), 693-702. <http://www.popgen.dk/software/index.php/NgsAdmix>

Taub, M. A., Bravo, H. C., & Irizarry, R. A. (2010). Overcoming bias and systematic errors in next generation sequencing data. *Genome medicine, 2*(12), 1-5.

Thomas, J. A., Frandsen, P. B., Prendini, E., Zhou, X., & Holzenthal, R. W. (2020). A multigene phylogeny and timeline for Trichoptera (Insecta). *Systematic entomology, 45*(3), 670-686.

Tsuruishi, T., Ketavan, C., Suwan, K., & Sirikajornjaru, W. (2006). Importance of water flow on larval growth and pupation of Himalopsyche acharai,(Malicky and Chantaramongkol, 1989)(Trichoptera: Rhyacophilidae). *Hydrobiologia, 563*(1), 537-540.

Valiente-Mullor, C., Beamud, B., Ansari, I., Francés-Cuesta, C., García-González, N., Mejía, L., Ruiz-Hueso, P., & González-Candelas, F. (2021). One is not enough: On the effects of reference genome for the mapping and subsequent analyses of short-reads. *PLoS computational biology, 17*(1), e1008678.

Vaser, R., Sović, I., Nagarajan, N., & Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome research, 27*(5), 737-746. <https://github.com/isovic/racon>

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., & Schatz, M. C. (2017). GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics, 33*(14), 2202-2204. <https://github.com/tbenavi1/genomescope2.0>

Waldvogel, A. M., Wieser, A., Schell, T., Patel, S., Schmidt, H., Hankeln, T., Feldmeyer, B., & Pfenninger, M. (2018). The genomic footprint of climate adaptation in Chironomus riparius. *Molecular ecology, 27*(6), 1439-1456. <https://github.com/schellt/autotrim>

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., & Young, S. K. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS one, 9*(11), e112963. <https://github.com/broadinstitute/pilon>

Wang, G.-d., Zhai, W., Yang, H.-c., Fan, R.-x., Cao, X., Zhong, L., Wang, L., Liu, F., Wu, H., & Cheng, L.-g. (2013). The genomics of selection in dogs and the parallel evolution between dogs and humans. *Nature Communications, 4*(1), 1-9.

Waterhouse, R. M., Seppey, M., Simão, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E. V., & Zdobnov, E. M. (2018). BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular biology and evolution, 35*(3), 543-548. <https://busco-archive.ezlab.org/>

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, 1358-1370.

Xu, J., Grumbine, R. E., Shrestha, A., Eriksson, M., Yang, X., Wang, Y., & Wilkes, A. (2009). The melting Himalayas: cascading effects of climate change on water, biodiversity, and livelihoods. *Conservation Biology, 23*(3), 520-530.

Yang, X., Lee, W.-P., Ye, K., & Lee, C. (2019). One reference genome is not enough. *Genome biology, 20*(1), 1-3.

Zhang, C., Rabiee, M., Sayyari, E., & Mirarab, S. (2018). ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics, 19*(6), 15-30. <https://github.com/smirarab/ASTRAL>

Zimin, A. V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S. L., & Yorke, J. A. (2013). The MaSuRCA genome assembler. *Bioinformatics, 29*(21), 2669-2677. <http://www.genome.umd.edu/masurca_compile.html>

Zimin, A. V., Puiu, D., Luo, M.-C., Zhu, T., Koren, S., Marçais, G., Yorke, J. A., Dvořák, J., & Salzberg, S. L. (2017). Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome research, 27*(5), 787-792.

# **Data Accessibility**

All the COI and CAD data for this research are available in the Genbank of National Center for Biotechnology Information (NCBI), the accession code of each individual are provided in the Supplementary Material 2.

The raw data and genome assemblies of the three references have been deposited at NCBI under the Bioproject ID PRJNA728835, the raw data of populations has been deposited at NCBI under the Bioproject ID PRJNA749154.

All BUSCO genes, annotation gffs and predicted proteins resulting from GEMOMA, blastp and BLAST2GO results, as well as repeatmodeler and -masker results are available at: https://doi.org/10.6084/m9.figshare.c.6033011.v1.

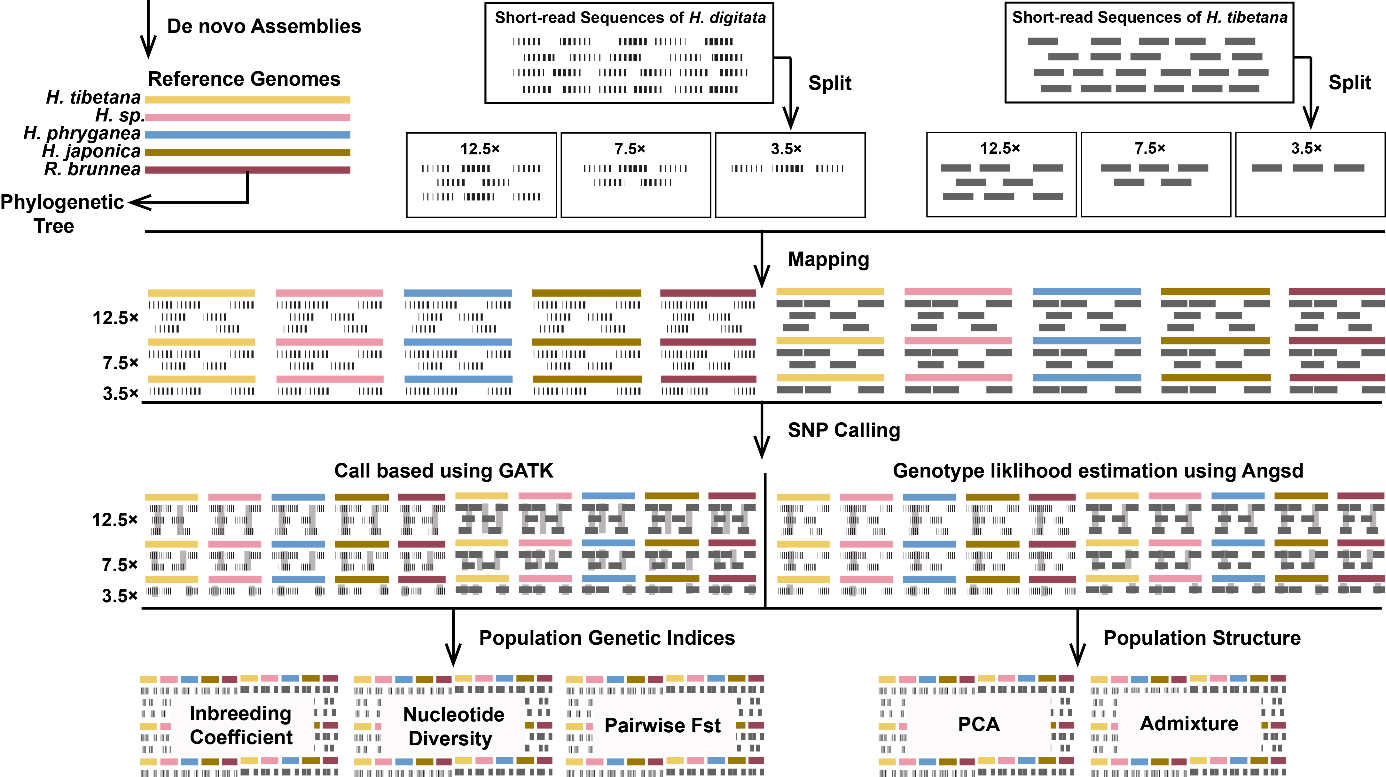
# **Author Contributions**

SUP and PBF designed the study. RDTS and DNS identified study areas and designed field work. SUP, PBF, RDTS, DNS performed the field work. JS and XD conducted the laboratory work. JH assembled and annotated the three *de novo* genomes*;* PBF and RBD conducted the phylogenetic analysis and cactus alignment; XD, JH performed the population genomic analyses. JH and XD drafted the manuscript. All authors contributed to the writing of the manuscript. All authors approved the final draft of the manuscript.

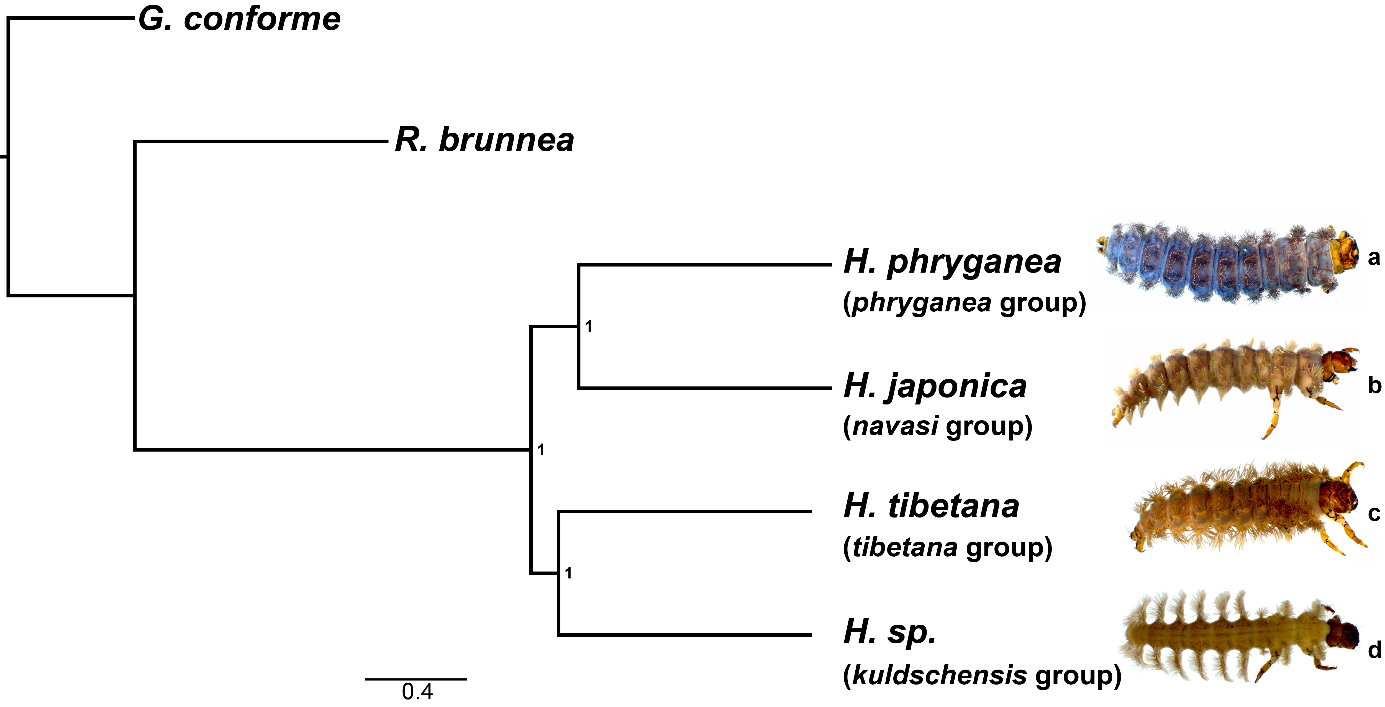
**Table 1: Assembly statistics of reference genomes used in this study.**

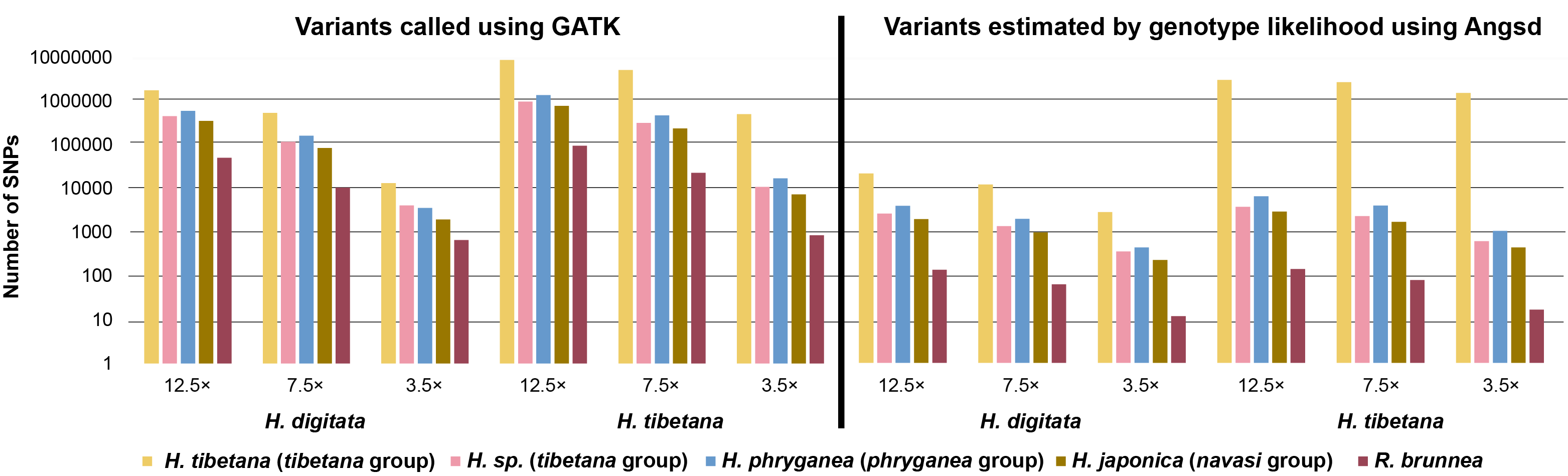
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Species** | **Accession number** | **Sequencing platform (coverage)\*\*\*\*** | **Assembly length bp** | **N50 (bp)** | **Np.of contigs** | **N's per 100kb** | **BUSCOS %\*\*\*** | **Number of proteins** |
| *Himalopsyche japonica*\* | JAHFWJ000000000 | Nanopore+Illumina (18x +170x) | 546,840,812 | 2,150,202 | 847 | 0 | C:97.2%[S:96.6%,D:0.6%],F:0.8%,M:2.0%,n:2124 | 9,983 |
| *Himalopsyche sp.*\* | JAHFWI000000000 | Nanopore+Illumina (26x +200x) | 592,402,457 | 7,599,818 | 528 | 0 | C:96.7%[S:96.3%,D:0.4%],F:1.0%,M:2.3%,n:2124 | 10,049 |
| *Himalopsyche phryganea* | JAGVSL000000000 | Nanopore+Illumina (36.8x +170x) | 633,785,554 | 4,634,010 | 710 | 0 | C:97.0%[S:96.5%,D:0.5%],F:1.0%,M:2.0%,n:2124 | 10,994 |
| *Himalopsyche tibetana*\* | JAHFWH000000000 | Nanopore+Illumina (24x +150x) | 665,312,086 | 949,059 | 1,853 | 0 | C:96.4%[S:95.6%,D:0.8%],F:1.0%,M:2.6%,n:2124 | 10,994 |
| *Rhyacophila brunne*a | JAGYXB000000000\*\* | Nanopore+Illumina (19x +154x) | 1,086,872,538 | 1,030,560 | 2,125 | 0.36 | C:96.0%[S:93.3%,D:2.7%],F:1.1%,M:2.9%,n:2124 | 10,846 |

\* This study; \*\* In this study, we used a previous version of this assembly for SNP calling; \*\*\* Based on the Endopterygota odb10 dataset (2124 genes), C: complete, S: single, D: duplicated, F: fragmented, M: missing; \*\*\*\* Based on Genomescope2 genome size estimation.

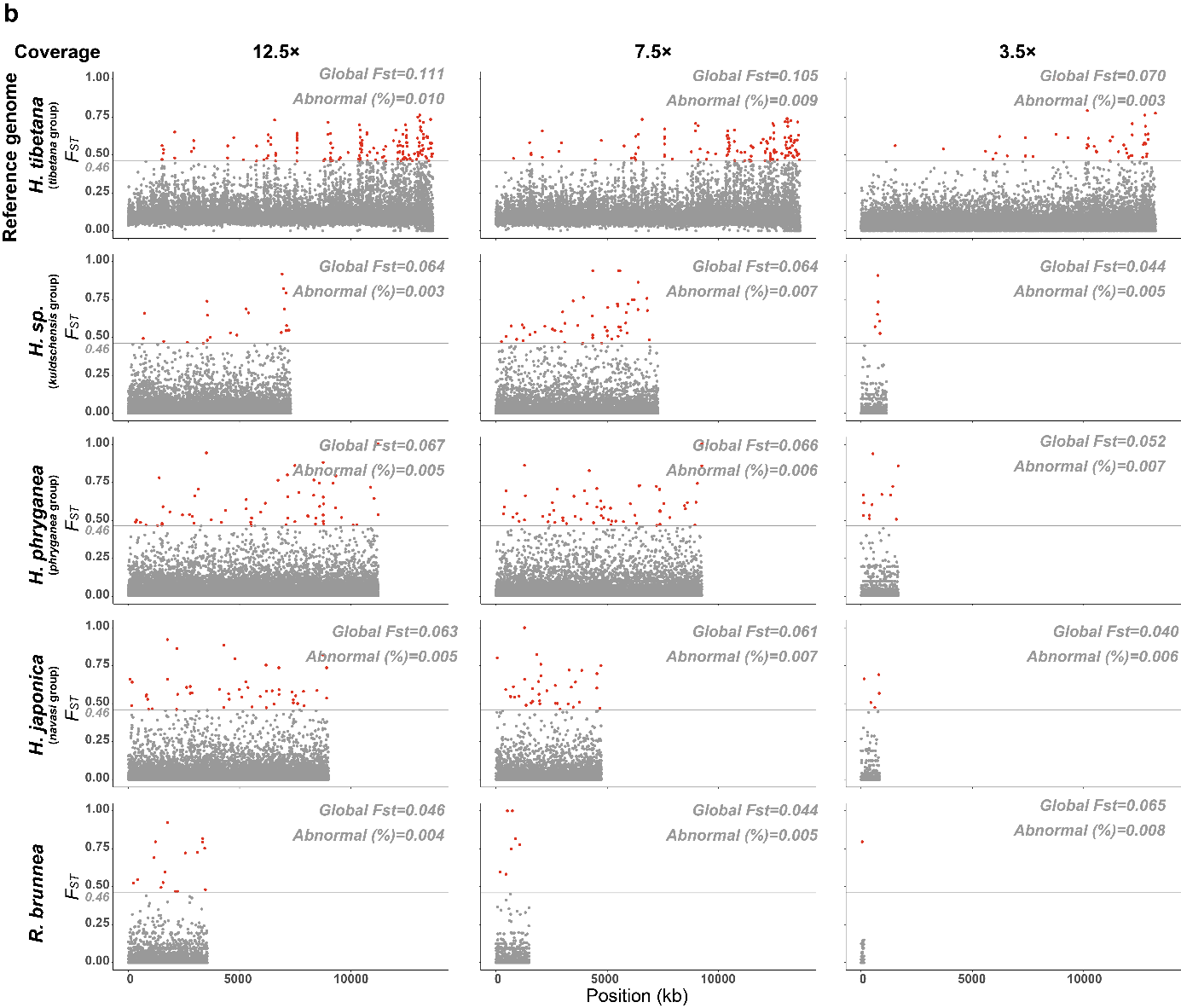


**Fig. 1.** **Workflow of data processing in this study showing the treatment of short read resequencing data from the two target species and subsequent mapping and variant calling with different reference genomes for assessing genetic diversity and population genetic structure** (see details in Section “Study Design”).

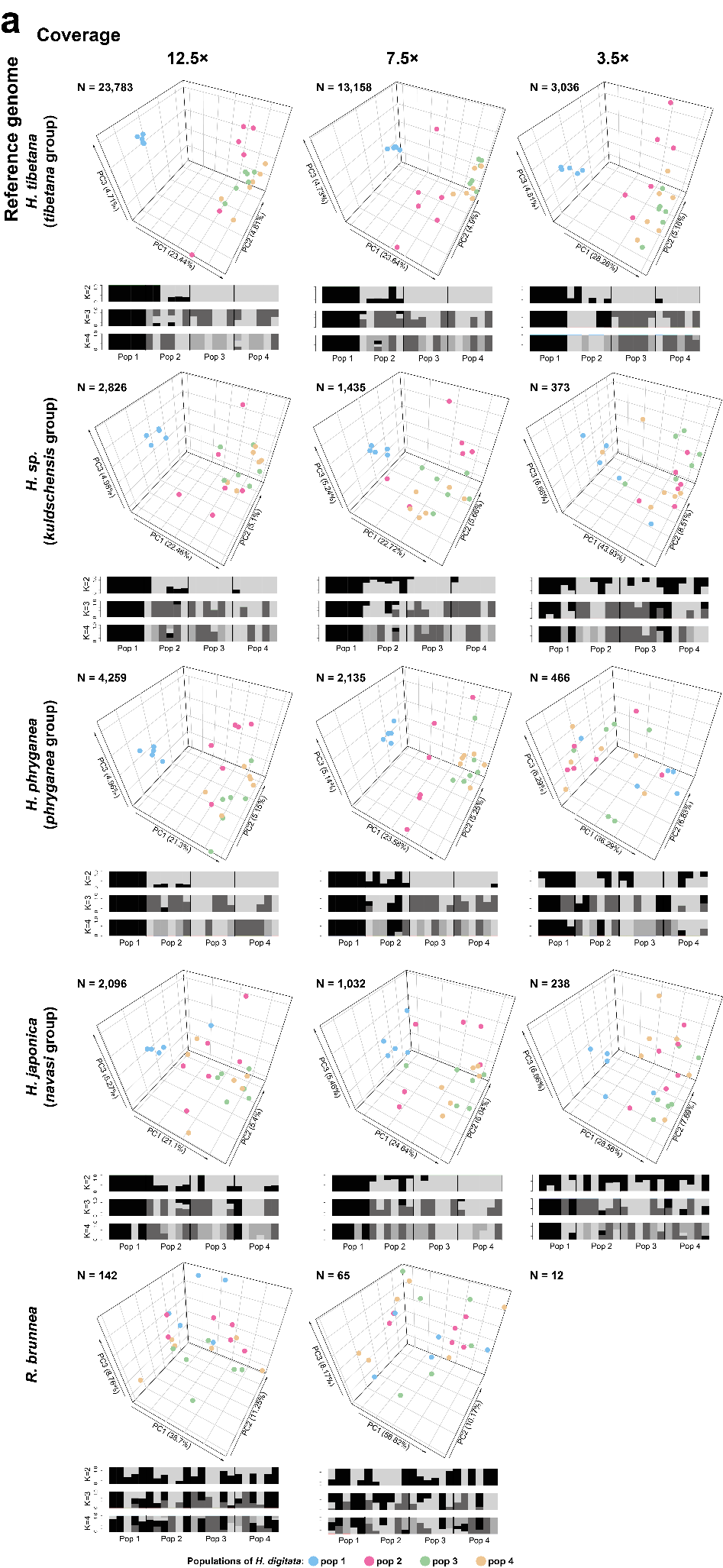
**Fig. 2.** **Astral tree of the five reference genomes generated from BUSCO genes.** Numbers on the nodes indicate local posterior probabilities. *G. conforme* was used as an outgroup. Images taken from Hjalmarsson et al. (2019), a) *H. phryganea*, b) *H. japonica*, c) *H. gregoryi* (identical with *H. tibetana* morphologically), d) *H. sylvicola* (identical with *H. sp.* (*kuldschensis* group)morphologically).

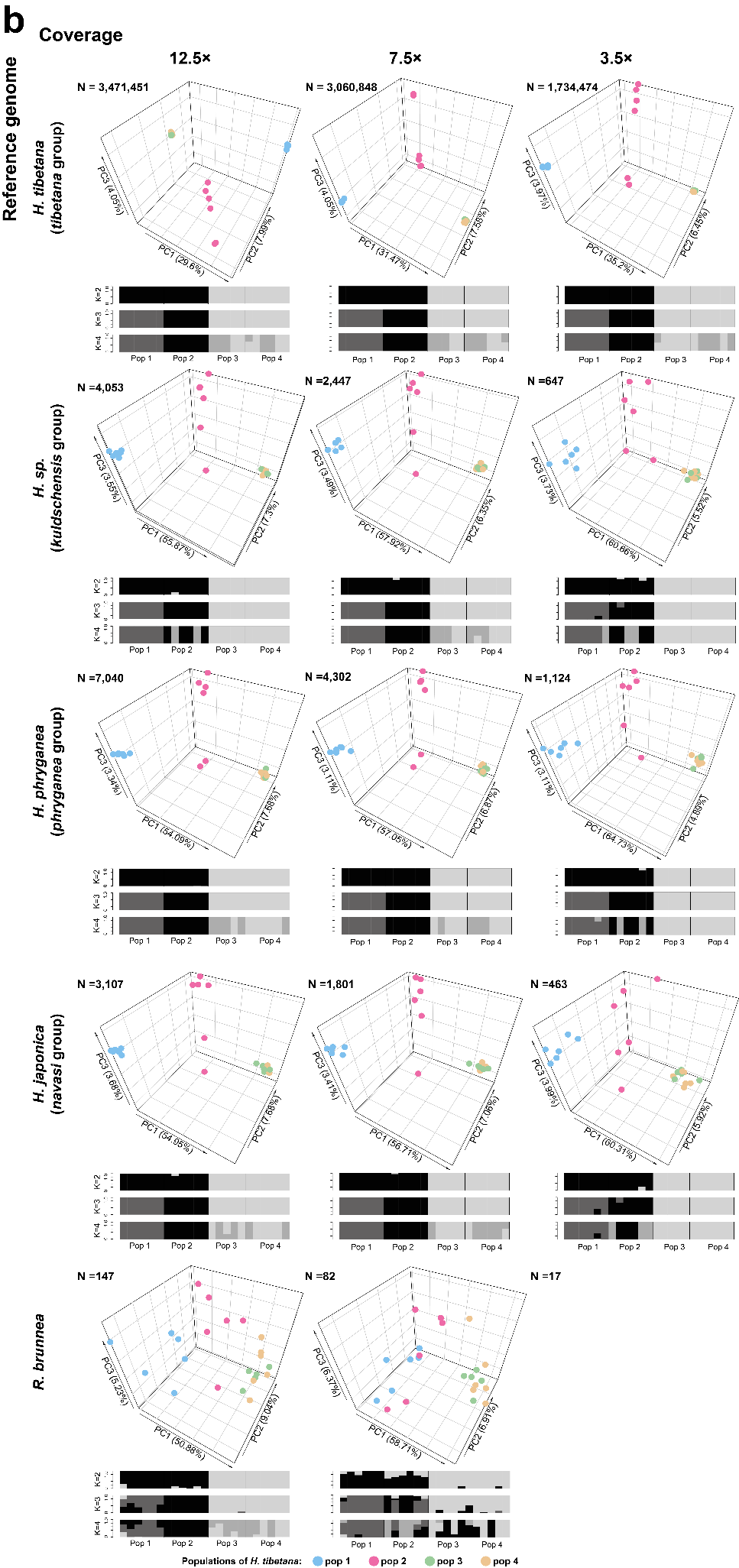


**Fig. 3.** **Number of variants called / estimated by two different strategies.** Numbers of the Y-axis were logarithmically scaled.

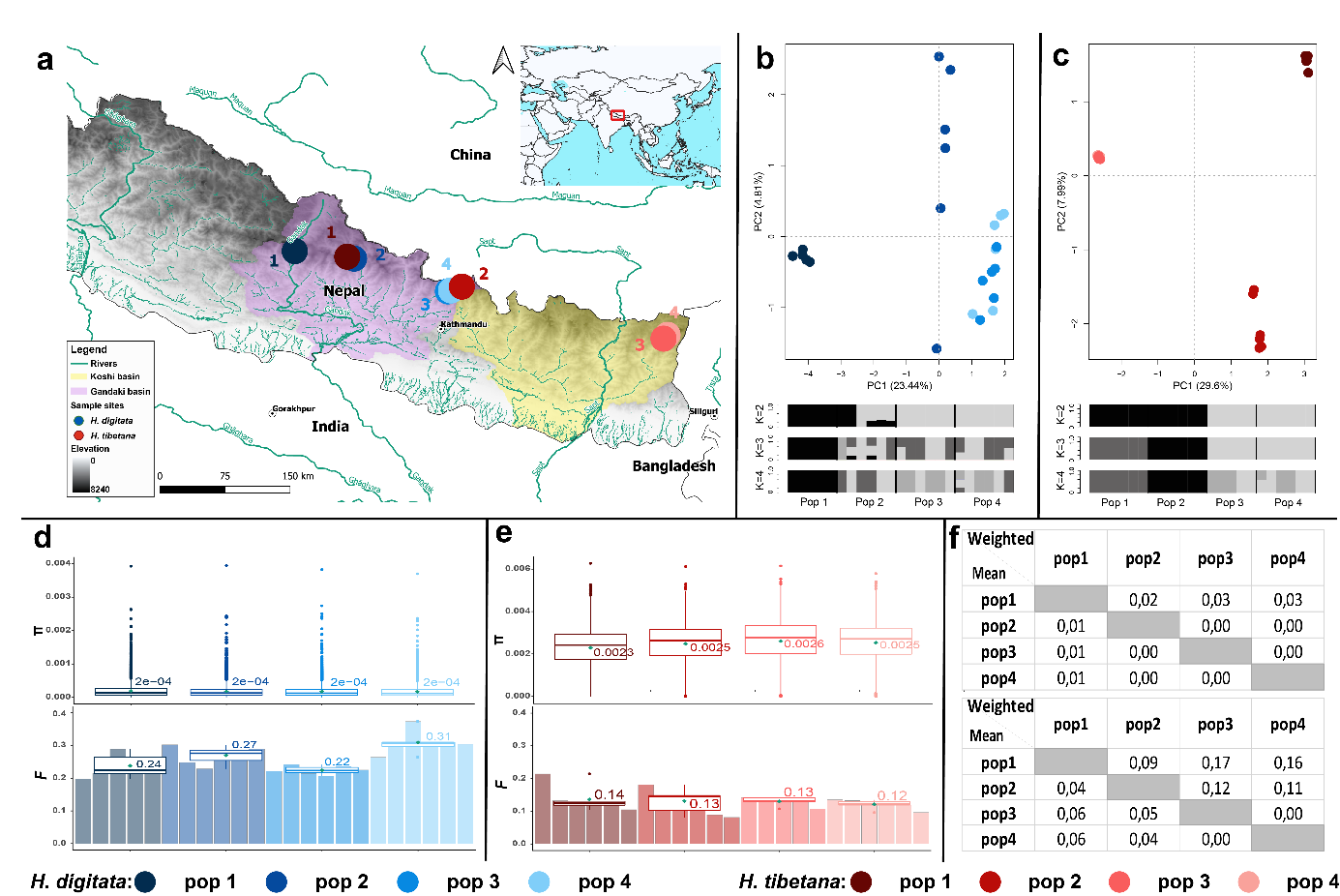


**Fig. 4.** **Genome-wide distribution of *F*ST values (weighted) of** **(a) the *H. digitata* populations and (b) *H. tibetana* populations.** *F*ST values were calculated in 50-kb windows across contigs obtained from each reference genome, thus the X-axis (position) is not comparable among datasets. The horizontal gray lines indicate a threshold that is used for selecting the abnormal *F*ST windows for each species. The threshold is the minimum value of the top 1% *F*ST windows on the whole genome when using *H. tibetana* as reference genome and 12.5× as coverage (0.16 for *H. digitata*, 0.28 for *H. tibetana*). The percentage of abnormal *F*ST windows is the number of *F*ST windows above the threshold over the total number of *F*ST windows. The global *F*ST is the genome-wide weighted *F*st value estimated based on Weir & Cockerham (1984).





**Fig. 5.** **3D scatter plots of all individuals derived from PCA by using *PCAngsd* and population structure (k = 2, 3, 4 respectively) inferred from *NgsAdmix* depending on different reference genome and different genome coverage of** **(a) *H. digitata* and (b) *H. tibetana* populations.** The explained variances are shown as percentages. Numbers on the top left of each plot show the number of SNPs used for the structure estimating. PCA and Admixture were not able estimate when using *R. brunnea* as reference genome with 3.5× coverage due to the limited number of variants. Colors in the plots represent the four populations of *H. digitata* and *H. tibetana* respectively.



**Fig. 6.** **Location of populations, genetic diversity and population structure of *H. digitata* and *H. tibetana*. a)** Map showing the 8 population sites of the two species across two main drainage basins in Nepal. Population numbers are signed on the sample sites; **b)** PCA plots and admixture proportions (k = 2, 3, 4) of *H. digitata*; **c)** PCA plots and admixture proportions (k = 2, 3, 4) of *H. tibetana*; **d)** nucleotide diversity (upper) and inbreeding coefficient (lower) of *H. digitata*, the mean values of each population were labeled with green dots and numbers; **e)** nucleotide diversity (upper) and inbreeding coefficient (lower) of *H. tibetana*, the mean values of each population were labeled with green dots and numbers; **f)** pairwise *F*ST of *H. digitata* (upper) and *H. tibetana*.