SUPPLEMENTARY INFORMATION

Physical mapping and refinement of the painted turtle genome (*Chrysemys picta*) inform amniote genome evolution and challenges turtle-bird chromosomal conservation

Daleen Badenhorst[1], LaDeanna Hillier[2], Robert Literman[1], Eugenia Elisabet Montiel[1], Srihari Radhakrishnan[1], Yingjia Shen[2], Patrick Minx[2], Daniel Janes[1,3], Wesley C. Warren[2], Scott V. Edwards[3], Nicole Valenzuela[1]

[1]Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames IA 50011.
[2]The Genome Institute at Washington University, St Louis, MO 63108, USA
[3]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge MA 02138

Refinement of the painted turtle genome assembly

The original genome assembly 3.0.1 of the painted turtle (*Chrysemys picta bellii* - CPI) was reported in and released to NCBI (Shaffer et al. 2013), and contained 81,642 scaffolds larger than 500 bases (with a total size of 2618 Mb), and an N50 scaffold size of 3.01 Mb (N50 number is 248, i.e. the number of ordered scaffolds starting from the longest that add up to half the total length of all scaffolds in the assembly). To refine the CPI assembly 3.0.1 (Shaffer et al. 2013) we attempted to close all possible scaffold gaps using a series of scripts that mimic a previous method (Tsai et al. 2010), but were modified to scale up for large genome (>1Gb) gap closing (https://drive.google.com/folderview?id=0By6uqdmCrxS_fmppRk9lbW02anZkNVFLd2tyVzh0RmZhRHQ5dVE0eHFpT1J5UDNmRC1OMEE&usp=sharing). Our integrative approach aligns Illumina reads at contig ends, performs local assembly of aligned reads into new contigs, and extends or merges reference contigs within scaffolds. We performed three iterations of gap closing for the 3.0.1 assembly. To further scaffold growth we aligned the gap filled 3.0.1 version to the *P. sinensis* (PSI) and *C. mydas* (CMY) assemblies (Wang et al. 2013) using NUCmer, part of the MUMmer 3.0 package (Kurtz et al. 2004). In both cases we used 3.0.1 as the query assembly and the other turtle species as the reference (Wang et al. 2013). We obtained the best one to one reciprocal matches by using the -r and -q options and the following NUCmer options: minInd=40; c=400; l=10; g=500; rearrange=" -r -q -o 1 ".

When scaffolds were identically ordered/oriented by alignments to PSI and CMY scaffolds the order/orientation provided by the alignments was introduced into the CPI assembly. When scaffolds were ordered/oriented by a CMY or PSI alignment and had the support of at least one BAC end sequence pair or by alignment to the chicken genome (galGal4), lizard (anoCar2), or alligator (allMis1) they were also introduced into the assembly. Similarly when a scaffold merge was suggested by alignment to the chicken, lizard or alligator genome and supported by at least a single BAC end pair, the scaffolds were joined into a single ultrascaffold. Alignments to finished BAC clones were also used to identify additional ordered/oriented scaffolds. Finally, we

use CEGMA (Core Eukaryotic Genes Mapping Approach) (Parra et al. 2007), to detect in the CPI genome assembly the set of 248 highly conserved core proteins that are present across eukaryotes to assess the completeness of our refined assembly and compare it with that of the original CPI genome assembly (Shaffer et al. 2013).

Using the approach described above we created 100 new ultra-scaffolds containing 259 scaffolds spanning 941 Mb of sequence (~40%) of the painted turtle genome assembly (scaffolds are sets of contigs ordered and oriented, while contigs are consensus sequences from overlapping sequence reads). The resulting assembly was designated 3.0.3. The final assembly had a scaffold N50 of 6.6Mb (contig N50 size of 21.3 Kb). This represents a significant improvement over the previous release assembly 3.0.1 (Table S1).

A total of 61 BACs were mapped to the CPI chromosomes as described in the main text methods. The resulting maps provided only order for the BACs and did not contribute orientation information unless there were BAC end pairs providing connectivity between neighboring scaffolds. Chromosomal AGPs were created and centromeres were positioned using the BAC maps that localized 461 Mb (~20% of the genome assembly) to 18 chromosomes. AGPs are “A Golden Path” description files of the components of each chromosome. Thus, chromosomal AGPs describe the assembly of the chromosomes by listing the order of the component BACs that were mapped to them. This is the first chromosomal AGP produced for a turtle and the second for non-avian reptiles (Alfoldi et al. 2011).

The CEGMA (Parra et al. 2007) results show a slight sequence improvement in contiguity of the 248 core eukaryotic genes (CEGs), while the total numbers of complete and partial CEGs present including putative orthologs increased considerably [e.g. 47 additional predicted proteins identified overall (an increase of 14%); ~15% representation of CEGs and up to 54% increase in the proportion of orthologs) Table 1] in comparison to assembly 3.0.1 (Shaffer et al. 2013). The improved genome sequence of CPI 3.0.3 was deposited in the DDBJ/EMBL/GenBank database (Accession Number AHGY00000000.2).

Table S3. Comparison of the *Chrysemys picta bellii* genome assemblies 3.0.1 and 3.0.3 deposited in NCBI.

	Shaffer et al. 2013	**This study**
NCBI Assembly Version	3.0.1	3.0.3
Total sequence length	2,589,745,704	2,365,749,696
Total assembly gap length	431,452,339	192,562,473
Gaps between scaffolds	0	56
Number of scaffolds	80,984	78,630
*Scaffold N50	5,212,367	6,605,846
Number of contigs	551,713	262,325
Contig N50	11,852	21,349
Total number of chromosomes and plasmids	1	18

*Scaffold N50 include singletons (single contig scaffolds)

CEGMA Results per Assembly Version	3.0.1		3.0.3	
	Complete	Partial	Complete	Partial
Proteins	231	247	232	246
% Completeness	93.15	99.6	93.55	99.19
Total	291	344	335	391
Average	1.26	1.39	1.44	1.59
% Ortho	21.21	29.96	32.76	40.24

Proteins = number of 248 ultra-conserved CEGs present in genome. % Completeness = percentage of 248 CEGs present. Total = total number of CEGs present including putative orthologs. Average = average number of orthologs per CEG. %Ortho = percentage of detected CEGS that have more than 1 ortholog. Complete = predicted proteins in the set of 248 CEGs that when aligned to the HMM for the KOG for that protein-family, give an alignment length that is 70% of the protein length. Partial = total of completely aligned proteins plus partially aligned proteins that still exceed a pre-computed minimum alignment score

REFERENCES

Alfoldi J, et al. 2011. The genome of the green anole lizard and a comparative analysis with birds and mammals. Nature. 477: 587-591. doi: 10.1038/nature10390

Kurtz S, et al. 2004. Versatile and open software for comparing large genomes. Genom Biol. 5. doi: 10.1186/gb-2004-5-2-r12

Parra G, et al. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genornes. Bioinformatics. 23: 1061-1067. doi: 10.1093/bioinformatics/btm071

Shaffer HB, et al. 2013. The western painted turtle genome, a model for the evolution of extreme physiological adaptations in a slowly evolving lineage. Genom Biol. 14: doi:10.1186/gb-2013-1114-1183-r1128.

Tsai IJ, et al. 2010. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. Genom Biol. 11. doi: 10.1186/gb-2010-11-4-r41

Wang Z, et al. 2013. The draft genomes of soft-shell turtle and green sea turtle yield insights into the development and evolution of the turtle-specific body plan. Nat Genet. 45: 701-706. doi: 10.1038/ng.2615
http://www.nature.com/ng/journal/v45/n6/abs/ng.2615.html#supplementary-information