The revised reference genome of the leopard gecko (*Eublepharis macularius*) provides insight into the considerations of genome phasing and assembly

Brendan J. Pinto[1,2,3,*], Tony Gamble[3,4,5,*], Chase H. Smith[6], Shannon E. Keating[4], Justin C. Havird[6], Ylenia Chiari[7,*]

1. School of Life Sciences, Arizona State University, Tempe, AZ USA
2. Center for Evolution and Medicine, Arizona State University, Tempe, AZ USA
3. Department of Zoology, Milwaukee Public Museum, Milwaukee, WI USA
4. Department of Biological Sciences, Marquette University, Milwaukee WI USA
5. Bell Museum of Natural History, University of Minnesota, St Paul, MN USA
6. Department of Integrative Biology, University of Texas at Austin, Austin, TX, USA
7. Department of Biology, George Mason University, Fairfax, VA, USA

*Correspondence emails:
brendanjohnpinto@gmail.com
tgamble@geckoevolution.org
ychiari@gmu.edu

Abstract

Genomic resources across squamate reptiles (lizards and snakes) have lagged behind other vertebrate systems and high-quality reference genomes remain scarce. Of the 23 chromosome-scale reference genomes across the order, only 12 of the ~60 squamate families are represented. Within geckos (infraorder Gekkota), a species-rich clade of lizards, chromosome-level genomes are exceptionally sparse representing only two of the seven extant families. Using the latest advances in genome sequencing and assembly methods, we generated one of the highest quality squamate genomes to date for the leopard gecko, *Eublepharis macularius* (Eublepharidae). We compared this assembly to the previous, short-read only, *E. macularius* reference genome published in 2016 and examined potential factors within the assembly influencing contiguity of genome assemblies using PacBio HiFi data. Briefly, the read N50 of the PacBio HiFi reads generated for this study was equal to the contig N50 of the previous *E. macularius* reference genome at 20.4 kilobases. The HiFi reads were assembled into a total of 132 contigs, which was further scaffolded using HiC data into 75 total sequences representing all 19 chromosomes. We identified that 9 of the 19 chromosomes were assembled as single contigs, while the other 10 chromosomes were each scaffolded together from two or more contigs. We qualitatively identified that percent repeat content within a chromosome broadly affects its assembly contiguity prior to scaffolding. This genome assembly signifies a new age for squamate genomics where high-quality reference genomes rivaling some of the best vertebrate genome assemblies can be generated for a fraction previous cost estimates. This new *E. macularius* reference assembly is available on NCBI at JAOPLA010000000. The genome version and its associated annotations are also available via this Figshare repository https://doi.org/10.6084/m9.figshare.20069273.

Introduction

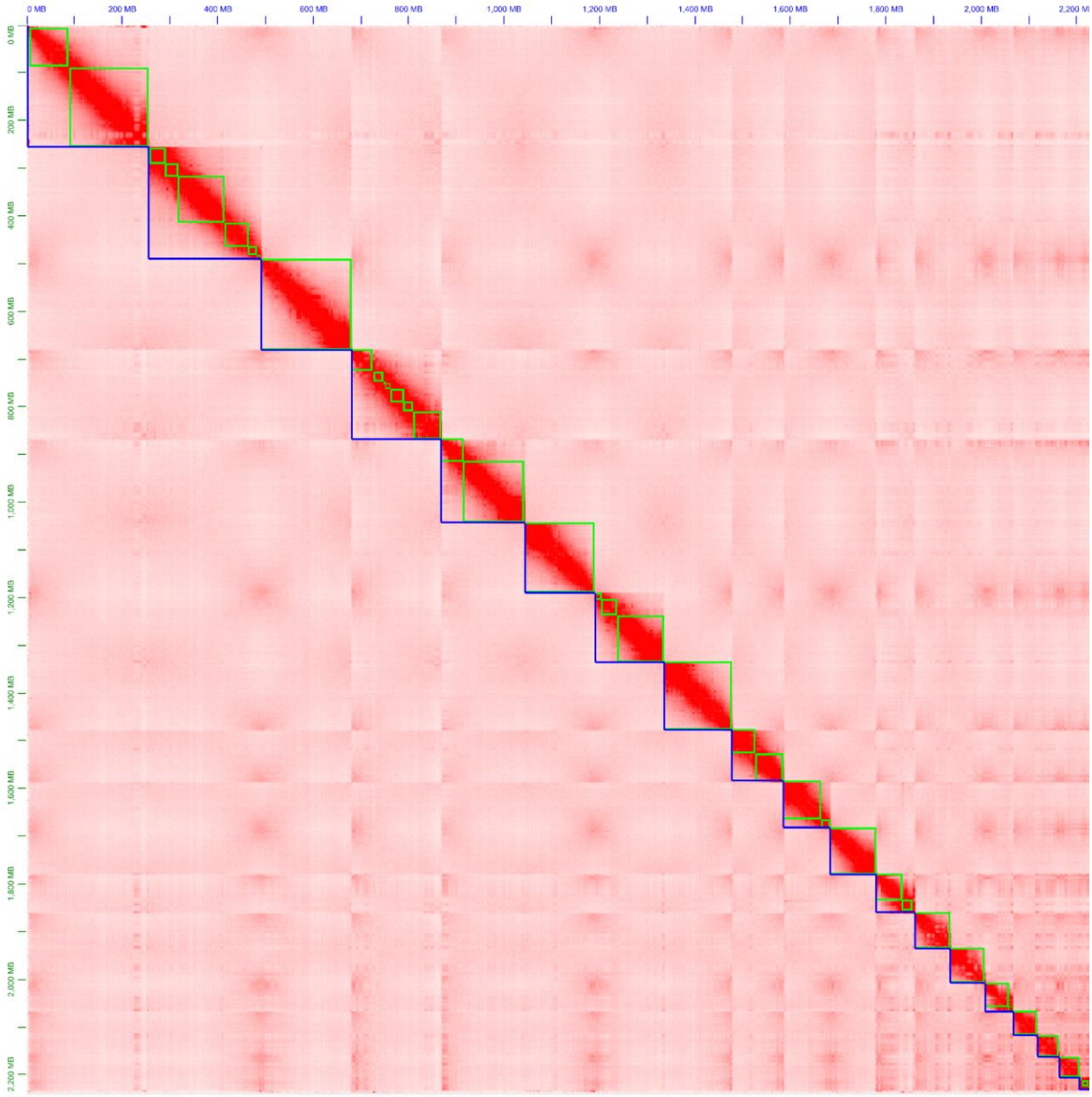
Genomic data in squamate reptiles (lizards and snakes) has lagged behind other vertebrate model systems, such as birds and mammals, and high-quality reference genomes remain scarce (Bravo et al. 2022; Hotaling et al. 2021; Pinto et al. 2023). Of the 23 previously published chromosome-scale squamate reference genomes, only 12 of the ~60 families are represented. Within geckos, chromosome-level genomes are exceptionally sparse, representing only two of the seven extant gecko families (Pinto et al. 2022; Yamaguchi et al. 2021). While there are also a handful of non-chromosome level gecko genomes, including the leopard gecko, *Eublepharis macularius* (Gekkota: Eublepharidae), draft genomes are limited in their utility to address many ecological and evolutionary hypotheses (Xiong et al. 2016). Adding to the current genomics resources in geckos, we used the latest advances in genome sequencing and assembly methods to generate one of the highest quality squamate reference genomes to date for *E. macularius*.

Investigating the evolution of genomes and phenotypes involves examining multiple species in a phylogenetic context. The foundation of integrative and comparative biology is that one can infer the likely ancestral condition for a specific trait, such as gene structure or function, by carefully choosing species that span the deepest bifurcations of a particular clade of interest (Felsenstein, 1985; Bryant and Russell, 1992; Witmer, 1995; Pagel et al. 2004). This idea is also justification of using model species (Wake, 2008; Hall, 2012; Dobzhansky, 1973; Sanger and Rajakumar, 2019). Consequently, gecko lizards (Infraorder Gekkota) are prime candidates to be a powerful model system for vertebrate genomics. Geckos are a species rich group of lizards—2,186 species as of December 2022—distributed in tropical and subtropical regions around the world (Uetz et al. 2021; Bauer, 2013). Geckos make up a large part of amniote diversity, representing ~8% of total species. They are the sister clade to all other lizards and snakes, with the possible exception of the poorly known, limbless, dibamids—whose phylogenetic position remains unresolved (Wiens et al. 2012; Zheng and Wiens, 2016; Townsend et al. 2004). Indeed, geckos diverged from all other lizards and snakes over 250 million years ago and extant geckos began to diversify ~120

million years ago (Gamble et al. 2011 & 2015a). For scale, this makes geckos as divergent from other squamates as humans are from a platypus (Kumar et al. 2017). Therefore, the inclusion of geckos in any evolutionary study of squamates is crucial to understanding genome evolution in lizards and snakes more broadly. As such, a high-quality genome assembly from a gecko is an important resource for investigating amniote genome evolution. Geckos are also interesting in their own right, for example, geckos possess unique biological traits, many of which have evolved repeatedly within the group, including adhesive toepads, sex determination systems, and photic activity patterns (Gamble et al. 2012; 2015a; 2015b; 2019; Pinto et al. 2019); and deep investigations into these, and other, aspects of gecko biology requires robust genomic resources.

Since its humble beginnings as a charismatic staple in the international pet trade, *Eublepharis macularius* has become a standard laboratory model system for studying a variety of biological questions surrounding tissue regeneration, coloration, sex determination, behavior, and cancer (Whimster, 1965; Viets et al., 1993; McLean & Vickaryous, 2011; Sakata et al., 2022; Delorme et al., 2012; Kiskowski et al., 2019; Szydłowski et al. 2020; Glimm et al. 2021; Guo et al. 2021; Agarwal et al. 2022; Katlein et al. 2022). However, more detailed investigations into genotype-phenotype associations in *E. macularius* have been hampered by modest genomic resources (Chernyavskaya et al. 2022; Gamble, 2019; Nurk et al. 2022). Thus, available genomic resources remain a research limitation and a high-quality reference genome for this model taxon will reduce potential error in downstream inference (Kim et al. 2019). Here, we generated a phased, chromosome-level genome assembly using a combination of Pacific Biosciences® High Fidelity (PacBio HiFi) and Dovetail® Omni-C (HiC) data. This assembly stands as one of the best primary assemblies for any squamate (132 contigs), and perhaps any vertebrate, standing alongside the highest-quality assemblies like the newest Telomere-to-Telomere assembly for humans (Nurk et al. 2022).

Figure 1: HiC contact map for the MPM_Emac_v1.0 assembly. Each blue segment indicates the delimitation of a chromosome-length scaffold, while the internal green squares indicate contigs. Approximately half of the assembled chromosomes are represented by a single contig, indicating the extreme contiguity of the primary assembly pre-scaffolding.



Methods

Data Generation

The reference genome individual was an unsexed, juvenile *E. macularius* (TG4126) incubated at a female temperature. This individual was homozygous for the recessive Tremper strain albino allele and heterozygous for the incomplete dominant Lemon Frost allele. We also sequenced the parents for phasing using the trios approach: the sire (TG4151) was homozygous for the recessive Tremper strain albino allele, heterozygous for the recessive Murphy patternless allele, and a Tremper giant – a phenotype with unknown genetics; the dam (TG4152) was homozygous for the recessive Tremper strain albino allele, heterozygous for the recessive Murphy patternless allele, and heterozygous for the incomplete dominant Lemon Frost allele (Supplemental Figure 1). We extracted high molecular weight DNA from blood of the offspring by Salting Out Phenol-Chloroform with an Ethanol Precipitation (SOP-CEPC; Pinto et al. 2021) and sent DNA to the Genomics & Cell Characterization Core Facility (GC3F) at the University of Oregon. A single PacBio HiFi library was sequenced across 3 SMRT cells. We generated a HiC library from the same individual using a DoveTail® Omni-C kit (Cantata Bio; Cambridge, MA, USA) and sequenced on an Illumina NovaSeq 6000 at the Texas A&M Agrilife Core Facility (College Station, TX, USA). DNA from the sire and dam were extracted using the Qiagen® DNeasy Blood and Tissue kit and sequenced on an Illumina NovaSeq 60000 by Novogene® (Beijing, China). All Illumina® sequencing was run using PE 150 bp (2x150) reads .

Genome Assembly

We generated five total assemblies from using offspring HiFi data using HiFiasm [v0.16.1-r375] (Cheng et al. 2021 & 2022). The assemblies were as follows (1) an aggregate contig assembly using all HiFi reads (this assembly was chosen for further scaffolding of the reference genome), (2-3) the trio phased assemblies (paternal and maternal) using parental short reads, (4-5) the HiC phased assemblies (haplotypes 1 and 2). Although there were subtle differences in contiguity between primary haplotype assemblies, they were overall very similar in quality (Table 1). After the initial contig assemblies, we scaffolded contig set (1) using the offspring HiC data in 3D-DNA

[v201008] (Dudchenko et al. 2017). We visualized the final HiC contact map for misassemblies and, with no large-scale misassemblies visible, we manually refined the contact map using Juicebox Assembly Tools [v1.11] (Durand et al. 2016).

Table 1: Assembly metrics comparing previous different phasing schemes for the assembled HiFi reads. * = merqury scores

Metric	Maternal hap	Paternal hap	HiC-hap1	HiC-hap2
Genome size (bp)	2,218,286,779	2,216,931,080	2,225,573,295	2,220,343,595
Total contigs	240	290	311	205
Largest contig (bp)	96,916,709	183,432,273	101,928,423	140,659,604
Contig N50 (bp)	33,011,424	37,464,150	30,021,720	32,760,080
Contig L50	22	17	24	21
GC content %	44.03%	44.03%	44.04%	44.04%
Completeness*	91.21%	91.20%	91.43%	91.45%
QV (Phred)*	64.517	64.077	64.210	64.163

Genome QC

We estimated metrics of genomic completeness using the raw sequencing reads and a database of conserved single-copy orthologs with merqury [v1.3.0] (Rhie et al. 2020) and Benchmarking Universal Single-Copy Orthologs (BUSCO) [v5.1.2] (Simão et al. 2015), respectively. We implemented all BUSCO analyses using the gVolante web server [v2.0.0] (Nishimura et al. 2017) with the Core Vertebrate Genes (CVG) and Sauropsida_odb10 databases. We also compared the relative ability of the two methods of phasing available in Hifiasm, parental data (trios) or chromatin-contact data (HiC) using merqury. We counted kmers for the offspring HiFi data, as well as the parental Illumina data using meryl [v1.3]. To calculate genomic heterozygosity of parental samples, we mapped reads to the reference using bwa-mem2 [v2.2.1] (Vasimuddin et al. 2019) and called SNPs using freebayes [v1.3.5] (Garrison and Marth, 2012). We

removed non-biallelic sites, sites with <30 quality score, and sites with a read depth <5 using vcftools [0.1.14-12] (Danecek et al. 2011).

Genome Annotation

We masked the assembly for repeats using a combination of RepeatModeler [v2.0.3] and RepeatMasker [v4.1.2] (Flynn et al. 2020; Smit et al. 2013). Later statistical analysis of the genomic repeat content used Mann-Whitney-Wilcoxon tests (Figure 2). For the initial release of the genome assembly, we chose to liftover the annotations from the previous reference genome (Xiong et al. 2016), given the absence of additional RNAseq data for *Eublepharis macularius*. Previous annotations were transferred using Liftoff [v1.6.3] (Li, 2018; Shumate & Salzberg, 2020). We diagnosed success by comparing the transferred annotations to the original annotation using BUSCO (Table 2).

Results and Discussion

The HMW DNA extraction optimized specifically for extremely small sample inputs (Pinto et al. 2021 & 2022), e.g. *Sphaerodactylus* gecko tissues, worked well in our *E. macularius* tissue samples. DNA extractions had an average molecule length of ~52 kilobase-pairs (kb) and 45% of the total extraction >50kb, much longer than the input for PacBio HiFi DNA sequencing allowed for pre-library preparation shearing optimization (Supplemental Figure 2). Post-circular consensus sequencing (CCS) correction, we recovered 66Gb of data (~30X coverage) with an average read length of 19.6kb and a read N50 of 20.4kb. With a read length N50 equal to the contig N50 of the previous *E. macularius* reference genome, the primary assembly generated by HiFiasm contained 132 total contigs with a N50 and L50 of 80,105,973bp and 9, respectively. The smallest contig was 2,547bp and the largest was 188,850,821bp. Scaffolding using HiC data added 61 gaps to produce the final assembly. The final *E. macularius* genome assembly contained 19 primary chromosome-level scaffolds and 56 unanchored contigs, ranging from ~400kb to ~11kb (75 total sequences), with a scaffold N50 and L50 of 145,573,841bp and 6, respectively.

Figure 2: Comparison between chromosomes assembled as a single contig ("Contigs", dark gray) and those composed of multiple contigs ("Scaffolds", light gray) displayed using vioplot (Adler et al. 2022). The violins represent the distribution of the underlying data points, while the internal bars represent a traditional bar graph representation. The mid-lines represent the median of the data. Our *a priori* hypothesis was that chromosomes assembled as a single contig would possess lower overall GC content and/or repeat content. However, neither GC content or repetitive element content were significantly different using Mann-Whitney-Wilcoxon tests. Qualitatively, there appears to be a difference in median repeat content between the two groups. It's possible that our ability to detect a true difference using frequentist methods lies in the low sample size (N=19).

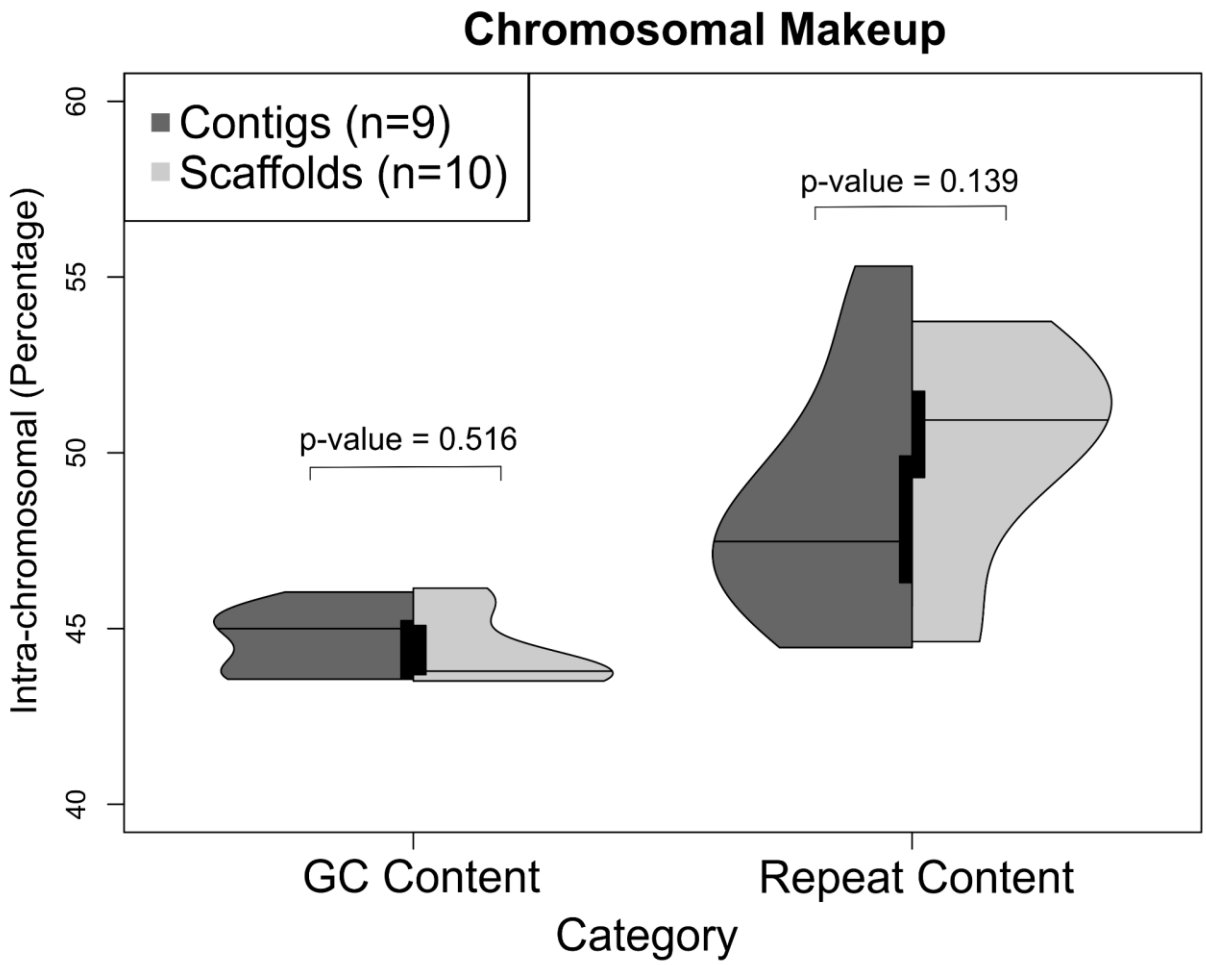


Table 2: Assembly metrics comparing previous *Eublepharis macularius* reference genome (Xiong et al. 2016) to MPM_Emac_v1.0 reference assembly. BUSCO scores abbreviated as follows, C=complete, S=complete single copy, D=complete multi-copy, F=fragmented, M=missing.
* missing data not included in calculation

Metric	Xiong et al. 2016	MPM_Emac_v.1.0
Contig N50 (bp)*	20,426	80,105,973
Contig L50*	27,671	9
Scaffold N50 (bp)	663,762	145,573,841
Scaffold L50	796	6
Repeat content %	42.1%	48.0%
GC content %	43.6%	44.1%
Total scaffolds	206,349	75
Genome Size (bp)*	1,981,651,937	2,237,374,058
BUSCO Sauropsida (Genome)	C:93.4%[S:91.4%,D:2.0%], F:2.0%,M:4.6%,n:7480	C:95.1%[S:93.0%,D:2.1%],F:0.9%,M:4.0%,n:7480
BUSCO Sauropsida (Annotation)	C:90.6%[S:88.7%,D:1.9%], F:3.2%,M:6.2%,n:7480	C:89.3%[S:87.6%,D:1.7%],F:3.6%,M:7.1%,n:7480
BUSCO CVG (Genome)	C:96.6%[S:95.7%,D:0.9%], F:2.1%,M:1.3%,n:233	C:99.2%[S:97.9%,D:1.3%],F:0.9%,M:0.1%,n:233

RepeatModeler identified 47.98% of the genome as repetitive (Table 2). Most repetitive elements in the genome remain unclassified (21.1%), followed by a majority being retroelements, either LINEs (14.31%), LTR elements (4.96%), or SINEs (4.06%). All other categories combined totaled <4% of the total repetitive elements, including DNA transposons (1.91%). We calculated the merqury completeness score at 91.5% using the PacBio HiFi reads used to generate the primary assembly, suggesting our assembly was largely complete. The BUSCO completeness scores were comparably valued at 99.2% and 95.1% using the Core Vertebrate Genome (CVG) and Sauropsida ortholog databases, respectively.

Nine of the 19 chromosomes were assembled as single contigs (Figure 1). With such high contiguity of the primary assembly prior to scaffolding, we further examined which chromosomes we assembled as a single contig. Centromeres are a typical assembly breakpoint (Peona et al. 2020) in genome assemblies but *E. macularius* has a karyotype consisting of 19 pairs of acrocentric chromosomes gradually decreasing in size (Gorman, 1973). Like other geckos, *E. macularius* chromosomes possess no sharp divide between macro- and micro-chromosomes (Figure 1; Pinto et al. 2023). Thus, greater contiguity may be due, in part, to this chromosomal arrangement. We aimed to identify any properties of individual chromosomes that might explain the contiguity of the assembled molecules. We binned chromosome-length scaffolds into single-contig and multi-contig categories and compared these groups with relation to their GC content and repetitive DNA content (Figure 2). We chose to use these statistics considering GC content is often considered a proxy for DNA stability and repetitive elements have been consistently demonstrated to cause assembly gaps (Eyre-Walker and Hurst, 2001; Peona et al. 2020). Interestingly, we did not find any significant differences between the two groups for either comparison. We did observe a higher repetitive element content in multi-contig scaffolds, as we expected, but the difference was not found to be significant ($p = 0.14$). However, the lack of significance could easily be explained by the small sample size (i.e. N=19, the number of chromosomes present) and qualitatively there may be a lower repeat content in those chromosomes assembled as a single contig than those scaffolded together as multiple contigs (Figure 2). Alternatively, these (rare) assembly gaps may be caused by additional genomic elements that are beyond the scope of the present study.

There is an inherent tradeoff to account for when planning a genome assembly and phasing experiment. Indeed, low heterozygosity tends to improve contiguity of the final assembly, but heterozygosity is a necessary component to successful phasing (e.g. Chin et al., 2016; Koren et al. 2018). We investigated the phasing capabilities of the parent/offspring trio approach to a single-individual with HiC data in *E. macularius*, an animal with low overall heterozygosity and no sex chromosomes . Perhaps

surprisingly, HiC outperformed the Trios method for phasing, where phase blocks are approximately equal to contig sizes (Figure 3A5-6, B-5-6). However, neither method provided complete haplotype-resolution with either (1) high switch error rates disrupting contig phasing (Trios, Figure A3-4) or (2) inconsistent assignment to maternal/paternal haplotypes (HiC, Figure B3-4).

We hypothesized that HiC outperformed the trios approach because of low levels of heterozygosity contained within this lab bred lineage of leopard gecko—originally sourced from the pet trade. To examine this further, we mapped reads from each parent to the reference, called SNPs (see methods), and estimated heterozygosity by dividing the total number of SNPs by the total genome size (father/TG4151 = 0.35%; mother/TG4152 = 0.38%). However, we acknowledge that only heterozygous sites that are not shared between parents are informative for phasing purposes. We identified sites that were not shared between parents using vcf-compare to calculate the informative heterozygosity rate for phasing (father/TG4151 = 0.15%; mother/TG4152 = 0.18%). Indeed, less than 50% of heterozygous sites in each parent are informative for phasing, which limits theoretically informative phasing sites to <2,500 SNPs per Mb on average. This constraint on informative sites was not observed in HiC phasing given that every heterozygous site in the genome is theoretically informative for HiC phasing – approximately doubling the number of informative sites when phasing with HiC data. In sum, this genome assembly experiment was conducted on a trio of animals with too little heterozygosity for successful offspring phasing, but HiC provided sufficient resolution for phasing. For future studies facing a similar situation, we suggest either planning the experiment around a single individual using HiC or outcrossing two individuals with different genetic backgrounds and sequencing this doubly heterozygous offspring trio to increase site informativeness, which are analogous to the established standards for traditional linkage mapping experiments (e.g. Amores et al. 2014).

Our annotation for this reference genome, MPM_Emac_v1.0, maintained a completeness of 89.3% using the sauropsida_odb10 dataset in BUSCO [v5.1.2] (Simão et al. 2015), nearly mirroring the Xiong et al. (2016) original reference genome

annotation of 90.6%. Of note, these numbers do not match those from the Xiong et al. (2016) manuscript due to changes in both software versions and query databases. We also compared other differences between the current assembly and the original reference assembly. MPM_Emac_v1.0 assembly size ~12% larger than Xiong et al. (2016)—2.24Gb vs. 2.02Gb, respectively. Interestingly, MPM_Emac_v1.0 is much closer to the kmer estimated genome size from Xiong et al. (2016) of 2.23Gb. There is also an increase of repetitive DNA content in MPM_Emac_v1.0 of ~6% (Table 2). However, the GC content deviated by 0.5% between the two assemblies, indicating that the GC content in gecko genomes may not be as biased with short-read based sequence data as might be anticipated *a priori* (Benjamini and Speed, 2012).

In conclusion, we present a chromosome-level genome assembly for the leopard gecko, *E. macularius*. This is simultaneously the first phased chromosome-level assembly and the first long-read based genome assembly available for any species of gecko. Further, this assembly is one of the most contiguous squamate genomes available and has achieved the second highest BUSCO score of any squamate genome (Pinto et al. 2023). The last hurdle for this assembly to overcome before this assembly can be considered a finished “telomere-to-telomere” assembly is placing the final 5.02Mb of unassembled sequence into the 19 primary scaffolds representing the 19 chromosomes of *E. macularius*. This would likely require generation of a modest number of ultra-long reads to fill gaps and complete centromeric/telomeric regions (e.g. Rautiainen et al. 2022). Nonetheless, our genome assembly represents the new ‘gold standard’ in squamate genomes at this ever-fleeting moment.

Data Availability

The raw data generated in this study has been deposited to NCBI SRA under Bioproject PRJNA884264. The genome version described in this study MPM_Emac_v1.0,annotation file, and all four phased assemblies are available via this Figshare repository https://doi.org/10.6084/m9.figshare.20069273. Additionally, genome assembly version MPM_Emac_v1.0.1 is archived as a Whole Genome Shotgun project deposited at DDBJ/ENA/GenBank under the accession JAOPLA000000000. The genome version and its associated annotations are also available via this Figshare repository https://doi.org/10.6084/m9.figshare.20069273.

Funding Sources

This work was funded piecemeal in author order: the University of Texas at Austin Office of the Vice President (OVPR) Special Research Grant to B.J.P.; National Science Foundation (USA) DEB1657662 to T.G.; the University of Texas at Austin Stengl-Wyer Endowment to C.H.S.; and National Institute of Health (USA) R35GM142836 to J.C.H.

Acknowledgements

The authors would like to thank: M. Weitzman, formally at the University of Oregon GC3F sequencing core, for her exceptional help and support of the HiFi sequencing portion of this project; R. Tremper for generously providing the parental geckos; and Gamble Lab animal care technicians. Gecko breeding and tissue sampling were conducted under Marquette University IACUC protocol AR-279.

References

Adler D, Kelly ST, Elliott T, Adamson J (2022). *vioplot: violin plot*. R package version 0.4.0, https://github.com/TomKellyGenetics/vioplot.

Agarwal, I., Bauer, A. M., Gamble, T., Giri, V. B., Jablonski, D., Khandekar, A., ... & Ramakrishnan, U. (2022). The evolutionary history of an accidental model organism, the leopard gecko Eublepharis macularius (Squamata: Eublepharidae). *Molecular Phylogenetics and Evolution*. 168:107414.

Amores, A., Catchen, J., Nanda, I., Warren, W., Walter, R., Schartl, M., & Postlethwait, J. H. (2014). A RAD-tag genetic map for the platyfish (*Xiphophorus maculatus*) reveals mechanisms of karyotype evolution among teleost fish. *Genetics*, 197(2), 625-641.

Bauer, A.M. (2013). Geckos: The Animal Answer Guide, (Baltimore, MD: Johns Hopkins University Press).

Benjamini, Y., & Speed, T. P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research*. 40(10):e72.

Bravo, G. A., Schmitt, C. J., & Edwards, S. V. (2022). What have we learned from the first 500 avian genomes? *Annu Rev Ecol Evol Syst*. 52(1):611-639.

Bryant, H.N., and Russell, A.P. (1992). The role of phylogenetic analysis in the inference of unpreserved attributes of extinct taxa. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*. 337:405-418.

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., & Li, H. (2021). Haplotype-resolved *de novo* assembly using phased assembly graphs with Hifiasm. *Nature Methods*. 18(2):170-175.

Cheng, H., Jarvis, E. D., Fedrigo, O., Koepfli, K. P., Urban, L., Gemmell, N. J., & Li, H. (2022). Haplotype-resolved assembly of diploid genomes without parental data. *Nature Biotechnology*. 1-4.

Chin, CS., Peluso, P., Sedlazeck, F. et al. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods.* 13, 1050–1054.

Chernyavskaya, Y., Zhang, X., Liu, J., & Blackburn, J. (2022). Long-read sequencing of the zebrafish genome reorganizes genomic architecture. *BMC genomics*. 23(1):1-13.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. *Bioinformatics*. 27: 2156-2158.

Delorme, S.L., Lungu, I.M., Vickaryous, M.K., 2012. Scar-free wound healing and regeneration following tail loss in the Leopard Gecko, *Eublepharis macularius*. Anatomical Record 295, 1575–1595.

Dobzhansky, T. (1973). Nothing in biology makes sense except in the light of evolution. *The American Biology Teacher*. 75:87-91.

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., Shamim, M. S., Machol, I., Lander, E. S., Aiden, A. P., & Aiden, E. L. (2017). De

novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*. 356(6333), 92–95.

Durand, N. C., Robinson, J. T., Shamim, M. S., Machol, I., Mesirov, J. P., Lander, E. S., & Aiden, E. L. (2016). Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Systems*. 3(1), 99–101.

Eyre-Walker, A., Hurst, L,D. (2001). The evolution of isochores. *Nat Rev Genet*. 2:549-555.

Felsenstein, J. (1985). Phylogenies and the comparative method. *Am Nat*. 125:1-15.

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *PNAS*. 117(17):9451-9457.

Gamble, T., Bauer, A.M., Colli, G.R., Greenbaum, E., Jackman, T.R., Vitt, L.J., and Simons, A.M. (2011). Coming to America: Multiple origins of New World geckos. *J Evol Biol*. 24:231-244.

Gamble, T., Greenbaum, E., Jackman, T. R., Russell, A. P., & Bauer, A. M. (2012). Repeated Origin and Loss of Adhesive Toepads in Geckos. *PLoS ONE*. 7(6):e39429.

Gamble, T., Greenbaum, E., Jackman, T.R., and Bauer, A.M. (2015a). Into the light: Diurnality has evolved multiple times in geckos. *Biol J Linn Soc*. 115:896-910.

Gamble, T., Coryell, J., Ezaz, T., Lynch, J., Scantlebury, D. P., & Zarkower, D. (2015b). Restriction Site-Associated DNA Sequencing (RAD-seq) Reveals an Extraordinary Number of Transitions among Gecko Sex-Determining Systems. *Molecular Biology and Evolution*, 32(5), 1296–1309.

Gamble, T. (2019). Duplications in corneous beta protein genes and the evolution of gecko adhesion. *Integrative and Comparative Biology*. 59(1):193-202.

Garrison, E., Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv*. 1207.3907.

Gorman, G.C. (1973). The chromosomes of the Reptilia, a cytotaxonomic interpretation. pp 349-424 in: Chiarelli & Capanna (eds) Cytotaxonomy and Vertebrate Evolution Academic Press, London, New York.

Glimm, T., Kiskowski, M., Moreno, N., & Chiari, Y. (2021). Capturing and analyzing pattern diversity: an example using the melanistic spotted patterns of leopard geckos. *PeerJ*, 9, e11829.

Guo, L., Bloom, J., Sykes, S., Huang, E., Kashif, Z., Pham, E., Ho, K., Alcaraz, A., Xiao, X. G., Duarte-Vogel, S., & Kruglyak, L. (2021). Genetics of white color and iridophoroma in “Lemon Frost” leopard geckos. PLOS Genetics, 17(6), e1009580.

Hall, B.K. (2012). Homology: The Hierarchical Basis of Comparative Biology, (Academic Press).

Hotaling, S., Kelley, J.L., and Frandsen, P.B. (2021). Toward a genome sequence for every animal: Where are we now? *PNAS*. 118:e2109019118.

Katlein, N., Ray, M., Wilkinson, A., Claude, J., Kiskowski, M., Wang, B., Glaberman, S., & Chiari, Y. (2022). Does colour impact responses to images in geckos? *Journal of Zoology*, 317(2), 138–146.

Kim, J., Lee, C., Ko, B. J., Yoo, D., Won, S., Phillippy, A., Fedrigo, O., Zhang, G., Howe, K., Wood, J., Durbin, R., Formenti, G., Brown, S., Cantin, L., Mello, C. V., Cho, S., Rhie, A., Kim, H., & Jarvis, E. D. (2021). False gene and chromosome losses affected by assembly and sequence errors. *BioRxiv*. doi:10.1101/2021.04.09.438906

Kiskowski, M., Glimm, T., Moreno, N., Gamble, T., Chiari, Y., 2019. Isolating and quantifying the role of developmental noise in generating phenotypic variation. PLoS Comput. Biol. 15 (4), e1006943.

Koren, S., Rhie, A., Walenz, B. P., Dilthey, A. T., Bickhart, D. M., Kingan, S. B., Hiendleder, S., Williams, J. L., Smith, T. P. L., & Phillippy, A. M. (2018). *De novo* assembly of haplotype-resolved genomes with trio binning. *Nature Biotechnology*, 36(12), 1174–1182.

Kumar, S., Stecher, G., Suleski, M., Hedges, S.B. (2017). TimeTree: a resource for timelines, timetrees, and divergence times. *Mol Biol Evol*. 34:1812-1819.

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

McLean, K. E., & Vickaryous, M. K. (2011). A novel amniote model of epimorphic regeneration: the leopard gecko, Eublepharis macularius. *BMC Developmental Biology*. *11*(1):1-25.

Nishimura O, Hara Y, Kuraku S. (2017). gVolante for standardizing completeness assessment of genome and transcriptome assemblies. *Bioinformatics*. 33:3635-3637.

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., … Phillippy, A. M. (2022). The complete sequence of a human genome. *Science*. 376(6588):44-53.

Pagel, M., Meade, A., and Barker, D. (2004). Bayesian estimation of ancestral character states on phylogenies. *Syst Biol*. 53:673-684.

Peona, V., Blom, M. P. K., Xu, L., Burri, R., Sullivan, S., Bunikis, I., Liachko, I., Haryoko, T., Jønsson, K. A., Zhou, Q., Irestedt, M., & Suh, A. (2020). Identifying the causes and consequences of assembly gaps using a multiplatform genome assembly of a bird-of-paradise. *Molecular Ecology Resources*. 21(1):263-286.

Pinto, B. J., Nielsen, S. V., & Gamble, T. (2019). Transcriptomic data confirms an ancient nocturnal bottleneck in gecko lizards. *Molecular Phylogenetics and Evolution*. 141:106639

Pinto, B. J., Weis, J. J., Gamble, T., Ode, P. J., Paul, R., & Zaspel, J. M. (2021). A chromosome-level genome assembly of the parasitoid wasp, *Cotesia glomerata* (Hymenoptera: Braconidae). *Journal of Heredity*. 112(6):558-564.

Pinto, B. J., Keating, S. E., Nielsen, S. V., Scantlebury, D. P., Daza, J. D., & Gamble, T. (2022). Chromosome-level genome assembly reveals dynamic sex chromosomes in Neotropical leaf-litter geckos (Sphaerodactylidae: Sphaerodactylus). *Journal of Heredity*. 113(3):272-287.

Pinto, B. J., Gamble, T., Smith, C. H., Wilson, M. A. (2023). A lizard is never late: squamate genomics as a recent catalyst for understanding microchromosome and sex chromosome evolution. *BioRxiv*.

Rautiainen M, Nurk S, Walenz BP, Logsdon GA, Porubsky D, Rhie A, Eichler EE, Phillippy AM, Koren S. (2022). Verkko: telomere-to-telomere assembly of diploid chromosomes. *BioRxiv*.

Rhie A, Walenz BP, Koren S, Phillippy AM. (2020). Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol*. 21:1-27.

Sakata, J.T., Gupta, A., Chuang, C.P., Crews, D., 2002. Social experience affects territorial and reproductive behaviours in male leopard geckos, *Eublepharis macularius.* Animal Behavior 63, 487–493.

Sanger, T.J., and Rajakumar, R. (2019). How a growing organismal perspective is adding new depth to integrative studies of morphological evolution. *Biol Rev*. 94:184-198.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. 31:3210-3212.

Shumate, A., & Salzberg, S. L. (2021). Liftoff: accurate mapping of gene annotations. *Bioinformatics*. 37(12):1639-1643.

Smit AFA, Hubley R, Green P. (2013). RepeatMasker Open-4.0. http://www.repeatmasker.org.

Szydłowski, P., Madej, J. P., Duda, M., Madej, J. A., Sikorska-Kopyłowicz, A., Chełmońska-Soyta, A., ... & Duda, P. (2020). Iridophoroma associated with the Lemon Frost colour morph of the leopard gecko (Eublepharis macularius). *Scientific Reports*. 10(1):1-6.

Townsend, T.M., Larson, A., Louis, E., and Macey, J.R. (2004). Molecular phylogenetics of Squamata: The position of snakes, amphisbaenians, and dibamids, and the root of the squamate tree. *Syst Biol*. 53:735-757.

Uetz, P., Koo, M.S., Aguilar, R., Brings, E., Catenazzi, A., Chang, A.T., and Wake, D.B. (2021). A quarter century of reptile and amphibian databases. *Herpetol Rev*. 52:246-255.

Vasimuddin, M., Misra, S., Li, H., & Aluru, S. (2019, May). Efficient architecture-aware acceleration of BWA-MEM for multicore systems. In 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS) (pp. 314-324). *IEEE*.

Viets, B.E., Tousignant, A., Ewert, M.A., Nelson, C.E., Crews, D., 1993. Temperature- dependent sex determination in the leopard gecko, *Eublepharis macularius*. J. Exp. Zool. 265, 679–683.

Xiong, Z., Li, F., Li, Q., Zhou, L., Gamble, T., Zheng, J., ... & Zhang, G. (2016). Draft genome of the leopard gecko, *Eublepharis macularius*. *Gigascience*. 5(1):s13742-016.

Wake, M.H. (2008). Integrative biology: Science for the 21st century. *Bioscience.* 58:349-353.

Whimster, I.W., 1965. An experimental approach to the problem of spottiness. Br. J. Dermatol. 77, 397–420.

Wiens, J.J., Hutter, C.R., Mulcahy, D.G., Noonan, B.P., Townsend, T.M., Sites, J.W., and Reeder, T.W. (2012). Resolving the phylogeny of lizards and snakes (Squamata) with extensive sampling of genes and species. *Biol Lett*. 8:1043-1046.

Witmer, L.M. (1995). Homology of facial structures in extant Archosaurs (birds and crocodilians), with special reference to paranasal pneumaticity and nasal conchae. *J Morph*. 225:269-327.

Yamaguchi, K., Kadota, M., Nishimura, O., Ohishi, Y., Naito, Y., & Kuraku, S. (2021). Technical considerations in Hi-C scaffolding and evaluation of chromosome-scale genome assemblies. *Molecular Ecology*. 30(23):5923-5934.

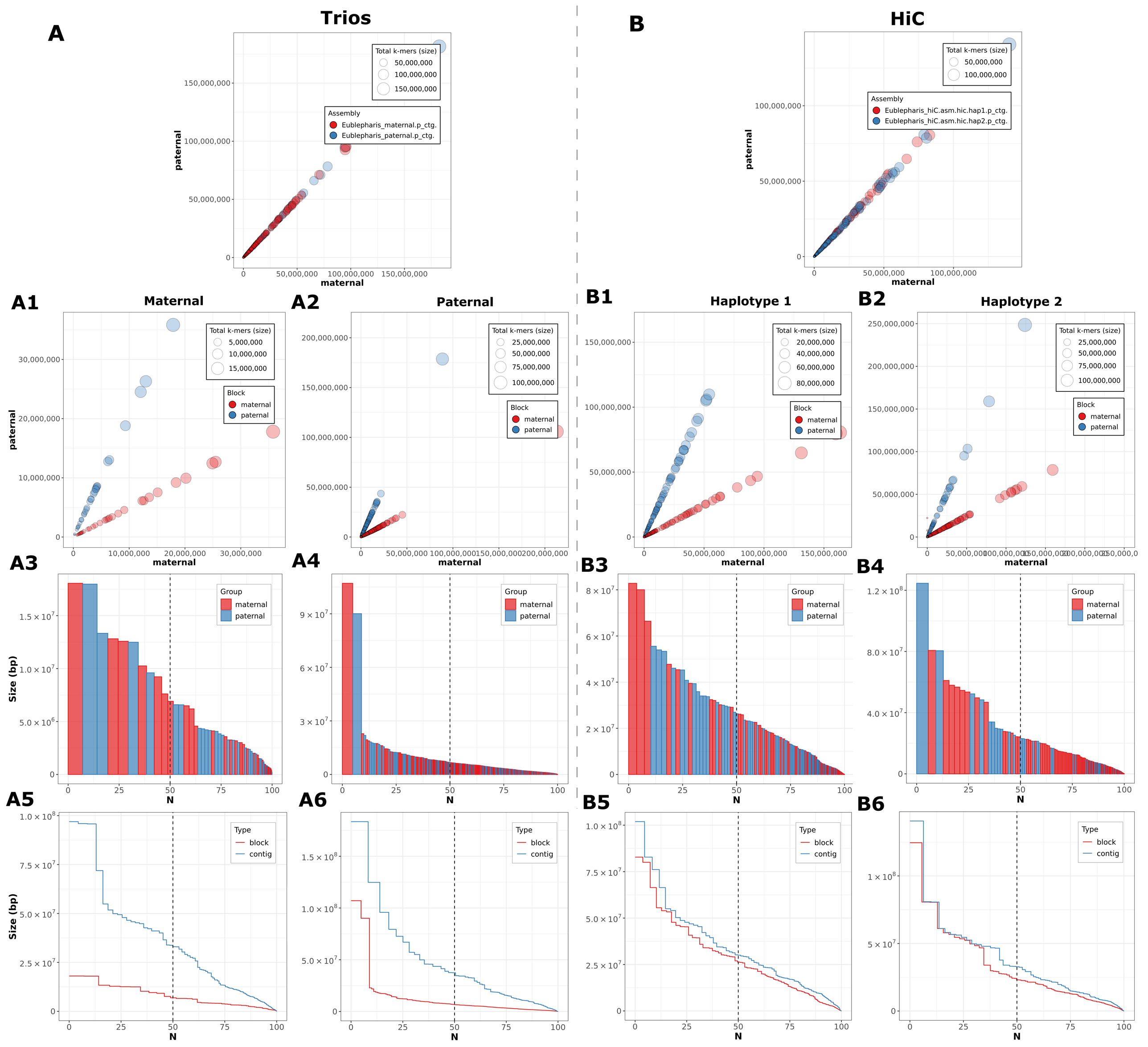
Zheng, Y., and Wiens, J.J. (2016). Combining phylogenomic and supermatrix approaches, and a time-calibrated phylogeny for squamate reptiles (lizards and snakes) based on 52 genes and 4162 species. *Mol Phylogenet Evol*. 94:537-547.

Figure Legends:

Figure 1: HiC contact map for the MPM_Emac_v1.0 assembly. Each blue segment indicates the delimitation of a chromosome-length scaffold, while the internal green squares indicate contigs. Approximately half of the assembled chromosomes are represented by a single contig, indicating the extreme contiguity of the primary assembly pre-scaffolding.

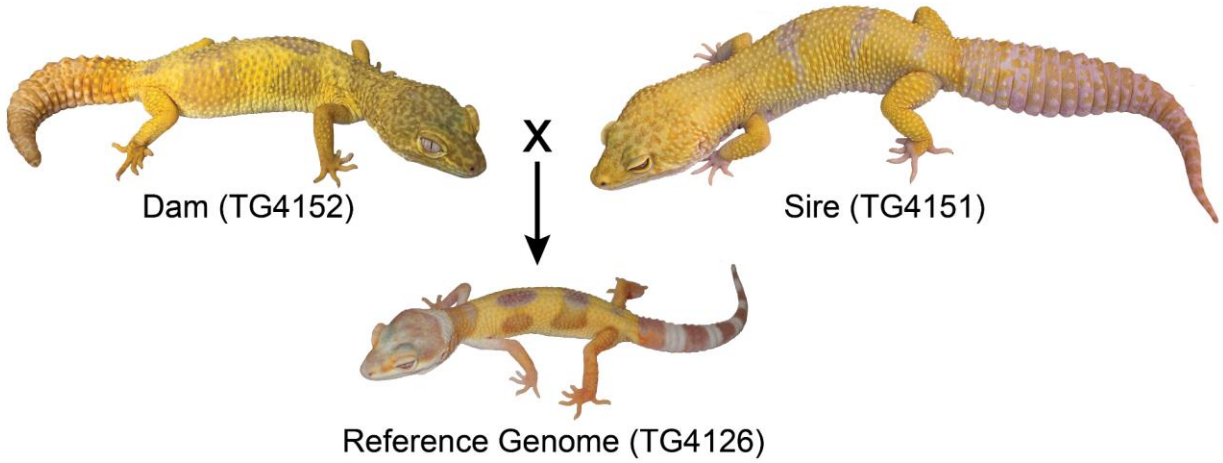
Figure 2: Comparison between chromosomes assembled as a single contig ("Contigs", dark gray) and those composed of multiple contigs ("Scaffolds", light gray) displayed using vioplot (Adler et al. 2022). The violins represent the distribution of the underlying data points, while the internal bars represent a traditional bar graph representation. The mid-lines represent the median of the data. Our *a priori* hypothesis was that chromosomes assembled as a single contig would possess lower overall GC content and/or repeat content. However, neither GC content or repetitive element content were significantly different using Mann-Whitney-Wilcoxon tests. Qualitatively, there appears to be a difference in median repeat content between the two groups. It's possible that our ability to detect a true difference using frequentist methods lies in the low sample size (N=19).

Figure 3: Comparative QC results from assembly phasing generated by merqury (Rhie et al. 2020) between (A) Trios phasing and (B) HiC phasing methods. HiC equaled or outperformed Trios phasing in all measured categories. Notably, Trios phasing appears to have suffered from high switch-error rates, which resulted in short phase block, relative to contig size. HiC phasing performed extremely well, however by definition; HiC phasing was unable to coordinate multiple phased contigs to their parent of origin.


A
Trios
Total k-mers (size)
50,000,000
100,000,000
150,000,000
Assembly
Eublepharis_maternal.p_ctg.
Eublepharis_paternal.p_ctg.
paternal
maternal
B
HiC
Total k-mers (size)
50,000,000
100,000,000
Assembly
Eublepharis_hiC.asm.hic.hap1.p_ctg.
Eublepharis_hiC.asm.hic.hap2.p_ctg.
paternal
maternal
A1
Maternal
Total k-mers (size)
5,000,000
10,000,000
15,000,000
Block
maternal
paternal
paternal
maternal
A2
Paternal
Total k-mers (size)
25,000,000
50,000,000
75,000,000
100,000,000
Block
maternal
paternal
maternal
B1
Haplotype 1
Total k-mers (size)
20,000,000
40,000,000
60,000,000
80,000,000
Block
maternal
paternal
maternal
B2
Haplotype 2
Total k-mers (size)
25,000,000
50,000,000
75,000,000
100,000,000
Block
maternal
paternal
maternal
A3
Group
maternal
paternal
Size (bp)
N
A4
Group
maternal
paternal
N
B3
Group
maternal
paternal
N
B4
Group
maternal
paternal
N
A5
Type
block
contig
Size (bp)
N
A6
Type
block
contig
N
B5
Type
block
contig
N
B6
Type
block
contig
N

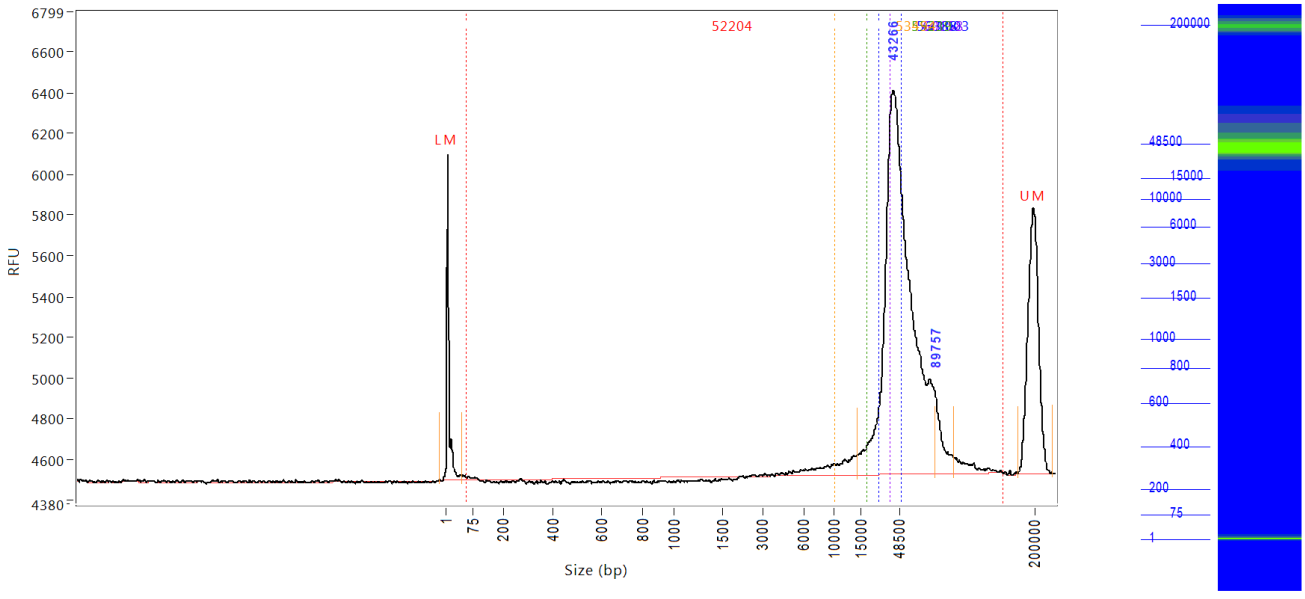
Supplemental Information

Supplemental Figure 1: Crossing scheme to generate the reference genome animal for *Eublepharis macularius*, TG4126.



Supplemental Figure 2: High-molecular weight DNA extraction Bioanalyzer results.

Data File: 2020 12 10 18H 31M.raw
Sample: Ylenia eublepharis rec'd gDNA (dil 1:40)
Well Location: E7



Peak	Size (bp)	Conc. (ng/uL)	From (bp)	To (bp)	Avg. Size (bp)	CV%	RFU	Corr. Peak Area
1	1 (LM)	0.0108	0	41	4	200.86	1599	10.391
2	43266	0.8602	14294	89757	50972	30.57	1884	69.207
3	89757	0.0433	89757	110724	97126	6.20	390	3.482
4	200000 (UM)	0.0177	183092	221516	199733	2.83	1303	17.094

TIC:	0.9035	ng/uL
TIM:	0.0285	nmole/L
Total Conc.:	0.9735	ng/uL

Smear Analysis					
50 bp to 165000 bp	0.9712 ng/ul	99.8 %Total	0.0306 nmole/L	52204 Avg. Size (bp)	43.28 %CV
10000 bp to 165000 bp	0.9445 ng/ul	97.0 %Total	0.0291 nmole/L	53494 Avg. Size (bp)	40.28 %CV
20000 bp to 165000 bp	0.9047 ng/ul	92.9 %Total	0.0270 nmole/L	55239 Avg. Size (bp)	36.78 %CV
30000 bp to 165000 bp	0.8719 ng/ul	89.6 %Total	0.0255 nmole/L	56388 Avg. Size (bp)	35.17 %CV
40000 bp to 165000 bp	0.7389 ng/ul	75.9 %Total	0.0202 nmole/L	60158 Avg. Size (bp)	32.27 %CV
50000 bp to 165000 bp	0.4377 ng/ul	45.0 %Total	0.0101 nmole/L	71003 Avg. Size (bp)	26.50 %CV

```
Sample Peak Width (sec): 10   Sample Min Peak Height: 50    Sample Baseline V to V?: Y    Sample Baseline V to V pts: 3
Sample Filter: Binomial       # of Pts for Filter: 3        Sample Start Region (min): 0  Sample End Region (min): 55
Manual Baseline Start (min): 12     Manual Baseline End (min): 52
Marker Peak Width (sec): 5    Marker Min Peak Height: 500   Marker Baseline V to V?: N    Marker Baseline V to V pts: 3
Lower Marker Selection: First Peak > 500 RFU                Upper Marker Selection: Last Peak > 500 RFU
Ladder Size (bp)1, 75, 200, 400, 600, 800, 1000, 1500, 3000, 6000, 10000, 15000, 48500, 200000
Quantification Using: Ladder          Final Concentration (ng/uL): 0.1250          Dilution Factor: 12.0
Min. RFU for Data Processing: 3
```


Copyright 2015 Advanced Analytical Technologies, Inc.