Received: 19 March 2018 | Revised: 11 June 2018 | Accepted: 6 July 2018

DOI: 10.1111/1755-0998.12933

NEWS AND VIEWS

WILEY MOLECULAR ECOLOGY RESOURCES

Opinion

How complete are “complete” genome assemblies?—An avian perspective

Valentina Peona[1]* | Matthias H. Weissensteiner[1,2]* | Alexander Suh[1]

[1]Department of Evolutionary Biology, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden

[2]Division of Evolutionary Biology, Faculty of Biology, Ludwig-Maximilian University of Munich, Planegg-Martinsried, Germany

Correspondence
Valentina Peona, Matthias H. Weissensteiner and Alexander Suh, Department of Evolutionary Biology, Evolutionary Biology Centre, Uppsala University, SE-752 36 Uppsala, Sweden.
Emails: valentina.peona@ebc.uu.se (V.P.), matthias.weissensteiner@ebc.uu.se (M.H.W.), alexander.suh@ebc.uu.se (A.S.)

Funding information
SciLifeLab Swedish Biodiversity Program, Grant/Award Number: 2015-R14; Swedish Science Foundation, Grant/Award Number: 2016-05139

Abstract

The genomics revolution has led to the sequencing of a large variety of nonmodel organisms often referred to as “whole” or “complete” genome assemblies. But how complete are these, really? Here, we use birds as an example for nonmodel vertebrates and find that, although suitable in principle for genomic studies, the current standard of short-read assemblies misses a significant proportion of the expected genome size (7% to 42%; mean 20 ± 9%). In particular, regions with strongly deviating nucleotide composition (e.g., guanine-cytosine-[GC]-rich) and regions highly enriched in repetitive DNA (e.g., transposable elements and satellite DNA) are usually underrepresented in assemblies. However, long-read sequencing technologies successfully characterize many of these underrepresented GC-rich or repeat-rich regions in several bird genomes. For instance, only ~2% of the expected total base pairs are missing in the last chicken reference (galGal5). These assemblies still contain thousands of gaps (i.e., fragmented sequences) because some chromosomal structures (e.g., centromeres) likely contain arrays of repetitive DNA that are too long to bridge with currently available technologies. We discuss how to minimize the number of assembly gaps by combining the latest available technologies with complementary strengths. At last, we emphasize the importance of knowing the location, size and potential content of assembly gaps when making population genetic inferences about adjacent genomic regions.

KEYWORDS
birds, genomics, hybrid assembly, long reads, multiplatform sequencing, repeats

Among vertebrates, birds currently exhibit one of the highest numbers of freely available genome assemblies. For example, as of May 2018, there were 170 mammalian and 101 avian genome assemblies present in GenBank (https://www.ncbi.nlm.nih.gov/genbank/). While the genomes of only three bird species (chicken, turkey and zebra finch) were sequenced by 2010 (Dalloul et al., 2010; Hillier et al., 2004; Warren et al., 2010), already over 50 were sequenced by 2014 (Ellegren et al., 2012; Poelstra et al., 2014; Zhang et al., 2014) and over 75 by mid-2016 (reviewed by Kapusta & Suh, 2017). Hundreds of additional genomes are currently being sequenced by the Bird 10,000 Genomes (B10K) project, with the ultimate aim of generating genome assemblies for all bird species (Jarvis, 2016). As of July 5, 2017, B10K has generated 334 genome assemblies of representatives from nearly all avian families (http://b10k.genomics.cn/).

*Equal contributions, alphabetical order

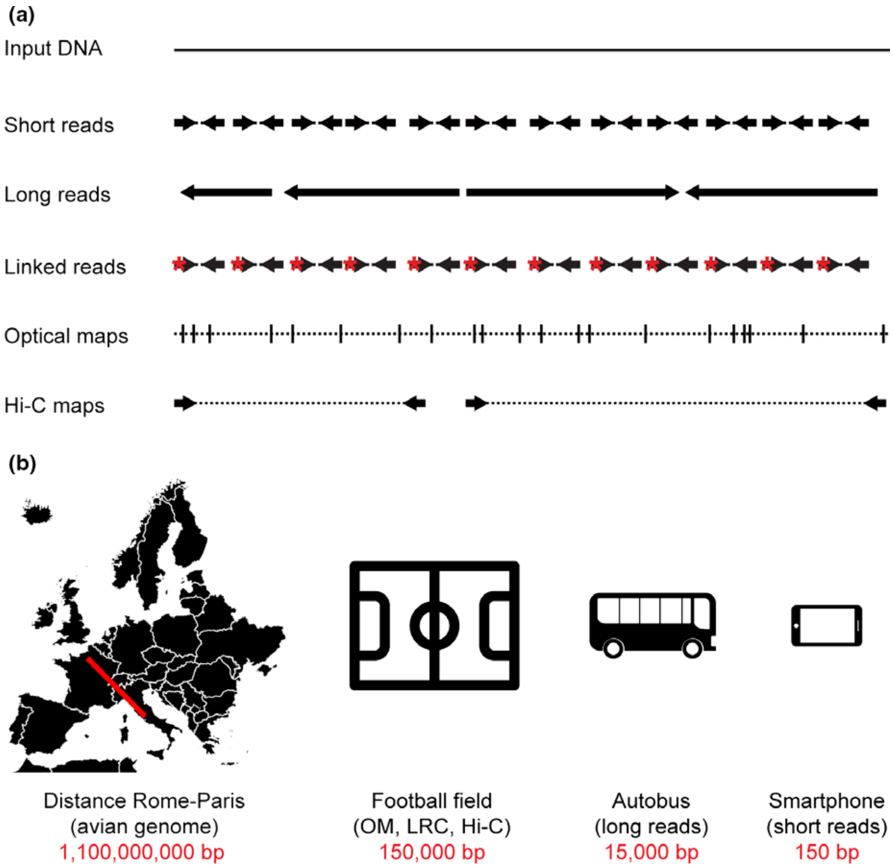

FIGURE 1 Currently available genomics technologies. (a) Schematic illustration of the data structure of these technologies produced from a hypothetical input DNA molecule. Short reads come in read pairs, long reads as single reads, linked-read clouds (LRC) as short-read pairs with a unique barcode (red asterisk) for each input molecule. Optical maps (OM) contain physical distances between short sequence motifs, and Hi-C maps are short-read pairs of 3D genome interactions obtained through chromatin conformation capture. (b) Schematic size relations of the data structure from panel (a). Examples are scaled by illustrating 1 base pair as 1 mm. Icons made by Freepik from www.flaticon.com [Colour figure can be viewed at wileyonlinelibrary.com]

In parallel to these quantity-focused efforts, others aim to improve the quality of already existing genome assemblies (chicken, Anna's hummingbird, zebra finch, hooded crow; Korlach et al., 2017; Warren et al., 2017; Weissensteiner et al., 2017).

Why are (avian) genome assemblies of varying quality? To date, no sequencing technology exists that is capable of sequencing entire avian chromosomes from one end to the other in a single read (Figure 1a). Instead, short-read sequencing technologies produce sequence information ("reads"; usually in "read pairs") of some hundreds of base pairs (bp), and long-read sequencing technologies yield reads of some tens of thousands of bp (Figure 1b; Goodwin, McPherson, & McCombie, 2016). Similar to a jigsaw puzzle, these reads are then assembled into contiguous sequences ("contigs") and linked contigs ("scaffolds") (Yandell & Ence, 2012). Scaffolds thus consist of contigs (all nucleotides determined) and assembly gaps (placeholders of undetermined "N" nucleotides), the latter usually containing repetitive elements such as interspersed repeats (transposable elements and endogenous viruses) and tandem repeats (microsatellites and satellites; Figure 2; Chaisson, Wilson, & Eichler, 2015b; Thomma et al., 2016). Like a puzzle piece occurring multiple times in a single puzzle game, repetitive elements are problematic for genome assembly because they contain ambiguous information about their exact position. If reads or read pairs are shorter than the repeat unit (an individual transposon or tandem repeat) and there are multiple identical repeat copies in the genome, this ambiguity will interfere with the assembly process and cause a loss of information (assembly gaps). This issue typically results in assembly gaps of known size (i.e., approximated by "N" nucleotides) when contigs are bridged into scaffolds by linkage information of read pairs (Figure 2, left; Chaisson et al., 2015b). On the other hand, repeat-rich regions (e.g., clusters of interspersed repeats or large arrays of tandem repeats) are usually not spanned by reads or read pairs at all and thus often lead to termination of scaffolds, that is, assembly gaps of unknown size (Figure 2, right; Chaisson et al., 2015b).

Nearly all currently available avian genome assemblies were generated using short-read sequencing (mostly using the Illumina platform; Kapusta & Suh, 2017). Considering that one can expect a positive correlation between read length and the ability to assemble individual repeat units or repeat-rich regions, we hypothesized that currently published avian genomes based only on short-read sequencing contain significant amounts of missing DNA (i.e., the sum of all DNA hidden in assembly gaps as defined in Figure 2). Although the "true" genome sizes of birds cannot be determined precisely, at least as long as read lengths are shorter than individual chromosomes, genome sizes of hundreds of bird species have been approximated through flow cytometry (Gregory, 2017) and we consider these estimates to be entirely independent of genome assembly sizes. We therefore quantified the amount of missing DNA and the numbers of assembly gaps by comparing assembly summary statistics, flow cytometric genome size estimates and haploid chromosome numbers. While we cannot determine whether these genome size estimates might be biased by genomic properties (such as a higher GC or repeat content) because "true" genome sizes are unknown, we note that the comparison of haploid chromosome numbers vs.

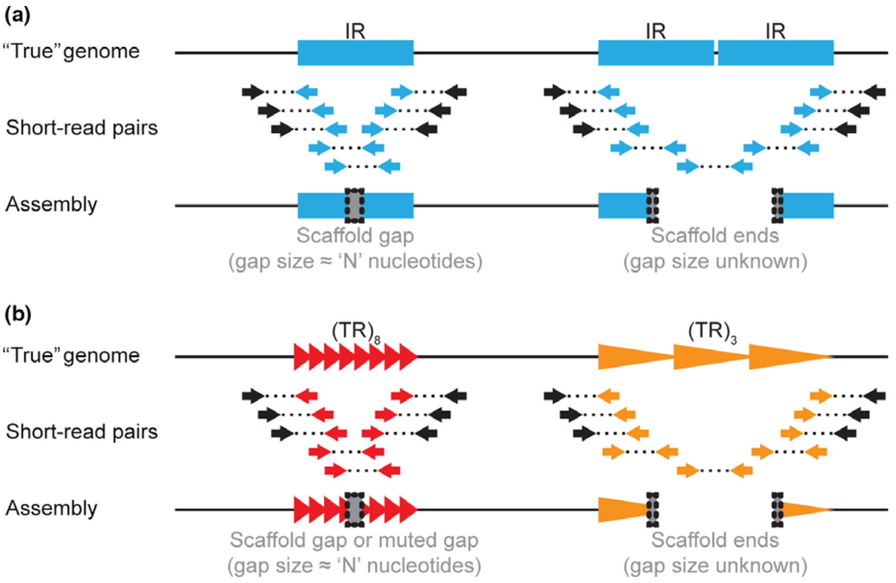

FIGURE 2 Schematic illustration of how repetitive elements may cause gaps in short-read genome assemblies. (a) Interspersed elements (IRs; transposable elements or endogenous viruses, both in blue) can lead to within-scaffold gaps of approximate size (left) or between-scaffold gaps of unknown size (right). (b) Tandem repeats (TRs; microsatellites in red or satellites in orange) can lead to within-scaffold gaps (left; alternatively, a muted gap, i.e., a sequence contraction) or between-scaffold gaps (right) [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 1 Quantification of missing DNA in 13 bird species from Zhang et al. (2014) where cytogenetic, flow cytometric and short-read genome assembly data are available

Name	Chromosomes (n)[a]	Assembly scaffolds[b]	Expected size (Gb)[a]	Assembly size (Gb)[b]	Missing (Mb)[c]	"N" gaps (Mb)[d]	% missing DNA[e]
Anas platyrhynchos (Pekin duck)	40	78,487	1.41	1.11	303.27	34.69	24.0
Balearica regulorum (grey-crowned crane)	40	125,353	1.49	1.14	347.70	5.31	23.7
Calypte anna (Anna's hummingbird)	37	122,597	1.41	1.11	293.70	38.67	23.6
Cariama cristata (red-legged seriema)	54	139,827	1.47	1.15	321.30	5.69	22.3
Columba livia (pigeon)	40	38,878	1.30	1.11	189.16	20.50	16.1
Corvus brachyrhynchos (American crow)	40	33,296	1.22	1.10	127.33	40.18	13.7
Falco peregrinus (peregine falcon)	25	21,224	1.42	1.17	244.05	18.54	18.5
Haliaeetus leucocephalus (bald eagle)	33	346,419	1.40	1.26	139.75	49.34	13.5
Melopsittacus undulatus (budgerigar)	29	25,212	1.30	1.12	183.38	30.75	16.5
Phoenicopterus ruber (American flamingo)	40	144,901	1.22	1.14	77.75	6.52	6.9
Struthio camelus (ostrich)	40	32,461	2.06	1.23	835.14	40.39	42.4
Tinamus guttatus (white-throated tinamou)	40	176,848	1.21	1.06	152.58	22.72	14.5
Tyto alba (barn owl)	46	166,074	1.50	1.14	357.53	11.27	24.6

Notes. n: Haploid chromosome number.
Weblinks to sampled genome assemblies are listed in Supporting Information Data S1.
[a]Chromosome number and genome size estimates from Gregory (2017) and Christidis (1990). Genome size estimates were converted from C-values into billion basepairs (Gb) assuming 1 pg = 0.978 Gb (Doležel, Bartoš, Voglmayr, & Greilhuber, 2003).
[b]Assembly metrics from Table S1 of Kapusta and Suh (2017).
[c]Assembly size subtracted from expected genome size.
[d]Sum of all "N" nucleotides present in the genome assembly.
[e]Percentage of the expected genome size either missing in the assembly or assembled as "N" nucleotides.

the number of scaffolds should provide an additional measure of genome completeness. From 45 bird species with short-read genome assemblies in Zhang et al. (2014), we were able to obtain genome size and karyotype data for 13 species (Table 1) which span most of the major groups within Neoaves, Galloanserae and Palaeognathae (sensu Jarvis et al., 2014; Suh, 2016).

In a "complete" assembly, the number of scaffolds (or ideally, contigs) should equal the haploid chromosome number. However, the haploid chromosome number of the sampled birds ranges from 25 to 54 and the number of scaffolds ranges from approximately 21,000 to 346,000 (mean 112,000 ± 91,000; Table 1). Therefore, there are tens of thousands to hundreds of thousands of gaps between scaffolds (i.e., of unknown size) in these genome assemblies (Table 1). Furthermore, there are significant amounts of within-scaffold gaps, given that the number of undetermined "N" nucleotides ranges from approximately six to 49 million base pairs (Mb; mean 25 ± 15 Mb; Table 1). We next calculated the total amount of missing DNA by subtracting the assembly size from the flow cytometric

TABLE 2 Quantification of missing DNA in the reference genomes of three model organisms

Name	Chromosomes (n)[a]	Assembly scaffolds	Expected size (Gb)[a]	Assembly size (Gb)	Missing (Mb)[b]	"N" gaps (Mb)[c]	% missing DNA[d]
Arabidopsis thaliana (arabidopsis) [TAIR10]	5	7	0.125	0.12	5.33	0.20	4.4
Drosophila melanogaster (fruit fly) [dm6]	4	1,870	0.17	0.14	30.00	1.10	18.0
Homo sapiens (human) [hg38]	23	594	3.42	3.25	162.00	161.00	10.3

Notes. *n*: Haploid chromosome number.
Weblinks to sampled genome assemblies are listed in Supporting Information Data S1.
[a] Chromosome number and genome size estimates from Gregory (2017). Genome size estimates were converted from C-values into billion basepairs (Gb) assuming 1 pg = 0.978 Gb (Doležel et al., 2003).
[b] Assembly size subtracted from expected genome size.
[c] Sum of all "N" nucleotides present in the genome assembly.
[d] Percentage of the expected genome size either missing in the assembly or assembled as "N" nucleotides.

genome size estimate and adding the number of "N" nucleotides in the assembly. The estimates range from approximately 7% to 42% missing DNA (mean 20 ± 9%). Even the lowest estimate is a significant proportion considering that analyses based on such genome assemblies are often referred to as "whole-genome" or "genome-wide" analyses. Note that hundreds of gaps are still unresolved in the human genome (Table 2), which is arguably the best vertebrate genome assembly available (Chaisson et al., 2015a). Even the well-curated reference genomes of important model organisms, such as *Drosophila melanogaster* and *Arabidopsis thaliana*, still contain missing DNA (Table 2). One may argue that this missing DNA almost entirely consists of repetitive DNA and is outside the scope or interest of most (avian) genomics studies. However, simply ignoring assembly gaps "may lead to false positives and over-optimistic findings," as shown in Domanska, Kanduri, Simovski, and Sandve (2018) where depletion of mapped reads in gap regions biased the inference of co-localization of genomic features. We currently lack a comprehensive understanding of the functional relevance of repetitive DNA even in the most-studied model organisms such as humans (Cordaux & Batzer, 2009; Kellis et al., 2014; Koonin, 2016) and *Drosophila* (Gallach, 2015; Joshi & Meller, 2017; Zhou et al., 2013); thus, it might be premature to label these regions as completely irrelevant in birds. Furthermore, short-read sequencing is known to be biased against highly GC-rich sequences, meaning that these will be largely underrepresented in the resulting assembly (Chaisson et al., 2015b). This problem might be particularly pronounced in birds because their smallest chromosomes ("microchromosomes") are highly GC-rich (Burt, 2002). It is therefore imaginable that many genes and other functionally important regions are hidden in the missing DNA due to their repetitiveness and/or nucleotide composition. To this end, a growing number of studies suggest that many genes previously declared as "missing" in bird genomes were in fact just "missed" due to their GC richness (Bornelöv et al., 2017; Botero-Castro, Figuet, M-k, Nabholz, & Galtier, 2017; Hron, Pajer, Pačes, Bartůněk, & Elleder, 2015). Overcoming the issue of GC underrepresentation requires long-read sequencing data (Chaisson et al., 2015b) or modified protocols for short-read library preparation (Tilak, Botero-Castro, Galtier, & Nabholz, 2018).

To further quantify missing DNA, we next analysed the genome assemblies of chicken and zebra finch (Table 3), two avian model systems where considerable efforts combining conventional Sanger sequencing, bacterial artificial chromosome libraries and cytogenetic methods were used to build chromosome models (Hillier et al., 2004; Warren et al., 2010). Thanks to the combination of all these techniques (including Sanger read lengths longer than those in short-read sequencing), these genome assemblies have lower amounts of missing DNA than the aforementioned short-read assemblies, but nevertheless contain tens of thousands of gaps between scaffolds (Table 3). With the recent release of the first avian genome assemblies using the Pacific Biosciences long-read sequencing platform (chicken, zebra finch, Anna's hummingbird and hooded crow), the sequence resolution of GC-rich and repeat-rich regions has been strongly improved (Korlach et al., 2017; Warren et al., 2017; Weissensteiner et al., 2017). Among these four birds with now available long-read assemblies (Table 3), the chicken long-read genome assembly (version galGal5) is the most complete and facilitates a direct comparison to the previous chicken Sanger genome (version galGal4). Strikingly, the total amount of missing DNA decreased from 14.1% to 2.4% and the number of "N" nucleotides decreased from approximately 58 to 12 Mb (Table 3). The total number of scaffolds is very similar between the galGal5 and galGal4 assemblies (approximately 23,000), despite the significant increase in assembly contiguity through long-read sequencing (Kapusta & Suh, 2017; Warren et al., 2017). It is likely that the high number of galGal5 scaffolds despite many closed gaps results from the fact that many GC-rich or repeat-rich regions (which were largely inaccessible with previous technologies) have been successfully sequenced and partially assembled, but remain unplaced on chromosomes (Warren et al., 2017). This would explain why sequences belonging to the three smallest chicken microchromosomes (36, 37 and 38) have still not been confidently assigned (Warren et al., 2017).

Does this mean that there are no complete avian genome assemblies? A truly "complete" assembly should contain each chromosome as a gap-free sequence from one chromosome end to another, including entire centromeres and telomeres. For the time being, this is not feasible for avian genomes. To continue with the puzzle metaphor, all current sequencing technologies rely on assembling puzzle pieces of DNA into a jigsaw puzzle where the end result is unknown. Although sequencing technologies are currently

TABLE 3 Quantification of missing DNA in four birds where both short-read (Illumina; except for Sanger in avian models) and long-read (PacBio) assemblies are available

Name	Chromosomes (*n*)[a]	Assembly scaffolds[b]	Expected size (Gb)[a]	Assembly size (Gb)[b]	Missing (Mb)[c]	"N" gaps (Mb)[d]	% missing DNA[e]
Calypte anna (Anna's hummingbird) [Illumina; from Table 1]	37	122,597	1.41	1.11	293.70	38.67	23.6
Calypte anna (Anna's hummingbird) [PacBio contigs]	37	1,076	1.41	1.00	410.00	0.00	29.1
Corvus cornix (hooded crow) [v1, Illumina]	80	1,299	1.19	1.04	152.00	30.64	15.3
Corvus cornix (hooded crow) [v2, PacBio]	80	145	1.19	1.05	144.00	9.55	12.9
Gallus gallus (chicken) [galGal4, Sanger]	39	23,870	1.25	1.11	114.02	57.94	14.1
Gallus gallus (chicken) [galGal5, PacBio]	39	23,474	1.25	1.23	18.31	11.76	2.4
Taeniopygia guttata (zebra finch) [taeGut2, Sanger]	40	35,422[f]	1.22	1.22	0.88	10.12	0.9
Taeniopygia guttata (zebra finch) [PacBio contigs]	40	1,159	1.22	1.14	81.20	0.00	6.7

n: Haploid chromosome number.

Weblinks to sampled genome assemblies are listed in Supporting Information Data S1.

[a] Chromosome number and genome size estimates from Gregory (2017) and Christidis (1990). Genome size estimates were converted from C-values into billion basepairs (Gb) assuming 1 pg = 0.978 Gb (Doležel et al., 2003).

[b] Assembly metrics from Table S1 of Kapusta and Suh (2017), except for galGal4 (Hillier et al., 2004), hooded crow (Weissensteiner et al., 2017) and Anna's hummingbird + zebra finch PacBio (present study).

[c] Assembly size subtracted from expected genome size.

[d] Sum of all "N" nucleotides present in the genome assembly.

[e] Percentage of the expected genome size either missing in the assembly or assembled as "N" nucleotides.

[f] Although the zebra finch assembly taeGut2 contains 64 chromosome-level scaffolds, one of these ("chrUn") is a concatenation of 35,359 unanchored contigs separated by "N" gaps.

undergoing massive improvements in read lengths leading to closure of particularly difficult assembly gaps (Jain et al., 2018a; Kuderna et al., 2018), some parts of avian genomes will likely remain "un-assemblable" (Figure 3) until read lengths of millions of base pairs can be achieved. Recent technological developments for long-range scaffolding, such as linked-read cloud sequencing (Weisenfeld, Kumar, Shah, Church, & Jaffe, 2017), nanochannel optical mapping (OM; Lam et al., 2012) and chromosome conformation capture (Hi-C; Dudchenko et al., 2017), offer complementary information to achieve chromosome-level scaffolds by minimizing the number of gaps between scaffolds and yield scaffolds spanning the entire length of individual chromosomes, potentially even for the notoriously difficult-to-assemble avian microchromosomes (Figure 3). Such an approach for multiplatform sequencing and assembly was recently successful in mammalian genomes (Bickhart et al., 2017; Seo et al., 2016) and is likely achievable for avian genomes (Cooke et al., 2017; V. Peona & A. Suh, unpublished data; M. H. Weissensteiner, unpublished data).

There is already the chance to get a glimpse into particularly difficult-to-assemble gaps such as centromeres in humans (Jain et al., 2018b). For avian genomes, Weissensteiner et al. (2017) recently demonstrated that optical mapping data provide an indirect means to estimate the size and potential sequence content of some assembly gaps. They could anchor candidate centromeric tandem repeat arrays with array lengths of over a million base pairs into the hooded crow genome assembly and illustrate an effect on genetic diversity and differentiation between populations in adjacent genomic regions. This approach was of importance to

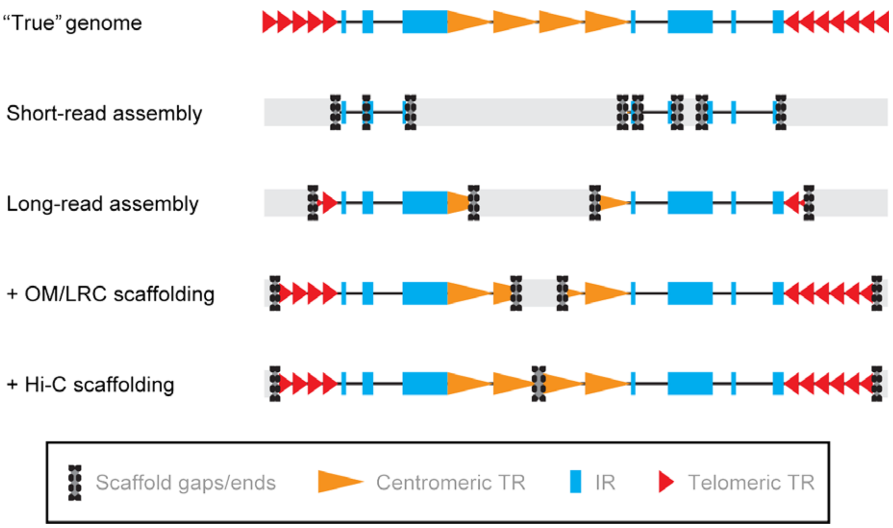

FIGURE 3 A road map for minimizing the number of assembly gaps using current technologies. Missing DNA is indicated by grey bars, interspersed repeats (IRs) are in blue, and tandem repeats (TRs) are in orange and red. OM: optical mapping; LRC: linked-read cloud sequencing; Hi-C: chromosome conformation capture [Colour figure can be viewed at wileyonlinelibrary.com]

chromosome 18 containing the previously identified "speciation island"—a region of particularly high genetic differentiation between European hooded and carrion crow populations presumably involved in reproductive isolation (Poelstra et al., 2014). Although chromosome 18 contains multiple assembly gaps, only the between-scaffold gap in the middle of the "speciation island" is large and contains a tandem repeat array which potentially is (part of) a centromere (Weissensteiner et al., 2017), showcasing the importance of incorporating information on genome structure into population genetic studies. While assembly gaps may bias results in co-localization analyses of genomic features (Domanska et al., 2018), fragmented assemblies may also lead to biased results when assessing the chromosomal landscape of population genetic statistics. For example, stretches of elevated differentiation ("F_{ST} peaks") are often used to detect genomic regions under selection or to infer gene flow (Wolf & Ellegren, 2017). However, in an overly fragmented assembly, consecutive stretches of elevated differentiation may be too short to be detected, or erroneous inferences may occur if stretches are considered across scaffold boundaries. Thus, it is likely that both false-positive and false-negative discoveries may occur more frequently in incomplete assemblies.

At last, it is important to keep in mind that birds are on the low end of repeat content among vertebrates (Sotero-Caio, Platt, Suh, & Ray, 2017). Given that difficulty of genome assembly increases with repeat content (Sedlazeck, Lee, Darby, & Schatz, 2018), our case study on avian genomes might be a good starting point to illustrate that even sequencing genomes with relatively low repeat content is far from trivial and should not be labelled as "complete" yet. While avian genomes show a repeat density of only 4%–10% with a maximum of 22% in the downy woodpecker (Zhang et al., 2014), other vertebrates, invertebrates and plants often reach a repeat density of more than 50% (e.g., human genome 50%–69%, Cordaux & Batzer, 2009; de Koning, Gu, Castoe, Batzer, & Pollock, 2011; *Locusta migratoria* ~59%, Wu, Twort, Crowhurst, Newcomb, & Buckley, 2017; *Fritillaria* spp. 90%, Ambrozová et al., 2011). This even more increases the need for caution when interpreting results than illustrated here for birds.

So, how complete are "complete" avian genome assemblies? For now, the answer is indeed that substantial parts are usually missing. However, we are confident that the true extent of genetic variation, hidden in repeat-rich and other tricky-to-assemble regions, will be more and more appreciated in the near future, a development spurred by rapid technological developments. Meanwhile, considering that not all gaps are equal in size or structure, our recommendation is this: Mind the gap!

ACKNOWLEDGEMENTS

We thank Mozes Blom, Anne-Marie Dion-Côté, Jan Engler, Per Ericson, Takeshi Kawakami, Cormac Kinsella and Robyn Womack for valuable comments on an earlier version of the manuscript. We also thank Shawn Narum and four anonymous reviewers for further improving this manuscript with their comments. A.S. was supported by grants from the Swedish Science Foundation (2016-05139) and the SciLifeLab Swedish Biodiversity Program (2015-R14).

CONFLICT OF INTEREST

No conflict of interests to declare.

AUTHOR CONTRIBUTIONS

V.P., M.H.W. and A.S. conceived the study, analysed the data and wrote the manuscript.

DATA ACCESSIBILITY

All the data used in this study were previously made available by the cited references.

ORCID

Valentina Peona http://orcid.org/0000-0001-5119-1837
Matthias H. Weissensteiner https://orcid.org/0000-0001-9302-798X
Alexander Suh http://orcid.org/0000-0002-8979-9992

REFERENCES

Ambrozová, K., Mandáková, T., Bures, P., Neumann, P., Leitch, I. J., Koblízková, A., … Lysak, M. A. (2011). Diverse retrotransposon families and an AT-rich satellite DNA revealed in giant genomes of *Fritillaria* lilies. *Annals of Botany*, *107*, 255–268. https://doi.org/10.1093/aob/mcq235

Bickhart, D. M., Rosen, B. D., Koren, S., Sayre, B. L., Hastie, A. R., Chan, S., … Smith, T. P. (2017). Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nature Genetics*, *49*, 643–650. https://doi.org/10.1038/ng.3802

Bornelöv, S., Seroussi, E., Yosefi, S., Pendavis, K., Burgess, S. C., Grabherr, M., … Andersson, L. (2017). Correspondence on Lovell et al.: Identification of chicken genes previously assumed to be evolutionarily lost. *Genome Biology*, *18*, 112. https://doi.org/10.1186/s13059-017-1231-1

Botero-Castro, F., Figuet, E., M-k, T., Nabholz, B., & Galtier, N. (2017). Avian genomes revisited: Hidden genes uncovered and the rates vs. traits paradox in birds. *Molecular Biology and Evolution*, *34*, 3123–3131. https://doi.org/10.1093/molbev/msx236

Burt, D. W. (2002). Origin and evolution of avian microchromosomes. *Cytogenetic and Genome Research*, *96*, 97–112. https://doi.org/10.1159/000063018

Chaisson, M. J. P., Huddleston, J., Dennis, M. Y., Sudmant, P. H., Malig, M., Hormozdiari, F., … Eichler, E. E. (2015a). Resolving the complexity of the human genome using single-molecule sequencing. *Nature*, *517*, 608–611. https://doi.org/10.1038/nature13907

Chaisson, M. J. P., Wilson, R. K., & Eichler, E. E. (2015b). Genetic variation and the *de novo* assembly of human genomes. *Nature Reviews Genetics*, *16*, 627–640. https://doi.org/10.1038/nrg3933

Christidis, L. (1990). *Animal cytogenetics, volume 4. Chordata 3. B.* Berlin, Stuttgart: Aves Gebrüder Borntraeger.

Cooke, T. F., Fischer, C. R., Wu, P., Jiang, T. X., Xie, K. T., Kuo, J., … Bustamante, C. D. (2017). Genetic mapping and biochemical basis of yellow feather pigmentation in budgerigars. *Cell*, *171*(427–439), e421.

Cordaux, R., & Batzer, M. A. (2009). The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics*, *10*, 691–703. https://doi.org/10.1038/nrg2640

Dalloul, R. A., Long, J. A., Zimin, A. V., Aslam, L., Beal, K., Le Ann, B., … Reed, K. M. (2010). Multi-platform next-generation sequencing of the domestic turkey (*Meleagris gallopavo*): Genome assembly and analysis. *PLoS Biology*, *8*, 1–21.

de Koning, A. P. J., Gu, W., Castoe, T. A., Batzer, M. A., & Pollock, D. D. (2011). Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genetics*, *7*, e1002384. https://doi.org/10.1371/journal.pgen.1002384

Doležel, J., Bartoš, J., Voglmayr, H., & Greilhuber, J. (2003). Nuclear DNA content and genome size of trout and human. *Cytometry Part A*, *51A*, 127–128. https://doi.org/10.1002/(ISSN)1097-0320

Domanska, D., Kanduri, C., Simovski, B., & Sandve, G. K. (2018). Mind your gaps: Overlooking assembly gaps confounds statistical testing in genome analysis. *biorxiv*, https://doi.org/10.1101/252973

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., … Aiden, E. L. (2017). *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*, *356*, 92–95. https://doi.org/10.1126/science.aal3327

Ellegren, H., Smeds, L., Burri, R., Olason, P. I., Backström, N., Kawakami, T., … Wolf, J. B. (2012). The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature*, *491*, 756–760. https://doi.org/10.1038/nature11584

Gallach, M. (2015). 1.688 g/cm^3 satellite-related repeats: A missing link to dosage compensation and speciation. *Molecular Ecology*, *24*, 4340–4347. https://doi.org/10.1111/mec.13335

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, *17*, 333–351. https://doi.org/10.1038/nrg.2016.49

Gregory, T. R. (2017). Animal Genome Size Database. http://www.genomesize.com.

Hillier, L. W., Miller, W., Birney, E., Olason, P. I., Backström, N., Kawakami, T., … Wolf, J. B. (2004). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, *432*, 695–716.

Hron, T., Pajer, P., Pačes, J., Bartůněk, P., & Elleder, D. (2015). Hidden genes in birds. *Genome Biology*, *16*, 164. https://doi.org/10.1186/s13059-015-0724-z

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., … Loose, M. (2018a). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*, *36*, 338. https://doi.org/10.1038/nbt.4060

Jain, M., Olsen, H. E., Turner, D. J., Stoddart, D., Bulazel, K. V., Paten, B., … Miga, K. H. (2018b). Linear assembly of a human centromere on the Y chromosome. *Nature Biotechnology*, *36*, 321–323.

Jarvis, E. D. (2016). Perspectives from the Avian Phylogenomics Project: Questions that can be answered with sequencing all genomes of a vertebrate class. *Annual Review of Animal Biosciences*, *4*, 45–59. https://doi.org/10.1146/annurev-animal-021815-111216

Jarvis, E. D., Mirarab, S., Aberer, A. J., Li, B., Houde, P., Li, C., … Zhang, G. (2014). Whole genome analyses resolve the early branches in the tree of life of modern birds. *Science*, *346*, 1320–1331. https://doi.org/10.1126/science.1253451

Joshi, S. S., & Meller, V. H. (2017). Satellite repeats identify X chromatin for dosage compensation in *Drosophila melanogaster* males. *Current Biology*, *27*, 1393–1402.e1392.

Kapusta, A., & Suh, A. (2017). Evolution of bird genomes—a transposon's-eye view. *Annals of the New York Academy of Sciences*, *1389*, 164–185. https://doi.org/10.1111/nyas.13295

Kellis, M., Wold, B., Snyder, M. P., Bernstein, B. E., Kundaje, A., Marinov, G. K., … Hardison, R. C. (2014). Defining functional DNA elements in the human genome. *Proceedings of the National Academy of Sciences*, *111*, 6131–6138. https://doi.org/10.1073/pnas.1318948111

Koonin, E. V. (2016). The meaning of biological information. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*, 20150065.

Korlach, J., Gedman, G., King, S., Chin, C. S., Howard, J. T., Audet, J. N., … Jarvis, E. D. (2017). *De Novo* PacBio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads. *GigaScience*, *6*, 1–16. https://doi.org/10.1093/gigascience/gix085

Kuderna, L. F. K., Lizano, E., Julia, E., Gomez-Garrido, J., Serres-Armero, A., Kuhlwilm, M., … Marques-Bonet, T. (2018). Selective single molecule sequencing and assembly of a human Y chromosome of African origin. *biorxiv*, https://doi.org/10.1101/342667

Lam, E. T., Hastie, A., Lin, C., Ehrlich, D., Das, S. K., Austin, M. D., … Kwok, P. Y. (2012). Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nature Biotechnology*, *30*, 771–776. https://doi.org/10.1038/nbt.2303

Poelstra, J. W., Vijay, N., Bossu, C. M., Lantz, H., Ryll, B., Müller, I., … Wolf, J. B. (2014). The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. *Science*, *344*, 1410–1414.

Sedlazeck, F. J., Lee, H., Darby, C. A., & Schatz, M. C. (2018). Piercing the dark matter: Bioinformatics of long-range sequencing and mapping. *Nature Reviews Genetics*, *19*, 329–346. https://doi.org/10.1038/s41576-018-0003-4

Seo, J.-S., Rhie, A., Kim, J., Lee, S., Sohn, M. H., Kim, C. U., … Kim, C. (2016). *De novo* assembly and phasing of a Korean human genome. *Nature*, *538*, 243–247. https://doi.org/10.1038/nature20098

Sotero-Caio, C., Platt, R. N. II, Suh, A., & Ray, D. A. (2017). Evolution and diversity of transposable elements in vertebrate genomes. *Genome Biology and Evolution*, *9*, 161–177. https://doi.org/10.1093/gbe/evw264

Suh, A. (2016). The phylogenomic forest of bird trees contains a hard polytomy at the root of Neoaves. *Zoologica Scripta*, *45*, 50–62. https://doi.org/10.1111/zsc.12213

Thomma, B. P. H. J., Seidl, M. F., Shi-Kunne, X., Cook, D. E., Bolton, M. D., van Kan, J. A. L., & Faino, L. (2016). Mind the gap; seven reasons to close fragmented genome assemblies. *Fungal Genetics and Biology*, *90*, 24–30. https://doi.org/10.1016/j.fgb.2015.08.010

Tilak, M.-K., Botero-Castro, F., Galtier, N., & Nabholz, B. (2018). Illumina library preparation for sequencing the GC-rich fraction of heterogeneous genomic DNA. *Genome Biology and Evolution*, https://doi.org/10.1093/gbe/evy022, evy022-evy022

Warren, W. C., Clayton, D. F., Ellegren, H., Arnold, A. P., Hillier, L. W., Künstner, A., … Wilson, R. K. (2010). The genome of a songbird. *Nature*, *464*, 757–762. https://doi.org/10.1038/nature08819

Warren, W. C., Hillier, L. W., Tomlinson, C., Minx, P., Kremitzki, M., Graves, T., … Cheng, H. H. (2017). A new chicken genome assembly provides insight into avian genome structure. *G3: Genes|Genomes|Genetics*, *1*, 109–117. https://doi.org/10.1534/g3.116.035923

Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M., & Jaffe, D. B. (2017). Direct determination of diploid genome sequences. *Genome Research*, *27*, 757–767. https://doi.org/10.1101/gr.214874.116

Weissensteiner, M. H., Pang, A. W. C., Bunikis, I., Höijer, I., Vinnere-Petterson, O., Suh, A., & Wolf, J. B. W. (2017). Combination of short-read, long-read and optical mapping assemblies reveals large-scale tandem repeat arrays with population genetic implications. *Genome Research*, *27*, 697–708. https://doi.org/10.1101/gr.215095.116

Wolf, J. B. W., & Ellegren, H. (2017). Making sense of genomic islands of differentiation in light of speciation. *Nature Reviews Genetics*, *18*, 87–100. https://doi.org/10.1038/nrg.2016.133

Wu, C., Twort, V. G., Crowhurst, R. N., Newcomb, R. D., & Buckley, T. R. (2017). Assembling large genomes: Analysis of the stick insect (*Clitarchus hookeri*) genome reveals a high repeat content and sex-biased genes associated with reproduction. *BMC Genomics*, *18*, 884. https://doi.org/10.1186/s12864-017-4245-x

Yandell, M., & Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics*, *13*, 329–342. https://doi.org/10.1038/nrg3174

Zhang, G., Li, C., Li, Q., Li, B., Larkin, D. M., Lee, C., … Wang, J. (2014). Comparative genomics reveals insights into avian genome evolution and adaptation. *Science*, *346*, 1311–1320. https://doi.org/10.1126/science.1251385

Zhou, Q., Ellison, C. E., Kaiser, V. B., Alekseyenko, A. A., Gorchakov, A. A., & Bachtrog, D. (2013). The epigenome of evolving *Drosophila* neo-sex chromosomes: Dosage compensation and heterochromatin formation. *PLoS Biology*, *11*, e1001711. https://doi.org/10.1371/journal.pbio.1001711

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Peona V, Weissensteiner MH, Suh A. How complete are “complete” genome assemblies?—An avian perspective. *Mol Ecol Resour*. 2018;18:1188–1195. https://doi.org/10.1111/1755-0998.12933