

1 **Improved draft of the Mojave Desert tortoise genome, *Gopherus agassizii*,**
2 **version 1.1**

3

4

5 Timothy H. Webster^{1†*}, Greer A. Dolby^{1†*}, Melissa Wilson Sayres^{1,2}, Kenro Kusumi¹

6

7 ¹School of Life Sciences, Arizona State University, Tempe, AZ 85287

8 ²Center for Evolution and Medicine, Arizona State University, Tempe, AZ 85287

9 † *authors contributed equally*

10

11 ***Correspondence to (*):***

12 Timothy Webster
13 School of Life Sciences
14 Arizona State University
15 PO Box 874501
16 Tempe, AZ 85287
17 Timothy.H.Webster@asu.edu

18

19 Greer Dolby
20 School of Life Sciences
21 Arizona State University
22 PO Box 874501
23 Tempe, AZ 85287
24 gadolby@asu.edu

25

26 ***Running title:*** Mojave desert tortoise genome v1.1

27

28 ***Key words:*** genome, assembly, tortoise, scaffold, contamination

29

30 ABSTRACT

31 Exogenous sequence contamination presents a challenge in first-draft genomes because it can
32 lead to non-contiguous, chimeric assembled sequences. This can mislead downstream analyses
33 reliant on synteny, such as linkage-based analyses. Recently, the Mojave Desert Tortoise
34 (*Gopherus agassizii*) draft genome was published as a resource to advance conservation efforts
35 for the threatened species and discover more about chelonian biology and evolution. Here, we
36 illustrate steps taken to improve the desert tortoise draft genome by removing contaminating
37 sequences—actions that are typically carried out after the initial release of a draft genome
38 assembly. We used information from NCBI's Vecscreen output to remove intra-scaffold
39 contamination and trim heading and trailing Ns. We then reordered and renamed scaffolds, and
40 transferred the gene annotation onto this assembly. Finally, we describe the tools developed for
41 this pipeline, freely available on Github
42 (https://github.com/thw17/G_agassizii_reference_update), which facilitate post-assembly
43 processing of other draft genomes. The new gopAga1.1 genome has an N50 of 251 KB, L50 of
44 2592 scaffolds, and its annotation retains 17,201 of the original 20,172 genes that were
45 unaffected by the scaffold processing.

46

47 INTRODUCTION

48 The Mojave Desert Tortoise, *Gopherus agassizii*, is a long-lived, xeric-adapted species
49 endemic to southern California, southern Nevada, southwestern Utah, and northwestern Arizona
50 (Morafka & Berry 2002; Murphy *et al.* 2011). One of six extant species in the genus *Gopherus*, it
51 is thought to have diverged from the lineage leading to *G. evgoodei* and *G. morafkai* between 5–
52 6 million years ago when the Colorado River first began draining into the Gulf of California

53 (Dorsey *et al.* 2011; Murphy *et al.* 2011; Edwards *et al.* 2016). These three species have since
54 differentially adapted to their respective habitats, with the differences between *G. agassizii* of the
55 Mojave Desert and *G. morafkai* of the Sonoran Desert being well-characterized (Edwards *et al.*
56 2015). Differences between these deserts based on seasonal rainfall, total annual precipitation,
57 vegetation, and other key environmental characteristics likely underlie the differential
58 adaptations in these species (Pianka 1970; Reynolds *et al.* 2004) .

59 Significant conservation efforts have targeted *Gopherus agassizii* since its Threatened
60 listing under the Endangered Species Act in 1990 (Smith 1990). However, populations continue
61 to decline due to a combination of habitat loss, changes in land use, invasive grasses (Drake *et al.*
62 2016), and upper respiratory tract disease (URTD; (Jacobson *et al.* 1991; Doak *et al.* 1994;
63 Brown *et al.* 1994). As part of this conservation effort, Tollis *et al.* (2017) published a draft
64 genome (version 1.0; gopAga1) of *G. agassizii*, which was the first for any tortoise species.
65 Analysis of the genome revealed putative genes under selection in *G. agassizii* relative to other
66 non-avian reptiles, confirmed slow mutation rates among chelonians (Shaffer *et al.* 2013), and
67 found evidence of gene structure more closely resembling chicken than other non-avian reptiles
68 (Tollis *et al.* 2017).

69 Development of the reference genome for this species enables new and promising
70 avenues of research that will aid its conservation. Here, we present genome version 1.1 for *G.*
71 *agassizii* (gopAga1.1), with the following improvements from initial release (gopAga1): **1)**
72 screening for and removal of exogenous contaminant sequences; **2)** reordering and renaming of
73 scaffolds within the assembly; and **3)** an updated annotation that converts the physical
74 positioning of genes and gene features under this new scaffolding. Draft genomes of non-model
75 organisms are rapidly becoming more common and they represent the foundation for future

76 research. Because many such assemblies contain contamination (Alkan *et al.* 2010), software
77 tools and workflows designed to handle the splitting, sorting, and processing of scaffolds are
78 needed. In addition to introducing genome version 1.1 for *G. agassizii*, we provide software tools
79 to manipulate early-generation genome assemblies such as this one, and aim to add transparency
80 to the steps involved in processing a draft assembly to meet the standards required for deposition
81 in public databases (e.g., NCBI).

82

83 MATERIALS & METHODS

84 After submitting `gopAga1` for processing and hosting, NCBI
85 (<https://www.ncbi.nlm.nih.gov>) identified adapter and exogenous sequence contamination using
86 their Vecscreen (<http://www.ncbi.nlm.nih.gov/VecScreen/VecScreen.html>) pipeline, an issue
87 common to many draft genome assemblies. As a part of their pipeline, NCBI removed
88 contaminant sequences from the beginning and ends of scaffolds and provided the locations of
89 remaining contaminants. We used the scripts presented in this manuscript, the processed
90 assembly file, and contamination file to: **1)** split scaffolds at the intra-scaffold sites of
91 contamination provided in the Vecscreen output; **2)** soft-clip scaffold ends that contained Ns
92 after splitting; **3)** reorder and rename v1.1 scaffolds by descending size; **4)** transfer the v1
93 annotation to v1.1 assembly under these newly processed scaffolds (Figure 1).

94

95 *Genome assembly version 1.1*

96 Within-scaffold regions of contamination likely resulted in misjoining non-contiguous
97 regions. To remove such effects, we wrote a Python script
98 (`Remove_and_split_contamination_NCBI.py`) to read NCBI output (with 1-based coordinates)

99 and identify contaminated regions in the assembly. We used the script to remove these
100 contaminant sequences and split scaffolds at locations of contamination. For example, a 100-base
101 scaffold with contamination from 15 through 30 would be split into two scaffolds—one 14 bases
102 long (corresponding to bases 1–14) and one 70 bases long (corresponding to bases 31–100;
103 Figure 2). We ran this script with the following command line:

104

```
105 python Remove_and_split_contamination_NCBI.py --fasta GopAga1.0_NCBIout.fasta --output  
106 GopAga1.1_nocontam.fasta --ncbi_tab RemainingContamination.txt --wrap_length 90 --  
107 delimiters “.” “,” --fasta_id_junk “|cl|”
```

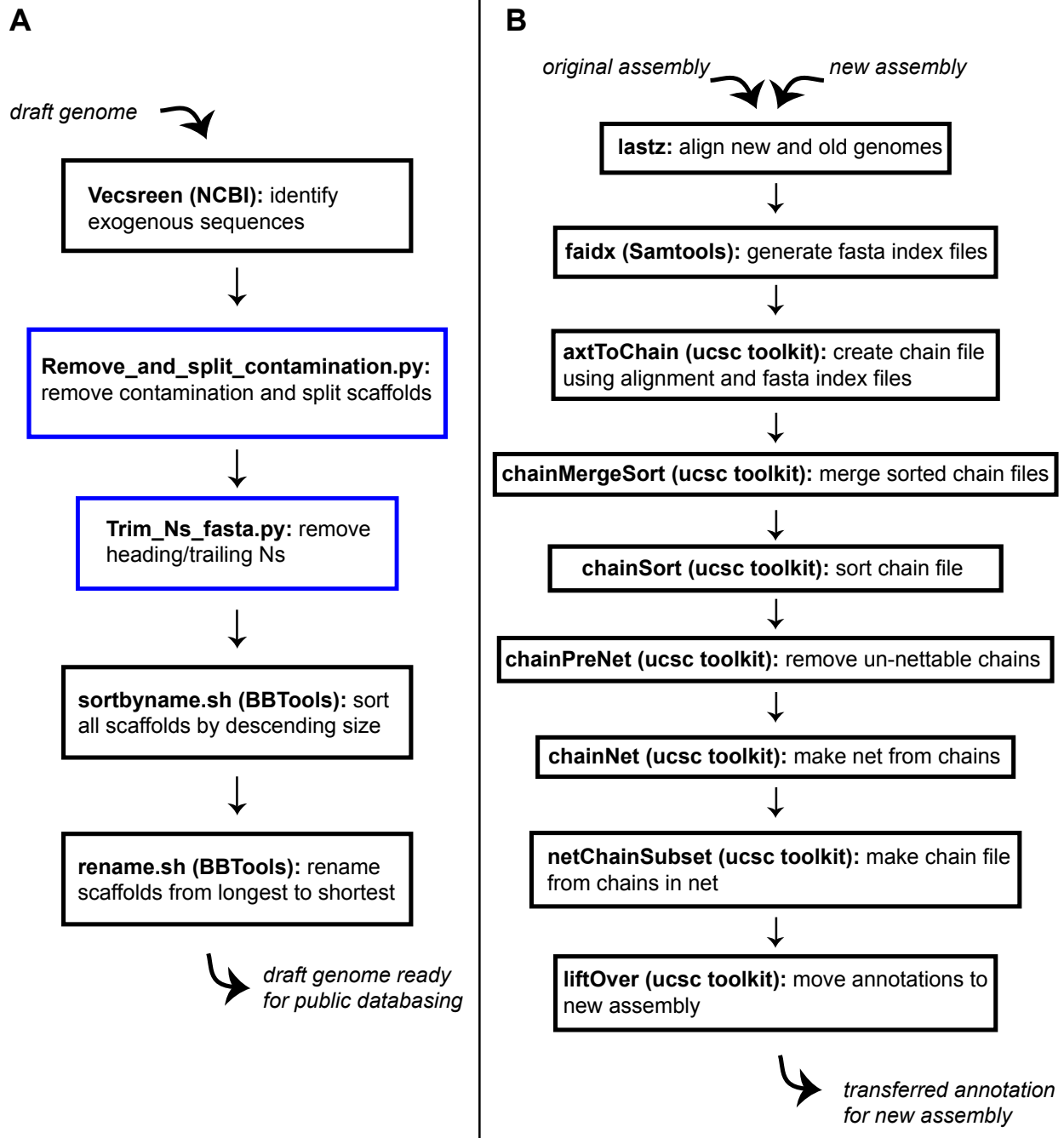


Figure 1. Overview of assembly (A) and annotation (B) processes used for *gopAga1.1*. Blue boxes are scripts presented here; black boxes are tools provided by other software packages. Descriptions of external tools are reproduced here from their original source documentation.

108

109

110 When the estimated physical distance between regions is known (e.g., through mate-pair
111 sequencing), Ns are often used to fill in unknown sequences between contigs. Many
112 contamination sites were adjacent to these strings of Ns, suggesting at least some contamination
113 was introduced during scaffolding steps. In these cases, splitting scaffolds at the sites of
114 contamination left the newly split scaffolds with long strings of either leading or trailing strings
115 of Ns. Using a second Python script (`Trim_Ns_fasta.py`), we dynamically trimmed these patterns
116 and removed any remaining contigs and scaffolds less than 100 bp in length with the following
117 command line:

118

```
119           python Trim_Ns_fasta.py --fasta INFILE.fasta --output_fasta OUTFILE.fasta --  
120 filtered_scaffolds removed_scaffolds.txt --wrap_length 60 --soft_buffer 10 --min_n 1 --  
121 minimum_length 100
```

(contamination: 15..30)

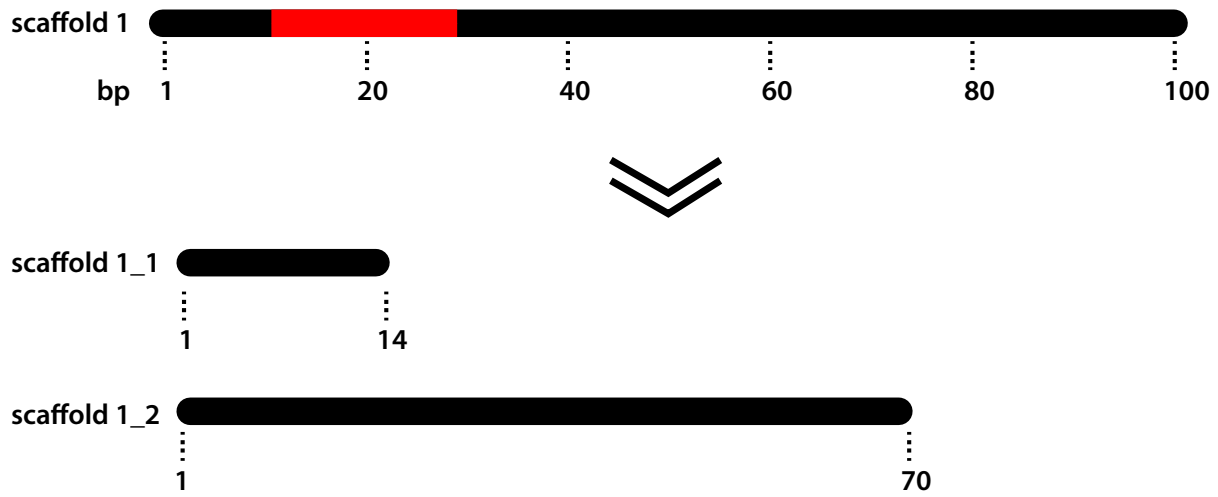


Figure 2. Schematic showing how the contaminant removal and scaffold splitting processes work. The contaminated regions (red) provided by Vecscreen output are inclusive, 1-based coordinates and are removed, leaving two new unassociated scaffolds. New scaffolds are named numerically using the original scaffold number (e.g., scaffold 1_1 and scaffold 1_2).

122

123 The *soft_buffer* parameter directs the program to remove up to and including the provided
 124 length before looking for Ns. However, no clipping will occur if an N is not discovered. A --
 125 *hard_buffer* option is also implemented in the program, which will instead hard-clip a sequence
 126 by a certain length, whether or not an N is discovered.

127 We then used *sortbyname.sh*, in the BBTools suite
 128 (<https://sourceforge.net/projects/bbmap>), to sort scaffolds by descending length with the
 129 command:

130

131 *sortbyname.sh in=GopAga1.1_unsorted.fasta out=GopAga1.1.sorted.fasta length*
 132 *descending*

133

134 We used *rename.sh*, also part of the BBTools suite
135 (<https://sourceforge.net/projects/bbmap>), to rename the split and sorted scaffolds for version 1.1.
136 When draft genomes are assembled using multiple software tools, it can result in subtly different
137 scaffold naming schemes that can cause confusion (i.e., scaffold_412 vs. scaffold412 vs.
138 Scaffold412). Here we used increasing numbers for scaffold names corresponding to decreasing
139 length. As such, the longest scaffold is named scaffold_0, the next longest is scaffold_1, and so
140 on. We achieved this using the following command:

```
141  
142     rename.sh in=GopAga1.1.sorted.fasta out=GopAga1.1.sorted.renamed.fasta  
143     prefix=scaffold
```

144
145 We performed manual quality control assessments at each step in these processes. Such
146 assessments included visual examination of split and excised scaffold regions in comparison to
147 Vecscreen output, comparing number of scaffolds pre- and post-splitting to the number of
148 contaminated regions, visual examination of pre- and post-soft-clipped scaffolds, comparing file
149 sizes before and after each step, and comparing checksums when moving files. We also used
150 standard UNIX tools to count scaffolds, changes in nucleotide composition, and check scaffold
151 names. Finally, we compared sequence statistics between the two assemblies using *stats.sh* in the
152 BBTools suite (<https://sourceforge.net/projects/bbmap>). We only modified the “minscf”
153 parameter to calculate statistics with different minimum scaffold sizes. Note that in this
154 manuscript, we use “scaffold” in a broad sense, to refer to any sequence in the assembly with an
155 identifier. This will include both scaffolds containing contigs joined during a scaffolding process
156 and unscaffolded contigs.

157

158 ***Annotation version 1.1***

159 The annotation for version 1 was generated using *ab initio* gene model predictions
160 combined with deep transcriptome mRNA transcription evidence from four adult tissues,
161 including blood, brain, lung, and skeletal muscle (Tollis *et al.* 2017). This original annotation
162 produced a similar number of protein-coding genes (20,172) to western painted turtle and
163 Chinese softshell turtle (21,796 and 19,327, respectively; (Shaffer *et al.* 2013; Wang *et al.* 2013).
164 As part of gopAga1.1 we lifted this *de novo* annotation for gopAga1 onto the gopAga1.1
165 assembly using the following methodology. First, we aligned genome assemblies for versions 1.0
166 and 1.1 using *lastz_32* (Harris 2007). After trying several different combinations of parameters,
167 the best results were produced using gapping, nochain, nogfextend, mismatch=(0,100), exact=20,
168 step=30, notransition, notwins, traceback=160.0M and seed=match12 with output format as .axt.
169 Importantly, the v1.1 assembly is a subset of v1.0 and has no nucleotide differences aside from
170 the removal of contamination and training Ns, which may be an uncommon scenario for these
171 alignment tools. We converted the output alignment (.axt) file to a chain file using the
172 *axtToChain* tool from ucsc toolkit (<http://genome.cse.ucsc.edu/index.html>). We sorted the chain
173 file by score using *chainSort* and removed chains that would not be netted using *chainPreNet*.
174 We performed netting with *netChainSubset* to create larger blocks of chains and used the –
175 skipMissing parameter because our chains were filtered. Using this final chain file, we lifted over
176 the annotation using *liftOver* from ucsc toolkit.

177

178 ***Removing small scaffolds***

179 NCBI submission requires assemblies to contain only scaffolds with sequence lengths
180 greater than or equal to 200 nucleotides. To filter the FASTA assembly itself, we used bioawk
181 (<https://github.com/lh3/bioawk>):

182

```
183 bioawk -c fastx '(length($seq) > 199) {print ">"$name"\n"$seq}'
```

```
184 GopAga1.1.sorted.renamed.fasta > GopAga1.1.sorted.renamed.min200.fasta
```

185

186 We then created a BED file of scaffolds removed in the above command, determined by
187 comparing fasta indexes generated with SAMtools faidx (Li *et al.*, 2009), and used BEDTools
188 (Quinlan & Hall 2010) to subtract annotations on these filtered scaffolds:

189

```
190 bedtools subtract -A -a GopAga1.1.annotation_final.gff -b GopAga1.1_min200.bed >
```

```
191 GopAga1.1.annotation_final_above200.gff
```

192

193 ***Data and Software availability***

194 The Python scripts described above, *Remove_and_split_contamination.py* and
195 *Trim_Ns_fasta.py*, are freely available on Github
196 (https://github.com/thw17/G_agassizii_reference_update) and in the Supporting Information. We
197 have deposited the fasta sequence and annotation files for gopAga1.1 in the Harvard Dataverse
198 (doi:10.7910/DVN/HUASUW).

199

200

201

202

203 **RESULTS & DISCUSSION**

204 GopAga1 and gopAga1.1—which we alternatively refer to in this manuscript as v1.0 and
205 v1.1, respectively—differ in a few important ways. First, we removed contaminant sequences
206 (primarily adapters) present in v1.0 and split scaffolds around sites of contamination. We
207 removed leading and trailing Ns from sequences before sorting and renaming scaffolds by size.
208 Finally, we removed all sequences smaller than 200 bases and lifted over the annotation to the
209 modified assembly. We outline the differences in resulting sequence statistics between v1.0 and
210 v1.1 below.

211

212 ***Genome assembly version 1.1***

213 While splitting scaffolds at sites of contamination initially increased the number of
214 scaffolds, removing scaffolds less than 200 bases led to a major overall reduction in the number
215 of scaffolds (v1.0: 863,216; v1.1: 172,559; Table 1). These procedures also led to a reduction in
216 assembly size, from 2.399 Gb in v1.0 to 2.184 Gb in v1.1 (Table 1). Of the removed sequences,
217 approximately 58% consisted of either contamination or Ns, while the remaining 42% were
218 removed because they were under the 200 bp threshold.

219 Filtering and trimming also affected other genome statistics. We measured N50 (more
220 than 50% of the genome is found in scaffolds this size scaffold or larger) and L50 (minimum
221 number of scaffolds containing 50% or more of the total sequence length) on scaffolds greater
222 than 200 bp in versions 1.0 and 1.1. In v1.0, the N50 was 251 KB and L50 was 2592 scaffolds
223 (Table 1); in v1.1, N50 was 228 KB and L50 was 2740 scaffolds (Table 1). This effect was
224 largely driven by splitting some larger scaffolds that may have been joined by contamination

225 because the differences between v1.0 and v1.1 are more pronounced when only considering
 226 scaffolds longer than 200 bp, which were the scaffolds primarily affected in the contaminant
 227 processing (v1.0: N50 = 265 KB , L50 = 2418; v1.1: N50 = 228 KB , L50 = 2740; Table 1).

228
 229
 230
 231

Table 1. Comparison of draft assembly statistics.

	<i>gopAga1.0_min0^a</i>	<i>gopAga1.0_min200^b</i>	<i>gopAga1.1^c</i>
Total Length^d	2,399,952,228	2,309,856,185	2,184,968,471
Num. Scaffolds^e	863,216	189,565	172,559
Longest Scaffold^f	2,046,553	2,046,553	1,743,037
L50/N50^g	2592/251 KB	2418/265 KB	2740/228 KB
L90/N90^h	13,331/19 KB	10,799/35 KB	10647/43 KB
%GCⁱ	43.85%	43.70%	43.62%
%N^j	1.55%	1.61%	1.51%

232 ^aVersion 1.0 of the *G. agassizii* genome containing all scaffolds.

233 ^bVersion 1.0 of the *G. agassizii* genome containing only scaffolds greater than 199 bp.

234 ^cVersion 1.1 of the *G. agassizii* genome (which contains only scaffolds greater than 199 bp).

235 ^dTotal length of the assembly, including Ns.

236 ^eTotal number of named scaffolds in the assembly.

237 ^fSequence length of the longest scaffold in the assembly.

238 ^gL50 is the minimum number of scaffolds containing 50% or more of the assembly. 50% of the
 239 genome is found in scaffolds of length N50 or greater.

240 ^hL90 is the minimum number of scaffolds containing 90% or more of the assembly. 90% of the
 241 genome is found in scaffolds of length N90 or greater.

242 ⁱPercent of total sequence that is G or C.

243 ^jPercent of total sequence that is N.

244
 245

246 **Annotation version 1.1**

247 Annotation of the draft genome v1.0 identified 20,172 genes, of which 17,201 are present
 248 in the v1.1 assembly. Of the 2,971 genes not lifted over in the v1.1 genome, 2,731 of those failed
 249 to lift over due to being split across scaffolds in the v1.1 assembly, 118 were partially deleted in

250 the v1.1 assembly, and 122 were fully deleted in the new scaffolds. These results indicate that
251 the assembly and/or annotation of some genic or gene-associated regions may have been
252 influenced by exogenous sequence in the v1.0 assembly, and may reflect a common challenge of
253 draft genomes.

254

255 CONCLUSIONS

256 The draft genome of *Gopherus agassizii*, the first tortoise sequenced, advances
257 conservation biology and management of this species and comparative genomic studies. Here we
258 improve the *G. agassizii* draft genome in version 1.1. Improvements include removing
259 contamination, splitting scaffolds at sites of internal contamination, reordering and renaming
260 scaffolds, and transferring the v1.0 annotation coordinates to v1.1. We include scripts and
261 detailed commands to aid in processing other draft genomes, which often require similar filtering
262 and restructuring, particularly for deposition into public databases. A particularly important
263 message is that adaptor contamination can present a major challenge for short read assemblers,
264 causing reads and contigs to misassemble (Alkan *et al.* 2010; Schmieder & Edwards 2011;
265 Bolger *et al.* 2014) and leading to errors in contiguity and/or synteny. Generally speaking, these
266 errors fell in intergenic regions, though a number of genes were impacted. Care must be taken to
267 include an exhaustive list of adaptors used by sequencing projects to ensure that trimming
268 programs are using all potentially relevant sequences.

269 We believe that the continued development of this resource will enable new, promising
270 directions in tortoise research and conservation. In particular, this resource allows
271 reconstructions of modern and historical demographic patterns with greater statistical power. It
272 enables researchers to disentangle the history of gene flow and ecological adaptations that

273 differentiate *G. agassizii* from *G. morafkai*. And finally, it can aid in the characterization of the
274 immune system of chelonians, leading to a better understanding of why URTD affects tortoise
275 species differently and development of better diagnostics for detection, which would benefit
276 management of the species.

277

278 ACKNOWLEDGEMENTS

279 We thank Marc Tollis, Dale DeNardo, John Cornelius, Taylor Edwards, Cristina Jones, and
280 Mariana Grizante Bortoletto for helpful conversations and ongoing collaborations.

281

282 REFERENCES

- 283 Alkan C, Sajjadian S, Eichler EE (2010) Limitations of next-generation genome sequence
284 assembly. *Nature Methods* 8:61–65.
- 285 Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence
286 data. *Bioinformatics* 30:2114–2120.
- 287 Brown MB, Schumacher IM, Klein PA *et al.* (1994) *Mycoplasma agassizii* causes upper
288 respiratory tract disease in the desert tortoise. *Infection and Immunity* 62:4580–4586.
- 289 Doak D, Kareiva P, Klepetka B (1994) Modeling population viability for the Desert Tortoise in
290 the western Mojave Desert. *Ecological Applications* 4:446–460.
- 291 Dorsey RJ, Housen BA, Janecke SU, Fanning CM, Spears ALF (2011) Stratigraphic record of
292 basin development within the San Andreas fault system: Late Cenozoic Fish Creek–
293 Vallecito basin, southern California. *Geological Society of America Bulletin* 123:771–793.
- 294 Drake KK, Bowen L, Nussear KE *et al.* (2016) Negative impacts of invasive plants on
295 conservation of sensitive desert wildlife. *Ecosphere* 7:e01531–20.
- 296 Edwards T, Berry KH, Inman RD *et al.* (2015) Testing taxon tenacity of tortoises: evidence for a
297 geographical selection gradient at a secondary contact zone. *Ecology and Evolution* 5:2095–
298 2114.
- 299 Edwards T, Tollis M, Hsieh P *et al.* (2016) Assessing models of speciation under different
300 biogeographic scenarios; an empirical study using multi-locus and RNA-seq analyses.
301 *Ecology and Evolution* 102:1–18.
- 302 Harris RS (2007) Improved pairwise alignment of genomic DNA. D. Phil. Thesis. The
303 Pennsylvania State University.
- 304 Jacobson ER, Gaskin JM, Brown MB *et al.* (1991) Chronic Upper Respiratory Tract Disease of
305 free-ranging desert tortoises (*Xerobates agassizii*). *Journal of Wildlife Diseases* 27:296–316.
- 306 Li H, Handsaker B, Wysoker A *et al.* (2009) The Sequence Alignment/Map format and
307 SAMtools. *Bioinformatics* 25:2078–2079.
- 308 Morafka DJ, Berry KH (2002) Is *Gopherus agassizii* a desert-adapted tortoise or an exaptive

- 309 opportunist? Implications for tortoise conservation. *Chelonian Conservation and Biology*
310 4:263–287.
- 311 Murphy R, Berry K, Edwards T *et al.* (2011) The dazed and confused identity of Agassiz’s land
312 tortoise, *Gopherus agassizii* (Testudines: Testudinidae) with the description of a new species
313 and its consequences for conservation. *Zookeys* 113:39–71.
- 314 Pianka ER (1970) Comparative autecology of the lizard *Cnemidophorus tigris* in different parts
315 of its geographic range. *Ecology* 51:703–720.
- 316 Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic
317 features. *Bioinformatics* 26:841–842.
- 318 Reynolds JF, Kemp PR, Ogle K, Fernández RJ (2004) Modifying the “Pulse-Reserve” paradigm
319 for deserts of North America: precipitation pulses, soil water, and plant responses. *Oecologia*
320 141:194–210.
- 321 Schmieder R, Edwards R (2011) Fast identification and removal of sequence contamination from
322 genomic and metagenomic datasets (F Rodriguez-Valera, Ed.). *PLoS ONE* 6:e17288–11.
- 323 Shaffer HB, Minx P, Warren DE *et al.* (2013) The western painted turtle genome, a model for
324 the evolution of extreme physiological adaptations in a slowly evolving lineage. *Genome*
325 *Biology* 14:R28.
- 326 Smith R (1990) Endangered and threatened wildlife and plants; determination of threatened
327 status for the Mojave population of the desert tortoise. *Federal Registrar* 55:12178–12191.
- 328 Tollis M, DeNardo DF, Cornelius JA *et al.* (2017) The Agassiz's desert tortoise genome provides
329 a resource for the conservation of a threatened species. *PLoS ONE* 12:e0177708.
- 330 Wang Z, Pascual-Anaya J, Zadissa A *et al.* (2013) The draft genomes of soft-shell turtle and
331 green sea turtle yield insights into the development and evolution of the turtle-specific body
332 plan. *Nature Genetics* 45:701–706.

333
334

335 AUTHOR CONTRIBUTIONS

336 THW and GAD conducted analyses. GAD and THW wrote the manuscript. KK and MWS
337 provided oversight and computing resources. THW, GAD, MWS, and KK edited the manuscript.

338

339 FUNDING

340 GAD and KK were supported by a US Geological Survey Cooperative Ecosystem Studies Units
341 (CESU) award, GAD and THW were supported by a Fostering Postdoctoral Research in the Life
342 Sciences seed grant from the School of Life Sciences at Arizona State University. All authors
343 were supported by a Heritage Grant from the Arizona Game and Fish Department. This work
344 was supported by funding from the College of Liberal Arts and Sciences at ASU to KK.

345 Computational analysis was supported by allocations from Research Computing at Arizona State
346 University.
347