



HAL
open science

A pangenome graph reference of 30 chicken genomes allows genotyping of large and complex structural variants

Edward Rice, Antton Alberdi, James Alfieri, Giridhar Athrey, Jennifer Balacco, Philippe Bardou, Heath Blackmon, Mathieu Charles, Hans Cheng, Olivier Fedrigo, et al.

► To cite this version:

Edward Rice, Antton Alberdi, James Alfieri, Giridhar Athrey, Jennifer Balacco, et al.. A pangenome graph reference of 30 chicken genomes allows genotyping of large and complex structural variants. BMC Biology, 2023, 21 (1), 10.1186/s12915-023-01758-0 . hal-04346503

HAL Id: hal-04346503

<https://hal.inrae.fr/hal-04346503v1>

Submitted on 15 Dec 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

A pangenome graph reference of 30 chicken genomes allows genotyping of large and complex structural variants

Edward S. Rice
Antton Alberdi
James Alfieri
Giridhar Athrey
Jennifer R. Balacco
Philippe Bardou
Heath Blackmon
Mathieu Charles
Hans H. Cheng
Olivier Fedrigo
Steven R. Fiddaman
Giulio Formenti
Laurent Frantz
M. Thomas P. Gilbert
Cari J. Hearn
Erich D. Jarvis
Christophe Klopp
Sofia Marcos
Deborah Velez-Irizarry
Luohao Xu
Wesley C. Warren (✉ warrenwc@missouri.edu)

Research Article

Keywords: Gallus gallus, K locus, IGLL1, ev21

Posted Date: June 21st, 2023

DOI: <https://doi.org/10.21203/rs.3.rs-3086064/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

1 **A pangenome graph reference of 30 chicken genomes allows genotyping of large and complex**
2 **structural variants.**

3

4 Edward S. Rice^{1,2}, Antton Alberdi³, James Alfieri⁴, Giridhar Athrey⁵, Jennifer R. Balacco⁶, Philippe
5 Bardou⁷, Heath Blackmon⁸, Mathieu Charles⁹, Hans H. Cheng¹⁰, Olivier Fedrigo⁶, Steven R.
6 Fiddaman¹¹, Giulio Formenti⁶, Laurent Frantz^{2,12}, M. Thomas P. Gilbert³, Cari J. Hearn¹⁰, Erich D.
7 Jarvis^{6,13}, Christophe Klopp¹⁴, Sofia Marcos^{3,15}, Deborah Velez-Irizarry¹⁰, Luohao Xu¹⁶, Wesley C.
8 Warren^{17*}

9

10 ¹ Bond Life Sciences Center, University of Missouri, Columbia, MO, USA

11 ² Faculty of Veterinary Medicine, Ludwig-Maximilians-Universität, München, Germany

12 ³ Center for Evolutionary Hologenomics, Globe Institute, University of Copenhagen (UCPH),
13 Copenhagen, Denmark

14 ⁴ Department of Ecology & Evolutionary Biology, Texas A&M University, College Station, TX, USA

15 ⁵ Department of Poultry Science, Texas A&M University, College Station, TX, USA

16 ⁶ Vertebrate Genome Laboratory, The Rockefeller University, New York, NY, USA

17 ⁷ Sigena, GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326, Castanet Tolosan, France

18 ⁸ Department of Biology, Texas A&M University, College Station, TX, USA

19 ⁹ INRA Centre de Toulouse Midi-Pyrénées: Castanet Tolosan, Midi-Pyrénées, FR

20 ¹⁰ USDA, ARS, USNPRC, Avian Disease and Oncology Laboratory, East Lansing, MI, USA

21 ¹¹ Department of Biology, University of Oxford, OX1 3SZ, UK

22 ¹² School of Biological and Behavioural Sciences, Queen Mary University of London, London E1
23 4DQ, UK

24 ¹³ The Howard Hughes Medical Institute, Chevy Chase, MD, USA

25 ¹⁴ Sigena, Genotoul Bioinfo, MIAT UR875, INRAE, Castanet Tolosan, France

26 ¹⁵ Applied Genomics and Bioinformatics, University of the Basque Country (UPV/EHU), Leioa,
27 Bilbao, Spain

28 ¹⁶ Key Laboratory of Freshwater Fish Reproduction and Development (Ministry of Education), Key
29 Laboratory of Aquatic Science of Chongqing, School of Life Sciences, Southwest University,
30 Chongqing 400715, China

31 ¹⁷ Department of Animal Sciences, University of Missouri, Columbia, MO, USA

32

33 * To whom correspondence should be addressed: warrenwc@missouri.edu

34

35 Keywords (at least 3): Gallus gallus, K locus, IGLL1, ev21

36 **Abstract**

37

38 **Background.** The red junglefowl, the wild progenitor of domestic chickens, has historically served as
39 a reference for genomic studies of domestic chickens. These studies have provided insight into the
40 etiology of traits of commercial importance. However, the use of a single reference genome does not
41 capture diversity present among modern breeds, many of which have accumulated molecular changes
42 due to drift and selection. While reference-based resequencing is well-suited to cataloging simple
43 variants such as single nucleotide changes and short insertions and deletions, it is mostly inadequate to
44 discover more complex structural variation in the genome.

45 **Results.** We present a pangenome for the domestic chicken consisting of thirty assemblies of chickens
46 from different breeds and research lines. We demonstrate how this pangenome can be used to catalog
47 structural variants present in modern breeds and untangle complex nested variation. We show that
48 alignment of short reads from 100 diverse wild and domestic chickens to this pangenome reduces
49 reference bias by 38%, which affects downstream genotyping results. This approach also allows for the
50 accurate genotyping of a large and complex pair of structural variants at the K ‘feathering’ locus using
51 short reads, which would not be possible using a linear reference.

52 **Conclusions.** We expect that this new paradigm of genomic reference will allow better pinpointing of
53 exact mutations responsible for specific phenotypes, which will in turn be necessary for breeding
54 chickens that meet new sustainability criteria and are resilient to quickly evolving pathogen threats.

55 **Introduction**

56 Accurately detecting sequence variation associated with traits of economic importance in the
57 domestic chicken is a major goal of genetic research into this globally widespread dietary protein
58 source [1]. Many groups are now genotyping chicken genomes to discover the underlying molecular
59 basis of specific traits [2–6], but current methods, both sequence- and array-based, have unquantified
60 limitations in assessing the underlying variation that connects many loci to studied traits.
61 Investigations in other species into the variant sets compiled by techniques relying on existing linear
62 references have revealed large gaps in variation discovery ability [7–10]. For the domestic chicken,
63 improved completeness and accuracy of bioinformatic queries into this variation are of vital
64 importance to the field, as computational experiments are rapidly becoming the venue of choice to
65 assess the potential of artificial selection to improve qualities such as growth, nutrient digestibility,
66 reproduction, and perhaps most importantly, immune resilience.

67 Current frequently employed methods for genotyping whole genomes mostly share the core
68 strategy of aligning short reads to a reference genome derived from a single individual [11]; these
69 references are usually compressed haploid representations of diploid genomes, with toggling of
70 haplotypes due to haploid compression, or chimeric haploblocks due to allele mixing [12,13]. While
71 these methods, given a reference genome of sufficient quality and reads of sufficient coverage, are able
72 to capture most single nucleotide variants (SNVs) and small insertions and deletions (indels) in
73 populations, they can lead to reference bias [14,15], and they consistently underestimate all types of
74 structural variants (SVs) [8]. Furthermore, for best performance, the most accurate genotyping
75 software [16] requires preexisting high-quality data about the distribution of polymorphic sites

76 throughout the genome for statistical calibration [17] or model training [18], information that does
77 not exist for most species. Large-scale long-read resequencing can mitigate some of these limitations
78 [19], but the high cost and low accuracy of long reads compared to short reads, and the large amount
79 of existing publicly available short-read sequencing data — for chicken, there are over 40,000 short
80 read data sets on the SRA at the time of writing but fewer than 500 long read experiments — make a
81 full transition to the use of long reads for resequencing studies unlikely in the near future.

82 The limitations of these approximations have led to the evolution of methods in the detection
83 of SVs [7]. Over the past decade, many algorithms have been developed to detect these segregating and
84 de novo SVs using short-read approaches [20]. The association of these SVs with traits has been
85 undeniable but usually overlooked in favor of the more straightforward use of single nucleotide
86 variants (SNVs) in GWAS studies [20]. The best-performing of these algorithms can detect around
87 11,000 SVs in humans, but face significant problems with both false positives and false negatives [7].
88 In contrast, when using long-read methodologies, this detection threshold nearly doubles [10,21]. In
89 chicken, several studies have called SVs or copy number variants (CNVs) using the aforementioned
90 short read algorithms [22,23], but each lacks an understanding of how SV detection could be limited
91 by not only the short reads used but also the reference to which sequences are aligned. In these studies
92 and others, content is not only missed but falsely called, with no regional genomic context in how this
93 variation is best classified.

94 To counter these limitations, several methods have been developed to create and use
95 pangenome graphs as references [24–28]. A pangenome graph is a data structure that encodes the
96 sequence and variation present among the genomes of multiple individuals [29]. While a linear

97 reference usually contains only the sequence of a single individual, a pangenome includes sequence
98 common to all individuals as well as information about the position, alleles, and frequencies of each
99 variant site. The recent publication of a draft pangenome for human demonstrated that this new
100 paradigm allows recovery of much sequence that appears with nonnegligible frequency in the genomes
101 of individuals across the species but is missing from even the telomere-to-telomere linear reference
102 [30].

103 An additional advantage of cataloging variation in the form of a pangenome instead of a list of
104 variants relative to a linear reference is that a pangenome can represent nested variation. For example, if
105 multiple non-reference individuals have an insertion, but the insertion happened far enough in the
106 past that the inserted sequence now contains segregating SNVs, variant calls against a linear reference
107 would treat each version of the insertion as an independent allele, while a pangenome shows the
108 variant as a single insertion containing additional nested variation within it. By contrast, a linear
109 representation of this variant would consider each possible path through the insertion as a separate
110 allele, rather than breaking it down into a single biallelic presence/absence variant along with a series of
111 smaller variants nested within the presence allele sequence. Furthermore, reads containing no reference
112 sequence can map to the insertion, whereas when aligning to a linear reference, such reads would either
113 remain unaligned, or, even worse, map to a different location in the genome with similar sequence.

114 Alignment of short reads to a pangenome reference instead of a linear reference has been
115 demonstrated in humans and other species, including birds, to recapitulate and improve downstream
116 genotype calling accuracy for both small variants (i.e., SNPs and small indels) and larger structural
117 variants [9,31,32]. Large insertions are nearly uncallable when using short-reads aligned to linear

118 references, with the recall of tools such as Delly [33] falling to zero for insertions larger than 400bp,
119 while graph-based tools such as VG and paragraph [25] are mostly unaffected by variant length [31].
120 The human pangenome’s demonstrations of improvements in read mapping, small variant
121 genotyping, novel variant discovery, SV genotyping, and representation of complex variants [30] show
122 the potential of this new paradigm for genome references.

123 In chicken, multiple alignments of reference-guided short-read assemblies [34] and *de novo*
124 assemblies of high-error PacBio CLR reads [35] have revealed sequences present among chickens
125 worldwide but missing from current references, as well as other previously unknown SVs. However,
126 although these whole-genome alignments were both described as pangenomes by their respective
127 authors, neither study generated a pangenome graph that can be used by other researchers as a
128 reference for alignment to overcome the limitations presented by reference bias and difficulty in
129 capturing SVs. They are further limited by their reliance on short reads or low-accuracy long reads,
130 respectively, for assembly.

131 In this study, we use current best practices to generate a pangenome graph of 30 highly
132 continuous genome assemblies of various chicken breeds, including broilers, layers, and research lines.
133 We use this pangenome to catalog variation present in the input assemblies, including variation that
134 was not detectable in studies using other methods, focussing on SVs in an immune system gene and a
135 feathering-related locus as illustrations. We then go on to align short reads from 100 chickens to the
136 graph, showing the improved performance of this method for alignment accuracy and genotyping
137 recall compared to linear reference alignment. We expect that adoption of this new resource will allow
138 better results in genotyping in future studies, with a goal to move toward more effective uses of

139 chicken genome references and in the process significantly improve researchers' ability to discover the
140 molecular mechanisms that determine bird healthiness.

141

142 **Results**

143 *Selection of chromosome-level assemblies*

144 To build assembly-based pangenome references, we used the five most continuous
145 chromosome-level assemblies of the domestic chicken currently available, along with alternate
146 haplotypes as applicable, and new contig-level assemblies of thirteen additional chickens, most of them
147 resolved into haplotypes. These chromosome-level assemblies have contig N50 values ranging from
148 5.47 to 91.3 Mb (see Table 1). This includes the current species reference assembly on NCBI RefSeq,
149 bGalGal1b, also known as GRCg7b (contig N50 = 18.8 Mb), a haplotype-resolved assembly of a
150 commercial broiler line created using the trio-binning method and an F1 cross between a
151 representative commercial broiler and a white leghorn layer [36]. bGalGal1b, as the current RefSeq
152 reference assembly, is fully annotated, so we use it as the source of annotations in this study. Because
153 this assembly was made using trio-binning, its creation also resulted in a haplotype-resolved assembly
154 of the genetic contribution of the other parent, a white leghorn layer. We refer to this assembly as
155 bGalGal1w, and it is also known as GRCg7w and we use both assemblies in our pangenome.

156 We sequenced and assembled to the chromosome level the genomes of two additional broilers
157 from the Ross (Aviagen) and Cobb (Cobb-Vantress) lines, to capture more of the diversity present
158 among commercial lines of domestic chickens, and to take advantage of advances in sequencing that
159 have occurred since the assembly of bGalGal1b and bGalGal1w, especially base-calling improvements

160 in PacBio’s HiFi/Circular Consensus Sequence (CCS) technology. HiFi reads are accurate enough to
161 allow the hifiasm algorithm to assemble contigs for both haplotypes [37], so although we only
162 assembled the primary contigs into chromosomes, we used the alternate contigs during pangenome
163 construction as well to take full advantage of their individual haploid diversity.

164 We also integrated the first nearly complete assembly of a chicken [38]. This assembly is of a
165 Huxu, a Chinese broiler breed, and we refer to it as “huxu”.

166 Finally, we sequenced and assembled both haplotypes of 13 additional chickens to a contig
167 level using HiFi sequencing (Supplemental table 1). These chickens include research lines bred to
168 study immune function as well as domestic breeds originating in Spain and Egypt. We produced
169 sequencing coverage of at least 25x (mean 35x) for each bird based on a genome size of 1.1Gb. We
170 successfully assembled both haplotypes of 10 out of 13 birds into contigs, and used partially phased
171 primary contig assemblies of the remaining 3, resulting in a total of 23 assemblies with a minimum
172 contig N50 of 11 Mb (mean 15 Mb).

173 Together, these 30 assemblies represent a diverse set of domestic chickens, including
174 commercial lines, research lines, and broiler and layer breeds originating on three continents. They also
175 were assembled using three different techniques: haplotype-resolved trio-binning of the F1 offspring
176 of a cross between two breeds (bGalGal1b and bGalGal1w), HiFi haplotype-resolved assembly
177 (bGalGal4, bGalGal5, and additional chickens), and the current best-practice de novo assembly
178 technique using a combination of HiFi and Oxford Nanopore Ultralong (ONT UL) reads (huxu)
179 [38]. While collectively these genomes do not come close to fully capturing the diversity of domestic

180 chickens worldwide, they provide a good working template of a first pangenome reference of the
181 domestic chicken genome.

ID	Assembled bird	Accession	Ref	Contig N50 (Mb)
bGalGal1b	Commercial broiler	GCA_016699485.1	[36]	18.8
bGalGal1w	White leghorn layer	GCA_016700215.2	[36]	17.7
bGalGal4	Ross broiler	GCA_027557775.1	N/A	5.47
bGalGal5	Cobb broiler	GCA_027408225.1	N/A	8.33
HuxuT2T	Huxu broiler	GCA_024206055.1	[38]	91.3

182 **Table 1:** The five chromosome-level assemblies used as a base for creation of pangenome references
183 for the domestic chicken.

184 *Creation of pangenome references*

185 We constructed pangenome references of the chicken genome using two different methods,
186 both used by the Human Pangenome Reference Consortium [30]: PanGenome Graph Builder
187 (PGGB) [30] and minigraph-cactus [39]. PGGB and minigraph-cactus both take multiple assemblies
188 as input, perform whole-genome alignments on them, and derive a pangenome graph from these
189 alignments. We made a preliminary graph using each method and five chromosome-level assemblies
190 (Table 1). For minigraph-cactus, we then created a 30-assembly graph using these five chromosome-
191 level assemblies as well as the contig-level alternate haplotype assemblies of bGalGal4 and bGalGal5
192 and assemblies of both haplotypes of thirteen additional chickens from HiFi data (Supplementary
193 Table 1). We did not create a 30-assembly graph with PGGB due to the computational intractability
194 of the 5-assembly PGGB graph for downstream applications, as described below. Therefore, the final
195 two graphs we tested were the 5-assembly PGGB graph and the 30-assembly minigraph-cactus graph.

196 The minigraph-cactus pangenome graph contains 49 million nodes and 67 million edges, and
197 therefore a mean degree, or the number of edges attached to a node, of 1.4. The total length of
198 sequence represented in the graph is 1.13 Gb. The combined length of nodes traversed by the most
199 complete assembly, Huxu, is 1.02 Gb. This is smaller than the 1.10 Gb total size of the assembly. This
200 difference is because a path can traverse the same sequence in the graph multiple times, for example, in
201 the case of a duplication. Therefore, there is in total 0.11 Gb (9.9%) of additional sequence in the
202 graph compared to the length of the most complete assembly. Of the other assemblies, bGalGal1b
203 contributes the most additional sequence, 55.6 Mb, to the graph, while some assemblies contribute as
204 little as 200 kb of additional sequence as a result of their relatedness to others (Supplementary Figure
205 S1).

206 The PGGB pangenome graph contains 33 million nodes and 45 million edges, and therefore
207 also a mean degree of 1.4. We found that parameter choice had a large effect on the numbers of nodes
208 and edges, as well as the maximum degree, although not the mean degree (Supplementary Figure S2).
209 Despite this pangenome being made up of only five assemblies instead of 30, it contains more
210 sequence than the minigraph-cactus pangenome: the total length of sequence represented in the
211 PGGB graph is 1.23 Gb, an additional 147 Mb or 12.0% of sequence compared to the total length of
212 graph nodes in the Huxu genome (1.09 Gb).

213 The 109 Mb of additional sequence is closer to previous estimates of total variation in diverse
214 groups of chickens [40–43] than 147 Mb, suggesting overestimation by PGGB. Comparative
215 examination of graph structures revealed that much of the additional sequence in the PGGB graph is
216 likely due to regions of assemblies that are homologous but were not properly aligned by the pipeline,

217 causing large bubbles and loose ends in the graph (Supplementary Figure S3). Another source of
218 additional sequence in this graph is duplications that are treated as simple insertions, leading to the
219 same sequence occurring twice in the graph; one example of this occurs in the K locus, which we
220 discuss below. For these reasons, we used only the minigraph-cactus graph for most subsequent
221 analyses.

222

223 *Cataloging of variants present in input assemblies*

224 A pangenome graph contains the variation present in the input assemblies, and can thus be
225 used to genotype the input assemblies compared to one chosen as a reference, based on deviations
226 from this reference path. We chose bGalGal1b for the reference as it is the highest-quality RefSeq-
227 annotated chicken reference genome currently available. In total, we found 15 million variants present
228 in at least one of the other 29 haplotypes compared to bGalGal1b. 12 million of these variants are
229 SNVs (Figure 1a). This is a smaller number of total SNVs than has been detected in large panel studies
230 [42,43], which is likely a result of the smaller sample size of our experiment, with 30 haplotypes
231 compared to 678 in [42]. We found a similar total length of deleted sequence, 19.2 Mb, as a previous
232 study based on long read alignments, 19.7 Mb [41]. However, we were able to recover 18.5 Mb of
233 inserted sequence, while the previous study recovered only 6.74 Mb [41] (Figure 1a). Although
234 distributions of lengths of deletions found previously by read alignment and by our pangenome
235 method were broadly similar, we found more long insertions than was possible with long-read
236 alignment (Figure 1b).

237 The B cell receptor gene *IGLL1*, which has been used as a marker for plasma B cells in chicken
238 [44], contains examples of these different kinds of variation. The overall structure of the pangenome
239 graph of *IGLL1* shows that there are many small variants, as well as two SVs (Figure 2). By encoding
240 the presence of small variants and their allele frequencies into the reference (Figure 2a), alignment to
241 pangenomes has been shown to reduce reference bias compared to a linear reference [24], which we
242 confirm below for our chicken pangenome. For example, for the SNV shown in Figure 2a, short reads
243 containing the non-reference allele are in less danger of mapping incorrectly as the aligner is aware of
244 the 17% chance of an A in this position of the genome.

245 The larger of the two SVs in the pangenome graph of *IGLL1* is a ~5kb deletion relative to
246 bGalGal1b present in only one haplotype of one chicken, UCD312 (Figure 2b). By recording this low-
247 frequency deletion in the reference, the pangenome method ensures that reads from resequenced
248 chickens containing the deletion are able to map to both flanking sequences through edge e1 without
249 splitting, which would introduce a potential source of error.

250 Finally, a ~300bp insertion relative to bGalGal1b demonstrates how a pangenome graph is able
251 to losslessly represent nested variation (Figure 2c). The SNVs and indels within the inserted sequence
252 are encoded in the exact same way as they would be in reference sequence, giving a full picture of the
253 variation present in this region.

254

255 *Disentangling a tandem repeat and viral insertion at the K locus*

256 The K locus, short for “short wing” (*kürzer Flügel*), is a region of chrZ with an early feathering
257 (EF) allele and a late feathering (LF) allele [45,46]. The EF allele contains single copies of the genes

258 *PRLR* and *SPEF2*. The LF allele contains a tandem duplication of parts of both genes [47], and often,
259 but not always [48,49], an insertion of the sequence of the avian leukosis virus ev21. The reference
260 genome bGalGal1b has the EF allele and no ev21 insertion, so genotyping the K locus in other
261 chickens using this reference is difficult as ev21 has a length of 9,679 bp [49], an order of magnitude
262 longer than the maximum insertion size that can be genotyped with short reads and a linear reference
263 [31]. As such, it is a region that can be more accurately genotyped with the use of a pangenome graph
264 approach.

265 We first created a one-dimensional representation of the minigraph-cactus pangenome graph
266 structure of the K locus colored by path coverage, as a node through which the same haplotype path
267 travels more than once indicates a duplication (Figure 3a). This representation shows that while most
268 of the haplotypes represented in the pangenome graph contain only one copy of this locus, Huxu has a
269 duplicated region and an insertion. The 2x path coverage region in Huxu covers parts of both *PRLR*
270 and *SPEF2*, consistent with the tandem duplication found by Elferink et al. [47]. We also found a
271 misassembly in bGalGal1w, with unassigned scaffolds containing the sequence (see Supplementary
272 Note 1 and Supplementary Figure S4). Furthermore, Huxu contains an insertion relative to the
273 reference sequence bGalGal1b. Alignment verified that the inserted sequence is the ev21 genome.

274 Next, to better understand the structure of the locus, we created a two-dimensional
275 representation of the graph at this locus (Figure 3b-d). This representation of the graph shows the
276 tandem duplication as a junction where a path can either leave the K locus or repeat it (Figure 3c), and
277 the insertion as a loop containing the ev21 genome covered only by Huxu (Figure 3d).

278 Finally, to view the alleles linearly, we used the “untangle” function of ODGI to lay out each
279 haplotype (Figure 3e). The resulting gene layout of the two alleles is consistent with previous
280 knowledge about the structure of the locus [47–49].

281 Repeating this process using the PGGB graph, we found that the PGGB graph did not contain
282 the ev21 insertion, and treated the tandem duplication as a simple insertion rather than a duplication.
283 Given the dependence of the PGGB output on good parameter choices, and previous demonstrations
284 that PGGB graphs of small complex regions such as the human MHC can accurately reconstruct the
285 structure of these loci, we hypothesize that choosing parameters specific to the level of divergence
286 present at complex loci is necessary for PGGB to accurately reconstruct them.

287

288 *Use as a reference for resequencing and genotyping*

289 Given the improvements in accuracy and recall of genotyping shown in other species by using
290 pangenome graph-based methods, we set out to demonstrate the usefulness of our pangenome
291 representations for alignment and genotyping. For this, we used simulated short reads as well as short
292 reads from 100 domestic and wild chickens (Supplementary Table 2). For comparison between linear
293 and graph-based methods, we called genotypes using both linear alignments to bGalGal1b as well as
294 graph alignments to our pangenomes.

295 For downstream use by existing short-read genotype callers, alignments must be converted
296 from graph coordinates to linear coordinates; this process is called surjection. Alignment of short reads
297 to the PGGB graph and surjection to bGalGal1b was infeasible, with a throughput of only 1.6 reads
298 per CPU-second on a test set of 10k paired-end reads, and inability to complete alignment of a larger

299 test set of 1M paired-end reads without running out of memory with 250 GB allocated to the job.
300 Further investigation revealed that surjection was the bottleneck, as graph alignment without
301 subsequent surjection had a throughput of 147 reads per CPU-second and a maximum memory usage
302 of 31GB for the 1M test set. By comparison, alignment of the 1M test set to the minigraph-cactus
303 graph followed by surjection to bGalGal1b had a throughput of 500 reads per CPU-second and a
304 maximum memory usage of 24GB, and minimap2 could align 1832 reads per CPU-second to
305 bGalGal1b with 5.4 GB memory (Figure 4a-b).

306 In order to compare accuracy of graph alignment to linear alignment, we simulated one
307 million pairs of paired-end reads and aligned them to both the cactus-minigraph pangenome with VG
308 giraffe and the linear bGalGal1b reference with minimap2. Giraffe performed better than minimap at
309 every level of stringency, based on what percentage of all reads were mapped correctly (Figure 4c).

310 To test the downstream effects of these differences in mapping accuracy, we genotyped 100
311 chickens from diverse breeds using both giraffe pangenome alignments and minimap linear alignments
312 of 10-15x coverage short reads, and compared the results between the two methods (Figure 5). While
313 the two methods found similar sets of SNVs (Figure 5a) and indels (Figure 5b), there were substantial
314 differences. Agreement was unsurprisingly higher for SNVs, although the pipeline using giraffe
315 alignments found a larger number with a quality score of at least 10 than the pipeline using minimap
316 (Figure 5a). For variants found by both methods, per-sample SNV concordance had a mean of 97.9%
317 with a standard deviation of 9.1% (Figure 5c). Indel concordance was lower, with a mean of 94.0% and
318 a standard deviation of 12.9% (Figure 5d).

319 To determine whether reference bias is a factor in the different genotyping results between the
320 two methods, we examined the proportion of mapped reads containing the reference allele at putative
321 heterozygous SNV sites. Reference bias across these sites, which we define as the difference between
322 the mean fraction of reads containing the alternate allele and the expected alternate allele fraction of
323 0.5, is lower for all of the 100 chickens when using pangenome alignment instead of linear alignment,
324 with a mean reference bias reduction of 38% (Figure 5e, Supplementary Figure S5).

325 Finally, we used the short read alignments to the pangenome graph to genotype the K locus
326 based on edge coverage (Figure 5f). All of these chickens are female and thus only have one copy of the
327 Z-linked K locus. Of the 100 chickens, 23 have the ev21 insertion (ev21+) and 24 have the tandem
328 repeat (late feathering/LF). As found in previous studies [48,49], the ev21 insertion and the tandem
329 duplication are not inextricably linked, although they do usually appear together: three chickens, all
330 standard Rhode Island breeds, have the ev21 insertion but not the tandem repeat, and four chickens,
331 two Silkies and two Cochins, have the tandem repeat but not the ev21 insertion.

332

333 **Discussion**

334 With the quickly accumulating numbers of haplotype-resolved genomes for many species, the
335 pangenome model of integrated presentation of within-species variation stands to become ubiquitous
336 [29,30]. Such resources already exist for other livestock such as swine [50] and cattle [51,52]. One of
337 the greatest advantages of pangenome references in other species has been the capture of sequences not
338 present in linear reference genomes. Compared to the nearly complete assembly of the Huxu chicken
339 genome, our pangenome graph contains 109 Mb of additional sequence. Some of this additional

340 sequence comes from SNVs or small indels that are relatively straightforward to represent in the
341 context of a linear reference, and some of it is made up of nodes whose sequences are similar to nodes
342 traversed by the Huxu assembly, but are represented separately. Thus, the true accessory genome
343 length is likely less than 109 Mb compared to Huxu. Nonetheless, the tripling of total insertion length
344 detectable using this pangenome compared to long read alignments shows that much of this additional
345 sequence is made up of variation that cannot be represented in a traditional linear reference genome,
346 and therefore, many reads from these regions of the genome cannot be mapped to a linear reference as
347 it does not contain the parts of the genome the reads came from. By adding additional assembled
348 chicken genomes of more diverse origins this amount of novel sequence will grow.

349 While other studies have presented multiple alignments of chickens as pangenomes [34,35],
350 our graph-based approach, which uses assemblies based on long and highly accurate PacBio HiFi reads
351 as well as one near-complete assembly, allows the pangenome to be used not just as a method for
352 cataloging variation present in the input assemblies, but also as a reference for future resequencing
353 studies. By comparing pipelines using linear versus pangenome alignments of short reads to genotype
354 100 chickens from diverse breeds, we demonstrated the improved alignment performance of
355 pangenome alignment over linear alignment, and showed the downstream effects of these
356 improvements on genotyping. Unfortunately, there does not yet exist a high-confidence truth set of
357 variant calls for chickens as there does for humans [16], so we cannot compare the accuracy of these
358 differing genotype calls. Nonetheless, given the improvements in alignment performance we have
359 shown in chicken with both simulated and real reads, and the improvements in genotyping
360 demonstrated in human and yeast by using the giraffe pangenome aligner [9,30], we predict that the

361 genotypes we inferred using giraffe pangenome alignment are substantially more accurate than those
362 we inferred using linear alignment.

363 Our determination of the structure of the K locus and subsequent genotyping demonstrates
364 the power of pangenome graphs in the study of loci containing complex structural variants. The initial
365 discovery of the insertion of an endogenous avian leukosis virus in the late feathering allele required
366 cell culture work [53], and a later study establishing the tandem repeat [47] necessitated extensive
367 quantitative PCR experiments targeted at 20 different segments of the locus. Although the latter was
368 performed after a linear reference genome was available, this reference, like all subsequent versions of
369 the reference genome for chicken, contains the early feathering allele and no ev21 insertion at the K
370 locus, and no current method can reliably genotype SVs of this size using short reads and a linear
371 reference [31]. More recent work on the relationship between the ev21 insertion and the late
372 feathering phenotype, though undertaken after improved reference genomes and large amounts of
373 public sequencing data from different breeds of chickens became available, also relied on targeted
374 PCR [48,49]. In contrast, we were able to replicate these findings using only existing short-read whole
375 genome sequencing data and pangenome methods. We expect that our pangenome, and future
376 pangenomes using telomere-to-telomere genome assemblies, which exist for increasing numbers of
377 species [54–58] but not yet chickens, will enable discoveries about complex structural variation at
378 important immune loci such as the major histocompatibility complex (MHC) and T cell receptor gene
379 (TCR), providing insight into the genetic diversity necessary to fight evolving pathogen threats in this
380 major worldwide source of protein, which also threaten wildlife with increasing frequency [59].

381 The current best-performing pangenome-based SV calling pipeline [9] subjects graph
382 alignments to linear coordinates, losing information in the process, and uses proprietary hardware and
383 software optimized for human data. Thus, at this stage we did not attempt to genotype all SVs in the
384 chickens with short-read data. Given the lack of ground truth SV data in chickens, we focussed on a
385 single locus that has already undergone targeted genotyping from a wide variety of chicken breeds, and
386 were able to confirm previous knowledge of the combinations of alleles present at this locus. We look
387 forward to the development of open-source SV calling pipelines that perform well on non-human
388 genomes and take full advantage of the information present in graph alignments.

389

390 **Conclusions**

391 In this paper, we have presented the first pangenome graph reference for the domestic chicken.
392 We show its utility as a catalog of variation, including structural variation too large or complex to be
393 detected using previous methods, and as a reference for the alignment of short reads. Given the
394 improvements we have demonstrated in this model over a linear reference, we expect this pangenome,
395 and new versions with additional broadly diverse chicken breeds incorporated, to serve as a resource to
396 the community for future resequencing studies as well as investigation of complex loci, especially in
397 immune-related genes.

398

399 **Methods**

400 *Sequencing and assembly of bGalGal4 and bGalGal5*

401 One female Ross 308 (Aviagen) and one female Cobb 550 (Cobb-Vantress), both commercial
402 broiler chickens, were euthanized in the framework of a research experiment at 38 days of age. Cardiac
403 puncture was immediately employed to collect 12 aliquots of 100 ul of blood in tubes with EDTA and
404 1 ml of ethanol >99.7% from each animal. Samples were frozen at -20 °C.

405 For both assemblies (bGalGal4 and bGalGal5), we followed the VGP 2.0 pipeline [12]. We
406 generated 32x Pacbio HiFi data on a Sequel IIe, and then used cutadapt [60] to trim off adapters that
407 were not trimmed in the Pacbio software processing. We assembled contigs using HiFiasm v0.14 [61],
408 generating a semi-haplotyped phased primary contig and alternate contig assembly. From the primary
409 assembly, we removed false haplotype duplication and placed them in the alternate using purge_dups
410 v1.2.5 [62]. We then scaffolded the contigs with Bionano Genomics optical maps (319x and 459x
411 respectively), generated on a Saphyr instrument using DLE label, with Bionano Solve. We then further
412 scaffolded with Arima Genomics Hi-C v2 (65x and 122x respectively), using salsa v2.2 [63]. The
413 primary assembly was then curated using gEVAL [64], structural errors corrected, and chromosomes
414 named according to their numbers in the bGalGal1 GRC7g reference. 10X Genomics data were also
415 generated, and used for orthogonal validation, but not scaffolding. The primary and alternate
416 assemblies were deposited in NCBI under accession numbers GCA_027557775.1 (bGalGal4) and
417 GCA_027408465.1 (bGalGal5), and all data are available in Genome Ark
418 https://genomeark.github.io/genomeark-all/Gallus_gallus/.

419

420 *Sequencing and assembly of additional chickens*

421 High molecular weight (HMW) DNA from blood of 13 juvenile male chickens
422 (Supplementary Table 1), maintained and bled under ADOL IACUC-approved Animal Use Protocol
423 #2019-15 for breeder management, was sequenced on the Pacific BioSciences Sequel IIe. HMW
424 samples were sheared using a Diagenode Megarupter3 shearing device targeting 18-22KB fragments.
425 Libraries were prepared with the PacBio SMRTbell Prep Kit 3.0. Library size distribution was
426 determined on the Agilent Femto Pulse and a Qubit fluorometer was used to measure concentration.
427 Sequencing polymerase was bound to the SMRTbell libraries with the Binding Kit 3.2, and run on
428 Sequel IIe with the Sequel II Sequencing Kit 2.0 and SMRT Cell 8M. HiFi data was collected with
429 Instrument Control Software Version 11.0 and Chemistry Bundle 11.0 with a movie time of 30 hours.
430 The On Plate Loading Concentration was 130pmolar.

431 HiFi reads for each of the chickens were assembled into contigs using hifiasm v0.18.9 [37]
432 with default options. Both haplotypes output by hifiasm were used in subsequent analyses.

433

434 *Creation of PGGB pangenome*

435 We constructed a pangenome reference from the five input assemblies bGalGal1b,
436 bGalGal1w, bGalGal4, bGalGal5, and HuxuT2T (Table “assemblies”). First, we extracted
437 chromosome sequences from the assemblies and gave them names according to the PanSN-spec, in the
438 format of “[assembly name]#[chromosome name]”, e.g., “bGalGal4#chr5”. We partitioned the
439 assemblies into 41 communities, one for each chromosome, and then constructed a pangenome graph
440 for each chromosome separately. Due to disagreements in the naming of microchromosomes among

441 the five assemblies, some of the communities contain chromosomes named differently in the different
442 assemblies (Supplementary Table 3).

443 For every chromosome, we constructed its pangenome graph using the Pangenome Graph
444 Builder (PGGB) v0.4.1 [30]. Briefly, this pipeline uses wfmash v0.9.1 [65] to align the input
445 assemblies, seqwish v0.7.6 [28] to build a graph from the alignments, smoothxg v0.6.5 [66] and gfa
446 v0.1.3 [67] to clean up the graph, and odgi v0.7.3 [27] to visualize the graph. We first ran pggb with
447 default parameters, except for parameter “-n” set to the number of assemblies being aligned for the
448 chromosome in question (this number is five for most chromosomes, with the exception of sex
449 chromosomes and some microchromosomes without full representation in all five assemblies) and “-G
450 3079,3559”. For postprocessing and optimal visualization, we redrew the 2D graph visualization using
451 the odgi draw command with parameters “-C -w1000”, and we redrew the 1D graph visualization by
452 first resorting the graph based on positions in the bGalGal5 path using the command odgi sort with
453 parameters ‘-H <(echo “bGalGal5# $\{\text{chromosome_name}\}$ ”) -Y’ and then drawing with the odgi viz
454 command with default parameters.

455 To find the optimal parameters for each chromosome, we performed a parameter sweep of the
456 segment length (-s), mapping percent identity (-p), and minimum match length (-k) options to the
457 pggb command. We tested every member of the cartesian product set of the parameter values $s=\{5k,$
458 $10k, 30k, 50k, 80k\}$, $p=\{85, 90, 94, 97\}$, and $k=\{10, 19, 50, 100, 150\}$. We evaluated the results as
459 suggested in PGGB documentation, using a combination of examination of graph statistics, especially
460 node count and maximum degree, with the odgi stats command and visual inspection of the graph
461 structure using the odgi viz output. For some microchromosomes, we made more granular

462 adjustments to the parameters to fine-tune their graphs. Supplementary Table 3 shows the final
463 parameters chosen for each chromosome.

464 Finally, we created a single pangenome graph containing the respective connected component
465 for each community using the `odgi squeeze` command with default parameters. This resulted in a
466 single graph file with extension “.og” that is easily convertible to other sequence graph formats such as
467 GFA and VG.

468

469 *Creation of minigraph-cactus pangenome*

470 We ran the minigraph-cactus pipeline [39] using the cactus v2.4.2 Docker image and a
471 nextflow pipeline built for this purpose [68]. As input, we used the five chromosome-level assemblies
472 in Table 1, the alternate haplotypes of bGalGal4 and bGalGal5, and both haplotype assemblies of an
473 additional 13 chickens listed in Supplementary Table 1. We specified bGalGal1b as the reference,
474 because although it is not the highest-quality assembly, it is the best RefSeq-annotated assembly on
475 NCBI, so we wanted to call variants against it downstream.

476

477 *Additional sequence analysis*

478 We determined the amount of additional sequence contributed to the graph by each sample
479 through an iterative process. First, we removed all nodes traversed by the Huxu assembly from the
480 graph as it is the most complete assembly. Then, for each remaining bird, we summed up the length of
481 all nodes traversed by either haplotype of this bird, found the bird with the largest sum, and removed
482 all nodes traversed by this bird’s haplotypes from the graph. We repeated this process until there were

483 no samples remaining. The python program we wrote for this purpose is included in the repository
484 cited in the Code Availability statement.

485

486 *Format conversions and subgraph extraction*

487 To convert GFAv1.1 format as output by minigraph-cactus to OG format, we used the
488 command “vg convert -gfW” to convert to GFAv1.0, and then “odgi build -g -Os” to build an OG
489 graph out of the GFAv1.0 file.

490 To convert GBZ format to HG format, we used the command “vg convert”.

491 To convert HG format to GFA format, we used the command “vg convert -f”.

492 To convert OG format to GFA format, we used the command “odgi view -a -g”.

493 To extract regions from graphs in HG format, we used the command “vg find -p
494 ‘bGalGal1b#[chromosome]:[start]-[end]”.

495 To extract regions from graphs in OG format, we used the command “odgi extract -d0 -E -r
496 ‘bGalGal1b#[chromosome]:[start]-[end]”.

497

498 *Genotyping input assemblies*

499 Both assembly-based graph construction pipelines, pggp and minigraph-cactus, can output vcf
500 files containing genotypes for the input assemblies relative to the reference, in our case bGalGal1b.

501 Minigraph-cactus does this by default; pggp does with the addition of the option “-V ‘bGalGal1b:#:’”.

502 Where necessary, we concatenated vcf files for each chromosome into a single genome-wide vcf using

503 the bcftools concat command v1.15.1 [69].

504

505 *Graph visualization*

506 To visualize specific regions of the pangenome graph, we first looked up coordinates relative to
507 bGalGal1b on RefSeq, extracted them from the graph, output in GFA format, and visualized using
508 bandage v0.8.1 [70]. Commands for extraction and conversion are given under the heading “Format
509 conversions and subgraph extraction.”

510

511 *Read simulation*

512 We simulated reads using the “vg sim” command with a nucleotide substitution error rate of
513 0.24% as estimated by Pfeiffer et al. [71] and an indel error rate of 0.029% as in [9].

514

515 *Sequencing of short read chickens*

516 We sampled 236 chickens from 62 breeding farms that specialize in heritage and rare chicken
517 breeds in May and December 2021. In short, we collected 0.5 - 2 mL of blood from each bird by
518 puncturing the brachial vein with a syringe (gauge size 18.5 - 28 depending on the size of the bird).
519 The blood was immediately expelled through the syringe into K2EDTA vacutainers and stored on dry
520 ice. Upon arrival at the lab, the blood samples were transferred to a -80 C freezer. DNA was extracted
521 using the QIAamp Fast DNA Tissue Kit. Library preparation and sequencing were performed at BGI
522 Group. Libraries were prepared using a DNA short-insert protocol for 150 bp paired-end reads and
523 sequenced on the DNBseq platform. 7 samples failed to be sequenced due to low quality, so were
524 excluded from further analyses. We chose a subset of 100 of these samples for the final analysis.

525

526 *Short read alignment*

527 To align short reads to the PGGB graph, we first converted the graph to GFA format using the
528 command “odgi view -g” and then converted the GFA format to GBZ format [72] and created giraffe
529 indices from the output with the command “vg autoindex -w giraffe”. The minigraph-cactus pipeline
530 outputs all indices necessary to run giraffe by default, so no further processing was necessary to prepare
531 it for alignment of reads with giraffe.

532 To test timing and memory usage, we arbitrarily chose a publicly available set of short reads
533 from a chicken (SRR9967588) and subsetted the first 1 million pairs. This test failed for alignment to
534 the PGGB graph due to running out of memory, but a smaller subset of 10,000 read pairs was
535 successful. We aligned the test set of reads to the graph using the command “vg giraffe” with
536 arguments “-o BAM”. Because the PGGB graph does not contain a reference sequence like the
537 minigraph-cactus graph, we additionally specified the reference chromosomes with the arguments “--
538 ref-paths bGalGal1b_paths.tsv”, where bGalGal1b_paths.tsv is a tab-separated file containing a list of
539 all chromosomes in bGalGal1b and their sizes. For comparison to alignment to a linear reference with
540 minimap2 v2.24 [73], we created a short-read minimap index of bGalGal1b with the command
541 “minimap2 -x sr -d” and then aligned reads to it with the command “minimap2 -a” piped to “samtools
542 view -bh” with samtools v1.16.1 [69] to convert to bam format for a fair comparison, since we ran
543 giraffe with bam output.

544 For alignment of short reads from 100 chickens, we ran `vg giraffe` with default options,
545 outputting the results in GAM format. We surjected the GAM files to BAM format with `bGalGal1b`
546 as the reference genome using the command “`vg surject`” with default options.

547

548 *Comparison of linear and graph alignments with simulated reads*

549 To compare the accuracy of alignments of simulated reads between linear and graph aligners,
550 we aligned the simulated reads both to the `bGalGal1b` linear reference using `minimap2` and to the
551 pangenome graph reference using `giraffe`, as described above. We converted the `minimap2` output to
552 GAM format using the command “`vg inject`”, and then compared the `minimap2` and `giraffe` GAMs to
553 the truth set using “`vg gamcompare`”, all as in [9].

554

555 *Genotyping*

556 We genotyped the 100 chickens based on these alignments using `elprep` [74] v5.1.2, a
557 multithreaded reimplementation of GATK. Briefly, we generated an `elfasta` sequence reference for
558 `bGalGal1b` using the command “`elprep fasta-to-elfasta`”, created a list of sites from the `minigraph-`
559 `cactus vcf` output with SVs larger than 1000bp filtered out using the command “`elprep vcf-to-elsites`”,
560 and ran the “`sfm`” command with settings as recommended in the manual to generate a `gvcf` for each
561 bird, which we then combined into a single `gvcf` with GATK `CombineGVCFs` and joint genotyped
562 with GATK `GenotypeGVCFs` [17]. The location of our scripts for genotyping, as well as all other
563 analyses in this paper, is given in the Data Availability section.

564

565 *Genotyping method comparison*

566 To compare the respective outputs of the giraffe- and minimap-based genotyping pipelines, we
567 used bcftools v1.17 [69] command “isec -c some” to create four vcf files: variants only detected by the
568 giraffe pipeline, variants only detected by the minimap pipeline, giraffe pipeline calls of variants
569 detected by both pipelines, and minimap pipeline calls of variants detected by both pipelines. We
570 counted variants with QUAL \geq 10 in all of these files, subsetting by variant type with “bcftools view -
571 v [snp|indel]”. To compare the per-sample calls made by the respective methods for variants detected
572 by both, we used “bcftools merge --force-samples” to create a single vcf containing calls made by both
573 methods, and then used a custom python script (included in code availability) to calculate the percent
574 agreement for each variant.

575

576 *Reference bias estimation*

577 We estimated the amount of reference bias by calculating the mean fraction of reads mapping
578 to putative heterozygous sites containing the alternate allele, and comparing to the expected value of
579 0.5. We define putative heterozygous sites as positions with coverage of at least 10x where the portion
580 of reads containing the minor allele is at least 25%, as in [15]. Briefly, we filtered low-quality mappings
581 and multimapping reads with “samtools view -F2304 -q10”, created pileups with “samtools mpileup -
582 d100 --no-BAQ”, and piped the results to a custom C program to find putative heterozygous sites and
583 calculate alternate allele frequencies at these sites. All code used to perform this analysis is in the
584 project’s code repository.

585

586 *K locus genotyping*

587 To genotype the K locus, we converted each GAM file to GAF format using the command “vg
588 convert -G” and counted reads covering the edges e1 through e7 as shown in Figure “K locus”. We
589 used binomial tests with p-value cutoffs of 0.05 to assign genotypes to each chicken for both the ev21
590 insertion and the tandem duplication; chickens with both $p(\text{insertion}) > 0.05$ and $p(\text{no insertion}) >$
591 0.05 were marked as inconclusive.

592

593 **Declarations**

594 *Ethics approval and consent to participate*

595 Chickens used for the bGalGal4 and bGalGal5 assemblies were euthanized according to the
596 procedures regulated in the Spanish Royal Decree RD 53/2013. Experimentation procedures were
597 approved by the Ethical Committee of Generalitat de Catalunya, Spain (Proceeding number 10226).

598 Chickens used for additional assemblies were maintained and bled under ADOL IACUC-
599 approved Animal Use Protocol #2019-15 for breeder management. SPF birds from each line were
600 grown in colony cages and provided food and water *ad libitum*.

601 For chickens used for short read sequencing, all handling and sample collection of animals
602 were performed in accordance with TAMU AUP 2022-0091.

603 *Consent for publication*

604 Not applicable.

605 *Availability of data and materials*

606 The datasets generated and/or analyzed in the current study are available in NCBI repositories
607 under BioProject accessions PRJNA838369, PRJNA838370, and PRJNA971225. The pangenome
608 graph, a vcf of variants present in the graph, and vcfs of the resequenced chickens genotyped using
609 both linear and pangenome methods are available in a Zenodo repository at [embargoed until final
610 publication]. The code used to perform the analysis in the current study is available on GitHub at
611 <https://github.com/WarrenLab/chicken-pangenome-paper>.

612 *Competing interests*

613 The authors declare that they have no competing interests.

614 *Funding*

615 This work was supported by USDA NIFA grants 2020-67015-31574 and 2022-67015-36218
616 and the European Union’s Horizon Research and Innovation Programme under grant agreement No.
617 817729 (Project HoloFood). Computation for this work was performed on the high performance
618 computing infrastructure provided by Research Computing Support Services and in part by the
619 National Science Foundation under grant number CNS-1429294 at the University of Missouri,
620 Columbia MO. The authors gratefully acknowledge the Gauss Centre for Supercomputing e.V.
621 (www.gauss-centre.eu) for funding this project by providing computing time on the GCS
622 Supercomputer SuperMUC-NG at Leibniz Supercomputing Centre (www.lrz.de).

623 *Authors’ contributions*

624 WCW and ESR conceived and designed the project. AA, JA, GA, HB, HHC, MTPG, CJH,
625 SM, and DV generated sequence data used in this project. ESR, JRB, OF, GF, EDJ, and LX assembled
626 genomes used to create the pangenome. ESR, PB, MC, SRF, LF, and CK genotyped chickens used in

627 this project. ESR constructed the pangenome. ESR and WCW wrote the manuscript. All authors
628 edited and approved the manuscript.

629 *Acknowledgements*

630 Not applicable.

631

632

633 **References**

- 634 1. Athrey G. Chapter 18 - Poultry genetics and breeding. In: Bazer FW, Lamb GC, Wu G, editors.
635 Animal Agriculture. Academic Press; 2020. p. 317–30.
- 636 2. Drobik-Czwaro W, Wolc A, Fulton JE, Arango J, Jankowski T, O’Sullivan NP, et al. Identifying
637 the genetic basis for resistance to avian influenza in commercial egg layer chickens. *Animal*.
638 2018;12:1363–71.
- 639 3. Xu L, He Y, Ding Y, Liu GE, Zhang H, Cheng HH, et al. Genetic assessment of inbred chicken
640 lines indicates genomic signatures of resistance to Marek’s disease. *J Anim Sci Biotechnol*. 2018;9:65.
- 641 4. Wang Q, Li D, Guo A, Li M, Li L, Zhou J, et al. Whole-genome resequencing of Dulong Chicken
642 reveal signatures of selection. *Br Poult Sci*. 2020;61:624–31.
- 643 5. Seifi Moroudi R, Ansari Mahyari S, Vaez Torshizi R, Lanjanian H, Masoudi-Nejad A.
644 Identification of new genes and quantitative trait locis associated with growth curve parameters in F2
645 chicken population using genome-wide association study. *Anim Genet*. 2021;52:171–84.
- 646 6. Perlas A, Argilaguet J, Bertran K, Sánchez-González R, Nofrarías M, Valle R, et al. Dual Host and
647 Pathogen RNA-Seq Analysis Unravels Chicken Genes Potentially Involved in Resistance to Highly
648 Pathogenic Avian Influenza Virus Infection. *Front Immunol*. 2021;12:800188.
- 649 7. Zhao X, Collins RL, Lee W-P, Weber AM, Jun Y, Zhu Q, et al. Expectations and blind spots for
650 structural variation detection from long-read assemblies and short-read genome sequencing
651 technologies. *Am J Hum Genet*. Elsevier; 2021;108:919–28.
- 652 8. Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. Structural
653 variant calling: the long and the short of it. *Genome Biol*. 2019;20:246.
- 654 9. Sirén J, Monlong J, Chang X, Novak AM, Eizenga JM, Markello C, et al. Pangenomics enables
655 genotyping of known structural variants in 5202 diverse genomes. *Science*. 2021;374:abg8871.
- 656 10. Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, Welch AE, et al.
657 Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell*. 2019;176:663–
658 75.e19.
- 659 11. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation
660 sequencing data. *Nat Rev Genet*. 2011;12:443–51.
- 661 12. Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, et al. Towards complete and
662 error-free genome assemblies of all vertebrate species. *Nature*. 2021;592:737–46.

- 663 13. Ko BJ, Lee C, Kim J, Rhie A, Yoo DA, Howe K, et al. Widespread false gene gains caused by
664 duplication errors in genome assemblies. *Genome Biol.* 2022;23:205.
- 665 14. Brandt DY, Aguiar VRC, Bitarello BD, Nunes K, Goudet J, Meyer D. Mapping Bias
666 Overestimates Reference Allele Frequencies at the HLA Genes in the 1000 Genomes Project Phase I
667 Data. *G3* . 2015;5:931–41.
- 668 15. Günther T, Nettelblad C. The presence and impact of reference bias on population genomic
669 studies of prehistoric human populations. *PLoS Genet.* 2019;15:e1008302.
- 670 16. Barbitoff YA, Abasov R, Tvorogova VE, Glotov AS, Predeus AV. Systematic benchmark of state-
671 of-the-art variant calling pipelines identifies major factors affecting accuracy of coding sequence
672 variant discovery. *BMC Genomics.* 2022;23:155.
- 673 17. Van der Auwera GA, O'Connor BD. *Genomics in the Cloud: Using Docker, GATK, and WDL in*
674 *Terra*. "O'Reilly Media, Inc."; 2020.
- 675 18. Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and
676 small-indel variant caller using deep neural networks. *Nat Biotechnol.* 2018;36:983–7.
- 677 19. De Coster W, Weissensteiner MH, Sedlazeck FJ. Towards population-scale long-read sequencing.
678 *Nat Rev Genet.* 2021;22:572–87.
- 679 20. Chiang C, Scott AJ, Davis JR, Tsang EK, Li X, Kim Y, et al. The impact of structural variation on
680 human gene expression. *Nat Genet.* 2017;49:692–9.
- 681 21. Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, et al. Multi-platform
682 discovery of haplotype-resolved structural variation in human genomes. *Nat Commun.* Nature
683 Publishing Group; 2019;10:1–16.
- 684 22. Rao YS, Li J, Zhang R, Lin XR, Xu JG, Xie L, et al. Copy number variation identification and
685 analysis of the chicken genome using a 60K SNP BeadChip. *Poult Sci.* 2016;95:1750–6.
- 686 23. Weng Z, Xu Y, Li W, Chen J, Zhong M, Zhong F, et al. Genomic variations and signatures of
687 selection in Wuhua yellow chicken. *PLoS [Internet]. journals.plos.org;* 2020; Available from:
688 <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0241137>
- 689 24. Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, et al. Variation graph toolkit
690 improves read mapping by representing genetic variation in the reference. *Nat Biotechnol.*
691 2018;36:875–9.
- 692 25. Chen S, Krusche P, Dolzhenko E, Sherman RM, Petrovski R, Schlesinger F, et al. Paragraph: a
693 graph-based structural variant genotyper for short-read sequence data. *Genome Biol. BioMed Central;*
694 2019;20:1–13.

- 695 26. Li H, Feng X, Chu C. The design and construction of reference pangenome graphs with
696 minigraph. *Genome Biol. BioMed Central*; 2020;21:265.
- 697 27. Guarracino A, Heumos S, Nahnsen S, Prins P, Garrison E. ODGI: understanding pangenome
698 graphs. *Bioinformatics*. 2022;38:3319–26.
- 699 28. Garrison E, Guarracino A. Unbiased pangenome graphs. *Bioinformatics*. Oxford Academic;
700 2022;39:btac743.
- 701 29. Eizenga JM, Novak AM, Sibbesen JA, Heumos S, Ghaffaari A, Hickey G, et al. Pangenome
702 Graphs. *Annu Rev Genomics Hum Genet*. 2020;21:139–62.
- 703 30. Liao W-W, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, et al. A Draft Human Pangenome
704 Reference. *Nature*. 2023;617:312–24.
- 705 31. Hickey G, Heller D, Monlong J, Sibbesen JA, Sirén J, Eizenga J, et al. Genotyping structural
706 variants in pangenome graphs using the vg toolkit. *Genome Biol*. 2020;21:35.
- 707 32. Secomandi S, Gallo GR, Sozzoni M, Iannucci A, Galati E, Abueg L, et al. A chromosome-level
708 reference genome and pangenome for barn swallow population genomics. *Cell Rep*. 2023;42:111992.
- 709 33. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant
710 discovery by integrated paired-end and split-read analysis. *Bioinformatics*. 2012;28:i333–9.
- 711 34. Wang K, Hu H, Tian Y, Li J, Scheben A, Zhang C, et al. The Chicken Pan-Genome Reveals Gene
712 Content Variation and a Promoter Region Deletion in IGF2BP1 Affecting Body Size. *Mol Biol Evol*.
713 2021;38:5066–81.
- 714 35. Li M, Sun C, Xu N, Bian P, Tian X, Wang X, et al. De Novo Assembly of 20 Chicken Genomes
715 Reveals the Undetectable Phenomenon for Thousands of Core Genes on Microchromosomes and
716 Subtelomeric Regions. *Mol Biol Evol* [Internet]. 2022;39. Available from:
717 <http://dx.doi.org/10.1093/molbev/msac066>
- 718 36. Smith J, Alfieri JM, Anthony N, Arensburger P, Athrey GN, Balacco J, et al. Fourth Report on
719 Chicken Genes and Chromosomes 2022. *Cytogenet Genome Res* [Internet]. 2023; Available from:
720 <http://dx.doi.org/10.1159/000529376>
- 721 37. Cheng H, Jarvis ED, Fedrigo O, Koepfli K-P, Urban L, Gemmell NJ, et al. Haplotype-resolved
722 assembly of diploid genomes without parental data. *Nat Biotechnol*. 2022;40:1332–5.
- 723 38. Huang Z, Xu Z, Bai H, Huang Y, Kang N, Ding X, et al. Evolutionary analysis of a complete
724 chicken genome. *Proc Natl Acad Sci U S A*. 2023;120:e2216641120.
- 725 39. Hickey G, Monlong J, Ebler J, Novak AM, Eizenga JM, Gao Y, et al. Pangenome graph

- 726 construction from genome alignments with Minigraph-Cactus. *Nat Biotechnol* [Internet]. 2023;
727 Available from: <http://dx.doi.org/10.1038/s41587-023-01793-w>
- 728 40. Li D, Li Y, Li M, Che T, Tian S, Chen B, et al. Population genomics identifies patterns of genetic
729 diversity and selection in chicken. *BMC Genomics*. 2019;20:263.
- 730 41. Zhang J, Nie C, Li X, Zhao X, Jia Y, Han J, et al. Comprehensive analysis of structural variants in
731 chickens using PacBio sequencing. *Front Genet*. 2022;13:971588.
- 732 42. Liu R, Xing S, Wang J, Zheng M, Cui H, Crooijmans RPMA, et al. A new chicken 55K SNP
733 genotyping array. *BMC Genomics*. 2019;20:410.
- 734 43. Malomane DK, Simianer H, Weigend A, Reimer C, Schmitt AO, Weigend S. The SYNBREED
735 chicken diversity panel: a global resource to assess chicken diversity at high genomic resolution. *BMC*
736 *Genomics*. 2019;20:345.
- 737 44. Warren WC, Rice ES, Meyer A, Hearn CJ, Steep A, Hunt HD, et al. The immune cell landscape
738 and response of Marek's disease resistant and susceptible chickens infected with Marek's disease virus.
739 *Sci Rep*. 2023;13:5355.
- 740 45. Hertwig P, Rittershaus T. Die Erbfaktoren der Haushühner. *Z Indukt Abstamm Vererbungsl*.
741 1929;51:354–72.
- 742 46. Siegel PB, Mueller CD, Craig JV. Some Phenotypic Differences Among Homozygous,
743 Heterozygous, and Hemizygous Late Feathering Chicks^{1,2}. *Poult Sci*. 1957;36:232–9.
- 744 47. Elferink MG, Vallée AAA, Jungerius AP, Crooijmans RPMA, Groenen MAM. Partial duplication
745 of the PRLR and SPEF2 genes at the late feathering locus in chicken. *BMC Genomics*. 2008;9:391.
- 746 48. Takenouchi A, Toshishige M, Ito N, Tsudzuki M. Endogenous viral gene ev21 is not responsible
747 for the expression of late feathering in chickens. *Poult Sci*. 2018;97:403–11.
- 748 49. Zhang X, Wang H, Zhang L, Wang Q, Du X, Ge L, et al. Analysis of a genetic factors contributing
749 to feathering phenotype in chickens. *Poult Sci*. 2018;97:3405–13.
- 750 50. Jiang Y-F, Wang S, Wang C-L, Xu R-H, Wang W-W, Jiang Y, et al. Pangenome obtained by long-
751 read sequencing of 11 genomes reveal hidden functional structural variants in pigs. *iScience*.
752 2023;26:106119.
- 753 51. Zhou Y, Yang L, Han X, Han J, Hu Y, Li F, et al. Assembly of a pangenome for global cattle reveals
754 missing sequences and novel structural variations, providing new insights into their diversity and
755 evolutionary history. *Genome Res*. 2022;32:1585–601.
- 756 52. Leonard AS, Crysanto D, Mapel XM, Bhati M, Pausch H. Graph construction method impacts

757 variation representation and analyses in a bovine super-pangenome. *Genome Biol.* 2023;24:124.

758 53. Bacon LD, Smith E, Crittenden LB, Havenstein GB. Association of the slow feathering (K) and an
759 endogenous viral (ev21) gene on the Z chromosome of chickens. *Poult Sci.* Elsevier; 1988;67:191–7.

760 54. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete
761 sequence of a human genome. *Science.* 2022;376:44–53.

762 55. Rautiainen M, Nurk S, Walenz BP, Logsdon GA, Porubsky D, Rhie A, et al. Telomere-to-
763 telomere assembly of diploid chromosomes with Verkko. *Nat Biotechnol* [Internet]. 2023; Available
764 from: <http://dx.doi.org/10.1038/s41587-023-01662-6>

765 56. Xue L, Gao Y, Wu M, Tian T, Fan H, Huang Y, et al. Telomere-to-telomere assembly of a fish Y
766 chromosome reveals the origin of a young sex chromosome pair. *Genome Biol.* 2021;22:203.

767 57. Belser C, Baurens F-C, Noel B, Martin G, Cruaud C, Istace B, et al. Telomere-to-telomere gapless
768 chromosomes of banana using nanopore sequencing. *Commun Biol.* 2021;4:1047.

769 58. Bliznina A, Masunaga A, Mansfield MJ, Tan Y, Liu AW, West C, et al. Telomere-to-telomere
770 assembly of the genome of an individual *Oikopleura dioica* from Okinawa using Nanopore-based
771 sequencing. *BMC Genomics.* 2021;22:222.

772 59. Stokstad E. Deadly bird flu establishes a foothold in North America. *Science.* 2022;377:912–912.

773 60. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads.
774 *EMBnet.journal.* 2011;17:10–2.

775 61. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using
776 phased assembly graphs with hifiasm. *Nat Methods.* 2021;18:170–5.

777 62. Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. Identifying and removing
778 haplotypic duplication in primary genome assemblies. *Bioinformatics.* 2020;36:2896–8.

779 63. Ghurye J, Pop M, Koren S, Bickhart D, Chin C-S. Scaffolding of long read assemblies using long
780 range contact information. *BMC Genomics.* BioMed Central; 2017;18:527.

781 64. Chow W, Brugger K, Caccamo M, Sealy I, Torrance J, Howe K. gEVAL - a web-based browser for
782 evaluating genome assemblies. *Bioinformatics.* 2016;32:2508–10.

783 65. Marco-Sola S, Moure JC, Moreto M, Espinosa A. Fast gap-affine pairwise alignment using the
784 wavefront algorithm. *Bioinformatics* [Internet]. academic.oup.com; 2021; Available from:
785 <https://academic.oup.com/bioinformatics/article-abstract/37/4/456/5904262>

786 66. smoothxg: linearize and simplify variation graphs using blocked partial order alignment [Internet].

787 Github; [cited 2023 Mar 9]. Available from: <https://github.com/pangenome/smoothxg>

788 67. GFAffix: GFAffix identifies walk-preserving shared affixes in variation graphs and collapses them
789 into a non-redundant graph structure [Internet]. Github; [cited 2023 Mar 9]. Available from:
790 <https://github.com/marschall-lab/GFAffix>

791 68. minigraph-cactus-nf: a nextflow pipeline for creating a pangenome with minigraph-cactus
792 [Internet]. Github; [cited 2023 Mar 10]. Available from: [https://github.com/WarrenLab/minigraph-](https://github.com/WarrenLab/minigraph-cactus-nf)
793 [cactus-nf](https://github.com/WarrenLab/minigraph-cactus-nf)

794 69. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of
795 SAMtools and BCFtools. *Gigascience* [Internet]. 2021;10. Available from:
796 <http://dx.doi.org/10.1093/gigascience/giab008>

797 70. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: interactive visualization of de novo genome
798 assemblies. *Bioinformatics*. 2015;31:3350–2.

799 71. Pfeiffer F, Gröber C, Blank M, Händler K, Beyer M, Schultze JL, et al. Systematic evaluation of
800 error rates and causes in short samples in next-generation sequencing. *Sci Rep*. 2018;8:10950.

801 72. Sirén J, Paten B. GBZ file format for pangenome graphs. *Bioinformatics*. 2022;38:5012–8.

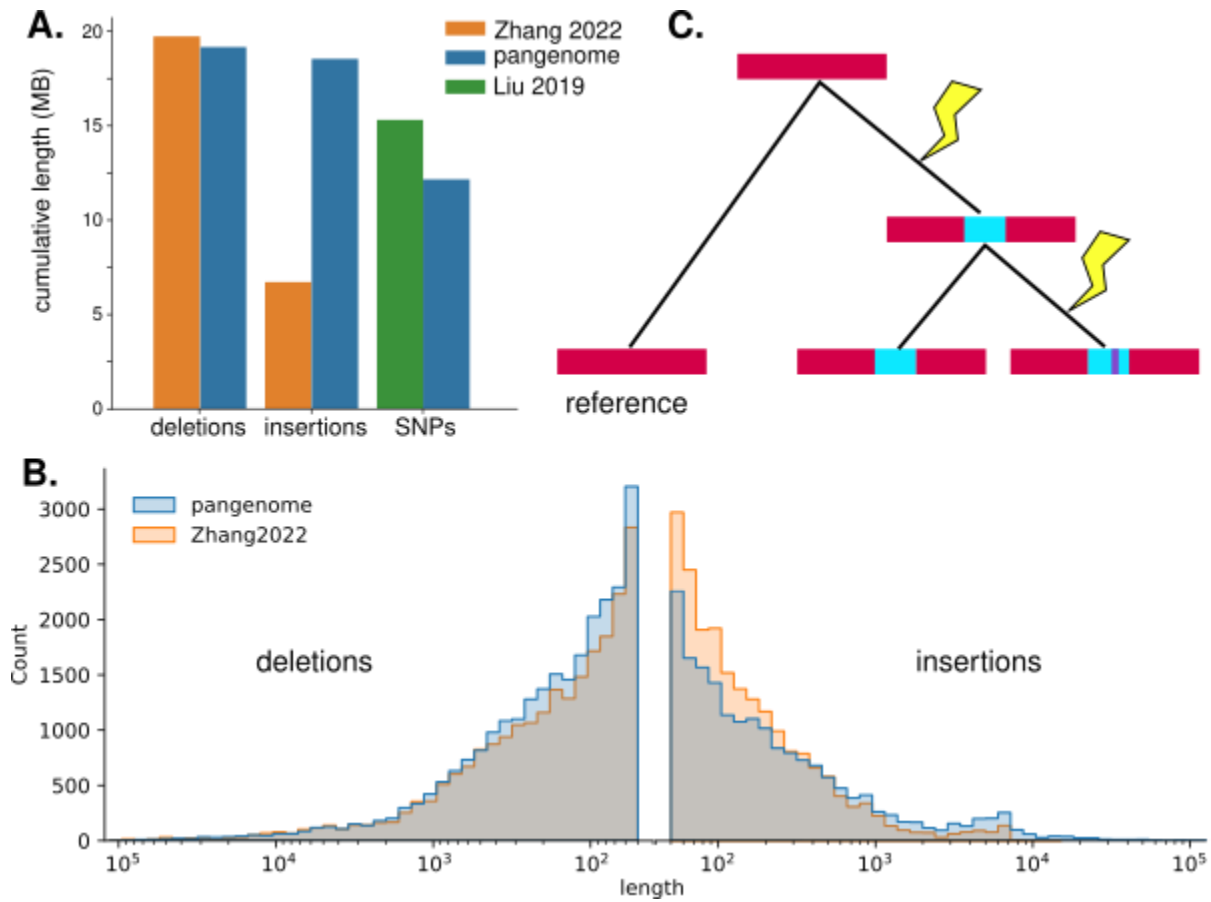
802 73. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018;34:3094–
803 100.

804 74. Herzeel C, Costanza P, Decap D, Fostier J, Wuyts R, Verachtert W. Multithreaded variant calling
805 in elPrep 5. *PLoS One*. 2021;16:e0244471.

806

807

808 **Figures**



809

810 **Figure 1: Cataloging variation in the pangenome graph.** (a) Total lengths of sequence contained

811 in insertions (INS), deletions (DEL), and SNVs, compared between this study (“pangenome”) and

812 read-alignment methods [41,42]. (b) Distribution of lengths of insertions and deletions found in this

813 study compared to those found by Zhang et al. [41] using long reads shows that while long-read

814 alignment finds more short insertions (< 1kb) than the pangenome, the larger cumulative length of

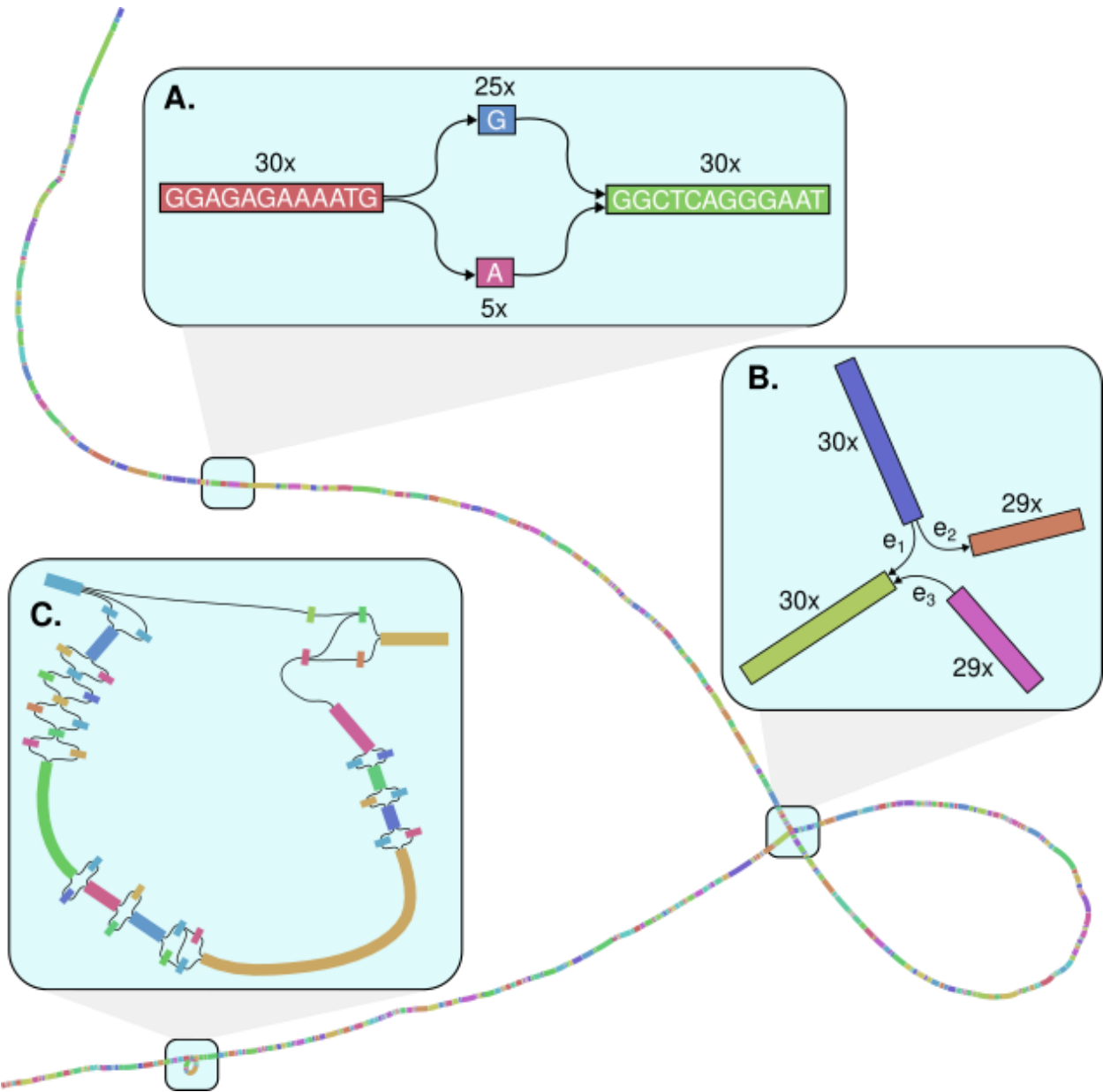
815 insertions found by our pangenome compared to Zhang as shown in (a) is driven by long insertions

816 (>1kb), which have a larger effect on cumulative length. (c) A hypothetical schematic of how nested

817 variation can evolve: an insertion mutation is followed by a later single nucleotide mutation, resulting

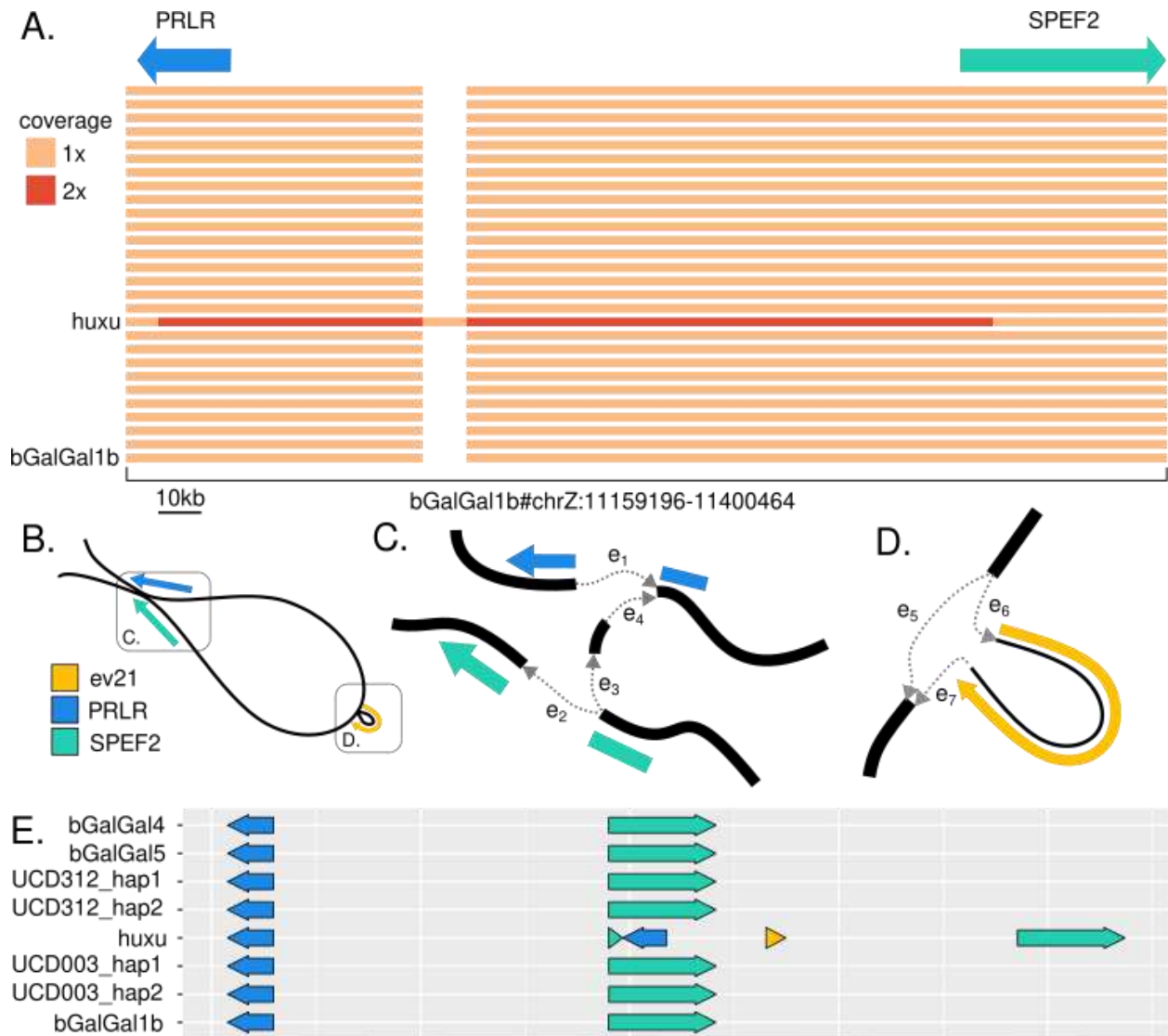
818 in an insertion relative to the reference that contains a segregating site. A genotype against a linear

819 reference would represent these as three different alleles, whereas a pangenome conserves the nested
820 structure of this variation.



823 **Figure 2: A visual representation of the pangenome graph for the gene *IGLL1*.** (a) *IGLL1*
 824 contains many SNVs, including one at bGalGal1b#chr15:7,955,357, in its coding sequence. The
 825 graph of this SNV shows that while all 30 haplotypes have the same sequence before and after the
 826 SNV, 25 haplotypes have G in this position and 5 have A. (b) The pangenome of *IGLL1* contains a
 827 ~5kb deletion compared to bGalGal1b in one haplotype of a single individual, UCD312. At the

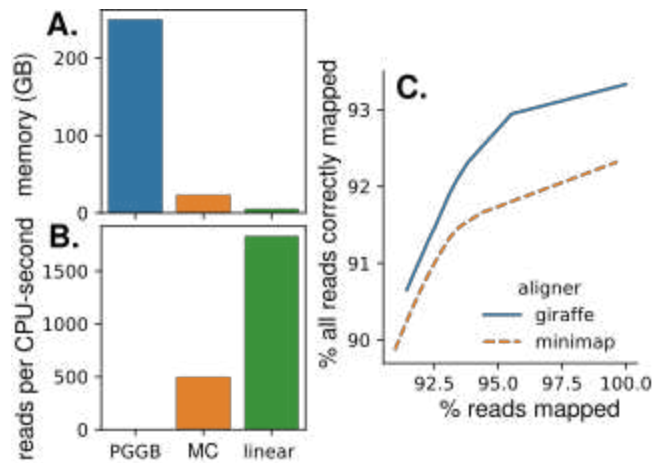
828 juncture in the pangenome graph where the deletion haplotype branches from the rest, this haplotype
829 follows edge e1 to skip the sequence in the loop, while the other 29 haplotypes follow edge e2 to
830 include the sequence, and then e3 to join back with the deletion haplotype afterwards. (c) *IGLL1* also
831 contains a ~300bp insertion compared to bGalGal1b in 22 haplotypes. The inserted sequence contains
832 SNVs, so while a linear representation of this insertion considers each version of the insertion as a
833 different allele, the pangenome graph is able to correctly record it as a biallelic variant (i.e., insertion or
834 no insertion) containing additional variable sites. Furthermore, reads can align to this sequence in the
835 pangenome but would be left unmapped when aligning to bGalGal1b as it does not contain this
836 sequence.



837

838 **Figure 3: Disentangling complex variation at the K locus with the pangenome graph.** (a) A
 839 one-dimensional view of the pangenome subgraph for the K locus, with nodes colored by path
 840 coverage (i.e., the number of times a haplotype path passes through them) and the locations of the
 841 genes *PRLR* and *SPEF2* denoted. Huxu shows double path coverage of part of the locus, as well as an
 842 insertion. Alignment verified that this insertion contains the sequence of the avian leukosis virus ev21.
 843 (b) A two-dimensional view of the same graph, showing both the tandem duplication and the ev21
 844 insertion. (c) At the junction where the paths containing the tandem duplication deviate from the

845 paths that do not, all paths begin by traversing edge e1 and moving through most of the sequence of
846 the K locus. However, at the e2/e3 fork, a path can either traverse e2 to leave the K locus, or traverse e3
847 and e4 to include a tandem duplication of parts of *PRLR* and *SPEF2*. (d) A more detailed view of the
848 ev21 insertion, showing the two possible paths at this juncture: a path can traverse edge e5 to skip the
849 insertion, or it can traverse edge e6, then the ev21 sequence, then e7, to include the insertion. (e)
850 Linear untangled view of the locus, confirming previous studies of the structure of the locus, with a
851 tandem duplication of parts of both genes and an insertion of the ev21 sequence.



852

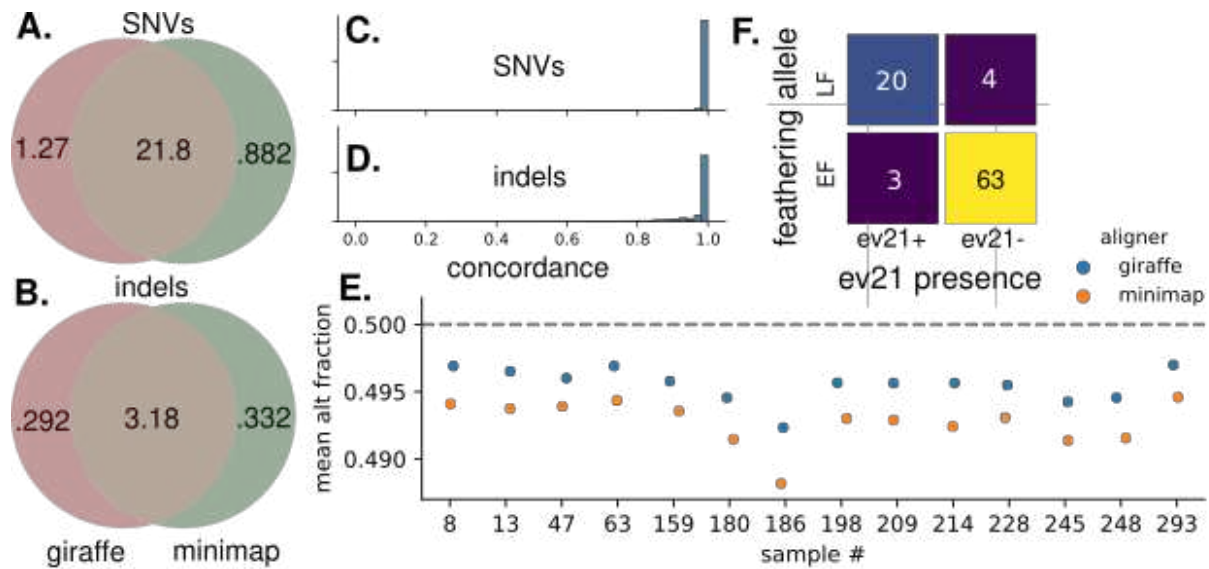
853 **Figure 4: Comparing pangenome and linear aligner performance for short reads. (a-b)**

854 Alignment of short reads with VG giraffe is more memory-efficient (a) and faster (b) when aligning to

855 the minigraph-cactus (MC) pangenome graph compared to the PGGB graph. Linear alignment with

856 minimap2 is the fastest and most memory-efficient. (c) A larger percentage of all simulated reads is

857 correctly aligned with giraffe regardless of how permissive the minimum map quality filter is.



858

859 **Figure 5: Genotyping 100 diverse chickens.** (a-b) Counts in millions of common and different
 860 SNVs (a) and indels (b) found by genotyping pipelines using giraffe vs. minimap as the aligner. Only
 861 variants with a quality score of at least 10 are considered. (c-d) Concordance distributions for SNVs (c)
 862 and indels (d) detected by both genotyping methods with $QUAL \geq 10$. (e) Mean fractions per sample
 863 of mapped reads containing the alternate allele at putative heterozygous sites show that giraffe
 864 alignments contain less reference bias for every chicken, as they deviate less from the expected value of
 865 0.5. Sample information in Supplementary Table 2 and full plot for all 100 chickens in Supplementary
 866 Figure S5. (f) Genotyping 100 chickens at the K locus reproduces previous results finding that while
 867 most chickens with the late feathering allele (LF) also have an ev21 insertion at the K locus (ev21+),
 868 some chickens have the late feathering allele without an ev21 insertion.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementarymaterial.docx](#)