A Fast, Reproducible, High-throughput Variant Calling Workflow for Population Genomics

Cade D. Mirchandani [1,2], Allison J. Shultz [3], Gregg W.C. Thomas,[4] Sara J. Smith,[4,5] Mara Baylis,[1,2] Brian Arnold,[6,7] Russ Corbett-Detig,[1,2,†] Erik Enbody [1,†] and Timothy B. Sackton[4,*,†]

[1]Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, CA 95064, USA
[2]Genomics Institute, University of California Santa Cruz, Santa Cruz, CA 95064, USA
[3]Ornithology Department, Natural History Museum of Los Angeles County, Los Angeles, CA 90007, USA
[4]Informatics Group, Harvard University, Cambridge, MA, USA
[5]Biology, Mount Royal University, Calgary, AB T3E 6K6, Canada
[6]Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ, USA
[7]Center for Statistics and Machine Learning, Princeton University, Princeton, NJ, USA
[†]These authors contributed equally.

***Corresponding author:** E-mail: tsackton@g.harvard.edu.

Associate editor: Andrey Rzhetsky

Abstract

The increasing availability of genomic resequencing data sets and high-quality reference genomes across the tree of life present exciting opportunities for comparative population genomic studies. However, substantial challenges prevent the simple reuse of data across different studies and species, arising from variability in variant calling pipelines, data quality, and the need for computationally intensive reanalysis. Here, we present snpArcher, a flexible and highly efficient workflow designed for the analysis of genomic resequencing data in nonmodel organisms. snpArcher provides a standardized variant calling pipeline and includes modules for variant quality control, data visualization, variant filtering, and other downstream analyses. Implemented in Snakemake, snpArcher is user-friendly, reproducible, and designed to be compatible with high-performance computing clusters and cloud environments. To demonstrate the flexibility of this pipeline, we applied snpArcher to 26 public resequencing data sets from nonmammalian vertebrates. These variant data sets are hosted publicly to enable future comparative population genomic analyses. With its extensibility and the availability of public data sets, snpArcher will contribute to a broader understanding of genetic variation across species by facilitating the rapid use and reuse of large genomic data sets.

Key words: **comparative population genomics, genomic workflow, conservation genomics, evolutionary genomics.**

Resources

Introduction

In the past decade, rapidly declining sequencing costs have led to a dramatic expansion in the availability of genomic resequencing data sets in diverse organisms, fueling a wide range of novel insights, including the prevalence of adaptive introgression between species (Huerta-Sánchez et al. 2014; Lamichhaney et al. 2015; Jones et al. 2018), the molecular basis of repeated local adaptation (Jones et al. 2012; Hill et al. 2019; Wooldridge et al. 2022), and the complex demographic histories of humans (Nielsen et al. 2017; Fan et al. 2023) and animals of conservation relevance (Robinson et al. 2018). In parallel, rapidly expanding efforts to generate high-quality reference genomes across the Tree of Life (Rhie et al. 2021; Lewin et al. 2022) are poised to empower population genetic inference across a wide diversity of organisms. The massive accumulation of existing genomic data sets facilitated by these advances can enable broad comparisons between diverse populations and uncover generalized principles that may explain processes that generate diversity across life. These questions include the determinants of molecular variation among species (Romiguier et al. 2014; Corbett-Detig et al. 2015; Buffalo 2021) and indirect estimates of the rates of loss of genetic variation among populations (Exposito-Alonso et al. 2022).

However, despite the rapid increase in accessibility of public sequencing data from diverse organisms, comparative population genetics and reuse of public data remain challenging for several reasons. In the absence of standardized variant calling pipelines for nonhuman species (Regier et al. 2018), computational batch effects introduced by differences in reference choice, alignment, and variant calling algorithms complicate efforts to jointly analyze existing variant calls across populations and species (Lek et al. 2016; Jia et al. 2020; Breton et al. 2021). Considerations must also be given to data quality prior to data processing, particularly in cases of low coverage

Received: June 22, 2023. **Revised:** October 27, 2023. **Accepted:** November 22, 2023

(Lou et al. 2021), and workflows must be flexible to accommodate these considerations. Because these computational and algorithmic choices can impact downstream analysis (Kulkarni and Frommolt 2017), comparative projects often must reanalyze raw data to produce comparable data sets, which can be computationally expensive.

Extensible, reproducible bioinformatic pipelines can help address these challenges, to facilitate both primary analysis of complex tasks such as variant calling and also allow for consistent reanalysis (Wratten et al. 2021). While reproducible workflows have had a major impact on human population genetics (Chen et al. 2022), the need for significant expertise to adapt pipelines optimized for human genetics to diverse species is a major technical hurdle for many researchers. Additionally, resequencing data sets are increasingly rapidly in scale (Ellegren 2014), driving a need for workflows optimized for computational efficiency and flexibility to be used across a variety of compute resources, including cloud resources that eliminate the need for costly on-site infrastructure (Mangul et al. 2019).

Due to the popularity and need for efficient and reproducible workflows, several solutions have already been proposed for variant calling pipelines (Czech and Exposito-Alonso 2022; Cullen and Friedenberg 2023). Here, we present snpArcher, a reproducible workflow for data set acquisition, variant calling, quality control (QC), and downstream analysis that is optimized for nonmodel organisms and comparisons across data sets (available at https://github.com/harvardinformatics/snpArcher). snpArcher implements a combination of several notable features not included in other existing solutions that address the challenges presented by the expanding scale of comparative population genomic studies. First, our workflow is optimized for nonmodel species, which often lack gene annotations, known variant sites, and other genomic information typically required for human-optimized pipelines. Second, we take advantage of the huge compute power available through cloud resources and large high-performance computing (HPC) clusters by highly parallelizing the workflow's variant calling step and thus greatly reducing analysis time. Finally, we designed snpArcher to be modular and easily extended. By providing module contribution guidelines and example analysis modules, we hope that users will be able to develop and contribute their own modules. This will enrich the snpArcher ecosystem and cater to a diverse range of genomic analyses.

To enable rapid analysis of a growing set of variant calls created in a functionally equivalent way, we apply this workflow to reanalyze public sequencing data from 26 focal species of nonmammalian vertebrates and make the resulting variant calls available for public use. Furthermore, we provide examples of analysis and visualization modules, and we use these to exemplify and enumerate a suite of criteria for future module contributions to this project. This new and immediately available toolset will enable highly reproducible comparative population genomic analyses for a broad range of taxa.

Results

Overview of snpArcher

We developed snpArcher, a comprehensive workflow for the analysis of polymorphism data sampled from nonmodel organism populations (Fig. 1). This workflow accepts short-read sequence data and a reference genome as input and ultimately produces a filtered, high-quality variant call format (VCF) genotype file for downstream analysis. It also accepts as input accession numbers for reads and reference genome, which are then automatically downloaded from public repositories. We largely follow the Genome Analysis Toolkit (GATK) best practices (Van der Auwera et al. 2013) to map reads to a reference genome, call individual-level variants, generate population-level consensus genotypes, apply filters, and generate QC metrics. This workflow is implemented as a Snakemake (Mölder et al. 2021) workflow, which enables scalable, reproducible, and efficient analysis of large-scale genomic data sets. Snakemake manages all aspects of running the workflow, such as the installation of software dependencies, creation of output directories, and execution of workflow steps, so that the user input required is minimal. To use snpArcher, users need only edit a configuration file to customize workflow settings and define their samples in a table. With these files in place, running snpArcher is as simple as running one command.

Example Data Sets Processed Using snpArcher

To thoroughly evaluate snpArcher and to provide a database of comparative population genomic data sets, we ran the workflow on 26 public resequencing data sets (supplementary table S1, Supplementary Material online). We identified 13 bird, 12 fish, and 1 reptile data sets that fit our criteria of whole-genome, multisample, moderate sequencing effort (see Materials and Methods) and have a reference genome available. Data sets vary by number of individuals from 6 to 306, all with a mean depth of coverage of at least 5. We recovered between 3.34 million and 83.83 million total single nucleotide polymorphisms (SNPs) on genomes ranging from 348 Mb to 1.6 Gb (supplementary table S1, Supplementary Material online). Nucleotide diversity (Watterson's θ) varies by an order of magnitude across these species, from 0.00126 in the cichlid *Amphilophus citrinellus* to 0.01568 in the zebra finch *Taeniopygia guttata*.

Benchmarking

Impact of Sequencing Depth

To evaluate the performance of snpArcher, we selected 10 individuals from a high-quality resequencing data set of zebra finch *T. guttata* (Singhal et al. 2015) and reanalyzed them using a range of approaches. First, we investigated the impacts of low sequencing depth by subsampling the initially high-depth data set (16.7× to 50.2× coverage) to uniform reduced coverage data sets (4×, 10×, and 20×). We ran each data set using the "low-coverage" and

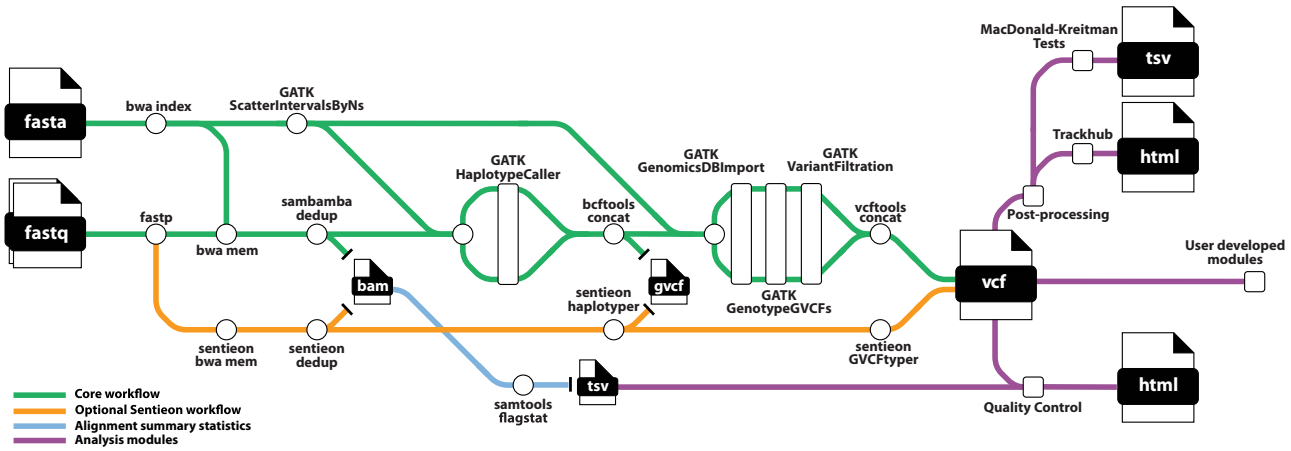

Fig. 1. snpArcher overview. snpArcher is an automated pipeline implemented in Snakemake (Mölder et al. 2021). It takes short-read whole-genome sequencing data (fastq) and a reference genome as input and produces a multisample variant callset (VCF). With the modules presented here, snpArcher produces basic QC statistics and visualizations.

"high-coverage" configurations of the pipeline; the "low-coverage" configuration alters certain GATK parameters to improve SNP calling in low-coverage data sets. After filtering for SNPs that passed all filters, we genotyped about 40, 55, and 50 million SNPs in the 4×, 10×, and 20× data sets, respectively, with about 1 million more SNPs recovered from the low-coverage pipeline at 4× coverage compared with the high-coverage version. There were negligible differences for the 2 pipeline versions at 10× and 20× (Fig. 2a). Assuming the 20× high-coverage pipeline produced the truth set of SNPs, the 4× data set was missing 35.6% (high-coverage pipeline) or 33.8% (low-coverage pipeline), and the 10× data set was missing 10.3% (high-coverage pipeline) or 11.3% (low-coverage pipeline) of SNPs. CPU time to run the low-coverage version of the pipeline was substantially higher compared with the high-coverage version and increased with sequencing depth (Fig. 2b). The percentage of heterozygous sites per individual was substantially reduced at low coverage, especially when using the high-coverage parameters, and slightly reduced at 10× coverage (Fig. 2c). Individual fixation indices measuring excess homozygosity (*F*-statistics) were correspondingly higher at lower sequencing depths, especially with the high-coverage parameters (Fig. 2d), indicating more heterozygous dropout. While heterozygous dropout is a substantial problem at low coverage (Nevado et al. 2014; Benjelloun et al. 2019), parameter tuning can partially mitigate its impact on genotype calls, at the cost of longer compute times.

Parallelization

We assessed the effectiveness of our parallelization method for variant calling with snpArcher on the 10× zebra finch data set by comparing our scatter-by-Ns approach to the traditional scatter-by-chromosome approach. Given that GATK HaplotypeCaller has limitations in efficiently utilizing multiple CPU cores, optimal parallelization requires a scatter–gather technique, processing each chromosome independently (Heldenbrand et al. 2019). However, as runtime scales with genomic interval size (Fig. 3a), using this approach will still result in potentially long execution times, especially for organisms with very large chromosomes. To address this, we employ a strategy of partitioning chromosomes at Ns (assembly gaps), creating smaller genomic intervals that can be processed in parallel. This approach shortens the run time per individual (Fig. 3b), as more intervals can be concurrently processed. Although the effectiveness of this approach is dependent on available compute resources, the wide availability of HPC clusters and affordable cloud compute resources renders this constraint generally acceptable.

Analysis Modules

QC and Data Visualization

An important component of any pipeline is QC and data visualization outputs. We have implemented a module in snpArcher, run by default, that produces an interactive QC dashboard, which can be used to evaluate individual-level sequencing quality (Fig. 4). This dashboard generates 10 figures that allow visualization of basic summary statistics relating to population structure, batch effects, sequencing depth, genetic relatedness, geography, and admixture. For speed, most of these summaries are based on a random sample of 100,000 SNPs from across the genome. Four panels at the top of the dashboard provide high-level summaries of the full variant data set (i.e. without random downsampling to 100,000 SNPs).

The use case for these simple visualizations is to quickly evaluate potential biases relating to individual-level sequencing variation. For example, in the principal component analysis (PCA) shown in the upper left panel of Fig. 4, it is possible to identify outliers that may represent cryptic genetic variation, batch effects, or otherwise problematic (or interesting) samples. By default, we identify 3 clusters based on PC1 and PC2 with k-means clustering (modifiable in the config file), and the remainder of the

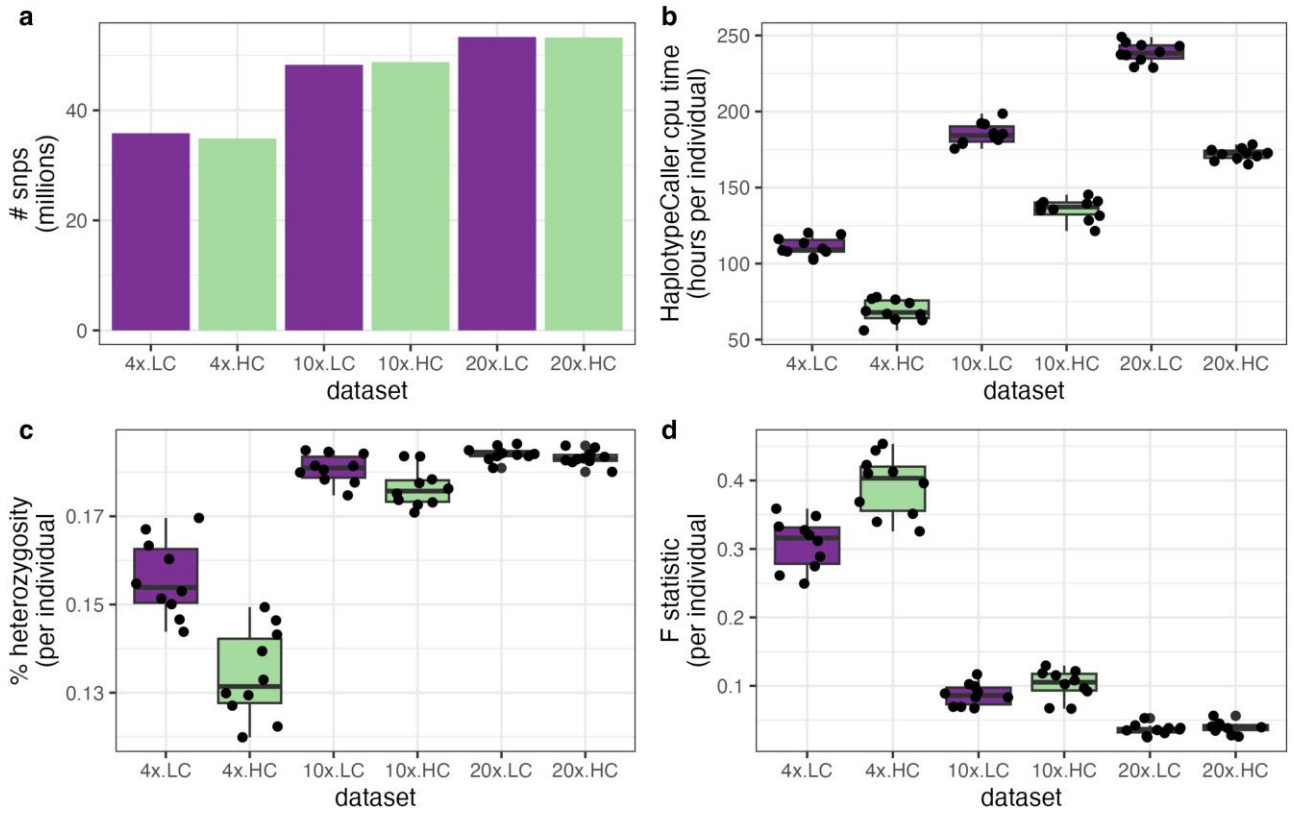

Fig. 2. Benchmarks for the 10 individual zebra finch coverage and pipeline testing. For each coverage data set (4×, 10×, 20×), we ran the low-coverage (LC) and high-coverage (HC) version of the pipeline, and calculated a) the overall number of SNPs following standard SNP filtering, b) the hours of CPU time to run HaplotypeCaller for each individual, c) the percentage of heterozygous sites for each individual, and d) the *F*-statistic calculated for each individual.

plots are colored according to these three clusters. Several metrics allow for the user to identify potential sequencing artifacts, for example by looking for associations between sequencing depth and PCA cluster (Fig. 4, upper right panel) or reference bias (Fig. 4, lower middle panel). An interactive heatmap of relatedness facilitates a rapid identification of close relatives in the data set that may have otherwise been overlooked. Finally, 2 maps project spatial data as an interactive plot and provide a first-pass visualization of the PCA clusters in space.

Postprocessing

By default, snpArcher produces a raw VCF file with only basic filters annotated. However, after viewing the individual-level QC visualizations as part of the QC module, users may wish to remove certain individuals from the analysis and apply additional filters on called variants. Additional postprocessing steps are implemented in a module, which runs if the user adds a column to the sample sheet header "SampleType." The postprocessing module will exclude from the filtered VCF any sample with "exclude" as the SampleType, retaining all other individuals. Following this sample filtering, this module implements additional user-configurable filters. By default, the postprocessing workflow removes sites that fall into regions of low mappability, regions with excess coverage, and regions with insufficient coverage (defined by the configuration file) and then removes sites with a minor allele frequency of <0.1 or missingness of >75%. These thresholds can be configured by the user. Finally, 2 clean variant files are produced for SNPs and indels separately.

MK Tests

To demonstrate the potential to extend snpArcher to incorporate downstream analysis, we developed a module to evaluate positive selection among a sample of individuals from a population (the ingroup) as well as one or more diverged samples (the outgroup) by computing MacDonald–Kreitman (MK) tests for each gene (McDonald and Kreitman 1991). This module is triggered when samples are annotated as "ingroup" and "outgroup" using the SampleType column in the sample sheet. Samples that do not have either designation will be excluded from the MK tests.

To facilitate the development of this module, we wrote a standalone Python program, degenotate (https://github.com/harvardinformatics/degenotate), that can retrieve coding sequences from an annotated genome, compute degeneracy across the genome, and calculate MK tables; degenotate can be installed via conda and run independently but is also incorporated into snpArcher's MK module. Briefly, degenotate assesses whether SNPs in the

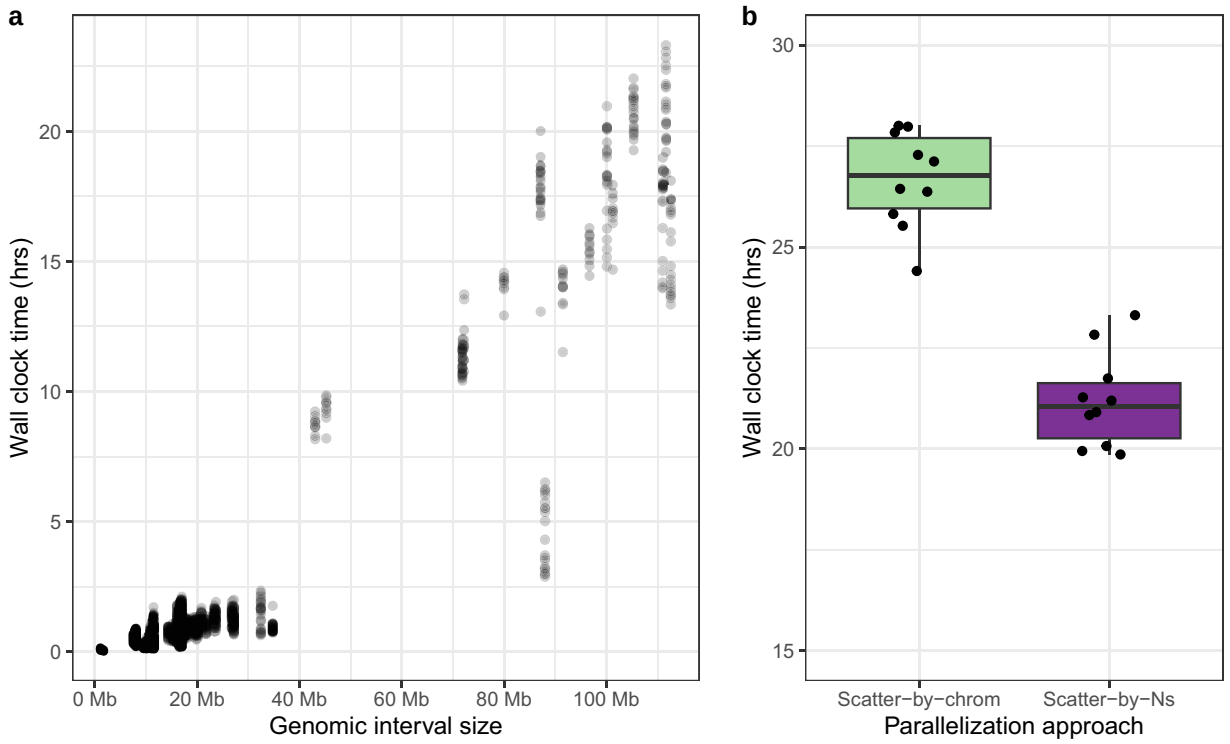

Fig. 3. Run time metrics from the HaplotypeCaller step of snpArcher. a) Wall clock time required to run HaplotypeCaller on different genomic interval sizes. b) Wall clock time elapsed per individual to complete the HaplotypeCaller step, using 2 parallelization approaches: scatter-by-chrom and scatter-by-Ns. Wall clock time per individual represents the maximum time taken of all intervals processed for that individual.

postprocessed VCF encode for polymorphic sites within the ingroup or fixed differences between the ingroup and the outgroup. It further classifies whether each SNP, whether polymorphic or fixed, is synonymous or nonsynonymous. Note that certain assumptions, detailed in the Materials and Methods, must be made about how to handle certain rare edge cases when doing this.

Based on these outputs, the MK module (or standalone degenotate) creates tables that are organized by gene and can be analyzed using the standard MK test statistic, using various extensions (Rand and Kann 1996; Stoletzki and Eyre-Walker 2011), or in aggregate to investigate genome-wide signatures of natural selection (Messer and Petrov 2013). This module will enable rapid application of population-genomic tests of selection (Fig. 5) and, in combination with the database of processed population data sets, provides a framework for comparing rates of adaptation to a range of species. Intriguingly, 3 collagen genes with potential roles in tooth and spine development across vertebrates (Jonsson et al. 1998; Bosse et al. 2017) are among the strongest outliers (Fig. 5a) and may be involved in the unique pufferfish morphology (Thiery et al. 2017).

UCSC Genome Browser Track Data Hub Generation

To facilitate downstream data exploration and as an example of the module development components of this work, we developed a module to generate UCSC Genome Browser track files to explore population variation data (see Materials and Methods). Briefly, this module computes and generates genome browser tracks for traditional population genomic summary statistics such as windowed estimates of Tajima's *D*, SNP density, pairwise nucleotide diversity (π), minor allele frequency, and SNP depth. The Genome Browser tracks allow for rapid analysis of common population genomic statistics along with other available genomic feature tracks in an easy-to-access and shareable format (Fig. 6).

Discussion

The production of high-quality and accurate genomic variation data sets for nonmodel species can be a challenging task, especially with the ever-increasing volume of genomic data that are being produced. The massive scale of population-scale whole-genome sequencing data sets presents significant hurdles in data management, processing, and analysis. In this manuscript, we introduce snpArcher, a powerful and user-friendly Snakemake workflow that addresses these challenges and enables the production of reliable and reproducible variation data sets. Crucially, our pipeline is parallelized, efficient, and scales well even up to modern population-scale data sets. snpArcher also provides an ideal tool for reanalyzing population-level data sets that are available on public databases and provides a consistent framework for comparative analyses across different data sets. By offering a reproducible and

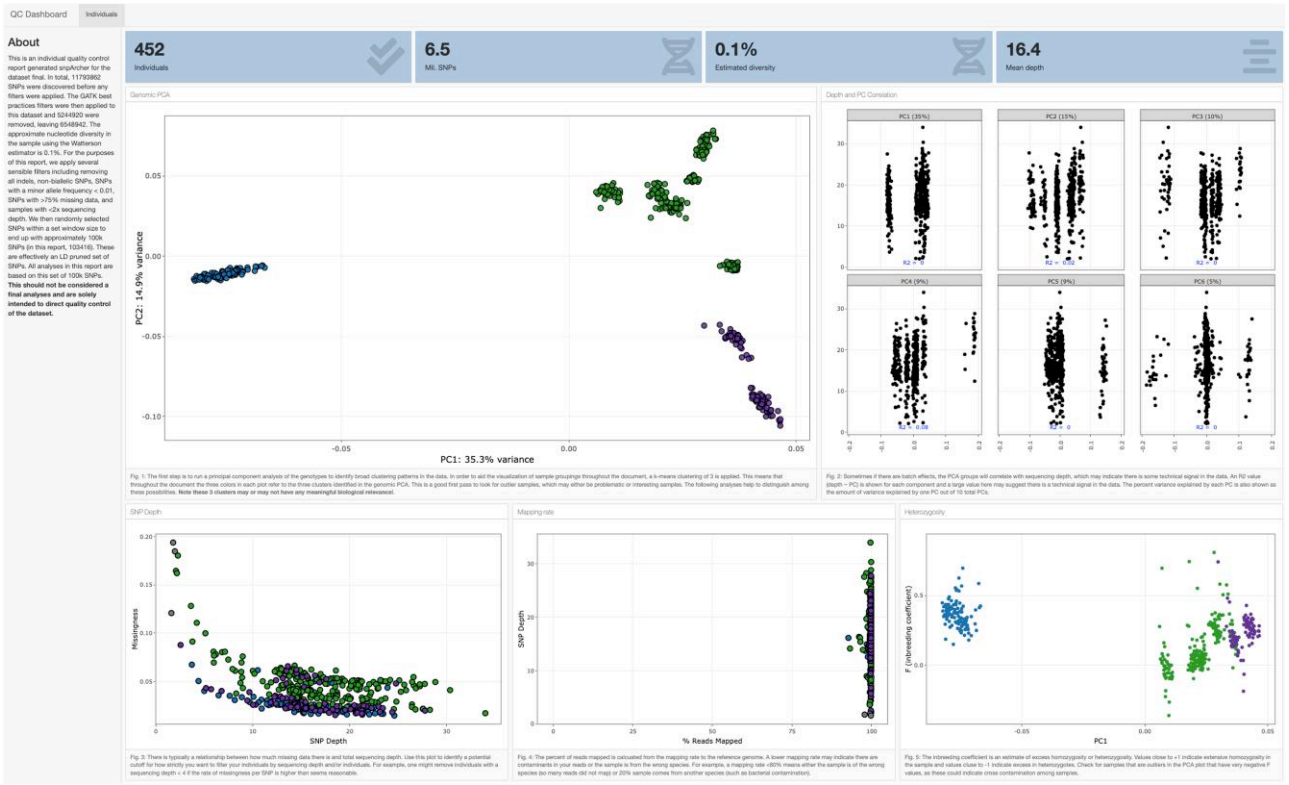

Fig. 4. Preview of QC dashboard for evaluating individual sequencing quality metrics. Shown here are genomic PCA, correlations between PCs and sequencing depth, relationship between missingness and SNP depth, percent mapped reads and SNP depth, and *FI* (inbreeding coefficient) and PC1. A complete interactive example can be found online (https://erikenbody.github.io/snpArcher/GCA_013435755.1_final_qc.html).

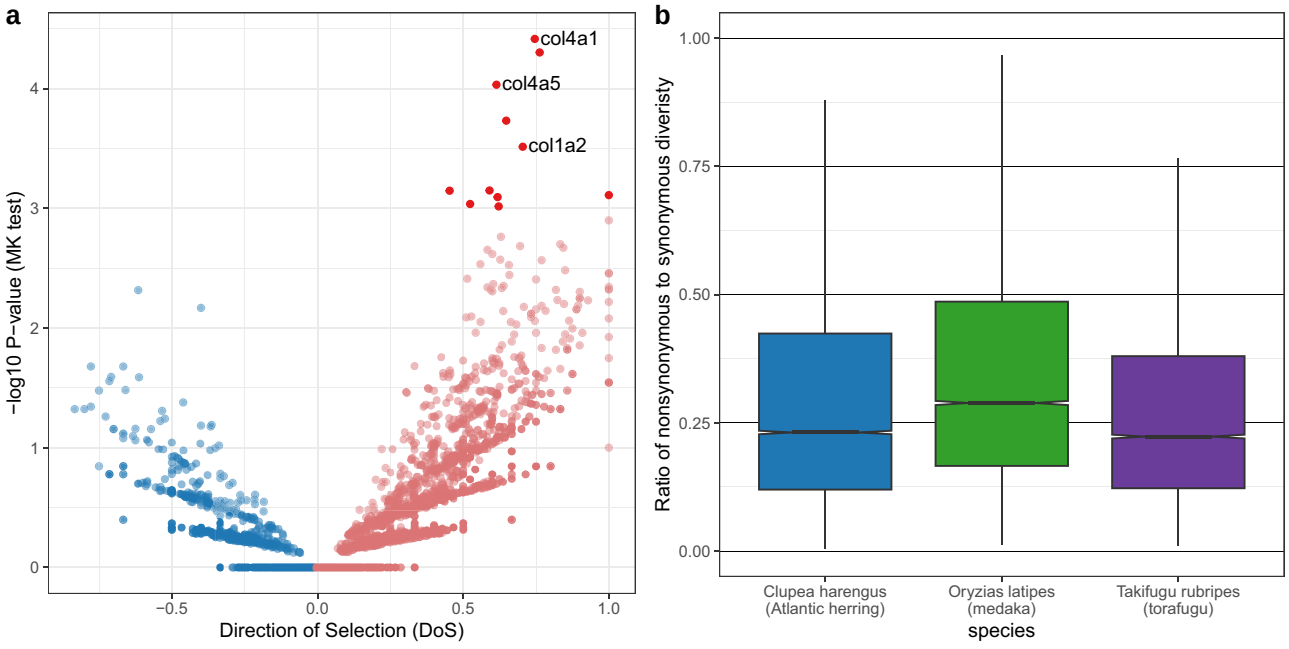

Fig. 5. Analysis of 3 fish data sets using snpArcher + degenotate, demonstrating the possible applications of this module. (a) MK results for *T. rubripes*, plotting the −log10 *P*-value of a Fisher's exact test of the MK table on the *y* axis, and the direction of selection on the *x* axis (Stoletzki and Eyre-Walker 2011); positive for an excess of nonsynonymous divergence, negative for an excess of nonsynonymous polymorphisms). Genes with nominal *P*-values of <0.001 are shaded dark, and 3 collagen genes with potential roles in tooth and spine development are highlighted. (b) Ratio of nonsynonymous to synonymous diversity (calculated based on number of segregating sites in each category) for 3 fish species. The species with the smallest population size, medaka, on average, has the largest values. Boxplot shows the median and interquartile range for protein-coding genes in the genomes of each species.

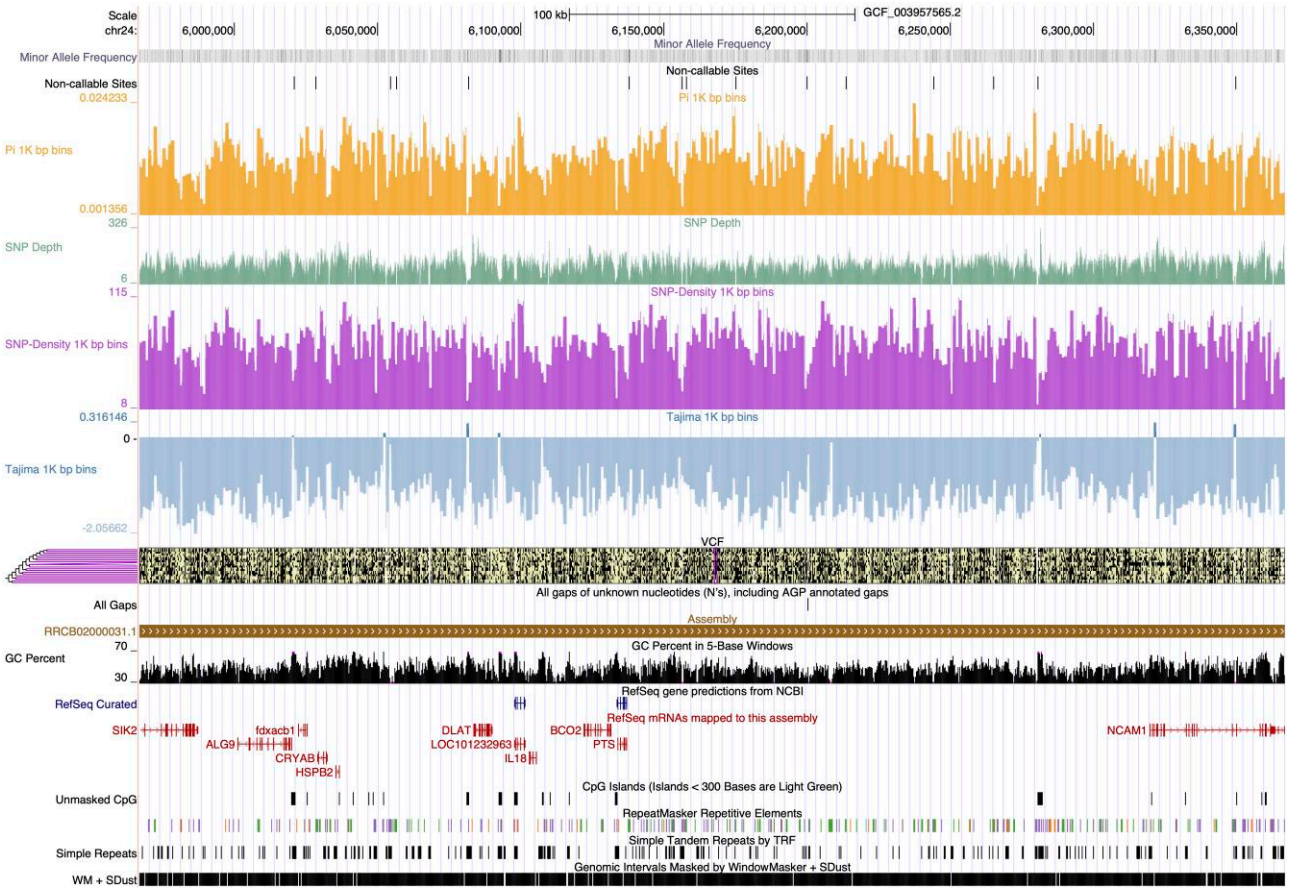

Fig. 6. Example genome browser trackhub created by the trackhub module. Tracks include minor allele frequency, noncallable sites, pairwise nucleotide diversity (π) in 1-kb bins, SNP depth, SNP density in 1-kb bins, Tajima's *D* in 1-kb bins, and a VCF track.

well-documented analysis pipeline, snpArcher ensures the reliability and consistency of results, empowering researchers to spend less time on complex data and workflow management and more time on analysis and discovery.

Extensibility of snpArcher

A key goal in the design of the snpArcher pipeline is to allow extensibility for subsequent analyses after the primary variant calling. We implement this using Snakemake modules, which allow additional rules to easily extend the main pipeline. To be added to snpArcher, a module only needs a way to indicate that it should be run, such as a flag in the config file or a column in the sample sheet, and for output files from snpArcher to be linked to input files of the workflow. As long as these constraints are met, any user-defined Snakemake workflow can be imported as a module. Furthermore, to enable seamless integration of user-developed modules, we define a set of guidelines for users to follow when developing and contributing modules (https://snparcher.readthedocs.io/en/latest/modules.html#module-contribution-guidelines). Finally, we present several modular extensions of snpArcher here, but we hope also that user-developed modules will grow the set of tools linked to snpArcher in order to facilitate diverse analysis.

Challenges and Prospects for Reuse of Public Data

Publically available data sets provide opportunities for comparative genomics and also present limitations inherent to data reuse. Metadata associated with genomic data is often fragmented or missing, meaning crucial information for QC of reusing data is not always available (Gonçalves and Musen 2019; Toczydlowski et al. 2021). A key function of the snpArcher pipeline is to produce metrics to evaluate potential biases in the data set for common population genomic issues. For example, pedigree information is typically not available for wild populations and likely to be missing from public data sets, but close relatives may bias many common population genomic analyses (Hendricks et al. 2018). Our QC module reports relatedness information, allowing rapid identification of related individuals. In the data sets we analyzed, 14% of all data sets considered included identical individuals by genotype and 47% of data sets included at least 1 first-degree relative in it. At the population scale, undetected population structure can bias population and quantitative genomic analysis, and the PCA and admixture reports in the QC module will give a first-pass assessment of known or unknown structuring. Sequencing data on public databases can contain contamination, either from other individuals or other species. These can be identified using measures of inbreeding (i.e. low inbreeding values may

suggest excess heterozygosity and cross-contamination) that are reported in the QC module. Outliers in sequencing depth, missingness, and mapping rate are all quickly identifiable using the interactive QC plots. Finally, data quality at short-scale genomic intervals can be visualized using the genome browser outputs, for example to evaluate sequencing depth and genetic diversity around regions of interest.

Together, we expect these improvements to variant calling for the average user will enable users to focus on analyzing their data, rather than processing it. This is made easier by the extensibility of snpArcher, where it is relatively simple to add on new analyses to the workflow. The standardization inherent to the workflow will further improve our ability to reuse large and unwieldy publically available data sets. We hope the ease of use and flexibility of the snpArcher workflow will enable a more rapid and reproducible workflow for researchers generating large population genomic data sets.

Materials and Methods

Configuration

Core workflow options in snpArcher are controlled by a YAML configuration file. This file controls options such as module selection, output prefix for final files, and temporary storage location. For complete instructions on setting up snpArcher, please refer to our documentation: https://snparcher.readthedocs.io/en/latest/setup.html.

In order to determine what outputs to create, snpArcher requires users to create a *sample sheet file*. This comma-separated file contains the required sample metadata about the user's samples in order to run the workflow. At a minimum, the snpArcher pipeline requires that each sample has a unique sample name, a reference genome accession or a path to a fasta file, and a Sequence Read Archive (SRA) accession or path to 2 paired-end fastq files (Table 1). We include with snpArcher a simple script, written in Python, to facilitate the generation of sample sheets from local data sets, and we include examples of how to create snpArcher sample sheets from SRA run tables in R.

Computer Resources and Cloud Configuration

Variant calling for large population-level sequencing data sets is computationally intensive and requires substantial computational resources to run. While it is possible to run snpArcher on a laptop for small data sets, such as the test data set included in the workflow or single samples, we have optimized it to run on HPC clusters and cloud compute platforms. We have tested snpArcher extensively on SLURM-based high-performance clusters and on the Google Cloud Life Sciences platform, and following Snakemake best practices, we provide configurable profiles that can be enabled depending on which computational resources you will use. The SLURM profile and associated bash script provide the basic configuration for running on a SLURM cluster, but the profile will need to be adjusted according to the configuration of the user's specific system.

Table 1 An example of the minimum required sample metadata to run snpArcher

BioSample	refGenome	Run
SAMEA3532857	GCF_003957565.2	ERR1013161
SAMEA3532860	GCF_003957565.2	ERR1013164
SAMEA3532862	GCF_003957565.2	ERR1013166
SAMEA3532864	GCF_003957565.2	ERR1013168
SAMEA3532865	GCF_003957565.2	ERR1013169

To run snpArcher on the Google Cloud Platform (GCP), the user must have a Google account linked to a billing account where charges for computational resources can be made. This Google configuration is set up outside of snpArcher, on the command line, and on the Google Cloud Console. Once this is set up in the local environment, snpArcher can be directed to run on Google Cloud instances using the GCP profile provided with the workflow. The user can define how many instances to create and also define the size of required resources in the resources.yaml file included in the workflow. The GCP profile also is configured to exploit preemptible instances, which are short-term compute instances that are offered at considerable cost savings, but can only run for 24 h and be bought out by other GCP users. The current defaults have been optimized for data sets of genome size of ~2 Gb, 150 individuals, and 10× sequencing depth with an estimated cost of $1/sample when a Sentieon license is available. Larger or smaller projects may need to tweak these resources to optimize cost/saving benefits best and prevent the preemption of long-running data sets.

Data Acquisition and Preprocessing

The first step of the workflow is the acquisition and preprocessing of raw sequence data and reference genomes. For each sample, 2 paired-end fastq files are required. The default behavior is to retrieve sequencing data from National Center for Biotechnology Information (NCBI) based on an SRA run accession (Leinonen et al. 2011) using *prefetch*. For various reasons, *prefetch* may fail. If this happens, *ffq* (Gálvez-Merchán et al. 2022) is used to generate a FTP link for the accession that is downloaded. Alternatively, users can supply paths to fastq files in the sample sheet, in which case snpArcher will operate on those locally stored files. Next, sequencing adapters are trimmed from the raw fastq files with *fastp* (Chen et al. 2018), and sequences with greater than 40% of bases with a phred score below Q15 were removed. Reference genomes are retrieved using the NCBI data sets tool (Sayers et al. 2021) if an NCBI accession is specified; otherwise, a path to the reference fasta must be included in the sample sheet. Once available, the reference fasta is processed using *bwa index* (Li and Durbin 2009), *samtools faidx*, and *samtools dict* (Li et al. 2009) to produce the indexes necessary for downstream processes.

Read Mapping

After the raw data are retrieved and preprocessed, the workflow aligns sequencing reads to the reference genome

using *bwa mem* (Li 2013) to produce per sample BAM files. For each sample, read groups are appended based on the sample sheet specification. We mark PCR duplicates using *sambamba markdup* (Tarasov et al. 2015) to exclude these technical artifacts from downstream analysis. Alignment statistics are calculated per sample using *samtools flagstat*.

Mappability and Coverage

Additionally, mappability statistics are computed on the reference genome using *genmap* (Pockrandt et al. 2020). Per-site coverage statistics are optionally computed and aggregated using *d4tools* (Hou et al. 2021), *mosDepth* (Pedersen and Quinlan 2018), and *bedtools* (Quinlan and Hall 2010). Mappability statistics for the reference genome, combined with per-site coverage statistics, can be used to generate a bed file delineating callable regions of the genome based on user-configurable thresholds.

Variant Calling

We use GATK (McKenna et al. 2010) for individual variant calling and joint genotyping. First, we employ GATK HaplotypeCaller to call SNPs and indels in each sample. If the user has selected the low-coverage configuration, we set the *--min-pruning* and *–min-dangling-branch-length* options equal to 1 (Hui et al. 2020); otherwise, defaults are used. Next, individual variant calls are aggregated into an efficient data structure via GATK GenomicsDBImport. This step is necessary to enable large cohort joint genotyping. Then, we use GATK GenotypeGVCFs to perform joint genotyping and produce a multisample VCF, retaining only high confidence variants. This approach is broadly adapted by the field as the standard for variant calling, as evidenced by nearly 20,000 citations of the flagship GATK paper to date. Finally, we apply filter annotations to the VCF according to the GATK best practices (Van der Auwera et al. 2013) using GATK VariantFiltration.

Parallelization

Processing even moderately sized data sets can be exceptionally slow with GATK. One solution is to parallelize each GATK step by splitting the reference genome into processing intervals for both the individual and joint genotyping steps. Optimally, this interval creation step divides the genome into shorter subchromosomal (or subscaffold) pieces so that each interval can finish in a shorter amount of time. In order to optimize runtime, we use a two-step interval creation process. We generate an initial set of calling intervals using the ScatterIntervalsByNs tool to divide the reference genome at large blocks of Ns. This is important because SNP calling with GATK Haplotype Caller is based on local reassembly, which can be adversely affected if, for example, reads that map across an interval boundary are discarded. However, for many reference genomes, this can result in thousands of intervals, which leads to inefficient workflows as the time to assess which jobs need to run becomes prohibitive. To create a balanced set of interval lists, we use the GATK SplitIntervals tool using the option *<-mode BALANCING_WITHOUT_INTERVAL_SUBDIVISION>*, which creates a set of interval lists (up to a maximum user-specified value) that all have approximately equal numbers of bases. For the joint genotyping step, each site is treated independently, so we can gain efficiency by creating additional intervals without the concern of splitting adjacent regions of the genome. Thus, for the second set of intervals, we use the option *<-mode INTERVAL_SUBDIVISION>* to produce a scalable number of intervals that can divide adjacent regions. These intervals are then used to parallelize GATK GenomicsDBImport for efficient multisample calling.

Sentieon Accelerated Variant Calling

In addition to the BWA/GATK mapping and variant calling pipeline, we include a Sentieon (Kendig et al. 2019) workflow. This software package is proprietary and produces identical results as GATK but has been much more efficiently parallelized, resulting in substantially reduced compute needs. The Sentieon workflow uses Sentieon's drop-in replacement tools for mapping, PCR duplicate removal, metrics, and variant calling. The use of this workflow is a user-specified option in snpArcher and requires a software license from Sentieon that can be specified in the config file.

Quality Control

snpArcher includes an optional QC module that aggregates various statistics from the workflow and produces preliminary analyses and plots in an interactive HTML file. We estimate the per-individual variant metrics SNP-depth, individual missingness, heterozygosity, and transition/transversions, using *vcftools* v0.1.16 (Danecek et al. 2011). We next generate a small subset of variant data for calculating several preliminary population genomic statistics. In order to generate this pruned data set, we use *bcftools* v1.12 (Danecek et al. 2021) to first remove all SNPs not passing the filters described above and remove indels, sites with minor allele frequency <0.01 (i.e. sites present in only 1% of the population), sites with >75% missing data, and any sites mapping to a previously annotated mitochondrial genome. We next calculate how large of a window to prune this filtered data set to retain 100k variant sites (i.e. $\text{WindowSize} = N_{\text{SNPs}}/100{,}000$) and use *bcftools* to select 1 SNP at random per window. This pruned variant file of 100k SNPs is used for all downstream QC calculations; however, several basic summaries (total number of SNPs, approximate theta, and number of individuals) are calculated from the full variant file and presented in the header of QC HTML file.

We used *Plink2* v2.00a2.3 (Chang et al. 2015; Purcell and Chang 2023) to perform genome PCA (Galinsky et al. 2016) and a KING relatedness matrix (Manichaikul et al. 2010). We also generate a distance matrix using *Plink v* 1.90b6.21 (Purcell et al. 2007). If geographic coordinates are provided, samples will be plotted on an interactive map. Lastly, we used *admixture* v1.3.0 (Alexander et al.

2009) to calculate admixture for $k = 2$ and $k = 3$ from the pruned variant file. The output of these analyses, tabulations of variant files, and mapping statistics are all summarized in a single interactive HTML dashboard. Briefly, we use R v4.1.3 (R Core Team 2022) and the following packages for building this summary: *tidyverse* v1.3.1 (Wickham et al. 2019) for data manipulation, *ggplot2* v3.3.5 (Wickham 2016) for graphics, *plotly* v4.9.4.1 (Sievert 2020) for interactive graphics, *ape* v5.5 (Paradis and Schliep 2019) and *ggtree* 3.2.0 (Yu et al. 2018) for phylogenetic tree visualization, reshape2 v1.4.4 (Wickham 2007) for data management, and *ggmap* v3.3.0 (Kahle and Wickham 2013) for terrain maps.

Postprocessing

In order to enable users to efficiently filter individuals from their VCF file after initially running snpArcher, we include the postprocessing module. Users can trigger this module by marking individuals for removal using the "SampleType" column in their sample sheet. The postprocessing module applies customizable filters, which by default remove sites in regions of low mappability and excessive or insufficient coverage (as defined in the configuration file) using *bedtools* and sites with a minor allele frequency of <0.1 or missingness of >75% using *bcftools* (after recalculating these metrics following sample removal). We also produce separate variant files for SNPs and small indels called by GATK.

Trackhubs

To display population genomic statistics calculated from the VCF generated by snpArcher, we include an optional module to generate a UCSC Genome Browser track data hub (Raney et al. 2014). At time of publication, this module calculates Tajima's *D* (Tajima 1989), SNP density, nucleotide diversity (π), and allele frequency. These statistics are calculated using VCFtools v0.1.15 and converted to bigBed format using *bedToBigBed* (Kent et al. 2010).

Annotating Codon Degeneracy and Inferring Synonymous and Nonsynonymous Variants

snpArcher also includes an optional module that annotates the degeneracy of all coding regions in the reference genome and implements the classic MK test for detecting selection acting in coding regions within a population (McDonald and Kreitman 1991). Briefly, this test compares the number of SNPs present within the population that either change (nonsynonymous) or do not change (synonymous) the amino acid encoded at that position. This is compared with similar counts of fixed differences in a diverged outgroup sample to see if and how the ratio of nonsynonymous to synonymous changes differs between them. While annotating degeneracy and computing tables for the MK test are common tasks in population genetics, we are not aware of any tools that automate these analyses at a genome-wide scale. To facilitate the integration of this functionality into snpArcher, we developed a standalone tool called degenotate (https://github.com/harvardinformatics/degenotate), which calculates MK tables, performs degeneracy annotation, and allows users to extract coding sequences from a genome by their degeneracy.

To implement the MK test across diverse organisms, we make some assumptions about how to classify polymorphic and divergent sites. We consider a polymorphic site to be any location where at least 1 individual within the ingroup possesses a nonreference allele and divergent sites to be only those where none of the outgroup alleles exist in the ingroup. Using these definitions, it is possible for a site to both be polymorphic and fixed if the outgroup alleles are different from the alleles segregating within the population. For quantifying variants, we also make some simplifying assumptions. First, if a codon has more than 1 variant segregating within a population (either because multiple positions at the codon have segregating sites or because 1 position has a multiallelic SNP), we treat each segregating variant as independent. For the outgroup, if there are multiple fixed differences in a single codon in the outgroup, we compute all possible mutational pathways between the ingroup codon and the outgroup codon and take the average number of nonsynonymous and synonymous changes across these paths, weighted equally. This means we can have fractional numbers of synonymous and nonsynonymous divergence. We also implement calculations of the neutrality index (Rand and Kann 1996) and direction of selection (Stoletzki and Eyre-Walker 2011) based on the MK test results.

Empirical Data Sets

In order to test our pipeline and provide a robust set of consistently processed variant calls for downstream applications, we ran snpArcher on a set of publicly available resequencing data sets (supplementary table S1, Supplementary Material online). The resulting VCF and genomic VCF files can be accessed via Globus (Foster 2011; Allen et al. 2012) in the "Comparative Population Genomics Data" public collection (link available at https://snparcher.readthedocs.io/en/latest/datasets.html). Of the data sets processed, 13 are multispecies samples mapped to a common reference genome, 7 are primarily a single species but with 1 or 2 outgroup samples, and 6 are purely a single species. We focus on nonmammalian vertebrates, as high-quality reference genomes are frequently available in this group, but genome sizes are manageable to limit the computational demands needed to process many large population samples. In total, we used the following 26 data sets: *A. citrinellus* (Kautt et al. 2020), *Anas platyrhynchos* (Zhou et al. 2018; Feng et al. 2021), *Anolis carolinensis* (Bourgeois et al. 2019), *Astatotilapia calliptera* (Malinsky et al. 2015, 2018; Weber et al. 2021), *Athene cunicularia* (Mueller et al. 2018; Feng et al. 2020), *Chaenogobius annularis* (Hirase et al. 2021), *Clupea harengus* (Feng et al. 2017; Lamichhaney et al. 2017; Han et al. 2020), *Corvus cornix* (Poelstra et al. 2014; Vijay et al. 2016), *Coturnix japonica* (Wu et al. 2018; Liu

et al. 2019), *Egretta garzetta* (Li et al. 2014; Feng et al. 2020), *Eopsaltria australis* (Gan et al. 2019), *Falco peregrinus* (Zhan et al. 2013; Gu et al. 2021), *Ficedula albicollis* (Burri et al. 2015; Kardos et al. 2016; Smeds et al. 2016), *Gasterosteus aculeatus* (Yoshida et al. 2014, 2020; Feulner et al. 2015; Liu et al. 2016; Ishikawa et al. 2017; Marques et al. 2017; Haenel et al. 2019; Miller et al. 2019; Verta and Jones 2019; Kirch et al. 2021), *Hippocampus comes* (Li et al. 2021), *Hirundo rustica* (Schield et al. 2021), *Hypoplectrus puella* (Hench et al. 2019), *Oryzias latipes* (Spivakov et al. 2014; Ansai et al. 2021), *Parus major* (Qu et al. 2015; Laine et al. 2016), *Passer domesticus* (Elgvin et al. 2017; Ravinet et al. 2018; Runemark et al. 2018), *Pungitius pungitius* (White et al. 2015; Dixon et al. 2019), *Sylvia atricapilla* (Delmore et al. 2020), *Symphodus melops* (Mattingsdal et al. 2020), *T. guttata* (Singhal et al. 2015), *Takifugu rubripes* (Zhang et al. 2020), and *Thunnus albacares* (Barth et al. 2017).

We used SRA to search for possible data sets for inclusion, limiting our search space to species with (i) a reference genome and (ii) at least 1 BioProject that contains a minimum of 10 BioSamples sequenced to at least 5× average coverage. The resulting list was then manually curated to identify publications associated with each BioProject, excluding from further consideration data sets for which a publication could not be identified. We then manually assessed the resulting plausible samples to identify a subset for further analysis. R notebooks are provided on GitHub that contain the code for initial and final assessments (https://github.com/sjswuitchik/compPopGen_ms).

Benchmarking

To investigate the impact of low sequencing depth on variant calling, we, first, subsampled the original high-depth data set zebra finch data set to 4×, 10×, and 20× coverage. We ran snpArcher on these subsampled data sets and filtered the resulting VCF files by removing sites not passing standard filters and calculated heterozygosity statistics using VCFtools v0.1.15 (Danecek et al. 2011). Second, we assessed the effectiveness of our variant calling parallelization (scatter-by-Ns) approach to the conventional (scatter-by-chromosome) approach using the 10× data set. We performed these benchmarking runs on Google Cloud compute instances, selecting the instance types for each rule to balance cost and runtime (supplementary table S2, Supplementary Material online).

Supplementary Material

Supplementary material is available at *Molecular Biology and Evolution* online.

Acknowledgments

We are grateful for the computational resources provided by the FASRC Cannon cluster, supported by the FAS Division of Science Research Computing Group at Harvard University. We thank the CCGP team (Brad Shaffer, Erin Toffelmier, Courtney Miller, Merly Escalona, and Anne Chambers) for their input during the development of snpArcher.

Funding

C.D.M., E.E., and M.B. were funded by the California Conservation Genomics Project, with funding provided to the University of California by the State of California, State Budget Act of 2019 (UC Award ID RSI-19-690224). T.B.S. and S.J.S. were funded by the National Science Foundation (NSF) Division of Biological Infrastructure award DEB-1754397 to T.B.S.

Conflict of interest statement. None declared.

Data Availability

The snpArcher source code is available at https://github.com/harvardinformatics/snpArcher. The Comparative Population Genomics Data public collection is freely available on Globus (link available at https://snparcher.readthedocs.io/en/latest/datasets.html), and SRA BioProject accessions for WGS data used to produce this resource are available in supplementary table S1, Supplementary Material online. Scripts used to assess public data sets for the Comparative Population Genomics Data public collection are available at https://github.com/sjswuitchik/compPopGen_ms. The zebra finch WGS data used to benchmark snpArcher is publicly available via the SRA BioProject accession PRJEB10586.

References

Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*. 2009:**19**(9):1655–1664. https://doi.org/10.1101/gr.094052.109.

Allen B, Bresnahan J, Childers L, Foster I, Kandaswamy G, Kettimuthu R, Kordas J, Link M, Martin S, Pickett K. Software as a service for data scientists. *Comm ACM*. 2012:**55**(2):81–88. https://doi.org/10.1145/2076450.2076468.

Ansai S, Mochida K, Fujimoto S, Mokodongan DF, Sumarto BKA, Masengi KWA, Hadiaty RK, Nagano AJ, Toyoda A, Naruse K, *et al.* Genome editing reveals fitness effects of a gene for sexual dichromatism in Sulawesian fishes. *Nat Commun*. 2021:**12**(1):1350. https://doi.org/10.1038/s41467-021-21697-0.

Barth JMI, Damerau M, Matschiner M, Jentoft S, Hanel R. Genomic differentiation and demographic histories of Atlantic and Indo-Pacific yellowfin tuna (*Thunnus albacares*) populations. *Genome Biol Evol*. 2017:**9**(4):1084–1098. https://doi.org/10.1093/gbe/evx067.

Benjelloun B, Boyer F, Streeter I, Zamani W, Engelen S, Alberti A, Alberto FJ, BenBati M, Ibnelbachyr M, Chentouf M, *et al.* An evaluation of sequencing coverage and genotyping strategies to assess neutral and adaptive diversity. *Mol Ecol Resour*. 2019:**19**(6):1497–1515. https://doi.org/10.1111/1755-0998.13070.

Bosse M, Spurgin LG, Laine VN, Cole EF, Firth JA, Gienapp P, Gosler AG, McMahon K, Poissant J, Verhagen I, *et al.* Recent natural selection causes adaptive evolution of an avian polygenic trait.

Science. 2017:**358**(6361):365–368. https://doi.org/10.1126/science.aal3298.

Bourgeois Y, Ruggiero RP, Manthey JD, Boissinot S. Recent secondary contacts, linked selection, and variable recombination rates shape genomic diversity in the model species anolis carolinensis. *Genome Biol Evol*. 2019:**11**(7):2009–2022. https://doi.org/10.1093/gbe/evz110.

Breton G, Johansson ACV, Sjödin P, Schlebusch CM, Jakobsson M. Comparison of sequencing data processing pipelines and application to underrepresented African human populations. *BMC Bioinformatics*. 2021:**22**(1):488. https://doi.org/10.1186/s12859-021-04407-x.

Buffalo V. Quantifying the relationship between genetic diversity and population size suggests natural selection cannot explain Lewontin's paradox. *eLife*. 2021:**10**:e67509. https://doi.org/10.7554/eLife.67509.

Burri R, Nater A, Kawakami T, Mugal CF, Olason PI, Smeds L, Suh A, Dutoit L, Bureš S, Garamszegi LZ, *et al.* Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of Ficedula flycatchers. *Genome Res*. 2015:**25**(11):1656–1665. https://doi.org/10.1101/gr.196485.115.

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*. 2015:**4**(1):7. https://doi.org/10.1186/s13742-015-0047-8.

Chen S, Francioli LC, Goodrich JK, Collins RL, Kanai M, Wang Q, Alföldi J, Watts NA, Vittal C, Gauthier LD, *et al.* A genome-wide mutational constraint map quantified from variation in 76,156 human genomes. *bioRxiv*. [preprint] https://www.biorxiv.org/content/10.1101/2022.03.20.485034v2.

Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. 2018:**34**(17):i884–i890. https://doi.org/10.1093/bioinformatics/bty560.

Corbett-Detig RB, Hartl DL, Sackton TB. Natural selection constrains neutral diversity across a wide range of species. *PLoS Biol*. 2015:**13**(4):e1002112. https://doi.org/10.1371/journal.pbio.1002112.

Cullen JN, Friedenberg SG. WAGS: user-friendly, rapid, containerized pipelines for processing, variant discovery, and annotation of short read whole genome sequencing data. *G3: Genes, Genomes, Genetics*. 2023:**13**:jkd117. https://doi.org/10.1093/g3journal/jkad117.

Czech L, Exposito-Alonso M. grenepipe: a flexible, scalable and reproducible pipeline to automate variant calling from sequence reads. *Bioinformatics*. 2022:**38**(20):4809–4811. https://doi.org/10.1093/bioinformatics/btac600.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, *et al.* The variant call format and VCFtools. *Bioinformatics*. 2011:**27**(15):2156–2158. https://doi.org/10.1093/bioinformatics/btr330.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, *et al.* Twelve years of SAMtools and BCFtools. *Gigascience*. 2021:**10**(2):giab008. https://doi.org/10.1093/gigascience/giab008.

Delmore K, Illera JC, Pérez-Tris J, Segelbacher G, Lugo Ramos JS, Durieux G, Ishigohoka J, Liedvogel M. The evolutionary history and genomics of European blackcap migration. *eLife*. 2020:**9**:e54462. https://doi.org/10.7554/eLife.54462.

Dixon G, Kitano J, Kirkpatrick M. The origin of a new sex chromosome by introgression between two stickleback fishes. *Mol Biol Evol*. 2019:**36**(1):28–38. https://doi.org/10.1093/molbev/msy181.

Elgvin TO, Trier CN, Tørresen OK, Hagen IJ, Lien S, Nederbragt AJ, Ravinet M, Jensen H, Sætre G-P. The genomic mosaicism of hybrid speciation. *Sci Adv*. 2017:**3**(6):e1602996. https://doi.org/10.1126/sciadv.1602996.

Ellegren H. Genome sequencing and population genomics in non-model organisms. *Trends Ecol Evol*. 2014:**29**(1):51–63. https://doi.org/10.1016/j.tree.2013.09.008.

Exposito-Alonso M, Booker TR, Czech L, Gillespie L, Hateley S, Kyriazis CC, Lang PLM, Leventhal L, Nogues-Bravo D, Pagowski V, *et al.* Genetic diversity loss in the Anthropocene. *Science*. 2022:**377**(6613):1431–1435. https://doi.org/10.1126/science.abn5642.

Fan S, Spence JP, Feng Y, Hansen MEB, Terhorst J, Beltrame MH, Ranciaro A, Hirbo J, Beggs W, Thomas N, *et al.* Whole-genome sequencing reveals a complex African population demographic history and signatures of local adaptation. *Cell*. 2023:**186**(5):923–939.e14. https://doi.org/10.1016/j.cell.2023.01.042.

Feng C, Pettersson M, Lamichhaney S, Rubin C-J, Rafati N, Casini M, Folkvord A, Andersson L. Moderate nucleotide diversity in the Atlantic herring is associated with a low mutation rate. *eLife*. 2017:**6**:e23907. https://doi.org/10.7554/eLife.23907.

Feng S, Stiller J, Deng Y, Armstrong J, Fang Q, Reeve AH, Xie D, Chen G, Guo C, Faircloth BC, *et al.* Dense sampling of bird diversity increases power of comparative genomics. *Nature*. 2020:**587**(7833):252–257. https://doi.org/10.1038/s41586-020-2873-9.

Feng P, Zeng T, Yang H, Chen G, Du J, Chen L, Shen J, Tao Z, Wang P, Yang L, *et al.* Whole-genome resequencing provides insights into the population structure and domestication signatures of ducks in eastern China. *BMC Genomics*. 2021:**22**(1):401. https://doi.org/10.1186/s12864-021-07710-2.

Feulner PGD, Chain FJJ, Panchal M, Huang Y, Eizaguirre C, Kalbe M, Lenz TL, Samonte IE, Stoll M, Bornberg-Bauer E, *et al.* Genomics of divergence along a continuum of parapatric population differentiation. *PLoS Genet*. 2015:**11**(2):e1004966. https://doi.org/10.1371/journal.pgen.1004966.

Foster I. Globus online: accelerating and democratizing science through cloud-based services. *IEEE Internet Comput*. 2011:**15**(3):70–73. https://doi.org/10.1109/MIC.2011.64.

Galinsky KJ, Bhatia G, Loh P-R, Georgiev S, Mukherjee S, Patterson NJ, Price AL. Fast principal-component analysis reveals convergent evolution of ADH1B in Europe and East Asia. *Am J Hum Genet*. 2016:**98**(3):456–472. https://doi.org/10.1016/j.ajhg.2015.12.022.

Gan HM, Falk S, Morales HE, Austin CM, Sunnucks P, Pavlova A. Genomic evidence of neo-sex chromosomes in the eastern yellow robin. *Gigascience*. 2019:**8**(9):giz111. https://doi.org/10.1093/gigascience/giz111.

Gálvez-Merchán Á, Min KH, Pachter L, Sina Booeshaghi A. Metadata retrieval from sequence databases with ffq. *bioRxiv* 2022. [preprint] https://www.biorxiv.org/content/10.1101/2022.05.18.492548v2.

Gonçalves RS, Musen MA. The variable quality of metadata about biological samples used in biomedical experiments. *Sci Data*. 2019:**6**(1):190021. https://doi.org/10.1038/sdata.2019.21.

Gu Z, Pan S, Lin Z, Hu L, Dai X, Chang J, Xue Y, Su H, Long J, Sun M, *et al.* Climate-driven flyway changes and memory-based long-distance migration. *Nature*. 2021:**591**(7849):259–264. https://doi.org/10.1038/s41586-021-03265-0.

Haenel Q, Roesti M, Moser D, MacColl ADC, Berner D. Predictable genome-wide sorting of standing genetic variation during parallel adaptation to basic versus acidic environments in stickleback fish. *Evol Lett*. 2019:**3**(1):28–42. https://doi.org/10.1002/evl3.99.

Han F, Jamsandekar M, Pettersson ME, Su L, Fuentes-Pardo AP, Davis BW, Bekkevold D, Berg F, Casini M, Dahle G, *et al.* Ecological adaptation in Atlantic herring is associated with large shifts in allele frequencies at hundreds of loci. *eLife*. 2020:**9**:e61076. https://doi.org/10.7554/eLife.61076.

Heldenbrand JR, Baheti S, Bockol MA, Drucker TM, Hart SN, Hudson ME, Iyer RK, Kalmbach MT, Kendig KI, Klee EW, *et al.* Recommendations for performance optimizations when using GATK3.8 and GATK4. *BMC Bioinformatics*. 2019:**20**(1):557. https://doi.org/10.1186/s12859-019-3169-7.

Hench K, Vargas M, Höppner MP, McMillan WO, Puebla O. Inter-chromosomal coupling between vision and pigmentation genes during genomic divergence. *Nat Ecol Evol*. 2019:**3**(4):657–667. https://doi.org/10.1038/s41559-019-0814-5.

Hendricks S, Anderson EC, Antao T, Bernatchez L, Forester BR, Garner B, Hand BK, Hohenlohe PA, Kardos M, Koop B, *et al.*

Recent advances in conservation and population genomics data analysis. *Evol Appl*. 2018:**11**(8):1197–1211. https://doi.org/10.1111/eva.12659.

Hill J, Enbody ED, Pettersson M, Sprehn CG, Bekkevold D, Folkvord A, Kleinau G, Scheerer P, Andersson L. Recurrent convergent evolution at amino acid residue 261 in fish rhodopsin. *Proc Natl Acad Sci U S A*. 2019:**116**(37):18473–18478. https://doi.org/10.1073/pnas.1908332116.

Hirase S, Tezuka A, Nagano AJ, Sato M, Hosoya S, Kikuchi K, Iwasaki W. Integrative genomic phylogeography reveals signs of mitonuclear incompatibility in a natural hybrid goby population. *Evolution*. 2021:**75**(1):176–194. https://doi.org/10.1111/evo.14120.

Hou H, Pedersen B, Quinlan A. Balancing efficient analysis and storage of quantitative genomics data with the D4 format and d4tools. *Nat Comput Sci*. 2021:**1**(6):441–447. https://doi.org/10.1038/s43588-021-00085-0.

Huerta-Sánchez E, Jin X, Asan , Bianba Z, Peter BM, Vinckenbosch N, Liang Y, Yi X, He M, Somel M, *et al.* Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature*. 2014:**512**(7513):194–197. https://doi.org/10.1038/nature13408.

Hui R, D'Atanasio E, Cassidy LM, Scheib CL, Kivisild T. Evaluating genotype imputation pipeline for ultra-low coverage ancient genomes. *Sci Rep*. 2020:**10**(1):18542. https://doi.org/10.1038/s41598-020-75387-w.

Ishikawa A, Kusakabe M, Yoshida K, Ravinet M, Makino T, Toyoda A, Fujiyama A, Kitano J. Different contributions of local- and distant-regulatory changes to transcriptome divergence between stickleback ecotypes. *Evolution*. 2017:**71**(3):565–581. https://doi.org/10.1111/evo.13175.

Jia T, Munson B, Lango Allen H, Ideker T, Majithia AR. Thousands of missing variants in the UK Biobank are recoverable by genome realignment. *Ann Hum Genet*. 2020:**84**(3):214–220. https://doi.org/10.1111/ahg.12383.

Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody MC, White S, *et al.* The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*. 2012:**484**(7392):55–61. https://doi.org/10.1038/nature10944.

Jones MR, Mills LS, Alves PC, Callahan CM, Alves JM, Lafferty DJR, Jiggins FM, Jensen JD, Melo-Ferreira J, Good JM. Adaptive introgression underlies polymorphic seasonal camouflage in snowshoe hares. *Science*. 2018:**360**(6395):1355–1358. https://doi.org/10.1126/science.aar5273.

Jonsson JJ, Renieri A, Gallagher PG, Kashtan CE, Cherniske EM, Bruttini M, Piccini M, Vitelli F, Ballabio A, Pober BR. Alport syndrome, mental retardation, midface hypoplasia, and elliptocytosis: a new X linked contiguous gene deletion syndrome? *J Med Genet*. 1998:**35**(4):273–278. https://doi.org/10.1136/jmg.35.4.273.

Kahle D, Wickham H. Ggmap: spatial visualization with ggplot2. *R J* 2013:**5**(1):144–161. https://journal.r-project.org/archive/2013-1/kahle-wickham.pdf. https://doi.org/10.32614/RJ-2013-014.

Kardos M, Husby A, McFarlane SE, Qvarnström A, Ellegren H. Whole-genome resequencing of extreme phenotypes in collared flycatchers highlights the difficulty of detecting quantitative trait loci in natural populations. *Mol Ecol Resour*. 2016:**16**(3):727–741. https://doi.org/10.1111/1755-0998.12498.

Kautt AF, Kratochwil CF, Nater A, Machado-Schiaffino G, Olave M, Henning F, Torres-Dowdall J, Härer A, Hulsey CD, Franchini P, *et al.* Contrasting signatures of genomic divergence during sympatric speciation. *Nature*. 2020:**588**(7836):106–111. https://doi.org/10.1038/s41586-020-2845-0.

Kendig KI, Baheti S, Bockol MA, Drucker TM, Hart SN, Heldenbrand JR, Hernaez M, Hudson ME, Kalmbach MT, Klee EW, *et al.* Sentieon DNASeq variant calling workflow demonstrates strong computational performance and accuracy. *Front Genet*. 2019:**10**:736. https://doi.org/10.3389/fgene.2019.00736.

Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*. 2010:**26**(17):2204–2207. https://doi.org/10.1093/bioinformatics/btq351.

Kirch M, Romundset A, Gilbert MTP, Jones FC, Foote AD. Ancient and modern stickleback genomes reveal the demographic constraints on adaptation. *Curr Biol*. 2021:**31**(9):2027–2036.e8. https://doi.org/10.1016/j.cub.2021.02.027.

Kulkarni P, Frommolt P. Challenges in the setup of large-scale next-generation sequencing analysis workflows. *Comput Struct Biotechnol J*. 2017:**15**:471–477. https://doi.org/10.1016/j.csbj.2017.10.001.

Laine VN, Gossmann TI, Schachtschneider KM, Garroway CJ, Madsen O, Verhoeven KJF, de Jager V, Megens H-J, Warren WC, Minx P, *et al.* Evolutionary signals of selection on cognition from the great tit genome and methylome. *Nat Commun*. 2016:**7**(1):10474. https://doi.org/10.1038/ncomms10474.

Lamichhaney S, Berglund J, Almén MS, Maqbool K, Grabherr M, Martinez-Barrio A, Promerová M, Rubin C-J, Wang C, Zamani N, *et al.* Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature*. 2015:**518**(7539):371–375. https://doi.org/10.1038/nature14181.

Lamichhaney S, Fuentes-Pardo AP, Rafati N, Ryman N, McCracken GR, Bourne C, Singh R, Ruzzante DE, Andersson L. Parallel adaptive evolution of geographically distant herring populations on both sides of the North Atlantic Ocean. *Proc Natl Acad Sci U S A*. 2017:**114**(17):E3452–E3461. https://doi.org/10.1073/pnas.1617728114.

Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database Collaboration. The Sequence Read Archive. *Nucleic Acids Res* 2011:**39**(Database):D19–D21. https://doi.org/10.1093/nar/gkq1019.

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, *et al.*Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016:**536**(7616):285–291. https://doi.org/10.1038/nature19057.

Lewin HA, Richards S, Lieberman Aiden E, Allende ML, Archibald JM, Bálint M, Barker KB, Baumgartner B, Belov K, Bertorelle G, *et al.* The Earth BioGenome Project 2020: starting the clock. *Proc Natl Acad Sci U S A*. 2022:**119**(4):e2115635118. https://doi.org/10.1073/pnas.2115635118.

Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv q-bio*. 2013. [preprint] http://arxiv.org/abs/1303.3997.

Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009:**25**(14):1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. *Bioinformatics* 2009:**25**(16):2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

Li S, Li B, Cheng C, Xiong Z, Liu Q, Lai J, Carey HV, Zhang Q, Zheng H, Wei S, *et al.* Genomic signatures of near-extinction and rebirth of the crested ibis and other endangered bird species. *Genome Biol*. 2014:**15**(12):557. https://doi.org/10.1186/s13059-014-0557-1.

Li C, Olave M, Hou Y, Qin G, Schneider RF, Gao Z, Tu X, Wang X, Qi F, Nater A, *et al.* Genome sequences reveal global dispersal routes and suggest convergent genetic adaptations in seahorse evolution. *Nat Commun*. 2021:**12**(1):1094. https://doi.org/10.1038/s41467-021-21379-x.

Liu S, Hansen MM, Jacobsen MW. Region-wide and ecotype-specific differences in demographic histories of threespine stickleback populations, estimated from whole genome sequences. *Mol Ecol*. 2016:**25**(20):5187–5202. https://doi.org/10.1111/mec.13827.

Liu Y, Liu S, Zhang N, Chen D, Que P, Liu N, Höglund J, Zhang Z, Wang B. Genome assembly of the common pheasant *Phasianus colchicus*: a model for speciation and ecological genomics. *Genome Biol Evol*. 2019:**11**(12):3326–3331. https://doi.org/10.1093/gbe/evz249.

Lou RN, Jacobs A, Wilder AP, Therkildsen NO. A beginner's guide to low-coverage whole genome sequencing for population genomics. *Mol Ecol*. 2021:**30**(23):5966–5993. https://doi.org/10.1111/mec.16077.

Malinsky M, Challis RJ, Tyers AM, Schiffels S, Terai Y, Ngatunga BP, Miska EA, Durbin R, Genner MJ, Turner GF. Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake. *Science*. 2015:**350**(6267):1493–1498. https://doi.org/10.1126/science.aac9927.

Malinsky M, Svardal H, Tyers AM, Miska EA, Genner MJ, Turner GF, Durbin R. Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. *Nat Ecol Evol*. 2018:**2**(12):1940–1955. https://doi.org/10.1038/s41559-018-0717-x.

Mangul S, Martin LS, Langmead B, Sanchez-Galan JE, Toma I, Hormozdiari F, Pevzner P, Eskin E. How bioinformatics and open data can boost basic science in countries and universities with limited resources. *Nat Biotechnol*. 2019:**37**(3):324–326. https://doi.org/10.1038/s41587-019-0053-y.

Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen W-M. Robust relationship inference in genome-wide association studies. *Bioinformatics*. 2010:**26**(22):2867–2873. https://doi.org/10.1093/bioinformatics/btq559.

Marques DA, Taylor JS, Jones FC, Di Palma F, Kingsley DM, Reimchen TE. Convergent evolution of SWS2 opsin facilitates adaptive radiation of threespine stickleback into different light environments. *PLoS Biol*. 2017:**15**(4):e2001627. https://doi.org/10.1371/journal.pbio.2001627.

Mattingsdal M, Jorde PE, Knutsen H, Jentoft S, Stenseth NC, Sodeland M, Robalo JI, Hansen MM, André C, Blanco Gonzalez E. Demographic history has shaped the strongly differentiated corkwing wrasse populations in Northern Europe. *Mol Ecol*. 2020:**29**(1):160–171. https://doi.org/10.1111/mec.15310.

McDonald JH, Kreitman M. Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature*. 1991:**351**(6328):652–654. https://doi.org/10.1038/351652a0.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010:**20**(9):1297–1303. https://doi.org/10.1101/gr.107524.110.

Messer PW, Petrov DA. Frequent adaptation and the McDonald–Kreitman test. *Proc Natl Acad Sci U S A*. 2013:**110**(21):8615–8620. https://doi.org/10.1073/pnas.1220835110.

Miller SE, Roesti M, Schluter D. A single interacting species leads to widespread parallel evolution of the stickleback genome. *Curr Biol*. 2019:**29**(3):530–537.e6. https://doi.org/10.1016/j.cub.2018.12.044.

Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, Forster J, Lee S, Twardziok SO, Kanitz A, *et al.* Sustainable data analysis with Snakemake. *F1000Res*. 2021:**10**:33. https://doi.org/10.12688/f1000research.29032.2.

Mueller JC, Kuhl H, Boerno S, Tella JL, Carrete M, Kempenaers B. Evolution of genomic variation in the burrowing owl in response to recent colonization of urban areas. *Proc Biol Sci*. 2018:**285**(1878):20180206. https://doi.org/10.1098/rspb.2018.0206.

Nevado B, Ramos-Onsins SE, Perez-Enciso M. Resequencing studies of nonmodel organisms using closely related reference genomes: optimal experimental designs and bioinformatics approaches for population genomics. *Mol Ecol*. 2014:**23**(7):1764–1779. https://doi.org/10.1111/mec.12693.

Nielsen R, Akey JM, Jakobsson M, Pritchard JK, Tishkoff S, Willerslev E. Tracing the peopling of the world through genomics. *Nature*. 2017:**541**(7637):302–310. https://doi.org/10.1038/nature21347.

Paradis E, Schliep K. Ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*. 2019:**35**(3):526–528. https://doi.org/10.1093/bioinformatics/bty633.

Pedersen BS, Quinlan AR. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics*. 2018:**34**(5):867–868. https://doi.org/10.1093/bioinformatics/btx699.

Pockrandt C, Alzamel M, Iliopoulos CS, Reinert K. GenMap: ultra-fast computation of genome mappability. *Bioinformatics*. 2020:**36**(12):3687–3692. https://doi.org/10.1093/bioinformatics/btaa222.

Poelstra JW, Vijay N, Bossu CM, Lantz H, Ryll B, Müller I, Baglione V, Unneberg P, Wikelski M, Grabherr MG, *et al.* The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. *Science*. 2014:**344**(6190):1410–1414. https://doi.org/10.1126/science.1253226.

Purcell S, Chang C. PLINK v2.00a2.3. 2023. http://www.cog-genomics.org/plink/2.0/.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007:**81**(3):559–575. https://doi.org/10.1086/519795.

Qu Y, Tian S, Han N, Zhao H, Gao B, Fu J, Cheng Y, Song G, Ericson PGP, Zhang YE, *et al.* Genetic responses to seasonal variation in altitudinal stress: whole-genome resequencing of great tit in eastern Himalayas. *Sci Rep*. 2015:**5**(1):14256. https://doi.org/10.1038/srep14256.

Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010:**26**(6):841–842. https://doi.org/10.1093/bioinformatics/btq033.

Rand DM, Kann LM. Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from *Drosophila*, mice, and humans. *Mol Biol Evol*. 1996:**13**(6):735–748. https://doi.org/10.1093/oxfordjournals.molbev.a025634.

Raney BJ, Dreszer TR, Barber GP, Clawson H, Fujita PA, Wang T, Nguyen N, Paten B, Zweig AS, Karolchik D, *et al.* Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics*. 2014:**30**(7):1003–1005. https://doi.org/10.1093/bioinformatics/btt637.

Ravinet M, Elgvin TO, Trier C, Aliabadian M, Gavrilov A, Sætre G-P. Signatures of human-commensalism in the house sparrow genome. *Proc Biol Sci*. 2018:**285**(1884):20181246. https://doi.org/10.1098/rspb.2018.1246.

R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing; 2022. https://www.r-project.org/.

Regier AA, Farjoun Y, Larson DE, Krasheninina O, Kang HM, Howrigan DP, Chen B-J, Kher M, Banks E, Ames DC, *et al.* Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. *Nat Commun*. 2018:**9**(1):4038. https://doi.org/10.1038/s41467-018-06159-4.

Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W, Fungtammasan A, Kim J, *et al.* Towards complete and error-free genome assemblies of all vertebrate species. *Nature*. 2021:**592**(7856):737–746. https://doi.org/10.1038/s41586-021-03451-0.

Robinson JA, Brown C, Kim BY, Lohmueller KE, Wayne RK. Purging of strongly deleterious mutations explains long-term persistence and absence of inbreeding depression in island foxes. *Curr Biol*. 2018:**28**(21):3487–3494.e4. https://doi.org/10.1016/j.cub.2018.08.066.

Romiguier J, Gayral P, Ballenghien M, Bernard A, Cahais V, Chenuil A, Chiari Y, Dernat R, Duret L, Faivre N, *et al.* Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature*. 2014:**515**(7526):261–263. https://doi.org/10.1038/nature13685.

Runemark A, Trier CN, Eroukhmanoff F, Hermansen JS, Matschiner M, Ravinet M, Elgvin TO, Sætre G-P. Variation and constraints in hybrid genome formation. *Nat Ecol Evol*. 2018:**2**(3):549–556. https://doi.org/10.1038/s41559-017-0437-7.

Sayers EW, Beck J, Bolton EE, Bourexis D, Brister JR, Canese K, Comeau DC, Funk K, Kim S, Klimke W, *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 2021:**49**(D1):D10–D17. https://doi.org/10.1093/nar/gkaa892.

Schield DR, Scordato ESC, Smith CCR, Carter JK, Cherkaoui SI, Gombobaatar S, Hajib S, Hanane S, Hund AK, Koyama K, *et al.* Sex-linked genetic diversity and differentiation in a globally distributed avian species complex. *Mol Ecol*. 2021:**30**(10):2313–2332. https://doi.org/10.1111/mec.15885.

Sievert C. *Interactive web-based data visualization with R, plotly, and shiny*. New York: CRC Press; 2020.

Singhal S, Leffler EM, Sannareddy K, Turner I, Venn O, Hooper DM, Strand AI, Li Q, Raney B, Balakrishnan CN, *et al.* Stable recombination hotspots in birds. *Science*. 2015:**350**(6263):928–932. https://doi.org/10.1126/science.aad0843.

Smeds L, Mugal CF, Qvarnström A, Ellegren H. High-resolution mapping of crossover and non-crossover recombination events by whole-genome re-sequencing of an avian pedigree. *PLoS Genet*. 2016:**12**(5):e1006044. https://doi.org/10.1371/journal.pgen.1006044.

Spivakov M, Auer TO, Peravali R, Dunham I, Dolle D, Fujiyama A, Toyoda A, Aizu T, Minakuchi Y, Loosli F, *et al.* Genomic and phenotypic characterization of a wild medaka population: towards the establishment of an isogenic population genetic resource in fish. *G3 Genes|Genomes|Genetics*. 2014:**4**(3):433–445. https://doi.org/10.1534/g3.113.008722.

Stoletzki N, Eyre-Walker A. Estimation of the neutrality index. *Mol Biol Evol*. 2011:**28**(1):63–70. https://doi.org/10.1093/molbev/msq249.

Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*. 1989:**123**(3):585–595. https://doi.org/10.1093/genetics/123.3.585.

Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics*. 2015:**31**(12):2032–2034. https://doi.org/10.1093/bioinformatics/btv098.

Thiery AP, Shono T, Kurokawa D, Britz R, Johanson Z, Fraser GJ. Spatially restricted dental regeneration drives pufferfish beak development. *Proc Natl Acad Sci U S A*. 2017:**114**(22):E4425-E4434. https://doi.org/10.1073/pnas.1702909114.

Toczydlowski RH, Liggins L, Gaither MR, Anderson TJ, Barton RL, Berg JT, Beskid SG, Davis B, Delgado A, Farrell E, *et al.* Poor data stewardship will hinder global genetic diversity surveillance. *Proc Natl Acad Sci U S A*. 2021:**118**(34):e2107934118. https://doi.org/10.1073/pnas.2107934118.

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics*. 2013:**43**(1):11.10.1–11.10.33. https://doi.org/10.1002/0471250953.bi1110s43.

Verta J-P, Jones FC. Predominance of cis-regulatory changes in parallel expression divergence of sticklebacks. *Elife*. 2019:**8**:e43785. https://doi.org/10.7554/eLife.43785.

Vijay N, Bossu CM, Poelstra JW, Weissensteiner MH, Suh A, Kryukov AP, Wolf JBW. Evolution of heterogeneous genome differentiation across multiple contact zones in a crow species complex. *Nat Commun*. 2016:**7**(1):13195. https://doi.org/10.1038/ncomms13195.

Weber A-T, Rajkov J, Smailus K, Egger B, Salzburger W. Diversification dynamics and (non-)parallel evolution along an ecological gradient in African cichlid fishes. *bioRxiv*. 2021. [preprint] https://www.biorxiv.org/content/10.1101/2021.01.12.426414v3.

White MA, Kitano J, Peichel CL. Purifying selection maintains dosage-sensitive genes during degeneration of the threespine stickleback Y chromosome. *Mol Biol Evol*. 2015:**32**(8):1981–1995. https://doi.org/10.1093/molbev/msv078.

Wickham H. Reshaping data with the reshape package. *J Stat Softw*. 2007:**21**(12):1–20. https://doi.org/10.18637/jss.v021.i12.

Wickham H. *Ggplot2: elegant graphics for data analysis*. New York: Springer-Verlag; 2016.

Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, Grolemund G, Hayes A, Henry L, Hester J, *et al.* Welcome to the tidyverse. *J Open Source Softw*. 2019:**4**(43):1686. https://doi.org/10.21105/joss.01686.

Wooldridge TB, Kautt AF, Lassance J-M, McFadden S, Domingues VS, Mallarino R, Hoekstra HE. An enhancer of *Agouti* contributes to parallel evolution of cryptically colored beach mice. *Proc Natl Acad Sci U S A*. 2022:**119**(27):e2202862119. https://doi.org/10.1073/pnas.2202862119.

Wratten L, Wilm A, Göke J. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. *Nat Methods*. 2021:**18**(10):1161–1168. https://doi.org/10.1038/s41592-021-01254-9.

Wu Y, Zhang Y, Hou Z, Fan G, Pi J, Sun S, Chen J, Liu H, Du X, Shen J, *et al.* Population genomic data reveal genes related to important traits of quail. *Gigascience*. 2018:**7**(5):giy059. https://doi.org/10.1093/gigascience/giy049.

Yoshida K, Makino T, Yamaguchi K, Shigenobu S, Hasebe M, Kawata M, Kume M, Mori S, Peichel CL, Toyoda A, *et al.* Sex chromosome turnover contributes to genomic divergence between incipient stickleback species. *PLoS Genet*. 2014:**10**(3):e1004223. https://doi.org/10.1371/journal.pgen.1004223.

Yoshida K, Ravinet M, Makino T, Toyoda A, Kokita T, Mori S, Kitano J. Accumulation of deleterious mutations in landlocked threespine stickleback populations. *Genome Biol Evol*. 2020:**12**(4):479–492. https://doi.org/10.1093/gbe/evaa065.

Yu G, Lam TTY, Zhu H, Guan Y. Two methods for mapping and visualizing associated data on phylogeny using GGTree. *Mol Biol Evol*. 2018:**35**(12):3041–3043. https://doi.org/10.1093/molbev/msy194.

Zhan X, Pan S, Wang J, Dixon A, He J, Muller MG, Ni P, Hu L, Liu Y, Hou H, *et al.* Peregrine and saker falcon genome sequences provide insights into evolution of a predatory lifestyle. *Nat Genet*. 2013:**45**(5):563–566. https://doi.org/10.1038/ng.2588.

Zhang H, Hou J, Liu H, Zhu H, Xu G, Xu J. Adaptive evolution of low-salinity tolerance and hypoosmotic regulation in a euryhaline teleost, *Takifugu obscurus*. *Mar Biol*. 2020:**167**(7):90. https://doi.org/10.1007/s00227-020-03705-x.

Zhou Z, Li M, Cheng H, Fan W, Yuan Z, Gao Q, Xu Y, Guo Z, Zhang Y, Hu J, *et al.* An intercross population study reveals genes associated with body size and plumage color in ducks. *Nat Commun*. 2018:**9**(1):2648. https://doi.org/10.1038/s41467-018-04868-4.