Canadian Science Publishing
NRC Research Press journals | Publishing Services

Genome / Génome

Whole-genome sequencing reveals the extent of heterozygosity in a preferentially self-fertilizing hermaphroditic vertebrate

Journal:	*Genome*
Manuscript ID	gen-2017-0188.R1
Manuscript Type:	Article
Date Submitted by the Author:	22-Nov-2017
Complete List of Authors:	Lins, Luana; Washington State University, School of Biological Sciences and Center for Reproductive Biology Trojahn, Shawn; Washington State University, School of Biological Sciences and Center for Reproductive Biology Sockell, Alexandra; Stanford University, Department of Genetics Yee, Muh-Ching; Stanford Functional Genomics Facility Tatarenkov, Andrey; University of California, Department of Ecology and Evolutionary Biology Bustamante, Carlos; Stanford University, Department of Genetics Earley, Ryan; University of Alabama, Department of Biological Sciences Kelley, Joanna; Washington State University, School of Biological Sciences
Is the invited manuscript for consideration in a Special Issue? :	Ecological Genomics (closed)
Keyword:	Kryptolebias marmoratus, Heterozygosity, self-fertilizing

SCHOLARONE™
Manuscripts

Title: Whole-genome sequencing reveals the extent of heterozygosity in a preferentially self-fertilizing hermaphroditic vertebrate

Authors:

Luana S. F. Lins[1#], Shawn Trojahn[1#], Alexandra Sockell[2], Muh-Ching Yee[3], Andrey Tatarenkov[4], Carlos D. Bustamante[2,5], Ryan L. Earley[6], Joanna L. Kelley[1*]

Addresses:

[1] School of Biological Sciences and Center for Reproductive Biology, Washington State University, 100 Dairy Road, Pullman, WA 99164, USA

[2] Department of Genetics, Stanford University, 300 Pasteur Dr., Stanford, CA 94305, USA

[3] Stanford Functional Genomics Facility, CCSR 0120, 269 Campus Drive, Stanford, CA 94305, USA

[4] Department of Ecology and Evolutionary Biology, University of California, Irvine, CA, 92697, USA

[5] Department of Biomedical Data Science, 365 Lasuen Street, Littlefield Center, Room 303, Stanford, CA, 94305 USA

[6] Department of Biological Sciences, University of Alabama, Tuscaloosa, AL 35487, USA

[#] Equal Contributions

*** Author for Correspondence**: Joanna L. Kelley, School of Biological Sciences, Washington State University, 100 Dairy Road, Pullman, WA 99164, USA, 509-335-0037 (phone), 509-335-4848 (fax), joanna.l.kelley@wsu.edu

Data deposition:

Sequence data has been deposited under NCBI BioProject PRJNA385014

Abstract:

The mangrove rivulus, *Kryptolebias marmoratus,* is one of only two self-fertilizing hermaphroditic fish and inhabits mangrove forests. While selfing can be advantageous, it reduces heterozygosity and decreases genetic diversity. Studies using microsatellites found that there are variable levels of selfing among populations of *K. marmoratus* but overall there is a low rate of outcrossing and therefore, low heterozygosity. In this study, we used whole-genome data to assess the level of heterozygosity in different lineages of the mangrove rivulus and infer the phylogenetic relationships among those lineages. We sequenced whole genomes from 15 lineages that were completely homozygous at microsatellite loci and used single nucleotide polymorphisms (SNPs) to determine heterozygosity levels. More variation was uncovered than in studies using microsatellite data due to the resolution of full genome sequencing data. Moreover, missense polymorphisms were found most often in genes associated with immune function and reproduction. Inferred phylogenetic relationships suggest that lineages largely group by their geographic distribution. The use of whole-genome data provided further insight into genetic diversity in this unique species. Although this study was limited by the number of lineages that were available, these data suggest that there is previously undescribed variation within lineages of *K. marmoratus* that could have functional consequences and/or inform us about the limits to selfing (e.g., genetic load, accumulation of deleterious mutations) and selection that might favor the maintenance of heterozygosity. These results highlight the need to sequence additional individuals within and among lineages.

Introduction:

Self-fertilization is a mode of reproduction employed by many plants and invertebrates, and comes with inherent advantages and disadvantages (reviewed in Shimizu and Tsuchimatsu 2015). Selfing assures reproduction when few mating partners are available. While selfing can result in coadapted suites of alleles that confer high fitness in a given environment (Allard 1975), it can also drive populations quickly toward extinction and dampen responses to selection (Noel et al. 2017). Many selfing species also can outcross (i.e., a mixed mating system), which introduces genetic diversity at the individual and population levels. Recent studies have shed light on the genomic and evolutionary consequences of selfing (Burgarella et al. 2015; Noel et al. 2017) but the extent to which genetic diversity is maintained in mixed-mating species remains an open question.

Mangrove rivulus fish (*Kryptolebias marmoratus*) are an excellent model in which to address such a question because individuals can exist as hermaphrodite or male (Mackiewicz et al. 2006a). The most common mode of hermaphrodite reproduction for *K. marmoratus* is self-fertilization (Harrington 1961), which can generate isogenic lineages characterized by complete homozygosity (at 32 neutral markers; Mackiewicz et al. 2006a) . Predominant self-fertilization in *K. marmoratus* has both disadvantages and advantages. Selfing can create barriers to gene flow between adjacent populations, resulting in varying levels of differentiation among populations. For example, a study utilizing microsatellite data found that the genetic differentiation among populations only 112 kilometers apart was high, with differentiation values (measured by F_{ST}) approaching 0.26 between some populations (Tatarenkov et al. 2012). Rare outcrossing events occur between males and hermaphrodites and result in a burst of heterozygosity, which is then reduced through subsequent generations of selfing (Mackiewicz et al. 2006b). Additionally,

previous studies have highlighted the impact self-fertilization has on the immune system and body size of these fish, with a reduction in heterozygosity (measured via microsatellites) leading to an increased parasite load (Ellison et al. 2011) and smaller adult male size (Molloy et al. 2011). Alternatively, selfing can maintain locally well-adapted genotypes (Avise and Tatarenkov 2012). Selfing can also potentially have a positive fitness effect in individuals by saving energy otherwise invested in courtship.

This species has a broad geographic distribution ranging from Florida, the Bahamas, the Caribbean and Central America (Davis et al. 1990). Preferentially inhabiting mangrove forests, mangrove rivulus are particularly well-adapted for this constantly changing habitat (Davis et al. 1990). Individuals are able to survive a wide range of environmental conditions, including high hydrogen sulfide levels, low oxygen levels, and a large range of salt concentrations (Taylor 2012). Individuals can also leave the water (emersion) for extended periods of time to avoid these extreme conditions (Abel et al. 1987). The ability to survive such extreme environmental fluctuations coupled with their preference for self-fertilization allows for genetic variation to be maintained within a location as subpopulations with distinct genotypes.

There is substantial evidence that outcrossing occurs in most mangrove rivulus populations (Lubinski et al. 1995; Mackiewicz et al. 2006a), which is likely driven by male-hermaphrodite matings. Crossing between hermaphrodites has not been observed either in the laboratory or in the wild and therefore, it is likely that males must be present for an outcrossing event to occur (Turner et al. 2006; Furness et al. 2015). Males can arise in a population of *K. marmoratus* in two ways: either via temperature-dependent sex determination during embryogenesis or via temperature-dependent sex change (Harrington 1961; Harrington 1967; Harrington 1968; Turner et al. 2006; Ellison et al. 2015). Both mechanisms can lead to the

spontaneous occurrence of males in a population. The percentage of males varies among wild populations, with males making up between 2% (Florida) to 25% (Twin Cayes, Belize) of the population (Davis et al. 1990). This difference in sex ratios may lead to varying outcrossing rates among populations (Turner et al. 2006; Tatarenkov et al. 2015). While outcrossing events appear to be rare in most populations, the amount of outcrossing varies drastically by site, with some sites having a selfing rate of 80-90% (Florida populations) and others roughly 40% (Twin Cayes population), which further contributes to the varying levels of differentiation among populations (Mackiewicz et al. 2006a).

There are high levels of homozygosity within populations of *K. marmoratus* but high levels of differentiation among lineages, as shown by previous studies utilizing DNA fingerprinting (Turner et al. 1990) and microsatellite data (Tatarenkov et al. 2010). To date, no study has combined both mitochondrial and full genome sequencing data to study genomic variation in this species. These types of data are essential to determine whether diversity seen in only a few markers (e.g. microsatellites) adequately describes genetic diversity and, ultimately, to explore whether and through which evolutionary mechanisms heterozygosity produced through male-hermaphrodite outcrossing is maintained in the genome. This study utilizes high-throughput sequencing data to infer genetic relationships among 14 laboratory-reared lineages of *K. marmoratus* and one laboratory-reared lineage of the sister species *Kryptolebias hermaphroditus*, the only other preferentially self-fertilizing vertebrate. Additionally, this study aims to determine the levels of intra-individual heterozygosity using single nucleotide polymorphisms (SNPs) in lineages that have been maintained in the laboratory for as short as one and as many as 11 generations.

Material and Methods

Sample Collection

One sample from each of 14 isogenic lineages of *Kryptolebias marmoratus* and one sample from the sister species *Kryptolebias hermaphroditus* were included in the study (Table S1). It should be noted that recent genetic and taxonomic analyses showed that earlier names –*K. bonairensis* and *K. heyei*– are available for Caribbean populations of species designated here as *K. hermaphroditus* (Tatarenkov et al. 2017a). As a result, the GITMO sample will ultimately bear one of these names. However, because taxonomy of this species is still in a state of flux, here we chose to use the recognized name *K. hermaphroditus*. Data from the lineage RHL is the same as that used by Kelley et al. (2016) to construct the reference genome used in this study. All samples were obtained from laboratory stocks in the Earley laboratory. The lineages were sampled from throughout the species range (Figure 1, Table S1). Representative samples from each of the isogenic lineages were genotyped for 32 microsatellites (Figure S1); note that samples genotyped were not the same individuals as those used in this study. Individuals were euthanized with a lethal dose of sodium bicarbonate-buffered Finquel® (MS-222, tricaine methanesulfonate) and muscle tissue was dissected and flash frozen at -80°C.

Library Preparation and Sequencing

DNA was extracted from ~50 mg flash frozen tissue using the Qiagen Gentra Puregene Tissue kit with the following modifications: samples were placed in Covaris TT1 bags, immersed in liquid nitrogen then pulverized using a Covaris CryoPrep system. The pulverized tissue was then incubated at 56°C for 3 hours in 300 µl cell lysis buffer with 1.5 µl Proteinase K to complete cell lysis. All steps following cell lysis were performed per the Gentra Puregene

protocol. Genomic DNA was checked for high molecular weight content by running on a 2% agarose Invitrogen E-gel, and concentration was determined using a Thermo Fisher Qubit Fluorometer.

Sequencing libraries were fragmented using either sonication or Illumina Nextera tagmentation technology (Table S2). For the libraries prepared using sonication, 500 nanograms of genomic DNA was sheared to an average size of ~500 base pairs (bp) using a Covaris sonicator. The KAPA BioSystems Library Preparation Kit for Illumina was used for end repair, adapter ligation and amplification. Libraries were amplified with eight cycles of PCR using KAPA's recommended standard cycling conditions. Size selection was performed using a 0.6X Agencourt AMPure bead cleanup to select for an average fragment size of 400 bp. For the libraries prepared using the Illumina Nextera tagmentation technology, the standard protocol was followed except that the tagmentation was followed by size selection on a PerkinElmer Labchip XT 750. The quality of all libraries was assessed using an Agilent Bioanalyzer, and concentration was determined by Qubit. Libraries were then pooled to achieve an equimolar concentration of each library prior to sequencing at the Stanford Genome Sequencing Service Center on a HiSeq 2000 with the paired read 101 bp option.

Nuclear Genome Analysis – Data Processing and Analysis

Prior to mapping, raw reads were inspected using FastQC (Andrews 2010). The adaptors were trimmed with a minimum overlap of 5 bp, and the reads were trimmed based on the quality with a minimum value of 28 PHRED score, additionally depending on base composition the reads were trimmed at the 5'end of both reads using TrimGalore! (Table S2) (Krueger 2015). Coverage was estimated based on the mapped reads using Picard Tools

(http://broadinstitute.github.io/picard) (Table S2). Reads from each sample were mapped to the reference genome (GCA_001663955.1, (Kelley et al. 2016)) including the mitochondrial genome (Tatarenkov et al. 2017b) using the Burrows-Wheeler aligner algorithm BWA-MEM (Li 2013). The resulting SAM files were converted to BAM format using SAMTOOLS 1.2 (Li et al. 2009), and read group information was added using Picard Tools. Variants were called on each sample using the Genome Analysis Toolkit (GATK version 3.7) (Li 2013) HaplotypeCaller module. The files were combined using the combineGVCFs module, and then joint genotyping was performed across samples using GATK module GenotypeGVCFs. Sites for which a genotype could be determined (callable loci) across the genome were identified utilizing the GATK module CallableLoci, with minimum coverage of 4 and maximum coverage of 250. Single nucleotide polymorphisms (SNPs) were extracted using the GATK module SelectVariants. The SNPs were flagged using VariantFiltration in GATK using the following criteria: QD < 2.0, FS > 60.0, MQ < 40.0, MQRankSum < -12.5, ReadPosRankSum < -8.0. Sites that passed the filter criteria were kept using vcftools (v0.1.15) (Danecek et al. 2011). Sites missing more than 15% of genotypes were excluded from the analysis, which is equivalent to excluding sites where at least two individuals are missing genotypes.

Summary statistics were calculated using vcf-stats (Danecek et al. 2011). To calculate the ratio of heterozygous to homozygous sites, the number of heterozygous sites were divided by the number of homozygous non-reference sites for each lineage. A custom SnpEff (Cingolani et al. 2012) database was built utilizing the reference genome and associated exon annotations (Kelley et al. 2016). The VCF containing filtered variable sites was reannotated using SNPeff with locations (e.g. coding, noncoding) and putative impacts of SNPs (e.g., silent, missense, nonsense). Additionally, the transition to transversion ratio (Ts/Tv) and the missense to silent

mutation ratio were determined using SnpEff. Nucleotide sequences for the reference coding sequences were obtained with gffread v0.9.9 (Trapnell et al. 2010). Sequences were then annotated by BLASTx (Altschul et al. 1990) to the Swiss-Prot database (accessed 11/2016) (The UniProt 2017) with an e-value of 10^{-5} and the top 20 hits were retained. Overrepresented Gene Ontology (GO) terms were identified using Blast2GO (Gotz et al. 2008) Fisher's Exact Test, with a false discovery rate (FDR) less than 0.01 and using only genes that had greater than one SNP resulting in a missense mutation.

Runs of homozygosity (ROH) using the whole genome data were calculated using vcftools (Danecek et al. 2011)The runs of homozygosity were separated in three classes using k-means clustering in R. Only the regions with runs longer than 526 bp (first quartile) were used for the inference of the clusters. The classes were defined as: class 1 (greater than 526 bp and less than 64,626 bp), class 2 (greater than 65,476 bp and less than 242,872 bp), and class 3 (greater than 244,848 bp to the maximum value of 974,281 bp). Only runs in class 3 (244,848 - 974,281 bp) were considered as long runs of homozygosity. Because our reference genome has a large number of contigs and some of them were smaller than the runs of homozygosity in class 3, we limited our runs of homozygosity analyses to the contigs that were longer than 244,848 bp in length (minimum length of class 3). Equation 1 from (Szpiech et al. 2013) was used to calculate the total fraction of the genome covered by any ROH in each of the classes for each individual using the sum of bp in contigs longer than 244,848 bp as the total length of the genome. We estimated the correlation between the proportion of heterozygous sites (total number of heterozygous sites divided by callable sites per sample) and the number of generations that lineages were maintained in the laboratory (Table S2). Additionally, long ROH were correlated with generations maintained in the laboratory. For analyses that rely on the number of

generations in the laboratory, we excluded DAN2K, SLC8E and UNK because we were not able to confirm the number of generations.

We also performed a principal component analysis (PCA) using PLINK (1.07) (Purcell et al. 2007) with data that were thinned to exclude SNPs that were within 5 kb of each other using vcftools (Danecek et al. 2011) to minimize the effect of linkage disequilibrium. A final set of 140,081 SNPs was included in the PCA analysis. Identity by decent relatedness was calculated using vcftools (Manichaikul et al. 2010; Danecek et al. 2011).

Inference of Population Splits and Migrations from the Nuclear Genome

The VCF file containing filtered SNPs was converted to plink format using vcftools (Danecek et al. 2011). PLINK was used to determine allele frequencies for each lineage at each site (Chang et al. 2015). The allele frequencies were used to create TreeMix formatted file utilizing the plink2treemix python script included in the TreeMix package (v1.13) (Pickrell and Pritchard 2012). TreeMix was run using SNPs grouped in windows of 500, sample size correction was turned off, and GITMO was specified as the root. A bootstrap analysis with 1000 replicates was performed; bootstrap support (bs) throught the manuscript is presented as percentage. The TreeMix analysis was also performed with samples grouped by location.

Mitochondrial Phylogenetic Analysis

Mitochondrial genomes were assembled for 14 lineages of *K. marmoratus* and 1 lineage of *K. hermaphroditus* with ARC (Hunter et al. 2015), subsampling the raw genomic sequence reads to achieve approximately 30x coverage as per the ARC user manual (Table S2). Sequences for the 13 protein-coding genes were identified using MitoAnnotator (Iwasaki et al. 2013).

Mitochondrial phylogenetic analyses were performed using the 13 protein coding genes. The 15 samples from this study were combined with 4 samples obtained from GenBank (Kim et al. (2016) (accession number: NC_032387.1), Lee et al. (2001) (accession number: AF283503), Rhee et al. (2017) (accession number: PRJNA317650), and Tatarenkov et al. (2017b) (accession number: KT893707)). Nucleotide sequences for each gene were aligned using the default options in MUSCLE (Edgar 2004). Alignment files were concatenated using FASCONcat (Kück and Meusemann 2010). Two *K. hermaphroditus* samples were used as an outgroup: a new sequence from this study (GITMO) and one sequence from GenBank (accession number: NC_032387.1) (Kim et al. 2016). We assigned an independent model of nucleotide substitution to each gene, chosen using PartitionFinder 2.1.1 (Lanfear et al. 2016): Model K81+I (for position 1 in all genes), Model HKY+I (for position 2 in all genes), and Model TRN+G (for position 3 in all genes). We performed both maximum likelihood and Bayesian analyses on the mitochondrial genome dataset. Maximum likelihood analysis was performed using RAxML 8.2.9 (Stamatakis 2014). Node support was estimated using 1000 rapid bootstrap replicates. Bayesian analysis was conducted in MrBayes 3.2.6 (Ronquist and Huelsenbeck 2003) using default priors. The Markov chain Monte Carlo was run for 10 million generations sampling every 1000 generations, with two parallel runs each with four chains (three hot and one cold). Convergence was considered reached on the basis of the standard deviation of split frequencies (<0.01). The first 10% of trees were discarded as burn-in.

Results

There were 2,106,131 SNPs in the entire whole-genome resequencing dataset, 1,168,538 of which are variable in *K. marmoratus*. Using these SNPs, we determined the level of

heterozygosity present in *K. marmoratus* lineages that were identified as being completely homozygous using 32 microsatellite markers. The percent of heterozygous sites per individual ranged from 0.0305% (RHL) to 0.0554% (LION2) per callable region of the genome (Table S3). For *K. marmoratus,* the count of private alleles (alleles that are only found in one lineage) varied between 7,279 (UNK) and 46,557 (R2) and for *K. hermaphroditus* (GITMO) there were 914,466 private alleles (Table S3). The heterozygous to homozygous ratio ranged from 0.83 (R2) to 1.54 (SLC8E) (Figure S2). As the data for RHL was used to assemble the reference genome (Kelley et al. 2016), there are very few homozygous non-reference sites and RHL was excluded from the analysis of heterozygous to homozygous ratios.

The correlation between the percent of heterozygous sites and the number of generations in the laboratory was not significant (Figure S3; $R^2 = 0.14$, $p = 0.23$). There were 1,468,846 predicted effects of the variants determined by SnpEff from the 1,168,538 SNPs specific to *K. marmoratus*. While 9.8% of the genome is coding, only 3.5% of the predicted effects were found within coding regions. Additionally, 54.4% of the SNPs found within coding regions were missense and 44.7% were silent. The number of heterozygous sites in each individual ranged from 126,649 (VOL) to 251,480 (UNK) and the percent of heterozygous sites found within coding regions ranged from 4.417% (RHL) to 3.970% (Vol). Finally, the Ts/Tv ratio was calculated to be 1.76, which is slightly lower than the expected 2.0. There were 13 overrepresented Gene Ontology (GO) terms associated with genes that had greater than one SNP resulting in a missense mutation present (Table S4). The lineage with highest proportion of its genome covered by long runs of homozygosity (class 3) was RHL (Figure S4). The proportion of long runs of homozygosity in the genome shows a positive correlation with the number of generations in the laboratory (Figure S5; $R^2 = 0.67$, $p = 0.025$).

To investigate the genetic structure of the lineages we performed principal component analysis (PCA) using the nuclear SNPs and phylogenetic analyses, with both nuclear SNPs and mitochondrial sequences. The principal component analysis (PCA) separated the sister species *K. hermaphroditus* (GITMO) from *K. marmoratus* on PC1 (Figure S6; percentage of variance explained by PC1 = 17.21%, PC2 = 2.47%). In the PCA performed with only populations of *K. marmoratus*, the populations clustered by geographic location except for DAN2K (Figure 2; PC1 = 2.57%, PC2 = 1.6%). Additionally, this study found one mislabeled lineage, a problem which has been previously noted (Tatarenkov et al. 2010); based on the PCA, the lineage clusters closely with the Belizean lineage DAN2K. Lineages from Florida clustered closely together in the PCA space and therefore we estimated relatedness among the Florida lineages. The identity by descent probability relatedness (Φ) among only Florida *K. marmoratus* lineages shows that BBSC, FDS1, LION2, and SLC8E are closely related (Figure S7).

The maximum likelihood topology for the genomic SNP data given by TreeMix (Figure S8) groups the Florida lineages with the Bahamas lineage (bootstrap (bs) = 100). The lineage labeled UNK grouped with the lineage from Belize DAN2K (bs = 100). The Honduran lineages did not group together; HON grouped with the lineages from Belize (DAN2K, FW2, BWN3) and the lineage labelled UNK (bs = 100). R2 appeared as an early branching lineage, however there was no support for this position. Additionally, when grouping individuals by sampling location, Belize and Honduran lineages clustered together (bs = 95) and the lineage from the Bahamas is the first branch to appear (bs = 100) (Figure S9).

The topologies of the maximum likelihood and the Bayesian analyses of the mitochondrial genomes were identical based on 13 protein-coding genes with 11,436 sites, of those 579 are variable and 487 are parsimony informative (Figure 3, Figure S10). There were

several samples that were identical to each other (FW2 and BWN3; UNK and DAN2K; FDS1 and FDS08). All seven lineages from Florida (including FDS08 from Tatarenkov et al. (2017b)) and one from the Bahamas (RHL) grouped together in a single clade (bs = 57, Posterior Probability (pp) = 0.78). Samples from the studies by Lee et al. (2001) and Rhee et al. (2017) grouped together (bs = 100, pp = 1). The lineage labeled UNK grouped with the lineage from Belize (DAN2K; bs = 100, pp = 1). The sister lineage of all above mentioned lineages (BBSC, FDS1, LION1, LION2, LK1, VOL, RHL, Tatarenkov et al. (2017b), Lee et al. (2001), Rhee et al. (2017), UNK, and DAN2K) is the clade with the Honduran lineage HON9 and the two Belizean lineages FW2 and BWN3. R2 did not group with the other Honduran lineage (HON9), and instead was an early branching lineage among *K. marmoratus* lineages.

Discussion

Kryptolebias marmoratus has always been regarded as having high rates of selfing in the wild meaning that outcrossing has been considered as a minor component of the mating strategy, with the exception of lineages in Twin Cayes, Belize (Lubinski et al. 1995). These conclusions have been drawn by the fact that males exist at very low frequencies (Vrijenhoek 1985) and that microsatellite markers are often highly homozygous in wild populations (Turner et al. 1990). Given the homozygous nature of the microsatellites for lineages used in this study, the heterozygosity results show that there is previously undescribed variation in individuals of *K. marmoratus*. This variation is rare, as most SNPs were found as singletons and were found mainly in intergenic regions of the genome, with only 3.5% of the SNPs found in coding regions even though 9.8% of the genome is made up of coding sequences. Looking closer at the SNPs within the coding region, specifically focusing on genes that have greater than one missense

mutation, we found that SNPs resulting in a missense mutation are more likely to fall within genes associated with the immune system. This is the first genomic evidence in *K. marmoratus* of the importance of genetic variation in genes associated with the immune system and supports the findings of Ellison et al. (2011), which showed that outcrossed fish had a lower parasitic load than selfing lineages. Other studies showed that there appears to be considerable heterozygosity in major histocompability complex (MHC) genes and maintenance of diverse MHC supertypes, despite persistent homozygosity at neutral markers and non-MHC loci (Sato et al. 2002; Ellison et al. 2012). Collectively, these data suggest the potential for parasite and/or pathogen-mediated selection on the maintenance of genetic diversity within some regions of the genome. Many of the remaining genes that were enriched for more than one SNP resulting in a missense mutation are associated with reproduction and merit further study to determine the role they play in this self-fertilizing species.

The ratio of heterozygous to homozygous sites for some lineages indicates that outcrossing may be frequent among individuals of *K. marmoratus* and that it may play a role in maintaining variation within and among lineages (Figure S2). Although the ratios in *K. marmoratus* were lower than in a randomly mating population (expected heterozygous to homozygous ratio in randomly mating populations is 2 (Jun et al. 2012)), the values for several of these lineages were greater than 1.25. A possible interpretation is that rare outcrossing events occur between distant lineages, which is supported by the findings of Lomax et al. (2017) showing that in some seasons, egg laying increased as a function of increased genetic dissimilarity in *K. marmoratus*. Additionally, Ellison et al. (2013) showed that males prefer genetically dissimilar hermaphrodites, which should promote more outcrossing between genetically distinct versus genetically similar lineages . The extensive number of variable sites

346 provided by whole-genome data allowed us to uncover this heterozygosity that was not identified
347 previously by studies using microsatellites.

348 Similarly, we found that inbreeding, when estimated by the percentage of the genome
349 covered by long runs of homozygosity (ROH), was lower than expected. Long ROH are the
350 result of processes that reduce effective population size and increase homozygosity (Szpiech et
351 al. 2013; Curik et al. 2014), such as selfing. Because *K. marmoratus* has high levels of selfing,
352 we expected to find a large portion of the genome covered by long ROH. However, only a
353 maximum of 9.5% of the genome that is contiguous enough to determine long ROH was covered
354 by long ROH, which is lower than the values found in human populations that range between 1-
355 19% of the genome (Szpiech et al. 2013). Our reference genome is fragmented, with an N50
356 scaffold length of 111,539 bp (Kelley et al. 2016), which inherently decreases our ability to
357 identify ROH in this species, especially long ROH.

358 This study shows that Florida was colonized in one event, and indicates that the lineage
359 present in the Bahamas is closely related to the populations in Florida. The lineages from the
360 geographically close locations, Belize and Honduras, did not form a monophyletic clade in either
361 mitochondrial or individual nuclear analysis, suggesting a complex colonization of the region.
362 Although the swimming abilities of adult *K. marmoratus* are limited, and they typically live their
363 entire life within a short distance from where they hatched (Davis et al. 1990; Taylor 2012),
364 adults have been found within log hollows that presumably can float to distant areas during
365 storms (Mackiewicz et al. 2006a; Tatarenkov et al. 2007). Additionally, eggs are resistant to
366 desiccation making them capable of long distance dispersal on floating material, which could
367 explain how *K. marmoratus* can colonize new areas and explains how movement between
368 Florida and Bahamas can occur.

The phylogenetic distribution of *K. marmoratus* lineages derived from protein-coding mitochondrial sequences is quite consistent with lineages having dispersed with the Caribbean and Florida currents (see Figure 1). Indeed, several studies have identified migration events that are likely explained by prevailing ocean currents (Tatarenkov et al. 2017b). Given the currents in the Gulf of Honduras, it is thus not surprising to find complex phylogenetic relationships among the Belizean and Honduran lineages. The strong Florida current might explain why lineages from the Florida Keys (LION1, LION2, LK1), Bahamas (RHL), and the east coast of Florida (SLC8E, VOL) are closely related; it seems quite reasonable that eggs and fish could disperse on flotsam from southern Florida to the eastern peninsular coast and Bahamas. Lineages from western Florida (BBSC, FDS1) are most derived and most closely related to lineages from the Florida Keys, which might be explained by the relatively weak surface currents that move northward from the Keys to places like Fort Myers (BBSC) and Tampa (FDS1).

Inferences of population splits from the nuclear data provide additional insights into the complexity of the phylogeographic relationships. The population split inference when grouping samples from the same location, showed Belize and Honduras in the same clade. Moreover, even with the individual nuclear analysis, relationships among Florida lineages do not follow the prevailing currents as clearly. Discrepancies between the mitochondrial and nuclear data could suggest biased migration, incomplete lineage sorting, or a combination of the two. The overall relationship among individuals sampled in Florida, Bahama, Honduras, and Belize are concordant with topologies estimated in other studies such as Tatarenkov et al. (2017b) that used different lineages to the ones in this study.

Earlier studies suggested that the low levels of heterozygosity were due to the fact that there are few naturally occurring males, however, there was evidence for multiple isogenic

lineages at a specific sampling site and the identity of those varied from year to year. As additional data have been collected, there is clear evidence for outcrossing in the wild. Although our study has only examined one specimen per lineage, we were able to uncover variation at a finer scale than has been achieved by microsatellites. This result and the discrepancies in our phylogenetic analyses are strong motivation for further collection of whole genome data from more specimens throughout the *K. marmoratus* range. This will enable genome-wide comparisons within and among populations, which will elucidate whether outcrossing is as rare as previously thought (Vrijenhoek 1985) and/or whether outcrossing preferentially occurs between genetically dissimilar lineages as suggested by Ellison et al. (2013).

Acknowledgements:

Fish maintenance and colony production was approved by the University of Alabama Institutional Animal Care and Use Committee Protocol #08-312-3/13-10-0048. We would also like to thank the Keys Marine Laboratory (Long Key, FL), Blake Ross and Sarah Edwards (Lighthouse Reef, Long Caye, Belize); Phillip Hughes (National Key Deer Refuge, Big Pine Key, FL); and Yvonne Wielhouwer for logistical support in the field. Portions of this work were approved by the following permits to Ryan Earley - FFWCC Special Activity License SAL-09-1132-SR; Florida State Parks Permit 03071220; National Key Deer Refuge Permit 2010-008; and Belize Fisheries Permit #000039-11.

412 **References:**

413

414 Abel DC, Koenig CC, Davis WP. 1987. Emersion in the mangrove forest fish *Rivulus*
415 *marmoratus*: a unique response to hydrogen sulfide. *Environmental Biology of Fishes* **18**:
416 67-72.
417 Allard RW. 1975. The mating system and microevolution. *Genetics* **79 Suppl**: 115-126.
418 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool.
419 *J Mol Biol* **215**: 403-410.
420 Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. Available
421 online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.
422 Avise JC, Tatarenkov A. 2012. Allard's argument versus Baker's contention for the adaptive
423 significance of selfing in a hermaphroditic fish. *Proc Natl Acad Sci USA* **109**: 18862-
424 18867.
425 Burgarella C, Gayral P, Ballenghien M, Bernard A, David P, Jarne P, Correa A, Hurtrez-Bousses
426 S, Escobar J, Galtier N et al. 2015. Molecular Evolution of Freshwater Snails with
427 Contrasting Mating Systems. *Mol Biol Evol* **32**: 2403-2416.
428 Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation
429 PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**: 7.
430 Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM.
431 2012. A program for annotating and predicting the effects of single nucleotide
432 polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118;
433 iso-2; iso-3. *Fly (Austin)* **6**: 80-92.
434 Curik I, Ferencakovic M, Solkner J. 2014. Inbreeding and runs of homozygosity: A possible
435 solution to an old problem. *Livest Sci* **166**: 26-34.
436 Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G,
437 Marth GT, Sherry ST. 2011. The variant call format and VCFtools. *Bioinformatics* **27**:
438 2156-2158.
439 Davis WP, Taylor DS, Turner B. 1990. Field observations of the ecology and habits of mangrove
440 rivulus (*Rivulus marmoratus*) in Belize and Florida (Teleostei: Cyprinodontiformes:
441 Rivulidae). *Ichthyol Explor of Freshwaters* **1**: 123-134.
442 Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high
443 throughput. *Nucleic Acids Res* **32**: 1792-1797.
444 Ellison A, Allainguillaume J, Girdwood S, Pachebat J, Peat KM, Wright P, Consuegra S. 2012.
445 Maintaining functional major histocompatibility complex diversity under inbreeding: the
446 case of a selfing vertebrate. *Proc Biol Sci* **279**: 5004-5013.
447 Ellison A, Cable J, Consuegra S. 2011. Best of both worlds? Association between outcrossing
448 and parasite loads in a selfing fish. *Evolution* **65**: 3021-3026.
449 Ellison A, Jones J, Inchley C, Consuegra S. 2013. Choosy males could help explain androdioecy
450 in a selfing fish. *The American naturalist* **181**: 855-862.
451 Ellison A, López CMR, Moran P, Breen J, Swain M, Megias M, Hegarty M, Wilkinson M,
452 Pawluk R, Consuegra S. 2015. Epigenetic regulation of sex ratios may explain natural
453 variation in self-fertilization rates. In *Proc R Soc B*, Vol 282, p. 20151900. The Royal
454 Society.
455 Furness AI, Tatarenkov A, Avise JC. 2015. A genetic test for whether pairs of hermaphrodites
456 can cross-fertilize in a selfing killifish. *J Hered* **106**: 749-752.

Gotz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talon M, Dopazo J, Conesa A. 2008. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res* **36**: 3420-3435.

Harrington RW, Jr. 1967. Environmentally controlled induction of primary male gonochorists from eggs of the selffertilizing hermaphroditic fish, Rivulus marmoratus Poey. *The Biological Bulletin* **132**: 174-199.

Harrington RW, Jr. 1968. Delimitation of the thermolabile phenocritical period of sex determination and differentiation in the ontogeny of the normally hermaphroditic fish *Rivulus marmoratus* Poey. *Physiological Zoology* **41**: 447-460.

Harrington RW, Jr. . 1961. Oviparous hermaphroditic fish with internal self-fertilization. *Science* **134**: 1749-1750.

Hunter SS, Lyon RT, Sarver BA, Hardwick K, Forney LJ, Settles ML. 2015. Assembly by Reduced Complexity (ARC): a hybrid approach for targeted assembly of homologous sequences. *bioRxiv*: 014662.

Iwasaki W, Fukunaga T, Isagozawa R, Yamada K, Maeda Y, Satoh TP, Sado T, Mabuchi K, Takeshima H, Miya M. 2013. MitoFish and MitoAnnotator: A mitochondrial genome database of fish with an accurate and automatic annotation pipeline. *Mol Biol Evol* **30**: 2531-2540.

Jun G, Flickinger M, Hetrick KN, Romm JM, Doheny KF, Abecasis GR, Boehnke M, Kang HM. 2012. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am J Hum Genet* **91**: 839-848.

Kelley JL, Yee M-C, Brown AP, Richardson RR, Tatarenkov A, Lee CC, Harkins TT, Bustamante CD, Earley RL. 2016. The genome of the self-fertilizing mangrove rivulus fish, *Kryptolebias marmoratus*: a model for studying phenotypic plasticity and adaptations to extreme environments. *Genome Biology and Evolution* **8**: 2145-2154.

Kim H-S, Hwang D-S, Hagiwara A, Sakakura Y, Lee J-S. 2016. Complete mitochondrial genome of the mangrove killifish Kryptolebias hermaphroditus (Cyprinodontiformes, Rivulidae). *Mitochondrial DNA Part B* **1**: 540-541.

Krueger F. 2015. Trim Galore. *A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files*.

Kück P, Meusemann K. 2010. FASconCAT: Convenient handling of data matrices. *Molecular Phylogenetics and Evolution* **56**: 1115-1118.

Lanfear R, Frandsen PB, Wright AM, Senfeld T, Calcott B. 2016. PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Mol Biol Evol* **34**: 772-773.

Lee J-S, Miya M, Lee Y-S, Kim CG, Park E-H, Aoki Y, Nishida M. 2001. The complete DNA sequence of the mitochondrial genome of the self-fertilizing fish *Rivulus marmoratus* (Cyprinodontiformes, Rivulidae) and the first description of duplication of a control region in fish. *Gene* **280**: 1-7.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:13033997*.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**: 2078-2079.

Lomax JL, Carlson RE, Wells JW, Crawford PM, Earley RL. 2017. Factors affecting egg production in the selfing mangrove rivulus (Kryptolebias marmoratus). *Zoology* **122**: 38-45.

Lubinski BA, Davis WP, Taylor DS, Turner BJ. 1995. Outcrossing in a natural population of a self-fertilizing hermaphroditic fish. *J Hered* **86**: 469-473.
Mackiewicz M, Tatarenkov A, Taylor DS, Turner BJ, Avise JC. 2006a. Extensive outcrossing and androdioecy in a vertebrate species that otherwise reproduces as a self-fertilizing hermaphrodite. *Proc Natl Acad Sci USA* **103**: 9924-9928.
Mackiewicz M, Tatarenkov A, Turner BJ, Avise JC. 2006b. A mixed-mating strategy in a hermaphroditic vertebrate. *Proc Biol Sci* **273**: 2449-2452.
Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. 2010. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**: 2867-2873.
Molloy PP, Nyboer EA, Côté IM. 2011. Male–Male Competition in a Mixed-Mating Fish. *Ethology* **117**: 586-596.
Noel E, Jarne P, Glemin S, MacKenzie A, Segard A, Sarda V, David P. 2017. Experimental Evidence for the Negative Effects of Self-Fertilization on the Adaptive Potential of Populations. *Curr Biol* **27**: 237-242.
Pickrell JK, Pritchard JK. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet* **8**: e1002967.
Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* **81**: 559-575.
Rhee J-S, Choi B-S, Kim J, Kim B-M, Lee Y-M, Kim I-C, Kanamori A, Choi I-Y, Schartl M, Lee J-S. 2017. Diversity, distribution, and significance of transposable elements in the genome of the only selfing hermaphroditic vertebrate *Kryptolebias marmoratus*. *Scientific Reports* **7**: 40121.
Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**: 1572-1574.
Sato A, Satta Y, Figueroa F, Mayer WE, Zaleska-Rutczynska Z, Toyosawa S, Travis J, Klein J. 2002. Persistence of Mhc heterozygosity in homozygous clonal killifish, Rivulus marmoratus: implications for the origin of hermaphroditism. *Genetics* **162**: 1791-1803.
Shimizu KK, Tsuchimatsu T. 2015. Evolution of Selfing: Recurrent Patterns in Molecular Adaptation. *Annual Review of Ecology, Evolution, and Systematics* **46**: 593-622.
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312-1313.
Szpiech ZA, Xu J, Pemberton TJ, Peng W, Zollner S, Rosenberg NA, Li JZ. 2013. Long runs of homozygosity are enriched for deleterious variation. *Am J Hum Genet* **93**: 90-102.
Tatarenkov A, Earley RL, Perlman BM, Taylor DS, Turner BJ, Avise JC. 2015. Genetic Subdivision and Variation in Selfing Rates Among Central American Populations of the Mangrove Rivulus, Kryptolebias marmoratus. *J Hered* **106**: 276-284.
Tatarenkov A, Earley RL, Taylor DS, Avise JC. 2012. Microevolutionary distribution of isogenicity in a self-fertilizing fish (*Kryptolebias marmoratus*) in the Florida Keys. *Integrative and comparative biology*: ics075.
Tatarenkov A, Gao H, Mackiewicz M, Taylor DS, Turner BJ, Avise JC. 2007. Strong population structure despite evidence of recent migration in a selfing hermaphroditic vertebrate, the mangrove killifish (Kryptolebias marmoratus). *Mol Ecol* **16**: 2701-2711.
Tatarenkov A, Lima SMQ, Earley RL, Berbel-Filho WM, Vermeulen FBM, Taylor DS, Marson K, Turner BJ, Avise JC. 2017a. Deep and concordant subdivisions in the self-fertilizing

mangrove killifishes (Kryptolebias) revealed by nuclear and mtDNA markers. *Biological Journal of the Linnean Society* **122**: 558-578.

Tatarenkov A, Mesak F, Avise JC. 2017b. Complete mitochondrial genome of a self-fertilizing fish *Kryptolebias marmoratus* (Cyprinodontiformes, Rivulidae) from Florida. *Mitochondrial DNA Part A* **28**: 244-245.

Tatarenkov A, Ring BC, Elder JF, Bechler DL, Avise JC. 2010. Genetic composition of laboratory stocks of the self-fertilizing fish *Kryptolebias marmoratus*: a valuable resource for experimental research. *PLoS One* **5**: e12863.

Taylor DS. 2012. Twenty-four years in the mud: what have we learned about the natural history and ecology of the mangrove rivulus, *Kryptolebias marmoratus*? *Integr Comp Biol* **52**: 724-736.

The UniProt C. 2017. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* **45**: D158-D169.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology* **28**: 511-515.

Turner BJ, Elder JF, Laughlin TF, Davis WP. 1990. Genetic variation in clonal vertebrates detected by simple-sequence DNA fingerprinting. *Proc Natl Acad Sci USA* **87**: 5653-5657.

Turner BJ, Fisher MT, Taylor DS, Davis WP, Jarrett BL. 2006. Evolution of 'maleness' and outcrossing in a population of the self-fertilizing killifish, *Kryptolebias marmoratus*. *Evol Ecol Res* **8**: 1475-1486.

Vrijenhoek R. 1985. Homozygosity and interstrain variation in the self-fertilizing hermaphroditic fish, *Rivulus marmoratus*. *J Hered* **76**: 82-84.

576 **Figure Captions**

577

578 **Figure 1.** Sampling locations for lineages of *Kryptolebias marmoratus* (BBSC, BWN3, DAN2K,
579 FDS1, FW2, HON9, LION1, LION2, LK1, R2, RHL, SLC8E, and VOL), and *Kryptolebias*
580 *hermaphroditus* (GITMO). Prevailing ocean currents are indicated by colored arrows.
581 The Caribbean current (green arrows) pushes northward from South America to the eastern tip of
582 Honduras and, further, to the Yucatan peninsula. The Gulf of Honduras lies to the west of the
583 Caribbean current and is characterized by a circular pattern of surface water movement that
584 heads southward along the coast and barrier islands of Belize and eastward toward Roatan and
585 Utila islands (black arrows). The Caribbean current extends northward and transitions to the
586 Florida current (red arrows), which rips along the Florida Keys and along the eastern coast of the
587 peninsula. The Gulf of Mexico current (blue arrows) also extends from the Caribbean current.

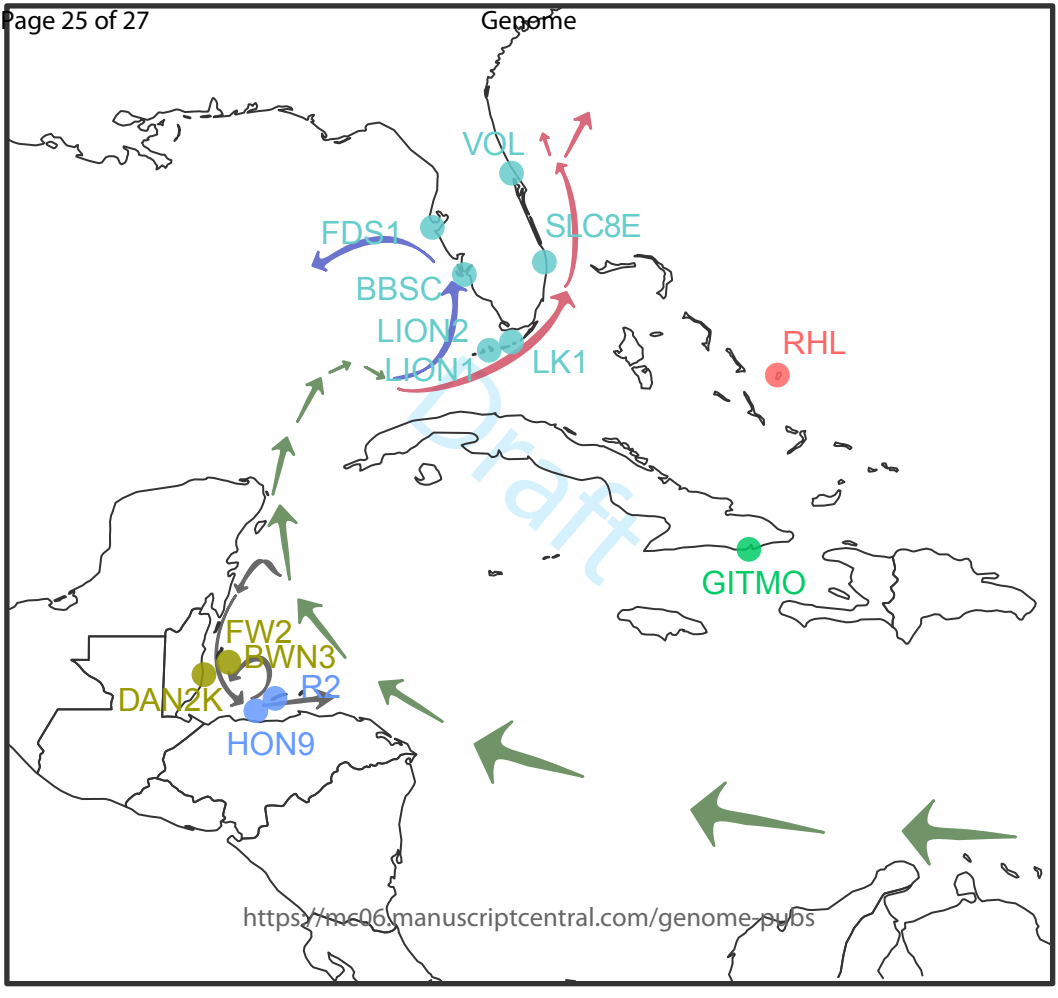
588

589 **Figure 2.** Principal Component Analysis (PCA) of genetic variation among samples of *K.*
590 *marmoratus*. PCA performed using PLINK (Purcell et al. 2007). To minimize the effects of
591 linkage disequilibrium, the data was thinned to exclude SNPs that were within 5 kb of each other
592 using vcftools (Danecek et al. 2011). A total of 140,081 SNPs were included in the analysis.
593 Principal component 1 explains 2.57 % of the variation and principal component 2 explains
594 1.60 % of the variation.

595

596 **Figure 3.** Mitochondrial maximum likelihood phylogenetic estimation of 13 protein coding
597 genes of the lineages of *K. marmoratus*; the sister species *K. hermaphroditus* was used as
598 outgroup (GITMO and Kim *et al*. 2016). Lineages are colored according to sampling location.

599 Node values are bootstrap and posterior probabilities, respectively. Posterior probabilities were
600 calculated for a Bayesian phylogentic inference and mapped to this maximum likelihood tree.
601 Bootstrap values below 50 and posterior probabilities below 0.5 are represented by *.
602


VOL
FDS1
SLC8E
BBSC
LION2
LION1
LK1
RHL
GITMO
FW2
BWN3
DAN2K
R2
HON9