Origin Matters: Using a Local Reference Genome Improves Measures in Population Genomics

Doko-Miles J. Thorburn[1,2+], Kostas Sagonas[1,3], Mahesh Binzer-Panchal[4], Frederic J.J. Chain[5], Philine G.D. Feulner[6,7], Erich Bornberg-Bauer[8], Thorsten BH Reusch[9], Irene E. Samonte-Padilla[10], Manfred Milinski[10], Tobias L. Lenz[11,12], Christophe Eizaguirre[1]

[1]School of Biological and Chemical Sciences, Queen Mary University of London, London, United Kingdom.

[2]Department of Life Sciences, Imperial College London, London, United Kingdom.

[3]Department of Zoology, School of Biology, Aristotle University of Thessaloniki, Panepistimioupoli, 54124 Thessaloniki, Greece.

[4]Department of Medical Biochemistry and Microbiology, National Bioinformatics Infrastructure Sweden (NBIS), Science for Life Laboratory, Uppsala University, Sweden.

[5]Department of Biological Sciences, University of Massachusetts Lowell, USA.

[6]Department of Fish Ecology and Evolution, Centre of Ecology, Evolution and Biogeochemistry, EAWAG Swiss Federal Institute of Aquatic Science and Technology, Kastanienbaum, Switzerland.

[7]Division of Aquatic Ecology and Evolution, Institute of Ecology and Evolution, University of Bern, Bern, Switzerland.

[8]Evolutionary Bioinformatics, Institute for Evolution and Biodiversity, Westfälische Wilhelms University, Münster, Germany.

[9]Marine Evolutionary Ecology, GEOMAR Helmholtz Centre for Ocean Research, Kiel, Germany.

[10]Department of Evolutionary Ecology, Max Planck Institute for Evolutionary Biology, Plön, Germany.

[11]Research Group for Evolutionary Immunogenomics, Max Planck Institute for Evolutionary Biology, Plön, Germany.

[12]Research Unit for Evolutionary Immunogenomics, Department of Biology, University of Hamburg, Hamburg, Germany.

[+]Corresponding author: d.m.j.thorburn@qmul.ac.uk

Abstract

Genome-level sequencing enables us to ask fundamental questions about the genetic basis of adaptation, population structure, and epigenetic mechanisms, but usually requires a suitable reference genome for mapping population-level re-sequencing data. In some model systems, multiple reference genomes are available, giving researchers the challenging task of determining which reference genome best suits their data. Here we compare the use of two different reference genomes for the three-spined stickleback (*Gasterosteus aculeatus*), one novel genome derived from a European gynogenetic individual and the published reference genome of a North American individual. Specifically, we investigate the impact of using a local reference versus one generated from a distinct lineage on several common population genomics analyses. Through mapping genome resequencing data of 60 sticklebacks from across Europe and North America, we demonstrate that genetic distance among samples and the reference impacts downstream analyses. Using a local reference genome increased mapping efficiency and genotyping accuracy, effectively retaining more and better data. Despite comparable distributions of the metrics generated across the genome using SNP data (i.e., π, Tajima's *D*, and F_{ST}), window-based statistics using different references resulted in different outlier genes and enriched gene functions. A marker-based analysis of DNA methylation distributions had a comparably high overlap in outlier genes and functions, yet with distinct differences depending on the reference genome. Overall, our results highlight how using a local reference genome decreases reference bias to increase confidence in downstream analyses of the data. Such results have significant implications in all reference-genome-based population genomic analyses.

Keywords: Reference genomes, population genomics, gynogenetic, genome assembly, read mapping, *Gasterosteus aculeatus*, stickleback, reference mapping bias

Introduction

Genome-level sequencing has revolutionized many biological fields including evolution, ecology, microbiology, and population genomics (M. R. Jones & Good, 2016; Kao et al., 2014; Stapley et al., 2010). Historically, scientists have relied on one or a small number of high-quality reference genomes to address their specific questions. Progressively, the availability of high-quality reference genomes from large-scale projects (e.g., Earth BioGenome and Vertebrate Genome Projects; Lewin et al., 2018; Scientists, 2009), the decreasing costs of sequencing, and the availability of curated variant databases (e.g., dbSNP and dbVar; Lappalainen et al., 2013; Sherry et al., 2001) have improved the breadth and depth of genomic research. However, limitations due to the availability and quality of these resources and the challenge of integrating and interpreting multiple sources of genetic variation still exist (Formenti et al., 2022). These problems are only exacerbated in non-model systems where a high-quality reference genome may not exist. Recently, a graph-based genome alignment methodology has been developed to index and incorporate variant databases, providing a practical and highly efficient method that better captures variation in a genome (Kim et al., 2019). However, databases of high-quality variants are not currently available for most study systems. Moreover, in highly variable regions graph-based approaches can incur significant computational overhead and have an increased change of false-positive alignments (Grytten et al., 2020; Pritt et al., 2018). Hence, until graph-based methods become common place, assembling a haploid reference is still a viable alternative to making and utilising a resource intensive genome graph.

Reference genomes are integral parts of the analytical frameworks of genomic research. Specifically, variants are identified and referred to by their loci in relation to their position mapped on a reference genome. Such an approach allows for direct comparisons among multiple individuals and organisms, therein enabling comprehensive research in various fields, including phylogenomics (Fang et al., 2018; Wang et al., 2013), comparative evolution (Palmer

& Kronforst, 2020; Parker et al., 2013), adaptive radiations (Ronco et al., 2020), and the genomics of domestication (Frantz et al., 2015).

Arguably one of the most important factors to consider when multiple reference genomes or assembly versions are available is their difference in quality. Whilst we are moving towards complete and error-free assemblies (Rhie et al., 2021), the continuing advances of methodologies can create significant differences among assemblies. For example, bioinformatic tools have been developed to resolve false gene duplications that stem from heterozygosity in homologous haplotypes (Guan et al., 2020; Roach et al., 2018). In some cases, such as in humans, reference genomes are continually updated alongside major advances, where choosing the most updated version will offer the most accurate analysis of human sequencing data (Guo et al., 2017; International Human Genome Sequencing Consortium, 2001, 2004; Pushkarev et al., 2009). However, when multiple similar quality and genetically diverse reference genomes are available from multiple populations or strains (e.g., Berner et al., 2019; Gan et al., 2011; Hirsch et al., 2016; Springer et al., 2018), genetic distance among samples and the reference may be an important factor to consider.

The detection of genomic polymorphisms is affected by the evolutionary time between the individuals being sequenced and the reference genome (Bohling, 2020; Prasad et al., 2022; B. N. Reid et al., 2021). This has implications on both the detection and the genotyping of single nucleotide polymorphisms (SNPs) and structural variants (SVs). For example, variant calling through pipelines such as GATK and freebayes use Bayesian inference to call genotypes (i.e., the likelihood of a genotype, given the data; Auwera et al., 2014; Garrison & Marth, 2012). When high differentiation between reference genome and the sampled individuals exists, a high proportion of segregating sites will emerge as fixed differences between the samples and the reference genome (i.e., homozygote non-reference in diploids), and therefore uncertain genotypes can have a higher likelihood of being called as homozygote non-reference, even if this is not the case. Analogously, most methods used to detect SVs (read-depth, split paired-read, breakpoints, and assembly) compare mapped reads to the reference genome (Pirooznia et al., 2015). Several studies on humans have demonstrated that ethnicity-specific reference genomes are beneficial (Ameur et al., 2017; Dewey et al., 2011; Fakhro et al., 2016; Lacaze et al., 2019). Specifically, the targeted reference genomes improved reliability of genetic and structural variation calls (Ameur et al., 2017; Fakhro et al., 2016). Moreover, recent studies have demonstrated that increasing phylogenetic distance between target species and reference genome decreases mapping efficiency and has strong effects on evolutionary inferences made from the data (Bohling, 2020; Prasad et al., 2022).

The three-spined stickleback (*Gasterosteus aculeatus*) is a supermodel in evolutionary biology (K. Reid et al., 2021), and research on this small teleost fish has pioneered discoveries related to the genomics of adaptation (Feulner et al., 2015; Haenel et al., 2019; F. C. Jones et al., 2012; Roesti et al., 2015), adaptive divergence (Feulner et al., 2015; Huang et al., 2016; Roesti et al., 2013) and development (Shapiro et al., 2004; Spitz et al., 2001). Genomic investigations using *G. aculeatus* have mostly relied on a single high-quality chromosome-level reference genome from an isolated population in Alaska (F. C. Jones et al., 2012), which has been updated with multiple improvements (Glazer et al., 2015; Nath et al., 2021; Peichel et al., 2017; Roesti et al., 2013). Recently, several additional *de novo G. aculeatus* contig-level genome assemblies have been made available, including European and marine derived assemblies (Berner et al., 2019). These additional genome assemblies may be appropriate given the wide geographic distribution of the species; there are Atlantic and Pacific clades of *G. aculeatus* that diverged an estimated 44.6 Kya with extensive phenotypic and genetic diversity across its range (Fang et al., 2018; McKinnon & Rundle, 2002).

In this study, we report the generation and *de novo* annotation of a European *G. aculeatus* genome assembly derived from a gynogen individual (Samonte-Padilla et al., 2011). The gynogenesis process in *G. aculeatus* produced a near complete homozygous diploid fish (Samonte-Padilla et al., 2011), which helps alleviate some of the genome assembly difficulties associated with heterozygosity. Then, using European and North American derived reference genomes, we investigated the effect of reference genome origin. We used high-quality genome-wide resequencing data from 60 *G. aculeatus* individuals from 5 recently diverged lake-river pairs distributed across the European (Atlantic clade; Fang et al., 2018) and North American (Western North America; Pacific clade) *G. aculeatus* ranges. The effect of using a local reference genome was investigated through comparing mapping efficiency (i.e., the percentage of reads that are successfully mapped to the reference genome), genotype calling, genome scans for F_{ST}, Tajima's *D*, and π, and identification of SVs. The resequenced genome data have been used to investigate the distribution of islands of differentiation across the genome (Feulner et al., 2015) and the role of copy-number variation in adaptation (Chain et al., 2014), offering baselines to evaluate specific metrics against a new local genome. We also compared how the different genomes affect DNA-methylation calling, focusing on an additional 50 European fish for which reduced-representation bisulfite sequencing (RRBS) was available (Sagonas et al., 2020). We hypothesise that the mapping and genotyping results will generally fall into 3 categories: (h_1) a local reference genome has a significant and putatively benficial effect, (h_2) one reference genome is better than the other, and (h_3) a local reference genome has an effect, but the effects are more substantial with local populations (i.e., a mixture of both h_1 and h_2). Our hypotheses can be observed in the outcome of statistical tests: h_1 presents as a significant interaction between reference origin and population origin, h_2 as a significant effect of reference origin, and h_3 presents similarly to h_1 with a significant interaction, but the post-hoc analysis indicates a clear bias towards one of the reference genomes.

Material & Methods

Reference Genome Assembly and Annotation

The induction of the diploid gynogen individual followed the protocol established for three-spined sticklebacks (Samonte-Padilla et al., 2011). In brief, fertilised stickleback eggs were mixed with UV-irradiated sperm (2 minutes of exposure) and then exposed to a heat-shock treatment of 34°C for 4 minutes, 5 minutes post-fertilisation. The treatment caused the genetic inactivation of the sperm, resulting in homozygote maternal offspring that lack paternal alleles (Samonte-Padilla et al., 2011). In order to increase the likelihood of embryo development, two siblings from the same family were used for this process. After 5 months post hatching, a fish was sacrificed. DNA was extracted using the Qiagen high molecular weight extraction kit following the manufacturer's protocol. Sequencing was then conducted at the Beijing Genomics Institute (BGI), taking place on a PacBio platform. In total, 2,805,993 reads were generated with a coverage of 44.1X. A total of 214,285 PacBio reads were discarded before further analysis given their short length. In addition, Illumina paired-end sequencing libraries in a HiSeq2500 platform were constructed with insert sizes of about 170 base pairs (bp), 500 bp and 800 bp.

Genome assembly was performed using Canu v1.6 assembler (Koren et al., 2017), followed by an internal polishing step using Quiver. To hybrid polish the PacBio assembly, a total of 316,797,342 high-quality Illumina reads were mapped to the contigs using BWA-MEM v0.7.15 (Li, 2013) and the alignment was then used for further polishing using Pilon v1.22 (Walker et al., 2014). Illumina raw reads were trimmed, and low quality and adaptor sequences were removed using Cutadapt v1.13 (Martin, 2011). To evaluate the PacBio *de novo* contig-level assembly and search for potential misassemblies, we used FRCbam v1.3.0 (Vezzi et al., 2012), whereas its completeness in terms of core orthologous genes was assessed using BUSCO v3.0.1 (Simão et al., 2015) and the 'actinopterygii' data set. The 32.1kb contig tig00001189_pilon mapped to the mitochondrial genome of the North American *G. aculeatus* (Peichel et al., 2017) and was trimmed and labelled the mitochondrial genome. The sequence was trimmed to only include the 15.8kb that aligned to the North American mitochondrial genome given the size of the mitochondrial genomes are moderately conserved even across long phylogenetic distances (Gissi et al., 2008). The excess 16.3kb of contig tig00001189_pilon were labelled and kept in the assembly, making up two additional contigs.

To scaffold the European *G. aculeatus* contig-level gynogen assembly into pseudo-chromosomes we used Chromosemble from the Satsuma2 package (Grabherr et al., 2010).

Here, we ordered and oriented the contigs based on synteny with the North American *G. aculeatus* chromosome-level genome (Peichel et al., 2017), excluding the unmapped scaffolds. We tested for the effect of contig size on alignment rates, and only retained contigs with an alignment rate of 70% or above for further assembly. Gynogen contigs not scaffolded onto a pseudo-chromosome were concatenated in size order into an unmapped scaffold for population genomic analyses, separating each contig by 1,000 N's. Synteny between the European and North American *G. aculeatus* assemblies was calculated using the SatsumaSynteny2 function (Grabherr et al., 2010), and plotted using the Circos v0.69 visualisation tool (Krzywinski et al., 2009). It should be noted that the reference genomes have been created using distinctly different methodologies. As such, significant reference origin by population origin interactions that do not clearly exhibit a reciprocal effect in the test statistic likely include a genome quality effect.

Repetitive sequences were identified *de novo* in the European pseudo-chromosome assembly using Repeat Modeler v.2.0.1 while Repeat Masker v.4.1.1 (Smit et al., 2015) was used to mask the genome using the three-spined stickleback and zebrafish libraries in two separately rounds. The results of each round were then analyzed together, complex repeats were separated, to produce the final repeat annotation. Genome annotation was performed on the European repeat-masked pseudo-chromosome genome assembly using MAKER2 v.2.31.9 (Holt & Yandell, 2011). Subsequent genome annotation was performed following a two-round approach. For the first round, the repeat annotation data (release 95) as well as *G. aculeatus* transcriptome and protein sequences from ENSEMBL and UniProt/SwissProt databases were used as evidence sets for the prediction of gene models, while *est2genome* and *protein2genome* setting were set as 1. For the second round, SNAP (Korf, 2004), with ADE of 0.25 and length of 50, and AUGUSTUS v.3.2.3 with default values (Stanke et al., 2008) were trained on the gene model predicted from the first round. Functional annotation was performed using BlastP against UniProt proteins with an E-value threshold of 1e-5, and InterProScan v.5.4-47 (P. Jones et al., 2014) was used for domain annotation. The resulting gene models were filtered to retain those with AED value of 0.5 or less, having PFAM annotations and significant hits to known proteins against UniProt DB (E-value 1e-5).

We identified orthologous and paralogous gene families among the North American and European *G. aculeatus* reference genomes using OrthoFinder v2.4.1 with the default parameters (Emms & Kelly, 2019). Protein sequences were extracted using the getfasta function in the BEDtools toolset v2.26.0, using the *-split* parameter to only include exonic regions (Quinlan & Hall, 2010). Where applicable, downstream analyses were restricted to only include the 7,529 1-to-1 orthologues identified between the two assemblies to remove any biases stemming from differences in gene number or functional annotation.

SNP Data Collection and Processing

Whole genome resequencing data from a total of 60 *G. aculeatus* individuals were used from 5 recently diverged freshwater lake (_L) and river (_R) population pairs (table S1; details on sampling, library preparation, sequencing, and original data processing up to adapter trimming can be found in; Feulner et al., 2015). The five population pairs were sampled from two sites in Germany (G1 and G2; European; Atlantic clade), and one site in Norway (No; European; Atlantic clade), the United States (US; North American; Pacific clade), and Canada (Ca; North American; Pacific clade).

For the genome scan using SNP data, raw data was processed following previous procedures (Feulner et al., 2013). Adapter-cleaned reads were trimmed using Trimmomatic v0.36 (Bolger et al., 2014) in paired-end mode, trimming read tails with a PHRED quality score below 20 and trimming to a maximum of 50bp. Using BWA-MEM v0.7.17 (Li, 2013), reads were independently mapped to both the anchored European and US reference genomes (Peichel et al., 2017). Mapping efficiency was calculated using Bamtools v2.4.1 (Barnett et al., 2011). All downstream processing of mapped reads and methodology for variant calling were identical for both reference genomes. Mapped reads were processed using Picard toolkit v2.18.7 (https://broadinstitute.github.io/picard/), applying FixMateInformation and CleanSam.

All reads belonging to the same individual from different lanes were combined using MergeSamFile, and then duplicate reads were flagged using MarkDuplicates. Variant calls were performed using GATK v4.0.6.1 (McKenna et al., 2010), calling variants in all genomes simultaneously, split by chromosome. The final set of SNPs was produced using hard filtering following the best practise workflow (see supplementary methods for filtering thresholds; (Depristo et al., 2011). Genome mapping and variant calls were conducted on the QMUL Apocrita High Performance Computing Cluster (King et al., 2017).

The gynogenesis process employed here aims to purge heterozygous variation, enabling us to reconstruct complex genomic regions. To assess the proportion of the genome that remains heterozygotic we used the same GATK variant calling and filtering pipeline detailed above with the paired-end Illumina libraries used to polish the gynogen assembly.

Phylogenetic Analyses

In order to assess the phylogenetic relationship between the reference genomes and the 60 resequenced genomes, we compared maximum-likelihood phylogenies based on each variant call with its respective reference genome. Maximum-likelihood phylogenies for both SNP calls were inferred using RAxML v8.2.11 (Stamatakis, 2014). We randomly sampled 1% of all segregating sites by setting the "—select-random-fraction" parameter to 0.01 in the GATK function SelectVariants. The resulting VCFs were converted to PHYLIP format (Felsenstein, 1989) using the vcf2phylip.py script (doi:10.5281/zenodo.1257058). Trees were constructed using the GTRGAMMA model with 1000 bootstraps. Phylogenetic trees were plotted using the R package *ggtree* package v2.2.4 (Yu et al., 2017).

Estimations of Genotype Bias

Calling of genotypes (i.e., heterozygote versus homozygote) may be affected by reference genome origin. Measures of genome-wide zygosity across all segregating sites were performed by parsing the genotype field in the VCF file using a custom R script. Genotypes were grouped into 5 distinct categories: homozygous reference, homozygous non-reference, heterozygous reference/non-reference, heterozygous non-reference/non-reference, and missing. This process was repeated to estimate genotypes for each individual mapped to both reference genomes.

Genome Scan Using SNP Data

Genome scans were performed in R v4.0.2 (R Development Core Team, 2019), and population genetics indices were calculated using the R package *PopGenome* v2.7.5 (Pfeifer et al., 2014). We measured Tajima's *D* and π in each population, and F_{ST} in each parapatric population pair. All metrics were calculated in non-overlapping 20kb windows across all 20 autosomal chromosomes and the unmapped scaffold. To obtain outlier genomic windows, we extracted the top 1% of the empirical distributions for each metric and population (or population pair for F_{ST}). This conservative criterion (e.g., Feulner et al., 2015; Lai et al., 2019; Stern & Lee, 2020) was chosen to increase our confidence in defining outliers. Finally, we defined outlier genes as any gene overlapping one or more outlier genomic windows from the SNP-based genome scans using the *foverlaps* function in the R package *data.table* v1.9.6 (Dowle et al., 2015).

Detection of Structural Variants

To investigate the impact of reference genome origin on the detection of structural variants (SVs), we used two independent SV callers, DELLY2 v0.8.3 (Rausch et al., 2012) and LUMPY v0.3.0 (Layer et al., 2014). Both LUMPY and DELLY were run using the default parameters. The LUMPY call was genotyped using SVTyper v0.7.1 (Chiang et al., 2015), and all SVs marked with the "LowQual" DELLY flag were removed. To ensure SVs found by both programs were the same, only SVs with an overlap of at least 50% were accepted and merged into a cohort level VCF file using SURVIVOR v1.0.3 (Jeffares et al., 2017). For downstream analyses, we included autosomal duplications, deletions, and inversions with a length of between 1kb and 1Mb supported by at least 6 split or discordant reads. All samples that were homozygote non-reference or heterozygote across all samples were analysed separately as these variants likely arise from the reference genome assemblies. Genes with coordinates entirely nested within SVs were defined as structurally variable genes (SVG) using the *foverlaps* function in the R package *data.table* v1.9.6.

Methylation Data Processing and Genome Scan

For the DNA methylation analysis, we used 50 methylomes of laboratory full-sib families of European *G. aculeatus* obtained from the study of Sagonas et al. (2020), were we investigated whether parasite infection alters genome-wide patterns and levels of DNA methylations. For each fish, a single-end library of 100 bp reads with an average of 11.5 million reads was produced. The raw data were quality checked using FASTQC v0.11.5 (Andrews, 2010). Cutadapt v1.9.1 (Martin, 2011) was used to trim and filter low-quality bases (-q 20), remove trimmed reads shorter than 10 bases and remove adapters using multiple adapter sequences (ATCGGAAGAGCACAC, AGATCGGAAGAGCACAC and NNAGATCGGAAGAGCACAC) with a minimum overlap of 1 bp between adapter and read. Trimmed reads were independently mapped against the European and US reference genomes (Peichel et al., 2017) to extract methylated cytosines, using Bismark v0.22.1 (Krueger & Andrews, 2011) with the Bowtie2 v2.3.2 aligner, allowing up to 2 mismatches. Similar to the SNP data, Bamtools v2.4.1 (Barnett et al., 2011) was used to calculate the mapping efficiency.

Cytosine methylation ratios in CpG sites was estimated for each fish and differentially methylated sites (DMS) were calculated between the two treatment groups (no parasite exposed or exposed to the nematode parasite *Camallanus lacustris*) respectively (for more information, see Sagonas et al., 2020) using the R package *MethylKit* v1.14.2. CpG methylation ratios were estimated by calculating the number of reads mapping to a given position carrying a cytosine divided by those reads carrying either C or T. CpG sites with coverage below 10X and sites that have more than 99.9th percentile of coverage were discarded in each sample from downstream analyses. We selected DMS between treatments using the following criteria: change in fractional methylation larger than 15%; q-values lower than 0.01 (SLIM method), and presence of methylated Cs in at least 50% of the samples within a treatment group. Differentially methylated genes were identified using the *genomation* R package v.1.1.0 (Akalin et al., 2015); a gene was considered differentially methylated if at least one DMSs was located no further than 1.5-kb upstream and 500 bases downstream of it.

GO Enrichment Analysis

To examine how the origin of the reference genome impacts functional enrichment of outlier genes, structurally variable genes, and differentially methylated genes, we performed gene ontology (GO) enrichment analyses. GO enrichment analyses were performed using the g:GOSt function in the g:Profiler v0.1.9 package (Raudvere et al., 2019). Outlier genes were grouped by all combinations of reference genome origin versus population of origin, resulting in 4 distinct combinations: Europe-Europe, Europe-North America, North America-Europe, and North America-North America. The lists of outlier genes or SVGs identified using the European assembly were assigned the orthologous North American gene identifiers for GO enrichment. *P*-values were corrected for multiple testing using FDR.

Statistical Analyses

Linear mixed effect models in the *lme4* package (Bates et al., 2015) were used to analyse how the origin of the reference genomes affected mapping efficiency, detection of genome-wise zygosity, and population-genetic indices (i.e., Tajima's *D*, π, and F_{ST}). For mapping efficiency, the proportion of mapped reads that were singletons or duplications were independently used as response variables. An interaction between reference genome origin (i.e., North America or Europe) and population origin (i.e., North America or Europe) were assigned as fixed effects, and population ID was set as a random factor. The same approach was used for all analyses, independently replacing the response variable with the zygosity categories or population-genetic indices, retaining the same fixed and random effects. *P*-values were inferred using Satterwaite's degree of freedom method in the *lmerTest* package (Kuznetsova et al., 2017). Tukey HSD post-hoc tests were performed using *emmeans* (Lenth et al., 2020).

Results

Reference Genome Assembly

The contig-level European gynogen assembly is 458.4Mb, making it 1.0% smaller than the North American reference genome (463.0Mb; Peichel et al., 2017). We achieved an N50 of 0.746Mb, comprised of 1,906 contigs (table 1). Guided by synteny, we conservatively placed 419.3Mb (91.5%) into 21 chromosomes (fig. 1), forming a pseudochromosome level assembly allowing for a more accurate comparison of the impact of the origin of reference genomes on population genomic metrics. A combination of gene evidence from the *G. aculeatus* North American reference genome (available cDNA and protein sequences) and *ab initio* gene predictions resulted in 22,739 genes annotated and a BUSCO completeness score of 95.2%. A total of 18,255 (90.2%) genes in the European gynogen assembly were orthologous to genes in the North American assembly. Additionally, we confirmed the European and North American derived reference genomes are nested within the phylogeny of populations sampled from the same geographic regions (supplementary fig. S1).

Next, we called SNPs in the paired-end libraries used to polish the European genome assembly to assess the remaining heterozygosity in the gynogen genome assembly, which is expected to be homozygous after the gynogenesis procedure. In total we identified 299,931 SNPs (average genome-wide read depth of 60.6) in the genome: 3,118 SNPs were homozygote non-reference, and 296,813 were heterozygote. Hence, after two generations of gynogenesis only 0.07% of the genome remains polymorphic, compared to 0.53% ± 0.01% (mean ± SE) across all samples mapped to both reference assemblies (supplementary table S1). These SNPs likely stem from paralogous sequences that were not identified by the genome assembler or from errors in the SNP calling process.

Mapping Efficiency

To assess the impact of reference genome origin on downstream analyses, we started by mapping whole genome resequencing reads from 60 individuals to the reference genomes. In total, we identified 10,672,162 SNPs using the North American reference genome, and 10,757,204 SNPs using the European reference genome (supplementary table S1). Overall, we achieved an average genome-wide depth of coverage of 20.40 (range; 11.09 - 35.46) using the North American reference and 20.99 (11.30 - 36.14; supplementary table S1) using the European reference. Reads mapped more efficiently (i.e., a higher proportion of reads mapped) to the European reference genome regardless of the population of origin with total reads mapping 2.31% ± 0.16% (European; estimate ± SE) and 1.13% ± 0.20% (North American) more efficiently, albeit with a greater increase in efficiency for local samples (LMER; Reference Origin:Population Origin, F_{108}=21.38, P<0.001; fig. 2A and supplementary table S2). This same result was also observed when only considering properly matched paired-end reads (i.e., paired reads where both reads map in the correct orientation; 0x2 flag), where reads mapped more efficiently to the European reference, but there was a greater increase in efficiency for local populations (European, 3.14% ± 0.22%; North American, 1.36% ± 0.27%; Reference Origin:Population Origin, F_{108}=25.94, P<0.001; fig. 2A). There were also 0.83% ± 0.08% (European) and 0.24% ± 0.09% (North American) fewer singletons (i.e., where only one of the paired-end reads mapped) when mapping European or North American populations to the European reference (LMER; Reference Origin:Population Origin, F_{108}=25.10, P<0.001; fig. 2B). Reference genome or population origin had no effect on mapping of duplicate reads (LMER; Reference Origin:Population Origin, F_{108}=0.01, P=0.942; Reference Origin, F_{108}=0.12, P=0.729; Population Origin, F_8=3.76, P=0.088; fig. 2B). Overall, using the European derived genome assembly noticeably improved mapping efficiency irrespective of the sample origins, although the effects were noticeably more efficient for European populations. Such results are consistent with hypothesis h_3, which describes that a local reference genome is beneficial, but the effects are generally more substantial when using the European reference genome.

Estimations of Genotype Performance

When using a local reference genome, there were 6.24% ± 0.38% (European; estimate ± SE) and 0.47% ± 0.47% (North American) fewer missing genotypes (LMER; Reference

Origin:Population Origin, F_{428}=123.97, P<0.001; fig. 3 and supplementary table S2). The same was true for the detection of heterozygote genotypes of both classes (i.e., heterozygote reference/non-reference and non-reference/non-reference), where a local reference genome decreases the number of calls by 0.13% ± 0.09% (European) and 0.31% ± 0.11% (North American) for heterozygote reference/non-reference, and 0.10% ± 0.003% (European) and 0.05% ± 0.004% (North American) for heterozygote non-reference/non-reference genotypes (LMER; all models report the Reference Origin:Population Origin interaction; reference/non-reference, F_{428}=9.86, P=0.002; non-reference/non-reference, F_{428}=872.85, P<0.001; fig. 3). Conversely, we identified approximately 4 times higher proportions of homozygote reference genotypes when using a local reference genome for European populations (12.19% ± 0.37%) in comparison to the North American populations (3.36% ± 0.35%; LMER; Reference Origin:Population Origin , F_{428}=691.30, P<0.001; fig. 3). Finally, approximately a two-fold decrease in the proportion of homozygote non-reference variants were identified when using a local reference genome for European populations (-5.72% ± 0.12%) over the North American populations (-2.53% ± 0.14%; LMER; Reference Origin:Population Origin , F_{428}=2012.29, P<0.001; fig. 3). Except for duplications which were not significantly different among calls, using a local reference genome offers a benefit in the detection of every class of genotype. Overall, these results are consistent with hypothesis h_1 which posited that a local reference genome would offer a benefit and would manifest as significant interactions.

Genome Scan Using SNP Data

The genome-wide distributions of metrics commonly used in population genomics were affected by the origin of the genome used (fig. 4 & S2-3). For a genome scan investigating patterns of differentiation, F_{ST} was 0.019 ± 8.4×10^{-4} (mean ± SE) higher for European populations and 0.005 ± 1.0×10^{-5} higher for North American populations when using a local reference genome (LMER; Reference Origin:Population Origin , F_{220053}=324.60, P<0.001; supplementary table S2). A similar pattern was observed using T_D, whereby T_D was 0.026 ± 2.95×10^{-3} (European) and 5.90×10^{-3} ± 3.61×10^{-3} (North American) higher when using a local reference genome (LMER; Reference Origin:Population Origin , F_{406709}=63.97, P<0.001). In addition, using a local reference genome led to significantly lower values of nucleotide diversity being detected (LMER; Reference Origin:Population Origin , F_{407698}=123.20, P<0.001), with low estimates for both European ($-3.28\times10^{-5} \pm 5.39\times10^{-6}$) and North American ($-9.99\times10^{-5} \pm 6.60\times10^{-6}$) populations. These results are consistent with hypothesis *h_1,* where a local reference genome has a significant impact on analyses.

Through sampling each metric individually and taking the top ~1% of the F_{ST}, π, or T_D distributions we generated multiple lists of outlier genes (table 2; supplementary table S3). Notably, the distributions of outlier genes across the genome were not significantly different when using either reference genome (two-sample Kolmogorov–Smirnov test; F_{ST}, D=0.140, P=0.988; π, D=0.121, P=0.998; T_D, D=0.157, P=0.962; fig. 4 & S2-3). Overall, higher number of outlier genes, including the subset of outlier 1-to-1 orthologous genes, were identified when using a local reference genome (table 2). The difference in genes among calls with different reference genomes putatively translated into no overlapping significantly enriched GO terms for the F_{ST}, π, and T_D analyses if any enrichment was detected (table 2). Complete GO enrichment tables are reported in supplementary table S4.

Genomic Structural Variants

We next addressed the question whether using a local reference genome affected the detection of structural variants. Firstly, we generated multiple distributions of SVs organised by the combination of reference genome and population origin (table 3). Overall, no fixed SVs were observed in all samples. The distribution of deletions significantly differed among SV calls (two-sample Kolmogorov–Smirnov test, D=0.278, P=0.008; fig. S4), whereby fewer SVs were detected when using the North American reference genome. The distribution of duplications and inversions did not significantly differ among SV calls with either reference genome (two-sample Kolmogorov–Smirnov test; duplications, D=0.109, P=0.890; inversions, D=0.140, P=0.754; fig. S4). Additionally, we identified less deletions when using a local reference genome, more deletions were identified overall when using the European reference genome (LMER; Reference Origin:Population Origin , F_{108}=61.813, P<0.001; table 3, supplementary table S5-S6). Analogously, we identified significantly fewer inversions when using a local reference genome, but more inversions were identified overall when using the North American reference genome (LMER; Reference Origin:Population Origin , F_{108}=34.52, P<0.001; table 3). Finally, a similar number of duplications were identified in European populations irrespective of reference origin, whereas fewer duplications were identified in North American populations when using a local reference genome (LMER, Reference Origin:Population Origin, F_{108}=3.99, P=0.048; supplementary table S5-S6). Overall, the results of the SV analysis are in line with hypothesis h_3 where the effects of a local reference genome are present, but unequal effects indicate one reference is better than the other.

Next, we investigated the role of a local reference genome on the detection of genes entirely nested within SVs, which we defined as structurally variable genes (SVG; table 3). This analysis was limited to 1-to-1 orthologs to ensure there was no bias arising from copy number variation among reference genome annotations. Here, we identified significantly fewer SVG-deletions for European populations using a local reference genome, however there was no significant difference in SVG-deletions in the North American populations when using either reference (LMER, Reference Origin:Population Origin , F_{108}=25.94, P<0.001; supplementary table S5-S6). On the other hand, we identified significantly more SVG-duplications when using the North American reference, regardless of the origin of the population (LMER, Reference Origin:Population Origin , F_{108}=9.40, P=0.003; supplementary table S5-S6). Finally, significantly fewer inversions were detected when using a local reference genome (LMER; Reference Origin:Population Origin , F_{108}=62.55, P<0.001; supplementary table S5-S6).

To identify whether a local reference genome correlated with the detection of functional enrichment in SVs, we investigated the SVGs GO enrichment. Similar to the overall SVG distribution analysis, this analysis was restricted to 1-to-1 orthologs. We identified no overlap in functional enrichment in the majority of the comparisons (table 3). The one exception was SVG-deletions in the North American populations, where 3 out of 13 significantly enriched terms (signaling receptor regulator activity, signaling receptor activator activity, and receptor ligand activity) were identified in both calls. Overall, the detection of SVs and SVGs were affected in different ways by the origin of the reference genome.

DNA Methylation Analysis

Finally, we conducted a DNA methylation analysis after mapping bisulfite sequencing reads to the two reference genomes. The inclusion of this analysis permits us to investigate the effect of reference genome origin on both DNA methylation analyses and on marker-based analyses, as opposed to the window-based genome scans reported above. The alignment of reads to the references showed significantly higher efficiency when using a local European reference genome (72.5%) compared to the North American reference (68.4%), resulting in an increase of 6% (paired t-test; t=-44.44, df=49, P<0.001). The more efficient mapping to a local reference produced a significantly higher calling of cytosine bases (6% increase, paired t-test; t=-28.31, df=49, P<0.001) and methylated Cs (8.2% increase, paired t-test; t=-33.18, df=49, P<0.001). Similarly, the comparison of the number of methylated sites per fish after filtering for low coverage sites revealed that using the local European reference genome resulted in identifying more methylated sites (560629.08 ± 12167.89) than with the North America reference (510240.14 ± 11037.78, paired t-test; t=-44.20, df=49, P<0.001). Additionally, the number of Differentially Methylated Sites (DMS) was higher when using a local European reference genome (N=2550 DMS) in comparison to the North American reference (N=2404 DMS). DMS overlapped with 711 and 712 genes in the European and North American assemblies, respectively, and specifically 298 and 299 genes were 1-to-1 orthologs between the two assemblies. A total of 204 of those genes with DMS (68%) were shared in both genomes. There were three significantly enriched GO terms (protein binding, calcium ion binding, and binding) shared in both analyses, and two enriched GO terms (cation binding and metal ion binding) only identified when using a divergent reference genome (Table 2).

Discussion

Until genome-graphs are more widely available, individual reference genomes remain an integral part of population genomic analyses. However, the reference genome can introduce mapping biases that significantly influence downstream analyses and inferences (Bohling, 2020; Prasad et al., 2022; Valiente-Mullor et al., 2021). Meanwhile, the effects of using a local or more differentiated reference genome remains understudied for ecological and evolutionary model species. To address this knowledge gap, we generated a *de novo* annotated synteny-guided assembly of a European *Gasterosteus aculeatus* fish. Using this novel genome and the established North American reference genome to map sequence data of samples from different populations in Europe and North America, we confirm the reference genome origin significantly impacts downstream analyses. Most notably, a local reference genome increased mapping and genotyping performance. Specifically, mapping efficiency was significantly better using the European reference genome, but the increase in performance was greater for local samples. When using a local reference, more genomic sites were genotyped, and genome window-based estimates of T_D and F_{ST} increased whilst π slightly decreased. Similarly, structural variants (SVs) analysis gave slightly different results based on the reference genome used. Consequently, most GO analyses resulted in only a minor proportion of matching enriched GO functions when using different reference genomes. In contrast to the window-based methods, the marker-based DNA methylation analysis pipeline was relatively less affected by reference genome origin, but still about one third of differentially methylated genes was uniquely identified by one but not both references.

Genome Assembly of a Gynogenetic Individual

Recent tools and techniques have increased the efficacy of reference genome assembly, such as utilising long and short read sequencing (Rhie et al., 2021), scaffolding with Hi-C (Peichel et al., 2017), optical or linkage mapping (Glazer et al., 2015), among the growing list of novel and effective techniques (Rhie et al., 2021). The generation of gynogenetic individuals purges genome-wide variation simplifying assembly of genomes (Christensen et al., 2018; Samonte-

Padilla et al., 2011). Here, after applying a previously established protocol for gynogenesis, we putatively removed more than 99.9% of genome-wide variation, likely aiding in scaffolding by long-read sequencing. The contiguity of the resulting contig-level gynogen *G. aculeatus* assembly is comparable to the top few assemblies published by the Fish10k project (Fan et al., 2020) in terms of number of contigs and contig N50. Notably, there are a few altered placements of contigs into chromosomes among assemblies. The largest was the 2.07Mb contig tig00002041_pilon, which aligns to both the middle of chrVIII (9.25Mb-10.87Mb) and the end of chrIX (17.89Mb-18.29Mb) and was placed in Gy_chrIX by Chromosemble. The Atlantic and Pacific *G. aculeatus* clades diverged an estimated 44.6Kya (Fang et al., 2018) leaving substantial time for large genomic rearrangements to occur. As such the correct placement of tig00002041_pilon is similarly likely in either chromosome which illustrates the need for further investigations to resolve the differences among assemblies. Overall, the new European gynogen reference genome is high quality, enabling us to test the effects of reference genome origin on downstream population genomic analyses.

Mapping and Calling Variants

Despite continuing advances in tools to assemble reference genomes and map sequenced reads, difficulty remains in correctly mapping reads to complex genomic regions enriched in heterozygosity, structural variation, or repetitive elements (Kajitani et al., 2014; Treangen & Salzberg, 2012). By using a reference genome with a longer evolutionary time to the most recent common ancestor (MRCA), complex variants have time to accumulate, likely decreasing mapping efficiency to genomic regions with arguably some of the most sought-after features (e.g., polymorphic regions associated with rapid evolutionary changes and adaptations). Here, we show that using a European stickleback reference genome that has a lower time to the MRCA with European populations increases mapping efficiency and decreases missing data. Conversely, the North American populations also showed increased efficiency when mapped to the European reference, but the difference was noticeably smaller than for European populations. Such a result indicates the removal of heterozygosity through gynogenesis (Samonte-Padilla et al., 2011) and the putative resolution of complex genomic regions in the European assembly improved mapping efficiency. These results are concordant with ethnicity-specific reference genome studies, which have demonstrated that local reference genomes increase depth of coverage resulting in increased sensitivity in variant calling (Ameur et al., 2017; Dewey et al., 2011; Fakhro et al., 2016).

Genome Scans

The effects of improved mapping efficiency and a decrease in missing data were observed to have significant impacts on the estimation of important population genomic metrics. Here, genome-wide estimates of F_{ST} and T_D were higher when using a local reference genome, whereas the opposite pattern was true for estimates of nucleotide diversity (π). Despite the differences in genome-wide estimates, the genome-wide distribution of outlier windows of π, F_{ST}, and T_D did not significantly differ across the genome for the same resequenced populations mapped to different reference genomes. Crucially, however, the detection of outlier genes was strongly impacted when using different reference genomes, even when conservatively limiting the analysis to 1-to-1 orthologs identified among the references. Specifically, genes overlapping π outliers had a low proportion of matching orthologs in the same populations when mapped to different reference genomes. Genes overlapping outliers from scans for F_{ST} and T_D were more consistent, but still revealed a large number of differences. In total, only 5.45% of the orthologs among assemblies were mapped to different chromosomes, clearly affecting the detection of outlier genes but only explaining a small proportion of differences observed. The difference in gene placement may instead highlight numerous resolved differences among assemblies that have accumulated since the divergence of Atlantic and Pacific *G. aculeatus* clades 44.6 Kya (Fang et al., 2018). For example, a small deletion in one reference assembly and not the other can result in a gene overlapping an outlier genomic window in only one scan. Overall improved mapping may help specific population genomic analyses including genome-wide representation sequencing where population structure may end up obscuring some SNPs present in only one or few populations (e.g., Baltazar-Soares et al., 2020).

Structural Variant Detection

Similar to the effect of reference genome origin on SNP-based scans, the detection of SVs appears to be affected by both reference genome origin and the assembly regardless of origin. Firstly, fewer deletions and inversions were identified when using a local reference genome. This result follows expectations, as there is less time between MRCA for SVs to build up between the sampled population and the reference genome. Secondly, more deletions and duplications and fewer inversions were detected when using the European reference, irrespective of population origin, suggesting differences in the genome assembly plays a role in the detection of SVs. However, the differences in SV detection did not translate into any significant differences in the number of genes overlapping deletions. Overall, our results suggest the detection of SVs may be affected by a combination of mapping efficiency, time to MCRA, and the methods used to assemble a reference genome.

Methylation

The most consistent analysis in terms of overlap among reference genomes was the marker-based DNA methylation test. Firstly, using a local reference genome significantly increased mapping efficiency, which resulted in more methylated sites and DMS being detected. The number of genes overlapping the DMS using either reference genomes was the most consistent among all analyses, with one fewer gene (0.14% of total genes) and ortholog (0.34% of total 1-to-1 orthologs) being identified when using a local reference. The DMS analyses recovered a relatively high proportion of overlapping outlier orthologs among reference genomes (68% of 1-to-1 orthologs), but still revealed an effect of the choice of reference. The higher proportion of overlap compared to window-based analyses may be due to the marker-based approach, which is less sensitive to genomic translocation, genome evolution, or assembly errors.

GO Enrichment

The largest effect of reference genome origin was in the GO enrichment analyses of outlier genes from genome scans, with only a minor proportion of enriched GO terms overlapping when using different reference genomes. Given the overall small proportion of overlapping outlier genes, these results were to be expected. GO enrichment analyses are particularly sensitive to minor changes in the number of genes or annotations in lists of genes (Gaudet & Dessimoz, 2017). It should be noted, however, that to allow for a direct comparison of the effects of the different reference genomes, we focused on 1-to-1 orthologs. GO enrichment analyses are common in population genomics (e.g., Chain et al., 2014; Feulner et al., 2015; Reimegård et al., 2017; Liu et al., 2018), and our results highlight reference genome origin strongly impacts such inferences.

Conclusions

Assembling reference genomes is a fast-moving field of research, which sees persistent updates and novel methodologies adopted (Rhie et al., 2021). Hence, variation seen in new reference genomes can reflect both the geographical range of the species distribution but also variation in methodologies used to sequence and assemble the genomes. For example, the two *G. aculeatus* genomes used here originate from samples representing two distinct lineages, the European and North American lineage but also differ significantly in how they were generated. Notably, the North American reference genome (Peichel et al., 2017) is part of a series of updates to the original *G. aculeatus* genome assembly (F. C. Jones et al., 2012). The original assembly used entirely Sanger sequence data (F. C. Jones et al., 2012), compared to the PacBio and Illumina sequence data used for the gynogen genome. As such, only when there is a significant interaction between population and reference origin and no obvious bias in the observations towards either reference can we exclude that the differences in sequencing and assembly methodologies are not the primary cause for the observed patterns.

The aim of this study was to investigate the effects of a local reference genome and its effects on downstream analyses. The reference-specific patterns of our results highlight that there is no simple solution. We suggest that the quality of the reference genome and annotations remains the single most important factor when choosing which reference to use. However, when multiple similar quality references are available, a local reference genome offers higher mapping efficiencies and decreases the proportion of missing data. The smallest reference effect among our analyses was for the marker-based methylation analysis, which had markedly more overlap among outliers in comparison to the window-based approach or the SV analysis, but still had over 30% of outliers that did not overlap. Taken together, using a local reference genome should increase the confidence of inferences made within a study, even if the difference is only minor.

Data Accessibility and Benefit-Sharing

The raw genomic sequences of the 60 *G. aculeatus* fish were obtained from previous publications (Chain et al., 2014; Feulner et al., 2015), and retrieved from the European Nucleotide Archive, accession number ERP004574. The raw methylome sequences were obtained from a previous publication (Sagonas et al., 2020), retrieved from the NIH genetic sequence database, accession number PRJNA605637.

The European-derived gynogen reference genome assembly and their annotations are available from the European Nucleotide Archive (ENA) accession number PRJEB54679. Please note, to comply with ENA submission guidelines the unmapped chromosome (chrUn) was broken into individual contigs, and annotations were lifted over using a custom R script. A total of 316 gene annotations spanned contig boundaries in chrUn and were removed. The original chrUn sequence and annotations are available upon request.

Scripts used in the data preparation and analysis pipelines are available online (https://github.com/dthorburn/Origin_Matters).

Acknowledgements

We thank Katie Peichel and Richard Nichols for their useful comments and discussion. We also thank Queen Mary University of London ITS Research support for technical assistance.

Author Contributions

D.M.J.T. and C.E. conceived and designed the study. F.J.J.C., P.G.D.F., C.E., T.L.L., M.B.P., and I.E.S.P. obtained the samples and oversaw the sequencing. T.B.H.R., E.B.B. and M.M. created the Big Screen Consortium with grant support from the Max Planck Society to M.M. K.S. and M.B.P. oversaw the initial genome assembly and annotation. D.M.J.T. analysed the data, advised by C.E. D.M.J.T. and C.E. drafted the manuscript. All co-authors contributed to the final version of the manuscript. C.E. acknowledges the support of the German Science Foundation (DFG, EI841/4-1 and EI 841/6-1). D.M.J.T. was supported by a studentship from Queen Mary University of London.

References

Akalin, A., Franke, V., Vlahoviček, K., Mason, C. E., & Schübeler, D. (2015). Genomation: A toolkit to summarize, annotate and visualize genomic intervals. *Bioinformatics*, *31*(7), 1127–1129. https://doi.org/10.1093/bioinformatics/btu775

Ameur, A., Dahlberg, J., Olason, P., Vezzi, F., Karlsson, R., Martin, M., Viklund, J., Kähäri, A. K., Lundin, P., Che, H., Thutkawkorapin, J., Eisfeldt, J., Lampa, S., Dahlberg, M., Hagberg, J., Jareborg, N., Liljedahl, U., Jonasson, I., Johansson, Å., … Gyllensten, U. (2017). SweGen: A whole-genome data resource of genetic variability in a cross-section of the Swedish population. *European Journal of Human Genetics*, *25*(11), 1253–1260. https://doi.org/10.1038/ejhg.2017.130

Andrews, S. (2010). FASTQC A Quality Control tool for High Throughput Sequence Data. *Babraham Institute*.

Auwera, G. A. van der, Carneiro, M. O., Chris Hartl, R. P., Angel, G. del, Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella1, K. v., Altshuler, D., Gabriel, S., & DePristo, M. A. (2014). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics*.

Baltazar-Soares, M., Klein, J. D., Correia, S. M., Reischig, T., Taxonera, A., Roque, S. M., dos Passos, L., Durão, J., Lomba, J. P., Dinis, H., Cameron, S. J. K., Stiebens, V. A., & Eizaguirre, C. (2020). Distribution of genetic diversity reveals colonization patterns and philopatry of the loggerhead sea turtles across geographic scales. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-74141-6

Barnett, D. W., Garrison, E. K., Quinlan, A. R., Str̈mberg, M. P., & Marth, G. T. (2011). Bamtools: A C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics*, *27*(12), 1691–1692. https://doi.org/10.1093/bioinformatics/btr174

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Berner, D., Roesti, M., Bilobram, S., Chan, S. K., Kirk, H., Pandoh, P., Taylor, G. A., Zhao, Y., Jones, S. J. M., & Defaveri, J. (2019). De novo sequencing, assembly, and annotation of four threespine stickleback genomes based on microfluidic partitioned DNA libraries. *Genes*, *10*(6), 10–15. https://doi.org/10.3390/genes10060426

Bohling, J. (2020). Evaluating the effect of reference genome divergence on the analysis of empirical RADseq datasets. *Ecology and Evolution*, *10*(14), 7585–7601. https://doi.org/10.1002/ece3.6483

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Chain, F. J. J., Feulner, P. G. D., Panchal, M., Eizaguirre, C., Samonte, I. E., Kalbe, M., Lenz, T. L., Stoll, M., Bornberg-Bauer, E., Milinski, M., & Reusch, T. B. H. (2014). Extensive Copy-Number Variation of Young Genes across Stickleback Populations. *PLoS Genetics*, *10*(12), 1–18. https://doi.org/10.1371/journal.pgen.1004830

Chiang, C., Layer, R. M., Faust, G. G., Lindberg, M. R., Rose, D. B., Garrison, E. P., Marth, G. T., Quinlan, A. R., & Hall, I. M. (2015). SpeedSeq: Ultra-fast personal genome analysis and interpretation. *Nature Methods*, *12*(10), 966–968. https://doi.org/10.1038/nmeth.3505

Christensen, K. A., Leong, J. S., Sakhrani, D., Biagi, C. A., Minkley, D. R., Withler, R. E., Rondeau, E. B., Koop, B. F., & Devlin, R. H. (2018). Chinook salmon (Oncorhynchus tshawytscha) genome and transcriptome. *PLoS ONE*, *13*(4), 1–15. https://doi.org/10.1371/journal.pone.0195461

Depristo, M. A., Banks, E., Poplin, R., Garimella, K. v., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., & Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–501. https://doi.org/10.1038/ng.806

Dewey, F. E., Chen, R., Cordero, S. P., Ormond, K. E., Caleshu, C., Karczewski, K. J., Whirl-Carrillo, M., Wheeler, M. T., Dudley, J. T., Byrnes, J. K., Cornejo, O. E., Knowles, J. W., Woon, M., Sangkuhl, K., Gong, L., Thorn, C. F., Hebert, J. M., Capriotti, E., David, S. P., … Ashley, E. A. (2011). Phased whole-genome genetic risk in a family quartet using a major allele reference sequence. *PLoS Genetics*, *7*(9). https://doi.org/10.1371/journal.pgen.1002280

Dowle, M., Srinivasan, A., Short, T., Lianoglou, S., Saporta, R., & Antonyan, E. (2015). *data.table: extension of data.frame. R package version 1.9.6*. https://cran.r-project.org/package=data.table

Emms, D. M., & Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*(238), 1–14. https://doi.org/10.1101/466201

Fakhro, K. A., Staudt, M. R., Ramstetter, M. D., Robay, A., Malek, J. A., Badii, R., Al-Marri, A. A. N., Khalil, C. A., Al-Shakaki, A., Chidiac, O., Stadler, D., Zirie, M., Jayyousi, A., Salit, J., Mezey, J. G., Crystal, R. G., & Rodriguez-Flores, J. L. (2016). The Qatar genome: A population-specific tool for precision medicine in the Middle East. *Human Genome Variation*, *3*(August 2015), 1–7. https://doi.org/10.1038/hgv.2016.16

Fan, G., Song, Y., Yang, L., Huang, X., Zhang, S., Zhang, M., Yang, X., Chang, Y., Zhang, H., Li, Y., Liu, S., Yu, L., Chu, J., Seim, I., Feng, C., Near, T. J., Wing, R. A., Wang, W., Wang, K., … He, S. (2020). Initial data release and announcement of the 10,000 Fish Genomes Project (Fish10K). *GigaScience*, *9*(8), 1–7. https://doi.org/10.1093/gigascience/giaa080

Fang, B., Merilä, J., Ribeiro, F., Alexandre, C. M., & Momigliano, P. (2018). Molecular Phylogenetics and Evolution Worldwide phylogeny of three-spined sticklebacks. *Molecular Phylogenetics and Evolution*, *127*(June), 613–625. https://doi.org/10.1016/j.ympev.2018.06.008

Felsenstein, J. (1989). PHYLIP—Phylogeny Inference Package (version 3.2). *Cladistics*, *5*, 164–166.

Feulner, P. G. D., Chain, F. J. J., Panchal, M., Eizaguirre, C., Kalbe, M., Lenz, T. L., Mundry, M., Samonte, I. E., Stoll, M., Milinski, M., Reusch, T. B. H., & Bornberg-Bauer, E. (2013). Genome-wide patterns of standing genetic variation in a marine population of three-spined sticklebacks. *Molecular Ecology*, *22*(3), 635–649. https://doi.org/10.1111/j.1365-294X.2012.05680.x

Feulner, P. G. D., Chain, F. J. J., Panchal, M., Huang, Y., Eizaguirre, C., Kalbe, M., Lenz, T. L., Samonte, I. E., Stoll, M., Bornberg-Bauer, E., Reusch, T. B. H., & Milinski, M. (2015). Genomics of Divergence along a Continuum of Parapatric Population Differentiation. *PLoS Genetics*, *11*(2), 1–18. https://doi.org/10.1371/journal.pgen.1004966

Formenti, G., Theissinger, K., Fernandes, C., Bista, I., Bombarely, A., Bleidorn, C., Ciofi, C., Crottini, A., Godoy, J. A., Höglund, J., Malukiewicz, J., Mouton, A., Oomen, R. A., Paez, S., Palsbøll, P. J., Pampoulie, C., Ruiz-López, M. J., Svardal, H., Theofanopoulou, C., … Zammit, G. (2022). The era of reference genomes in conservation genomics. *Trends in Ecology and Evolution*, *37*(3), 197–202. https://doi.org/10.1016/j.tree.2021.11.008

Frantz, L. A. F., Schraiber, J. G., Madsen, O., Megens, H. J., Cagan, A., Bosse, M., Paudel, Y., Crooijmans, R. P. M. A., Larson, G., & Groenen, M. A. M. (2015). Evidence of long-term gene flow and selection during domestication from analyses of Eurasian wild and domestic pig genomes. *Nature Genetics*, *47*(10), 1141–1148. https://doi.org/10.1038/ng.3394

Gan, X., Stegle, O., Behr, J., Steffen, J. G., Drewe, P., Hildebrand, K. L., Lyngsoe, R., Schultheiss, S. J., Osborne, E. J., Sreedharan, V. T., Kahles, A., Bohnert, R., Jean, G., Derwent, P., Kersey, P., Belfield, E. J., Harberd, N. P., Kemen, E., Toomajian, C., … Mott, R. (2011). Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature*, *477*(7365), 419–423. https://doi.org/10.1038/nature10414

Garrison, E., & Marth, G. (2012). *Haplotype-based variant detection from short-read sequencing*. 1–9. http://arxiv.org/abs/1207.3907

Gaudet, P., & Dessimoz, C. (2017). *Gene ontology: pitfalls, biases, and remedies* (C. Dessimoz & N. Škunca, Eds.). Humana Press. https://doi.org/10.1007/978-1-4939-3743-1_14

Gissi, C., Iannelli, F., & Pesole, G. (2008). Evolution of the mitochondrial genome of Metazoa as exemplified by comparison of congeneric species. *Heredity*, *101*(4), 301–320. https://doi.org/10.1038/hdy.2008.62

Glazer, A. M., Killingbeck, E. E., Mitros, T., Rokhsar, D. S., & Miller, C. T. (2015). Genome assembly improvement and mapping convergently evolved skeletal traits in sticklebacks with genotyping-by-sequencing. *G3: Genes, Genomes, Genetics*, *5*(7), 1463–1472. https://doi.org/10.1534/g3.115.017905

Grabherr, M. G., Russell, P., Meyer, M., Mauceli, E., Alföldi, J., di Palma, F., & Lindblad-Toh, K. (2010). Genome-wide synteny through highly sensitive sequence alignment: Satsuma. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btq102

Grytten, I., Rand, K. D., Nederbragt, A. J., & Sandve, G. K. (2020). Assessing graph-based read mappers against a baseline approach highlights strengths and weaknesses of current methods. *BMC Genomics*, *21*(1), 1–9. https://doi.org/10.1186/s12864-020-6685-y

Guan, D., Guan, D., McCarthy, S. A., Wood, J., Howe, K., Wang, Y., Durbin, R., & Durbin, R. (2020). Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*, *36*(9), 2896–2898. https://doi.org/10.1093/bioinformatics/btaa025

Guo, Y., Dai, Y., Yu, H., Zhao, S., Samuels, D. C., & Shyr, Y. (2017). Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. *Genomics*, *109*(2), 83–90. https://doi.org/10.1016/j.ygeno.2017.01.005

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

Hirsch, C. N., Hirsch, C. D., Brohammer, A. B., Bowman, M. J., Soifer, I., Barad, O., Shem-Tov, D., Baruch, K., Lu, F., Hernandez, A. G., Fields, C. J., Wright, C. L., Koehler, K., Springer, N. M., Buckler, E., Buell, C. R., de Leon, N., Kaeppler, S. M., Childs, K. L., & Mikel, M. A. (2016). Draft assembly of elite inbred line PH207 provides insights into genomic and transcriptome diversity in maize. *Plant Cell*, *28*(11), 2700–2714. https://doi.org/10.1105/tpc.16.00353

Holt, C., & Yandell, M. (2011). MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, *12*(1). https://doi.org/10.1186/1471-2105-12-491

Huang, Y., Chain, F. J. J., Panchal, M., Eizaguirre, C., Kalbe, M., Lenz, T. L., Samonte, I. E., Stoll, M., Bornberg-Bauer, E., Reusch, T. B. H., Milinski, M., & Feulner, P. G. D. (2016). Transcriptome profiling of immune tissues reveals habitat-specific gene expression between lake and river sticklebacks. *Molecular Ecology*, *25*(4), 943–958. https://doi.org/10.1111/mec.13520

International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. https://doi.org/10.1038/35057062

International Human Genome Sequencing Consortium. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, *431*(7011), 931–945.

Jeffares, D. C., Jolly, C., Hoti, M., Speed, D., Shaw, L., Rallis, C., Balloux, F., Dessimoz, C., Bähler, J., & Sedlazeck, F. J. (2017). Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nature Communications*, *8*(14061), 1–11. https://doi.org/10.1038/ncomms14061

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., Swofford, R., Pirun, M., Zody, M. C., White, S., Birney, E., Searle, S., Schmutz, J., Grimwood, J., Dickson, M. C., Myers, R. M., Miller, C. T., Summers, B. R., Knecht, A. K., … Kingsley, D. M. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, *484*(7392), 55–61. https://doi.org/10.1038/nature10944

Jones, M. R., & Good, J. M. (2016). Targeted capture in evolutionary and ecological genomics. *Molecular Ecology*, *25*(1), 185–202. https://doi.org/10.1111/mec.13304

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S. Y., Lopez, R., & Hunter, S. (2014). InterProScan 5: Genome-scale protein function classification. *Bioinformatics*, *30*(9), 1236–1240. https://doi.org/10.1093/bioinformatics/btu031

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., Yabana, M., Harada, M., Nagayasu, E., Maruyama, H., Kohara, Y., Fujiyama, A., Hayashi, T., & Itoh, T. (2014). Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Research*, *24*(8), 1384–1395. https://doi.org/10.1101/gr.170720.113

Kao, R. R., Haydon, D. T., Lycett, S. J., & Murcia, P. R. (2014). Supersize me: How whole-genome sequencing and big data are transforming epidemiology. *Trends in Microbiology*, *22*(5), 282–291. https://doi.org/10.1016/j.tim.2014.02.011

Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, *37*(8), 907–915. https://doi.org/10.1038/s41587-019-0201-4

King, T., Butcher, S., & Zalewski, L. (2017). *Apocrita - High Performance Computing Cluster For Queen Mary University Of London*. https://doi.org/10.5281/ZENODO.438045

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. *Genome Research*, *27*(5), 722–736. https://doi.org/10.1101/gr.215087.116

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics*, *5*, 1–9. https://doi.org/10.1186/1471-2105-5-59

Krueger, F., & Andrews, S. R. (2011). Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, *27*(11), 1571–1572. https://doi.org/10.1093/bioinformatics/btr167

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., & Marra, M. A. (2009). Circos: An information aesthetic for comparative genomics. *Genome Research*. https://doi.org/10.1101/gr.092759.109

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lacaze, P., Pinese, M., Kaplan, W., Stone, A., Brion, M. J., Woods, R. L., McNamara, M., McNeil, J. J., Dinger, M. E., & Thomas, D. M. (2019). The Medical Genome Reference Bank: a whole-genome data resource of 4000 healthy elderly individuals. Rationale and cohort design. *European Journal of Human Genetics*, *27*(2), 308–316. https://doi.org/10.1038/s41431-018-0279-z

Lai, Y. T., Yeung, C. K. L., Omland, K. E., Pang, E. L., Hao, Y., Liao, B. Y., Cao, H. F., Zhang, B. W., Yeh, C. F., Hung, C. M., Hung, H. Y., Yang, M. Y., Liang, W., Hsu, Y. C., Yao, C. te, Dong, L., Lin, K., & Li, S. H. (2019). Standing genetic variation as the predominant source for adaptation of a songbird. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1813597116

Lappalainen, I., Lopez, J., Skipper, L., Hefferon, T., Spalding, J. D., Garner, J., Chen, C., Maguire, M., Corbett, M., Zhou, G., Paschall, J., Ananiev, V., Flicek, P., & Church, D. M. (2013). DbVar and DGVa: Public archives for genomic structural variation. *Nucleic Acids Research*, *41*, 936–941. https://doi.org/10.1093/nar/gks1213

Layer, R. M., Chiang, C., Quinlan, A. R., & Hall, I. M. (2014). LUMPY: A probabilistic framework for structural variant discovery. *Genome Biology*, *15*(6), 1–19. https://doi.org/10.1186/gb-2014-15-6-r84

Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2020). emmeans : Estimated Marginal Means, aka Least-Squares Means. In *R package version 1.15-15*. https://doi.org/10.1080/00031305.1980.10483031>.License

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv*, 1–3. http://arxiv.org/abs/1303.3997

Liu, D., Hunt, M., & Tsai, I. J. (2018). Inferring synteny between genome assemblies: A systematic evaluation. *BMC Bioinformatics*, *19*(1), 1–13. https://doi.org/10.1186/s12859-018-2026-4

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1), 1–3.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Nath, S., Shaw, D. E., & White, M. A. (2021). Improved contiguity of the threespine stickleback genome using long-read sequencing. *G3 Genes|Genomes|Genetics*, *0*(0), 1–9. https://doi.org/10.1093/g3journal/jkab007

Palmer, D. H., & Kronforst, M. R. (2020). A shared genetic basis of mimicry across swallowtail butterflies points to ancestral co-option of doublesex. *Nature Communications*, *11*(1), 1–10. https://doi.org/10.1038/s41467-019-13859-y

Parker, J., Tsagkogeorga, G., Cotton, J. A., Liu, Y., Provero, P., Stupka, E., & Rossiter, S. J. (2013). Genome-wide signatures of convergent evolution in echolocating mammals. *Nature*, *502*(7470), 228–231. https://doi.org/10.1038/nature12511

Peichel, C. L., Sullivan, S. T., Liachko, I., & White, M. A. (2017). Improvement of the Threespine Stickleback Genome Using a Hi-C-Based Proximity-Guided Assembly. *Journal of Heredity*, *108*(6), 693–700. https://doi.org/10.1093/jhered/esx058

Pfeifer, B., Wittelsbürger, U., Ramos-Onsins, S. E., & Lercher, M. J. (2014). PopGenome: An efficient swiss army knife for population genomic analyses in R. *Molecular Biology and Evolution*, *31*(7), 1929–1936. https://doi.org/10.1093/molbev/msu136

Pirooznia, M., Goes, F., & Zandi, P. P. (2015). Whole-genome CNV analysis: Advances in computational approaches. *Frontiers in Genetics*, *6*(MAR), 1–9. https://doi.org/10.3389/fgene.2015.00138

Prasad, A., Lorenzen, E. D., & Westbury, M. v. (2022). Evaluating the role of reference-genome phylogenetic distance on evolutionary inference. *Molecular Ecology Resources*, *22*(1), 45–55. https://doi.org/10.1111/1755-0998.13457

Pritt, J., Chen, N. C., & Langmead, B. (2018). FORGe: Prioritizing variants for graph genomes 06 Biological Sciences 0604 Genetics. *Genome Biology*, *19*(1), 1–16. https://doi.org/10.1186/s13059-018-1595-x

Pushkarev, D., Neff, N. F., & Quake, S. R. (2009). Single-molecule sequencing of an individual human genome. *Nature Biotechnology*, *27*(9), 847–850. https://doi.org/10.1038/nbt.1561

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

R Development Core Team. (2019). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.* https://doi.org/10.1007/978-3-540-74686-7

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., & Vilo, J. (2019). g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research*, *47*(1), 191–198. https://doi.org/10.1093/nar/gkz369

Rausch, T., Zichner, T., Schlattl, A., Stütz, A. M., Benes, V., & Korbel, J. O. (2012). DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, *28*(18), 333–339. https://doi.org/10.1093/bioinformatics/bts378

Reid, B. N., Moran, R. L., Kopack, C. J., & Fitzpatrick, S. W. (2021). Rapture-ready darters: Choice of reference genome and genotyping method (whole-genome or sequence capture) influence population genomic inference in Etheostoma. *Molecular Ecology Resources*, *21*(2), 404–420. https://doi.org/10.1111/1755-0998.13275

Reid, K., Bell, M. A., & Veeramah, K. R. (2021). *Threespine Stickleback: A Model System For Evolutionary Genomics*. 1–27.

Reimegård, J., Kundu, S., Pendle, A., Irish, V. F., Shaw, P., Nakayama, N., Sundström, J. F., & Emanuelsson, O. (2017). Genome-wide identification of physically clustered genes suggests chromatin-level co-regulation in male reproductive development in

Arabidopsis thaliana. *Nucleic Acids Research*, *45*(6), 3253–3265. https://doi.org/10.1093/nar/gkx087

Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., Kim, J., Lee, C., Ko, B. J., Chaisson, M., Gedman, G. L., Cantin, L. J., Thibaud-Nissen, F., Haggerty, L., Bista, I., Smith, M., … Jarvis, E. D. (2021). Towards complete and error-free genome assemblies of all vertebrate species. In *Nature* (Vol. 592, Issue April, pp. 737–746). https://doi.org/10.1101/2020.05.22.110833

Roach, M. J., Schmidt, S., & Borneman, A. R. (2018). Purge Haplotigs: Synteny Reduction for Third-gen Diploid Genome Assemblies. *BMC Bioinformatics*, *19*(460), 1–10. https://doi.org/10.1101/286252

Roesti, M., Kueng, B., Moser, D., & Berner, D. (2015). The genomics of ecological vicariance in threespine stickleback fish. *Nature Communications*, *6*(1), 1–14. https://doi.org/10.1038/ncomms9767

Roesti, M., Moser, D., & Berner, D. (2013). Recombination in the threespine stickleback genome - Patterns and consequences. *Molecular Ecology*, *22*(11), 3014–3027. https://doi.org/10.1111/mec.12322

Ronco, F., Matschiner, M., Böhne, A., Boila, A., Büscher, H. H., Indermaur, A., el Taher, A., Malinsky, M., Ricci, V., Kahmen, A., Jentoft, S., & Salzburger, W. (2020). Drivers and dynamics of a massive adaptive radiation in African cichlid fish. *Nature*, 1–6. https://doi.org/10.1038/s41586-020-2930-4

Sagonas, K., Meyer, B. S., Kaufmann, J., Lenz, T. L., Häsler, R., & Eizaguirre, C. (2020). Experimental parasite infection causes genome-wide changes in DNA methylation. *Molecular Biology and Evolution*, *37*(8), 2287–2299. https://doi.org/10.1093/molbev/msaa084

Samonte-Padilla, I. E., Eizaguirre, C., Scharsack, J. P., Lenz, T. L., & Milinski, M. (2011). Induction of diploid gynogenesis in an evolutionary model organism, the three-spined stickleback (Gasterosteus aculeatus). *BMC Developmental Biology*, *11*, 1–11. https://doi.org/10.1186/1471-213X-11-55

Shapiro, M. D., Marks, M. E., Peichel, C. L., Blackman, B. K., Nereng, K. S., Jónsson, B., Schluter, D., & Kingsley, D. M. (2004). Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature*, *428*, 717–723. https://doi.org/10.1038/nature04500

Sherry, S. T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, *29*(1), 308–311. http://www.ncbi.nlm.nih.gov/SNP.

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://doi.org/10.1093/bioinformatics/btv351

Smit, A., Hubley, R., & Green, P. (2015). RepeatMasker Open-4.0. In *RepeatMasker Open-4.0*.

Spitz, F., Gonzalez, F., Peichel, C., Vogt, T. F., Duboule, D., & Zákány, J. (2001). Large scale transgenic and cluster deletion analysis of the HoxD complex separate an ancestral regulatory module from evolutionary innovations. *Genes and Development*, *15*(17), 2209–2214. https://doi.org/10.1101/gad.205701

Springer, N. M., Anderson, S. N., Andorf, C. M., Ahern, K. R., Bai, F., Barad, O., Barbazuk, W. B., Bass, H. W., Baruch, K., Ben-Zvi, G., Buckler, E. S., Bukowski, R., Campbell, M. S., Cannon, E. K. S., Chomet, P., Kelly Dawe, R., Davenport, R., Dooner, H. K., Du, L. H., … Brutnell, T. P. (2018). The maize w22 genome provides a foundation for functional genomics and transposon biology. *Nature Genetics*, *50*(9). https://doi.org/10.1038/s41588-018-0158-0

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Stanke, M., Diekhans, M., Baertsch, R., & Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*, *24*(5), 637–644. https://doi.org/10.1093/bioinformatics/btn013

Stapley, J., Reger, J., Feulner, P. G. D., Smadja, C., Galindo, J., Ekblom, R., Bennison, C., Ball, A. D., Beckerman, A. P., & Slate, J. (2010). Adaptation genomics: The next generation. *Trends in Ecology and Evolution*, *25*(12), 705–712. https://doi.org/10.1016/j.tree.2010.09.002

Stern, D. ben, & Lee, C. E. (2020). Evolutionary origins of genomic adaptations in an invasive copepod. *Nature Ecology and Evolution*, *4*(8), 1084–1094. https://doi.org/10.1038/s41559-020-1201-y

Treangen, T. J., & Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: Computational challenges and solutions. *Nature Reviews Genetics*, *13*(1), 36–46. https://doi.org/10.1038/nrg3117

Valiente-Mullor, C., Beamud, B., Ansari, I., Frances-Cuesta, C., Garcia-Gonzalez, N., Mejia, L., Ruiz-Hueso, P., & Gonzalez-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), 1–29. https://doi.org/10.1371/JOURNAL.PCBI.1008678

Vezzi, F., Narzisi, G., & Mishra, B. (2012). Reevaluating Assembly Evaluations with Feature Response Curves: GAGE and Assemblathons. *PLoS ONE*, *7*(12), 1–11. https://doi.org/10.1371/journal.pone.0052210

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., & Earl, A. M. (2014). Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE*, *9*(11). https://doi.org/10.1371/journal.pone.0112963

Wang, Z., Pascual-anaya, J., Zadissa, A., Li, W., Niimura, Y., Huang, Z., Li, C., White, S., Xiong, Z., Fang, D., Wang, B., Ming, Y., Chen, Y., Zheng, Y., Kuraku, S., Pignatelli, M., Herrero, J., Beal, K., Nozawa, M., … Wang, J. (2013). The draft genomes of soft-shell turtle and green sea turtle yield insights into the development and evolution of the turtle-specific body plan. *Nature Genetics*, *45*(6), 701–706. https://doi.org/10.1038/ng.2615

Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T. Y. (2017). Ggtree: an R Package for Visualization and Annotation of Phylogenetic Trees With Their Covariates and Other Associated Data. *Methods in Ecology and Evolution*, *8*(1), 28–36. https://doi.org/10.1111/2041-210X.12628

Tables and Figures

Table 1. Assembly Statistics

	Contig Assembly	Anchored Assembly
Number of contigs	1906	877[a]
Total size of contigs	458064828	458376166
Longest contig	5194525	33176498
N50	746270	18858959
Assembly validation		
Complete BUSCOs		4363 (95.2%)
Complete single-copy BUSCOs		3913 (85.4%)
Complete duplicated BUSCOs		450 (9.8%)
Fragmented BUSCOs		93 (2.0%)
Missing BUSCOs		128 (2.8%)
Total BUSCO groups searched		4584

[a] 21 LGs, chrM, and 855 unplaced contigs.

Table 2. Distribution of outlier windows and differentially methylated sites (DMS), overlapping genes, and their functional enrichment.

Population Origin	Reference Origin	Metric	Outlier Windows	Outlier Genes	1-1 Ortholog Outlier Genes	Shared Outlier Orthologs	Percentage Overlapping Outliers	Significant GO Terms	Overlapping Terms
North America	North America	Tajima's *D*	878	1019	335	221	65.97%	3	0
North America	Europe	Tajima's *D*	874	916	333		66.37%	0	
Europe	North America	Tajima's *D*	1308	1321	342	214	62.57%	0	0
Europe	Europe	Tajima's *D*	1309	1449	424		50.47%	3	
North America	North America	F_{ST}	446	541	194	99	51.03%	0	NA
North America	Europe	F_{ST}	440	534	174		56.90%	0	
Europe	North America	F_{ST}	672	703	240	101	42.08%	0	0
Europe	Europe	F_{ST}	660	788	243		41.56%	2	
North America	North America	π	888	696	212	29	13.68%	0	NA
North America	Europe	π	880	558	153		18.95%	0	
Europe	North America	π	1332	581	143	36	25.17%	0	NA
Europe	Europe	π	1320	630	199		18.09%	0	
Europe	North America	DMS	2404	712	299	204	68.23%	5	3
Europe	Europe	DMS	2550	711	298		68.45%	3	

Table 3. Distribution of structural variants and structurally variable genes and their functional enrichment.

Population Origin	Reference Origin	SV Type	Number of SVs	Number of SVGs	1-1 Ortholog SVGs	Shared Outlier Orthologs	Percentage Overlapping	Significant GO Terms	Overlapping Terms
North America	North America	Deletion	3915	3034	920	330	35.87%	3	3
North America	Europe	Deletion	5720	2613	877		37.63%	13	
Europe	North America	Deletion	4049	2968	893	296	33.15%	1	0
Europe	Europe	Deletion	5027	2113	716		41.34%	16	
North America	North America	Duplication	2684	3908	1195	500	41.84%	0	NA
North America	Europe	Duplication	3402	3180	1125		44.44%	0	
Europe	North America	Duplication	4300	4070	1256	524	41.72%	2	0
Europe	Europe	Duplication	4425	3072	1074		48.79%	0	
North America	North America	Inversion	794	4128	1358	626	46.10%	20	0
North America	Europe	Inversion	757	3382	1234		50.73%	10	
Europe	North America	Inversion	828	4031	1352	522	38.61%	15	0
Europe	Europe	Inversion	685	2991	1081		48.29%	0	

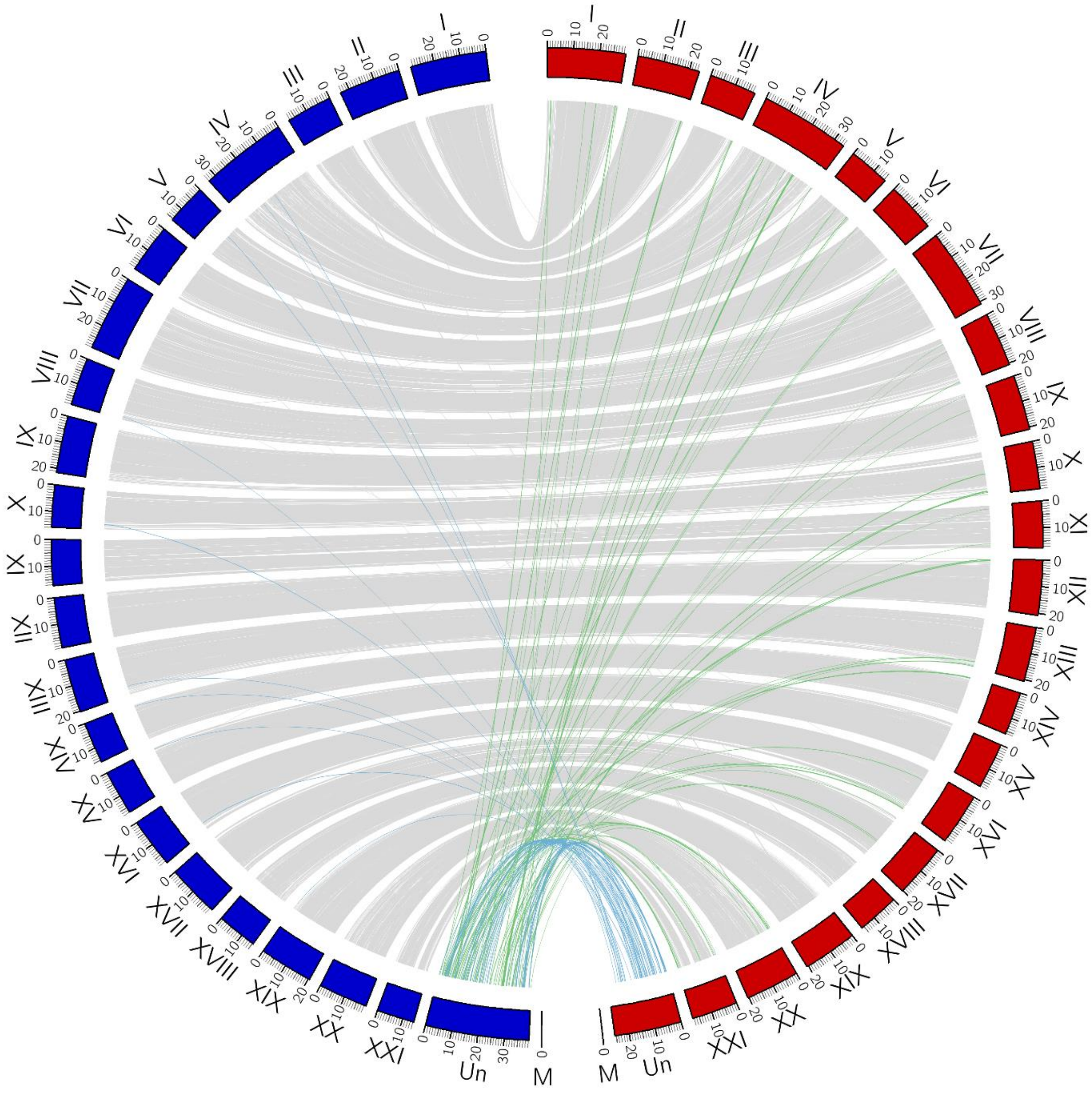
Figure Legends

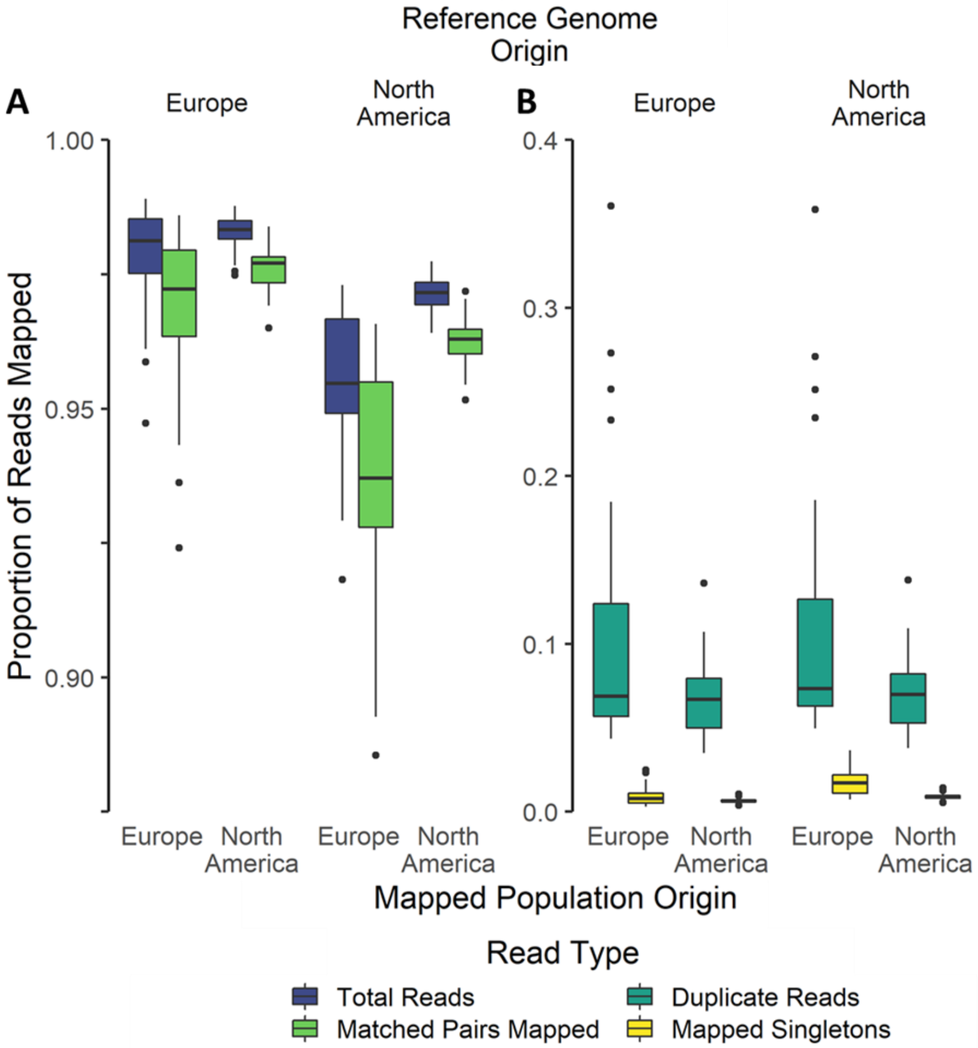
Figure 1. Synteny plot between the two *G. aculeatus* genome assemblies. Grey lines between the autosomes represent +99% syntenic blocks greater than 1kb between the European gynogen assembly (left, blue blocks) and the North American *G. aculeatus* assembly (right, red blocks). Coloured lines represent synteny between resolved regions and the unmapped scaffold in each assembly. Specifically, blue and green lines represent 1kb +99% syntenic blocks between the unmapped scaffolds and the alternate reference autosome.

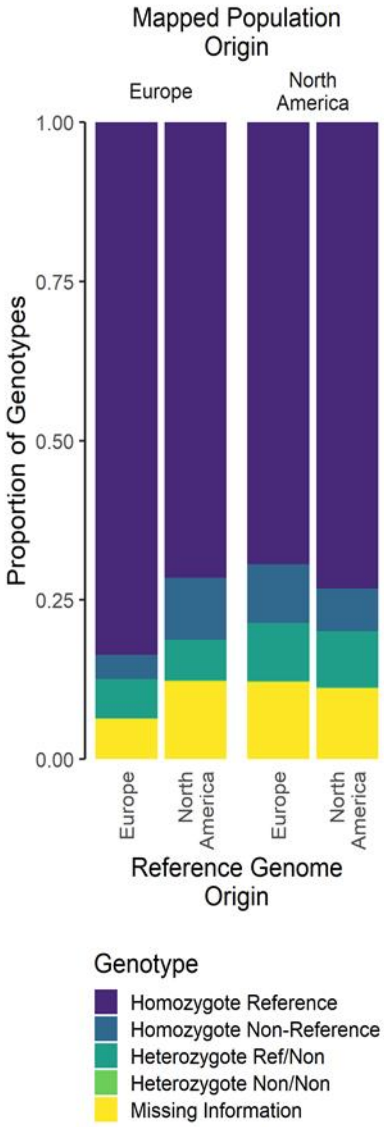
Figure 2. The effect of reference genome origin on mapping efficiency. Scales differ among panels, but the units are the same (*A*-*B*). There was significantly higher mapping efficiency when using the European reference genome regardless of sample origin for total reads (matched and singleton), matched pair reads (*A*) as well as significantly lower singletons (*B*). There was no difference in the number of duplicate reads identified. Reference genome origin is labelled at the top of the plot, and mapped population origin at the bottom.

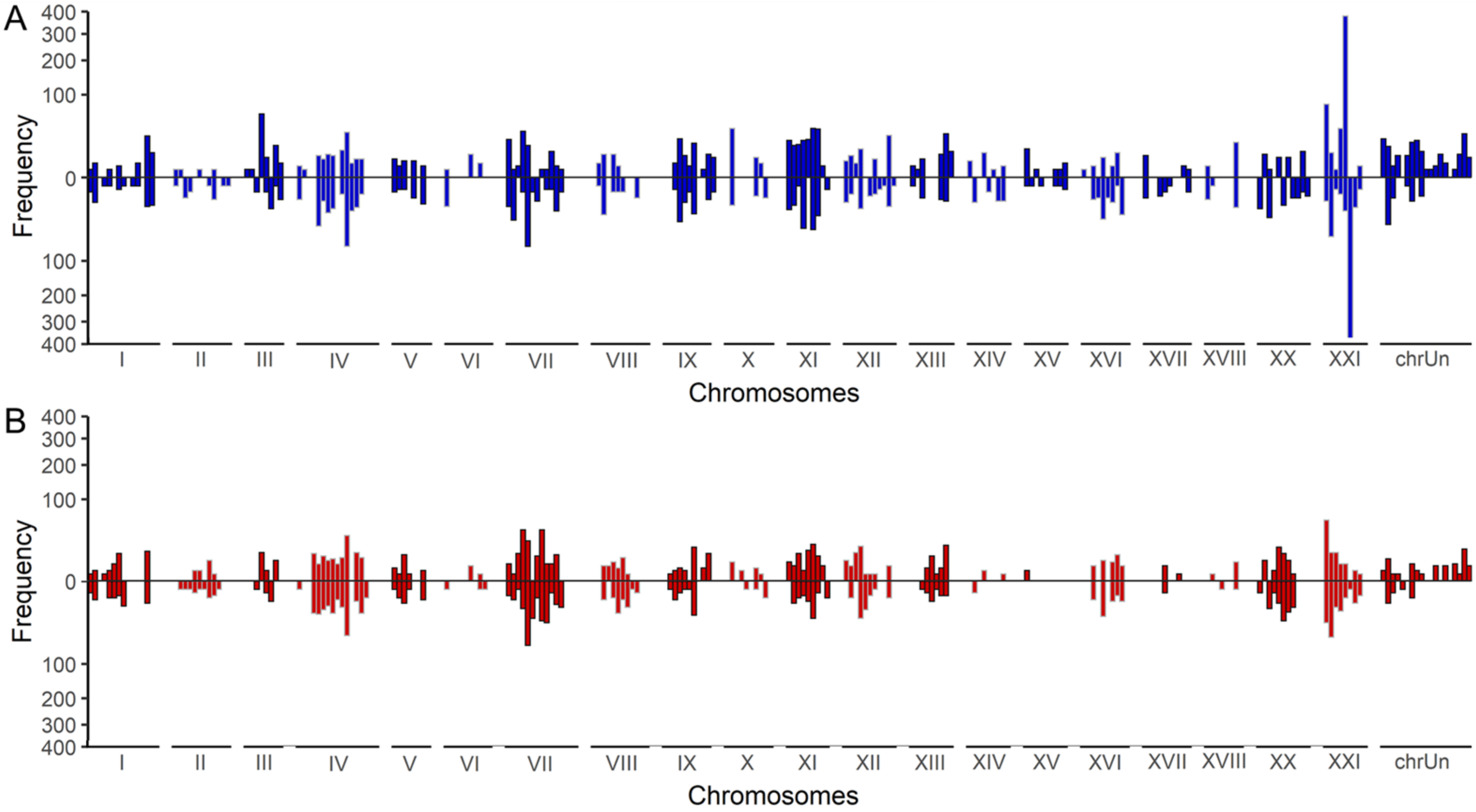
Figure 3. The proportion of SNP classes among all segregating sites. Segregating sites with no coverage or with a SNP that was removed during filtering are defined as missing information.

Figure 4. Comparing distributions of π outliers. Windows are compared across the genome for (top) European and (bottom) North American populations mapped to the (*A*) European or (*B*) North American reference genome. Axes are square root transformed.


I
II
III
IV
V
VI
VII
VIII
IX
X
XI
XII
XIII
XIV
XV
XVI
XVII
XVIII
XIX
XX
XXI
Un
M
0
10
20
30


Reference Genome
Origin
A
Europe
North America
B
Europe
North America
1.00
0.95
0.90
Proportion of Reads Mapped
0.4
0.3
0.2
0.1
0.0
Europe
North America
Europe
North America
Europe
North America
Europe
North America
Mapped Population Origin
Read Type
Total Reads
Matched Pairs Mapped
Duplicate Reads
Mapped Singletons


Mapped Population Origin
Europe
North America
1.00
0.75
0.50
0.25
0.00
Proportion of Genotypes
Europe
North America
Europe
North America
Reference Genome Origin
Genotype
Homozygote Reference
Homozygote Non-Reference
Heterozygote Ref/Non
Heterozygote Non/Non
Missing Information


A
B
400
300
200
100
0
Frequency
I
II
III
IV
V
VI
VII
VIII
IX
X
XI
XII
XIII
XIV
XV
XVI
XVII
XVIII
XX
XXI
chrUn
Chromosomes