Improved genome assembly and annotation for the rock pigeon (*Columba livia*)

Carson Holt[*,†], Michael Campbell[*,1], David A. Keays[‡], Nathaniel Edelman[‡], Aurélie Kapusta[*,†], Emily Maclary[§], Eric Domyan[§,**], Alexander Suh[††], Wesley C. Warren[‡‡], Mark Yandell[*,†], M. Thomas P. Gilbert[§§,***], Michael D. Shapiro[§]

* Department of Human Genetics, University of Utah, Salt Lake City, UT, USA

† USTAR Center for Genetic Discovery, University of Utah, Salt Lake City, UT, USA

‡ Research Institute of Molecular Pathology, Vienna, Austria

§ Department of Biology, University of Utah, Salt Lake City, UT, USA

** Department of Biology, Utah Valley University, Orem, UT, USA

†† Department of Evolutionary Biology (EBC), University of Uppsala, Uppsala, Sweden

‡‡ Genome Institute at Washington University, St. Louis, MO, USA

§§ Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark

*** Norwegian University of Science and Technology, University Museum, 7491 Trondheim, Norway

1 Current address: Division of Plant Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA

Accession numbers:

BioProject: PRJNA167554

Genome: This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession AKCR00000000. The version described in this paper is version AKCR02000000.

RNAseq data: SAMN07417936, SAMN07417937, SAMN07417938, SAMN07417939, SAMN07417940, SAMN07417941, SAMN07417942, SAMN07417943

Running title:

"HiRise genome assembly of rock pigeon"

Keywords:

Columba livia, rock pigeon, HiRise assembly, MAKER annotation

Author for Correspondence:

Michael D. Shapiro, Department of Biology, University of Utah, 257 S 1400 E, Salt Lake City, UT 84108, shapiro@biology.utah.edu, +1 801 581 5690

Abstract

The domestic rock pigeon (*Columba livia*) is among the most widely distributed and phenotypically diverse avian species. This species is broadly studied in ecology, genetics, physiology, behavior, and evolutionary biology, and has recently emerged as a model for understanding the molecular basis of anatomical diversity, the magnetic sense, and other key aspects of avian biology. Here we report an update to the *C. livia* genome reference assembly and gene annotation dataset (Cliv_1.0). Greatly increased scaffold lengths in the updated reference assembly, along with an updated annotation set, provide improved tools for evolutionary and functional genetic studies of the pigeon, and for comparative avian genomics in general.

Introduction

Intensive selective breeding of the domestic rock pigeon (*Columba livia*) has resulted in over 350 breeds with extreme differences in morphology and behavior (Levi 1986; Domyan and Shapiro 2017). The large phenotypic differences among different breeds make them a useful model for studying the genetic basis of radical phenotypic changes, which are more typically found among different species rather than within a single species.

In genetic and genomic studies of *C. livia*, linkage analysis is important for identifying genotypes associated with specific phenotypic traits of interest (Domyan and Shapiro 2017); however, short scaffold sizes in the Cliv_1.0 draft reference assembly (Shapiro et al. 2013) hinder computationally-based comparative analyses. Short scaffolds also make

61 it more difficult to identify structural changes, such as large insertions or deletions, that
62 are responsible for traits of interest (Domyan et al. 2014; Kronenberg et al. 2015).
63
64 Here we present the Cliv_2.0 reference assembly and an updated gene annotation set. The
65 new assembly greatly improves scaffold length over the previous draft reference
66 assembly, and updated gene annotations show improved concordance with both
67 transcriptome and protein homology evidence.
68
69 **Methods & Materials**
70 *Genome sequencing and assembly*
71 Genomic DNA from a female Danish tumbler pigeon (full sibling of the male bird used
72 for the original Cliv_1.0 assembly (Shapiro et al. 2013)) was used to produce long-range
73 sequencing libraries using the "Chicago" (Putnam et al. 2016) method by Dovetail
74 Genomics (Santa Cruz, CA). Two Chicago libraries were prepared and sequenced on the
75 Illumina HiSeq platform to a final physical coverage (1-50 kb pairs) of 390x (see Table
76 1).
77
78 Scaffolding was performed by Dovetail Genomics using HiRise assembly software and
79 the Cliv_1.0 assembly as input. Briefly, Chicago reads were aligned to the input assembly
80 to identify and mask repetitive regions, and then a likelihood model was applied to
81 identify mis-joins and score prospective joins for scaffolding. The final assembly was
82 then filtered for length and gaps according to NCBI submission specifications.
83

84 *Genome annotation*

85 The pre-existing reference Gnomon (Souvorov et al. 2010) derived gene models for the
86 Cliv_1.0 assembly (GCA_000337935.1) were mapped onto the updated Cliv_2.0
87 reference assembly using direct alignment of transcript FASTA entries. This was done
88 using the alignment workflow of the genome annotation pipeline MAKER (Cantarel et al.
89 2008; Holt and Yandell 2011), which first seeds alignments using BLASTN (Altschul et
90 al. 1990) and then polishes the alignments around splice sites using Exonerate (Slater and
91 Birney 2005). Results were then filtered to remove alignments that had an overall match
92 of less than 90% of the original model (match is calculated as percent identity multiplied
93 by percent end-to-end coverage).

94

95 For final annotation, MAKER was allowed to identify *de novo* gene models that did not
96 overlap the aligned Gnomon models. Protein evidence sets used by MAKER included
97 annotated proteins from *Pterocles gutturalis* (yellow-throated sandgrouse) (Zhang et al.
98 2014) and *Gallus gallus* (chicken) (International Chicken Genome Sequencing 2004)
99 together with all proteins from the UniProt/Swiss-Prot database (Bairoch and Apweiler
100 2000; UniProt 2007). The transcriptome evidence sets for MAKER included Trinity
101 (Grabherr et al. 2011) mRNA-seq assemblies from multiple *C. livia* breeds and tissues
102 (methods for transcriptome assembly are described below). Gene predictions were
103 produced within MAKER by Augustus (Stanke and Waack 2003; Stanke et al. 2008)
104 trained against the Cliv_1.0 Gnomon gene models. Repetitive elements in the genome
105 were identified using a custom repeat library.

106

107 *Custom repeat library*

108 A repeat library for *C. livia* was built by combining libraries from existing avian species
109 (Zhang et al. 2014) together with with repeats identified *de novo* for the Cliv_2.0
110 assembly. *De novo* repeat identification was performed using RepeatScout (Price et al.
111 2005) with default parameters (>3 copies) to generate consensus repeat sequences.
112 Identified repeats with greater than 90% sequence identity and a minimum overlap of 100
113 bp were assembled using Sequencher (Yokouchi et al. 1993). Repeats were classified into
114 transposable element (TE) families using multiple lines of evidence, including homology
115 to known elements, presence of terminal inverted repeats (TIRs), and detection of target
116 site duplications (TSDs). Homology-based evidence was obtained using RepeatMasker
117 (Smit et al. 1996), as well as the homology module of the TE classifying tool RepClass
118 (Feschotte et al. 2009). RepClass was also used to identify signatures of transposable
119 elements (TIRs, TSDs). We then eliminated non-TE repeats (simple repeats or gene
120 families), using custom Perl scripts (available at https://github.com/4ureliek/ReannTE).
121

122 In our custom repeat analysis, using the script ReannTE_FilterLow.pl, consensus
123 sequences were labeled as simple repeats or low complexity repeats if 80% of their length
124 could be annotated as such by RepeatMasker (the library was masked with the option -
125 noint). Next using the ReannTE_Filter-mRNA.pl script, consensus sequences were
126 interrogated against RefSeq (Pruitt et al. 2007) mRNAs (as of March 7th 2016) with
127 TBLASTX (Altschul et al. 1990). Sequences were eliminated from the library when: (i)
128 the e-value of the hit was lower than 1E-10; (ii) the consensus sequence was not
129 annotated as a TE; and (iii) the hit was not annotated as a transposase or an unclassified

130 protein. The script ReannTE_MergeFasta.pl was then used to merge our library with a
131 library combining RepeatModeler (Smit and Hubley 2008) outputs from 45 bird species
132 (Kapusta et al. 2017) and complemented with additional avian TE annotations
133 (International Chicken Genome Sequencing 2004; Warren et al. 2010; Bao et al. 2015).
134 Merged outputs were then manually inspected to remove redundancy, and all DNA and
135 RTE class transposable elements were removed and replaced with manually curated
136 consensus sequences.
137
138 *Transcriptomics Methods*
139 RNA was extracted from adult tissues (brain, retina, subepidermis, cochlear duct, spleen,
140 olfactory epithelium) of the racing homer breed, and one whole embryo each of a racing
141 homer and a parlor roller (approximately embryonic stage 25 (Hamburger and Hamilton
142 1951)). RNA-seq libararies were prepared and sequenced using 100-bp paired-end
143 sequencing on the Illumina HiSeq 2000 platform at the Research Institute of Molecular
144 Pathology, Vienna (adult tissues), and the Genome Institute at Washington University, St.
145 Louis (embryos). RNA-seq data generated for the Cliv_1.0 annotation were also
146 downloaded from the NCBI public repository for *de novo* re-assembly. Accession
147 numbers for the public data are SRR521357 (Danish tumbler heart), SRR521358 (Danish
148 tumbler liver), SRR521359 (Oriental frill heart), SRR521360 (Oriental frill liver),
149 SRR521361 (Racing homer heart), and SRR521362 (Racing homer liver).
150
151 Each FASTQ file was processed with FastQC (http://www.bioinformatics.babraham.ac.
152 uk/projects/fastqc/) to assess quality. When FastQC reported overrepresentation of

153 Illumina adapter sequences, we trimmed these sequences with fastx_clipper from the
154 FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). We used FASTX-Toolkit for
155 two additional functions: runs of low quality bases at the start of reads were trimmed with
156 fastx_trimmer when necessary (quality cutoff of -Q 33), and reads were then trimmed
157 with fastq_quality_trimmer (-Q 33). Finally, each pair of sequence files was assembled
158 with Trinity (Grabherr et al. 2011) version r20131110 using the --jaccard_clip option.
159
160 *Linkage map construction and anchoring to current assembly*
161 Genotyping by sequencing (GBS) data was generated, trimmed, and filtered as previously
162 described (Domyan et al. 2016). Reads were mapped to the Cliv_2.0 assembly using
163 Bowtie2 (Langmead and Salzberg 2012). Genotypes were called using Stacks (Catchen et
164 al. 2011), with a minimum read-depth cutoff of 10. Thresholds for automatic corrections
165 were set using the parameters –min_hom_seqs 10, --min_het_seqs 0.01, --max_het_seqs
166 0.15. Sequencing coverage and genotyping rate varied between individuals, and birds
167 with genotyping rates in the bottom 25% were excluded from map assembly.
168
169 Genetic map construction was performed using R/qtl (www.rqtl.org) (Broman et al.
170 2003). For autosomal markers, markers showing segregation distortion (Chi-square, $p <$
171 0.01) were eliminated. Sex-linked scaffolds were assembled and ordered separately, due
172 to differences in segregation pattern for the Z-chromosome. Z-linked scaffolds were
173 identified by assessing sequence similarity and gene content between pigeon scaffolds
174 and the Z-chromosome of the annotated chicken genome (Ensembl Gallus_gallus-5.0).
175

176 Pairwise recombination frequencies were calculated for all autosomal and Z-linked
177 markers. Missing data were imputed using "fill.geno" with the method "no_dbl_XO".
178 Duplicate markers were identified and removed. Within individual scaffolds, R/Qtl
179 functions "droponemarker" and "calc.errorlod" were used to assess genotyping error.
180 Markers were removed if dropping the marker led to an increased LOD score, or if
181 removing a non-terminal marker led to a decrease in length of >10 cM that was not
182 supported by physical distance. Individual genotypes were removed if they showed with
183 error LOD scores >5 (Lincoln and Lander 1992). Linkage groups were assembled from
184 2960 autosomal markers and 232 Z-linked markers using the parameters (max.rf 0.1,
185 min.lod 6). In the rare instance that single scaffolds were split into multiple linkage
186 groups, linkage groups were merged if supported by recombination frequency data; these
187 instances typically reflected large physical gaps between markers on a single scaffold.
188 Scaffolds in the same linkage group were manually ordered based on calculated
189 recombination fractions and LOD scores.
190
191 To compare the linkage map to the prior assembly (Cliv_1.0), each 90-bp locus
192 containing a genetic marker was parsed from the Stacks output file
193 "catalogXXX_tags.tsv" and queried to the Cliv_1.0 assembly using Nucleotide-
194 Nucleotide blast (v2.6.0+) with the parameters –max_target_seqs 1 – max hsps 1. 3175 of
195 the 3192 loci (99.47%) from the new assembly had a BLAST hit with an E-value < 4e-24
196 and were retained.
197
198 *Data availability*

199 This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under
200 the accession AKCR00000000. The version described in this paper is version
201 AKCR02000000. RNA-seq data are deposited in the SRA database with the BioSample
202 accession numbers SAMN07417936-SAMN07417943. Assembly and RNA-seq data are
203 publicly available in NCBI databases under BioProject PRJNA167554.
204

205 **Results and Discussion**

206 The final reference assembly is 1,108,534,737 base pairs in length and consists of 15,057
207 scaffolds (Table 1). A total of 1,015 scaffolds contain a gene annotation. Completion
208 analysis of the assembly using BUSCO (Simao et al. 2015) suggests that Cliv_2.0 is
209 72.9-86.2% complete which is nearly identical to the Cliv_1.0 assembly estimate of 72.3-
210 86.4% (Table 2). Thus, we found no significant changes to assembly completeness
211 between the two assemblies. The major improvement to the Cliv_2.0 assembly is rather
212 an increase in scaffold length (Fig. 1a). Overall, the N50 scaffold length increased to 14.3
213 megabases compared to 3.15 megabases for the previous reference assembly, a greater
214 than 4-fold increase. Recently, Damas et al. (Damas et al. 2017) used computational
215 methods and universal avian bacterial artificial chromosome (BAC) probes to achieve
216 chromosome-level scaffolding using the Cliv_1.0 assembly as input material; however,
217 this assembly is currently unannotated.
218

219 The new assembly joins scaffolds that we knew were adjacent but were separated
220 previously (see Table S1 for full catalog of positions of the original assembly in the new
221 assembly, and Table S2 for full catalog of breaks in the original assembly to form the

222 new assembly). For example, we previously determined that Cliv_1.0 Scaffolds 70 and
223 95 were joined based on genetic linkage data from a laboratory cross (Domyan et al.
224 2016). These two sequences are now joined into a single scaffold in the Cliv_2.0
225 assembly (see Table S3 for positions of genetic markers in Cliv_1.0 and Cliv_2.0). At
226 least one gene model (RefSeq LOC102093126), which was previously split across two
227 contigs, has now been unified into a single model on a single scaffold.
228
229 The updated annotation set contains 15,392 gene models encoding 18,966 transcripts (see
230 Table 3). This represents only a minor update of the reference annotation set as 94.7% of
231 previous models were mapped forward nearly unmodified (90% exact match for 14,898
232 out of 15,724 previous gene models) and only 494 new gene models were added to the
233 Cliv_2.0 annotation set (see Table 4).
234
235 The updated annotation set shows a modest improvement in concordance with aligned
236 evidence datasets from mRNA-seq and cross species protein homology evidence relative
237 to the Cliv_1.0 set as measured by Annotation Edit Distance (AED) (Eilbeck et al. 2009;
238 Holt and Yandell 2011). As a result, transcript models in the Cliv_2.0 annotation tend to
239 have lower AED values than the Cliv_1.0 set (Figure 2; the CDF curve is shifted to the
240 left). Lower AED values indicate greater model concordance with aligned transcriptome
241 and protein homology data. Furthermore, the Cliv_2.0 dataset displays greater transcript
242 counts in every AED bin despite having slightly fewer transcripts overall compared to the
243 Cliv_1.0 dataset (Table S4). The higher bin counts indicate that lower AED values are

244 not solely a result of removing unsupported models from the annotation set, but rather
245 suggest that evidence concordance has improved overall.
246
247 The improved scaffold lengths as well as updated gene model annotations should further
248 empower ongoing studies to identify genes responsible for phenotypic traits of interest
249 and improve detection of regions under selection due to longer scaffolds. We also expect
250 to be able to better identify large deletions and other structural variants responsible for
251 specific phenotypes now that they can be more clearly mapped to longer scaffolds.
252 Finally, the new transcriptomic data provides tissue-specific expression profiles for
253 several adult tissue types and an important embryonic stage for the morphogenesis of
254 limbs, craniofacial structures, skin, and other tissues.
255
256 **Acknowledgements**
257 We thank Dovetail Genomics for their aid in scaffolding the assembly, and Julia
258 Carleton, and Anna Vickrey for technical support. This work was supported by National
259 Science Foundation grant DEB1149160 and National Institutes of Health (NIH) grant
260 R01GM115996 to MDS; NSF EAGER grant IOS 1561337 to MY; a European Research
261 Council starting grant 336724 and Austrian Science Fund (FWF) grant Y726 to DAK;
262 and European Research Council Consolidator grant 681396 to MTPG. We gratefully
263 acknowledge research support from Boehringer Ingelheim at the Research Institute of
264 Molecular Pathology, and support and resources from the Center for High Performance
265 Computing at the University of Utah.
266

References

Altschul, S. F., W. Gish, W. Miller, E. W. Meyers, and D. J. Lipman. 1990. Basic Local Alignment Search Tool. Journal of Molecular Biology 215:403-410.

Bairoch, A. and R. Apweiler. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucl. Acids Res. 28:45-48.

Bao, W., K. K. Kojima, and O. Kohany. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA 6:11.

Broman, K., H. Wu, S. Sen, and G. Churchill. 2003. R/qtl: QTL mapping in experimental crosses. Bioinformatics 19:889-890.

Cantarel, B. L., I. Korf, S. M. C. Robb, G. Parra, E. Ross, B. Moore, C. Holt, A. Sanchez Alvarado, and M. Yandell. 2008. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 18:188-196.

Catchen, J. M., A. Amores, P. Hohenlohe, W. Cresko, and J. H. Postlethwait. 2011. Stacks: building and genotyping loci de novo from short-read sequences. G3 1:171-182.

Damas, J., R. O'Connor, M. Farre, V. P. E. Lenis, H. J. Martell, A. Mandawala, K. Fowler, S. Joseph, M. T. Swain, D. K. Griffin, and D. M. Larkin. 2017. Upgrading short-read animal genome assemblies to chromosome level using comparative genomics and a universal probe set. Genome Res 27:875-884.

Domyan, E. T., M. W. Guernsey, Z. Kronenberg, S. Krishnan, R. E. Boissy, A. I. Vickrey, C. Rodgers, P. Cassidy, S. A. Leachman, J. W. Fondon, 3rd, M. Yandell, and M. D. Shapiro. 2014. Epistatic and combinatorial effects of pigmentary gene mutations in the domestic pigeon. Curr Biol 24:459-464.

Domyan, E. T., Z. Kronenberg, C. R. Infante, A. I. Vickrey, S. A. Stringham, R. Bruders, M. W. Guernsey, S. Park, J. Payne, R. B. Beckstead, G. Kardon, D. B. Menke, M. Yandell, and M. D. Shapiro. 2016. Molecular shifts in limb identity underlie development of feathered feet in two domestic avian species. eLife 5:e12115.

Domyan, E. T. and M. D. Shapiro. 2017. Pigeonetics takes flight: Evolution, development, and genetics of intraspecific variation. Dev Biol 427:241-250.

Eilbeck, K., B. Moore, C. Holt, and M. Yandell. 2009. Quantitative measures for the management and comparison of annotated genomes. BMC Bioinformatics 10:67.

Feschotte, C., U. Keswani, N. Ranganathan, M. L. Guibotsy, and D. Levine. 2009. Exploring repetitive DNA landscapes using REPCLASS, a tool that automates the classification of transposable elements in eukaryotic genomes. Genome Biol Evol 1:205-220.

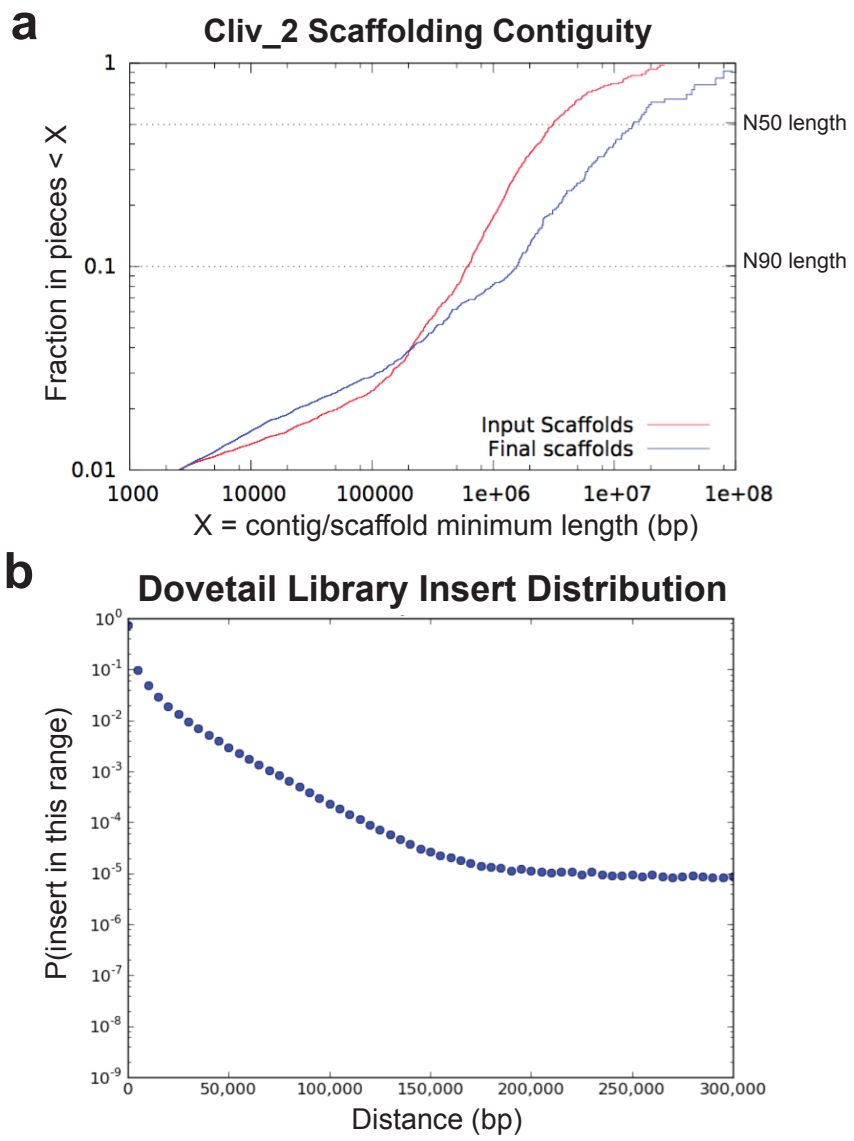
Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature biotechnology 29:644-652.

Hamburger, V. and H. L. Hamilton. 1951. A series of normal stages in the development of the chick embryo. Journal of Morphology 88:49-92.

310 Holt, C. and M. Yandell. 2011. MAKER2: an annotation pipeline and genome-database
311 management tool for second-generation genome projects. BMC Bioinformatics
312 12:491.
313 International Chicken Genome Sequencing, C. 2004. Sequence and comparative analysis
314 of the chicken genome provide unique perspectives on vertebrate evolution.
315 Nature 432:695-716.
316 Kapusta, A., A. Suh, and C. Feschotte. 2017. Dynamics of genome size evolution in birds
317 and mammals. Proceedings of the National Academy of Sciences 114:E1460-
318 E1469.
319 Kronenberg, Z. N., E. J. Osborne, K. R. Cone, B. J. Kennedy, E. T. Domyan, M. D.
320 Shapiro, N. C. Elde, and M. Yandell. 2015. Wham: Identifying Structural
321 Variants of Biological Consequence. PLoS Comput Biol 11:e1004572.
322 Langmead, B. and S. L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. Nat
323 Methods 9:357-359.
324 Levi, W. M. 1986. The Pigeon (Second Revised Edition). Levi Publishing Co., Inc.,
325 Sumter, S.C.
326 Lincoln, S. E. and E. S. Lander. 1992. Systematic detection of errors in genetic linkage
327 data. Genomics 14:604-610.
328 Price, A. L., N. C. Jones, and P. A. Pevzner. 2005. De novo identification of repeat
329 families in large genomes. Bioinformatics 21 Suppl 1:i351-358.
330 Pruitt, K. D., T. Tatusova, and D. R. Maglott. 2007. NCBI reference sequences (RefSeq):
331 a curated non-redundant sequence database of genomes, transcripts and proteins.
332 Nucleic Acids Res:D61 - 65.
333 Putnam, N. H., B. L. O'Connell, J. C. Stites, B. J. Rice, M. Blanchette, R. Calef, C. J.
334 Troll, A. Fields, P. D. Hartley, C. W. Sugnet, D. Haussler, D. S. Rokhsar, and R.
335 E. Green. 2016. Chromosome-scale shotgun assembly using an in vitro method
336 for long-range linkage. Genome Res 26:342-350.
337 Shapiro, M. D., Z. Kronenberg, C. Li, E. T. Domyan, H. Pan, M. Campbell, H. Tan, C. D.
338 Huff, H. Hu, A. I. Vickrey, S. C. Nielsen, S. A. Stringham, H. Hu, E. Willerslev,
339 M. T. Gilbert, M. Yandell, G. Zhang, and J. Wang. 2013. Genomic diversity and
340 evolution of the head crest in the rock pigeon. Science 339:1063-1067.
341 Simao, F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov.
342 2015. BUSCO: assessing genome assembly and annotation completeness with
343 single-copy orthologs. Bioinformatics 31:3210-3212.
344 Slater, G. and E. Birney. 2005. Automated generation of heuristics for biological
345 sequence comparison. BMC Bioinformatics 6:31.
346 Smit, A. F. and R. Hubley. 2008. RepeatModeler Open-1.0
347 http://www.repeatmasker.org/.
348 Smit, A. F., R. Hubley, and P. Green. 1996. RepeatMasker Open-3.0
349 http://www.repeatmasker.org/.
350 Souvorov, A., Y. Kapustin, B. Kiryutin, V. Chetvernin, T. Tatusova, and D. Lipman.
351 2010. Gnomon – NCBI eukaryotic gene prediction tool. NCBI.
352 Stanke, M., M. Diekhans, R. Baertsch, and D. Haussler. 2008. Using native and
353 syntenically mapped cDNA alignments to improve de novo gene finding.
354 Bioinformatics 24:637-644.

355 Stanke, M. and S. Waack. 2003. Gene prediction with a hidden Markov model and a new
356 intron submodel. Bioinformatics 19:ii215-225.
357 UniProt, C. 2007. The Universal Protein Resource (UniProt). Nucleic Acids Res:D193 -
358 197.
359 Warren, W. C., D. F. Clayton, H. Ellegren, A. P. Arnold, L. W. Hillier, A. Kunstner, S.
360 Searle, S. White, A. J. Vilella, S. Fairley, A. Heger, L. Kong, C. P. Ponting, E. D.
361 Jarvis, C. V. Mello, P. Minx, P. Lovell, T. A. Velho, M. Ferris, C. N.
362 Balakrishnan, S. Sinha, C. Blatti, S. E. London, Y. Li, Y. C. Lin, J. George, J.
363 Sweedler, B. Southey, P. Gunaratne, M. Watson, K. Nam, N. Backstrom, L.
364 Smeds, B. Nabholz, Y. Itoh, O. Whitney, A. R. Pfenning, J. Howard, M. Volker,
365 B. M. Skinner, D. K. Griffin, L. Ye, W. M. McLaren, P. Flicek, V. Quesada, G.
366 Velasco, C. Lopez-Otin, X. S. Puente, T. Olender, D. Lancet, A. F. Smit, R.
367 Hubley, M. K. Konkel, J. A. Walker, M. A. Batzer, W. Gu, D. D. Pollock, L.
368 Chen, Z. Cheng, E. E. Eichler, J. Stapley, J. Slate, R. Ekblom, T. Birkhead, T.
369 Burke, D. Burt, C. Scharff, I. Adam, H. Richard, M. Sultan, A. Soldatov, H.
370 Lehrach, S. V. Edwards, S. P. Yang, X. Li, T. Graves, L. Fulton, J. Nelson, A.
371 Chinwalla, S. Hou, E. R. Mardis, and R. K. Wilson. 2010. The genome of a
372 songbird. Nature 464:757-762.
373 Yokouchi, Y., M. Yamamoto, T. Toyota, H. Sasaki, and A. Kuroiwa. 1993. Regulatory
374 interaction of positional signalings on coordinate expression of homeobox genes
375 in developing limb buds. Limb Development and Regeneration. Wiley-Liss, Inc.
376 Zhang, G., B. Li, C. Li, M. T. Gilbert, E. D. Jarvis, J. Wang, and C. Avian Genome.
377 2014. Comparative genomic data of the Avian Phylogenomics Project.
378 Gigascience 3:26.
379

380 **FIGURES**



381

382 **Figure 1.** Assembly scaffolding contiguity and scaffolding library insert size

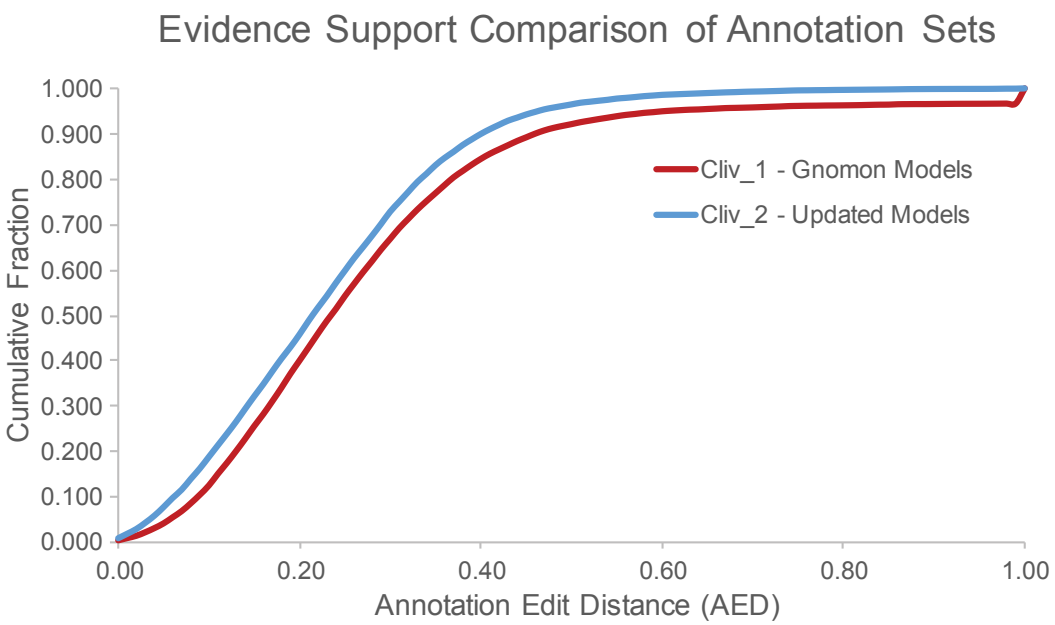
383 distributions. (a) Scaffolding comparison between Cliv_1.0 (input scaffolds) and Cliv_2.0

384 (final scaffolds) assemblies. (b) Distribution of Dovetail Genomics "Chicago" library

385 inserts.

386

387



388

389

390 **Figure 2.** Evidence support comparison of annotation sets. Annotation edit distance

391 (AED) support for gene models in Cliv_2.0 (red line) is improved over Cliv_1.0 (NCBI

392 Gnomon annotation, blue line).

393

394

395

396

397 **TABLES**

Table 1. Assembly statistics for Cliv_2.0

Estimated Physical Coverage	389.7x
Total Length	1,108,534,737bp
Total scaffolds	15,057
Total scaffolds >1kb	4,062
Total scaffolds >10kb	848

398

399

Table 2. Assembly version comparison

	Cliv_1.0	Cliv_2.0
Total Length	1110.8Mb	1110.9Mb
N50 Length	3.15Mb and 82 scaffolds	14.3Mb and 17 scaffolds
N90 Length	0.618Mb and 394 scaffolds	1.56Mb and 113 scaffolds
Completeness Estimate	72.3-86.4%	72.9-86.2%

400

401

402

403

Table 3. Annotation statistics

	Genes	Transcripts
Total	15,392	18,966
match[a]	14,898	18,472
new	494	494

[a] Count that match Cliv_1.0 annotations with a value of at least 90% (match is calculated as % identity multiplied by % end-to-end coverage)

404

Table 4. Annotation version comparison

	Cliv_1.0	Cliv_2.0
Total Gene Models	15,724	15,392
coding	15,022	14,683
non-coding	702	709
Total Transcripts	19,585	18,966
coding	18,569	18,148
non-coding	1016	818

405 **SUPPLEMENTAL TABLES**

406 **Table S1.** Tab-delimited table describing positions of Cliv_1.0 scaffolds in the Cliv_2.0
407 scaffolds. The table has the following format: column 1, Cliv_2.0 scaffold name; column
408 2, Cliv_1.0 sequence name; column 3, starting base (zero-based) of the Cliv_1.0
409 sequence; column 4, ending base of the Cliv_1.0 sequence; column 5, orientation of the
410 Cliv_1.0 sequence in the Cliv_2.0 scaffold, where (-) indicates that the Cliv_2.0 scaffold
411 sequence is reverse complemented relative to the Cliv_1.0 assembly; column 6, starting
412 base (zero-based) in the Cliv_2.0 scaffold; column 7, ending base in the Cliv_2.0
413 scaffold.

414

415 **Table S2.** Tab-delimited table describing positions of breaks made in the Cliv_1.0
416 assembly to create the Cliv_2.0 assembly. Data fields follow the same format that is used
417 in Supplemental Table 1.

418

419 **Table S3.** Table describing the linkage map assembled from genotype-by-sequencing
420 markers aligned to the Cliv_2.0 assembly, and relative positions of aligned markers
421 within the Cliv_2.0 and Cliv_1.0 genomes. The table has the following format: column 1,
422 Linkage map marker ID; column 2, Linkage group ID; column 3, Linkage map position;
423 column 4, Cliv_2.0 scaffold name; column 5, starting base (zero-based) of the alignment
424 in the Cliv_2.0 scaffold; column 6, alignment orientation in the Cliv_2.0 scaffold;
425 column 7, Cliv_1.0 scaffold name; column 8, starting base (zero-based) of the alignment
426 the Cliv_1.0 scaffold; column 9, alignment orientation in the Cliv_1.0 scaffold.

427

428 **Table S4.** Tab-delimited table describing transcript count and CDF binned by Annotation
429 Edit Distance (AED) values. AED is a modified sensitivity/specificity metric used to
430 compare annotation datasets to each other or to aligned transcriptome and protein
431 homology datasets. For calculating AED, sensitivity is defined as the fraction of a given
432 reference overlapping a prediction and measures false negative rates. For our purposes,
433 the prediction is a transcript model and the reference (or truth set) is a set of aligned
434 transcriptome and protein homology evidence. We calculate sensitivity using the formula
435 $SN = |p \cap r|/|r|$; where $|p \cap r|$ represents the number overlapping nucleotides between the
436 prediction and reference, and $|r|$ represents the total number of nucleotides in the
437 reference. Specificity is then defined as the fraction of a prediction overlapping a given
438 reference, and it measures false positive rates. We calculate specificity using the formula
439 $SP = |p \cap r|/|p|$. We then define concordance to be the average of sensitivity and specificity
440 ($C = (SN+SP)/2$), and AED is 1 minus the concordance ($AED = 1- C$). Transcript models
441 that have high AED values then show little concordance to aligned experimental
442 evidence, and models with low AED values show high concordance.
443