

But where did the centromeres go in the chicken genome models?

Benoît Piégu · Peter Arensburger · Florian Guillou · Yves Bigot

Received: 17 July 2018 / Revised: 31 August 2018 / Accepted: 3 September 2018 / Published online: 17 September 2018
© Springer Nature B.V. 2018

Abstract The chicken genome was the third vertebrate to be sequenced. To date, its sequence and feature annotations are used as the reference for avian models in genome sequencing projects developed on birds and other *Sauropsida* species, and in genetic studies of domesticated birds of economic and evolutionary biology interest. Therefore, an accurate description of this genome model is important to a wide number of scientists. Here, we review the location and features of a very basic element, the centromeres of chromosomes in the galGal5 genome model. Centromeres are elements that are not determined by their DNA sequence but by their epigenetic status, in particular by the accumulation of the histone-like protein CENP-A. Comparison of data from several public sources (primarily marker probes flanking centromeres using fluorescent in situ hybridization done on giant lampbrush chromosomes and CENP-A ChIP-seq datasets) with galGal5 annotations revealed that centromeres are likely inappropriately mapped in 9 of the 16 galGal5 chromosome models in

which they are described. Analysis of karyology data confirmed that the location of the main CENP-A peaks in chromosomes is the best means of locating the centromeres in 25 galGal5 chromosome models, the majority of which (16) are fully sequenced and assembled. This data re-analysis reaffirms that several sources of information should be examined to produce accurate genome annotations, particularly for basic structures such as centromeres that are epigenetically determined.

Keywords Centromere · Bioinformatics · Genome · Repeats · C-value

Abbreviations

CENP-A	centromere protein A
ChIP	chromatine immuno-precipitation
DNA	deoxyribonucleic acid
G	Giga
bp	base pairs
seq	sequencing

Responsible Editor: Beth A. Sullivan

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s10577-018-9585-0>) contains supplementary material, which is available to authorized users.

B. Piégu · F. Guillou · Y. Bigot (✉)
PRC, UMR INRA0085, CNRS 7247, Centre INRA Val de Loire,
37380 Nouzilly, France
e-mail: yves.bigot@inra.fr

P. Arensburger
Biological Sciences Department, California State Polytechnic
University, Pomona, CA 91768, USA

Introduction

During the last two decades, an increasing number of viral, prokaryotic, and eukaryotic genomes have been released to public databases thanks to progress in sequencing and analysis technologies (Sedlazeck et al. 2018). Second-generation sequencing technologies allow high-throughput production of small reads (50 to 400 nucleotides long) of excellent quality, while third-

generation sequencing technologies produce long sequencing reads from single DNA molecules. Furthermore, modern genome assemblers can integrate information from several sources. The constant decrease in costs of per base pair DNA sequencing that accompanies these technological innovations widens the range of questions that can be addressed in genomic studies.

At the same time, as these new research opportunities are becoming available, it is important to remember that these new studies rely on genome sequences and annotations that are just models. These models depend on the quality of the DNA sequences, the mapping technology, the automated genome assembly and on annotation pipelines, both automated and manual. The quality of these genome models is mainly evaluated from the point of view of DNA sequencing and assembly quality. These are evaluated through a series of metrics related to the size of the model genome, the sizes of the contigs and scaffolds, features of misassembled contigs, and how completely functional elements (mostly genes) have been assembled (Gurevich et al. 2013; Khiste and Ilie 2015). The location and sequence content of some genomic regions, such as centromeres, have acquired a reputation as being difficult to sequence and assemble because of their repeat content (Copenhaver 2003; Kapusta and Suh 2017; Kapusta et al. 2017; Aldrup-MacDonald and Sullivan 2014; Khost et al. 2017). Because of this, they have received less attention than other genomic regions, outside of the most studied model organisms such as the baker yeast (*Saccharomyces cerevisiae*), fruit flies (*Drosophila melanogaster*), thale cress (*Arabidopsis thaliana*), mice (*Mus musculus*), and humans (*Homo sapiens*). In consequence, the location and size of centromeres in chromosome models is often unknown and represented by long tracts of Ns.

Centromeres are specialized chromosomal regions that are involved in chromosome segregation during mitosis and meiosis. They are not determined by their DNA sequence but by epigenetic mechanisms that partially involve the deposition of centromere-specific histone H3 variant CENP-A (so-called CENH3) within centromeric nucleosomes and an enrichment in the histone modification H4K20me1 (Hori et al. 2014, 2017). The location of centromeres in a genome model can be identified by the position of CENP-A enrichment peaks on chromosomes using ChIP-seq data obtained with anti-CENP-A antibodies. A complementary approach

is to use in situ hybridization with marker probes surrounding the centromeres (Kretschmer et al. 2018).

Centromere regions can be very variable in size and in sequence between species, but also between chromosomes in each species (for review, see Plohl et al. 2014). There may be a single locus (centric) in each of all chromosomes or centromeres may be all diffuse (holocentric). The latter type arose independently at least 13 times during the evolution of both plants and animals, but the DNA sequences of such centromeres remain poorly described. There are two main types of centric centromeres, repeat-based and repeat-free. The repeat-based centromeres are generally composed of large arrays of tandem repeats (so-called satellite DNA) in which transposable elements are interspersed and have likely accumulated over time. Species can generally be classified via their centromeres as (i) all their chromosomes display repeat-based centromeres (e.g., *H. sapiens* and *D. melanogaster*), (ii) all chromosomes display repeat-free centromeres (e.g., *S. cerevisiae*), or (iii) centromeres may be both repeat-based and repeat-free, depending on the chromosome (e.g., *Equus caballus* and *Solanum tuberosum*).

The chicken genome was the third vertebrate to be sequenced (International Chicken Genome Sequencing Consortium 2004), and to date, avian genomes have been one of the most investigated groups of vertebrates in genome-sequencing projects (Zhang et al. 2014). However, avian genomes remain a technical challenge to sequence and assemble due in part to their high GC content. Recently, significant discrepancies between expected versus assembled genome size in eukaryotes have been reported (Peona et al. 2018). One of the most striking examples is the ostrich genome (*Struthio camelus*) for which the genome model has a size of 1.23 Gbp (Zhang et al. 2015) while its estimated genome size based on more classic methods is 2.16 Gbp (Eden et al. 1978). The chicken genome is organized into 10 macrochromosomes (1 to 9, plus Z) and 29 microchromosomes (10 to 38, plus W) and is reluctant to deliver “all its secrets,” especially those of certain microchromosomes. In the current galGal5 genome model (Warren et al. 2017), 6 microchromosomes (29 and 34 to 38) are not represented. In addition, the GC-rich outer arm ends and subtelomeric regions of macrochromosomes (Federico et al. 2005) were recently found to possibly harbor genes that are absent from the current chicken model (Seroussi et al. 2017; Mello and

Lovell 2018). Centromeres are also poorly described in the different versions of the chicken genome model.

In the previous galGal4 chicken genome model, macrochromosome centromere sizes were arbitrarily assigned as 1,500,000 Ns and those of microchromosomes as 500,000 Ns in the absence of any evidence of their true lengths (International Chicken Genome Sequencing Consortium 2004). In the galGal5 model, centromeres were again arbitrarily assigned a stretch of Ns (500,000 this time) in all 16 chromosomes where they were annotated (see the UCSC genome browser, <https://genome.ucsc.edu/cgi-bin/hgGateway>). Some authors have argued that their absence was due to the difficulty in sequencing and assembling centromeres (Kapusta and Suh 2017; Kapusta et al. 2017). However, this is not always the case since the organization of some chicken centromeres have previously been described, and these were identified as being of at least two types (Shang et al. 2010). The DNA sequence of centromeres in chromosomes 1, 2, 3, 4, 7, 8, and 11 was found to consist of chromosome-specific, tandem repeat arrays that span several hundred kilobases. By contrast, the DNA sequence of centromeres in chromosomes 5, 27, and Z do not contain tandem repeat sequences and span regions of about 30 kb. Therefore, one would expect both large and small centromeres in the chicken genome depending on the chromosome type. Furthermore, small centromeres should, a priori, not present any particular difficulties for sequencing and assembly.

Here, we review current knowledge regarding centromere localization in chicken chromosomes by comparing their features in the galGal5 model with three sources of published information: (i) one dataset of Illumina reads (SRA archive DRR018430) obtained from a ChIP-seq using chromatin of DT40 cells (a chicken line of *bursa lymphoma*) and using anti-CENP-A antibodies for immunoprecipitation (Shang et al. 2013); (ii) sequence markers close to centromeres with a location previously verified by fluorescent in situ hybridization mapping on giant lampbrush chromosomes from growing chicken oocytes (Krasikova et al. 2006, 2012; Zlotina et al. 2010, 2012); (iii) sequences assembled using Illumina reads obtained from a ChIP-seq experiment performed using the chromatin of DT40 cells transfected with a plasmid vector expressing a flag-CENP-A protein and using anti-flag antibodies for immunoprecipitation (Shang et al. 2010).

Results

Location of N-tracts and centromeres in galGal5

We reviewed karyology studies to verify the centromere location in each chicken chromosome, categorizing them as metacentric, submetacentric, acrocentric, subtelocentric, or telometric (Table 1, columns 2 to 5; Fehchheimer 1990). Using a custom-written Perl script, the N-tracts were inventoried in galGal5 chromosomes (Online Resource ESM_1.xlsx) and compared to the annotation of centromeres available on the UCSC website (<http://hgdownload.cse.ucsc.edu/goldenPath/galGal5/database/cytoBandIdeo.txt.gz>). In the assembled chromosomes, centromeres are located by tract of 500,000 Ns, but there are some (Table 1, columns 4 to 6) long N-tracts corresponding to regions that were difficult to sequence and assemble but do not correspond to centromeres (e.g., in chromosome 27 between positions 1,073,340 and 1,173,806 [100,466 bp]).

Location of CENP-A-enriched peaks in galGal5

The locations of the 500,000 N-tracts were first compared to those of CENP-A enrichment peaks in galGal5 calculated as described by Shang et al. (2013) from a ChIP-seq dataset based on DT40 cell chromatin, anti-CENP-A antibodies, and immunoprecipitation. Briefly, the DRR018430 SRA dataset was downloaded (<https://www.ncbi.nlm.nih.gov/sra/?term=DRR018430.sra>), filtered, and aligned with bowtie2 (Langmead and Salzberg 2012) to the galGal5 model. The resulting BAM file was then transformed into a bedgraph file using a window of 10,000 nucleotides, and the presence of peaks was visualized in galGal5 chromosomes using the Integrative Genomics Viewer (IGV; James et al. 2011; Thorvaldsdóttir et al. 2013). Four different types of outcomes were observed: (1) cases where both CENP-A peaks placed the centromere in the same region as N-tracts, that is, chromosomes in which the putative centromere (represented by 500,000 Ns, Fig. 1, boxes in gray) was flanked by the main peaks of enrichment of CENP-A (Fig. 1a–d); (2) chromosomes in which the putative centromere was flanked by peaks of enrichment in CENP-A that were not the main peaks of enrichment, (e.g., chromosome 3 Fig. 1b); (3) chromosomes in which the putative centromere

Table 1 Features of centromeres in galGal5 and location of the most enriched CENP-A regions

Chr.	Chromosome category ^a	Centromeres in galGal5 (N gap)				Most enriched CENP-A regions		
		Size in galGal5	Start	End	Size in bp	Start	End	Size in bp
1	Submetacentric ^b	196,200,000	<i>74,626,887</i>	<i>75,126,886</i>	<i>500,000</i>	<i>74,608,323</i>	<i>75,189,703</i>	<i>581,380</i>
2	Acrocentric	149,560,000	<i>52,321,885</i>	<i>52,821,884</i>	<i>500,000</i>	<i>52,306,100</i>	<i>52,850,700</i>	<i>544,600</i>
3 ^d	Subtelocentric	111,300,000	11,717,421	12,217,420	500,000	2,460,000	2,480,000	20,000
4	Acrocentric	91,280,000	<i>18,856,550</i>	<i>19,356,549</i>	<i>500,000</i>	<i>18,820,000</i>	<i>19,360,000</i>	<i>540,000</i>
5 ^d	Subtelocentric	59,830,000	5,808,101	6,308,100	500,000	3,020,000	3,080,000	60,000
6 ^d	Subtelocentric	35,470,000	1	500,000	500,000	30,640,000	30,660,000	20,000
7	Acrocentric	36,950,000	<i>7,340,971</i>	<i>7,840,970</i>	<i>500,000</i>	<i>7,340,000</i>	<i>7,850,000</i>	<i>510,000</i>
8	Acrocentric	29,960,000	<i>10,505,701</i>	<i>11,005,700</i>	<i>500,000</i>	<i>10,480,000</i>	<i>11,040,000</i>	<i>560,000</i>
9	Telocentric	24,090,000	1	500,000	500,000	n.a.	n.a.	n.a.
10 ^d	Acrocentric	20,440,000	1	500,000	500,000	1,620,000	1,660,000	40,000
11	Acrocentric	20,220,000	<i>2,804,945</i>	<i>3,304,944</i>	<i>500,000</i>	<i>2,800,000</i>	<i>3,330,000</i>	<i>530,000</i>
12 [#]	Subtelocentric	19,950,000	n.a.	n.a.	n.a.	1,070,000	1,090,000	20,000
13	Subtelocentric	18,410,000	0	500,000	500,000	1,650,000	1,680,000	30,000
14 ^c	Subtelocentric	15,600,000	<i>15,034,743</i>	<i>15,534,742</i>	<i>500,000</i>	<i>15,530,000</i>	<i>15,590,000</i>	<i>60,000</i>
15 ^d	Subtelocentric	12,791,000	n.a.	n.a.	n.a.	12,730,000	12,760,000	30,000
16 ^d	Subtelocentric	652,338	n.a.	n.a.	n.a.	1	20,000	20,000
17	Punctiform	10,960,000	1	500,000	500,000	1	600,000	600,000
18 ^d	Punctiform	11,050,000	n.a.	n.a.	n.a.	1	20,000	20,000
19 ^d	Punctiform	9,980,000	n.a.	n.a.	n.a.	50,000	60,000	10,000
20 ^d	Punctiform	14,110,000	n.a.	n.a.	n.a.	5,630,000	5,700,000	70,000
21 ^d	Punctiform	6,860,000	n.a.	n.a.	n.a.	6,830,000	6,860,000	30,000
22	Punctiform	4,730,000	3,007,279	3,507,278	500,000	2,990,000	3,007,000	20,000
23	Punctiform	5,790,000	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
24	Punctiform	6,280,000	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
25	Punctiform	2,910,000	277,762	777,761	500,000	n.a.	n.a.	n.a.
26 ^d	Punctiform	5,310,000	n.a.	n.a.	n.a.	5,290,000	5,300,000	10,000
27 ^d	Punctiform	5,660,000	1,073,341	1,173,806	500,000	1	30,000	30,000
28 ^d	Punctiform	4,970,000	n.a.	n.a.	n.a.	4,960,000	4,970,000	10,000
30	Punctiform	24,927	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
31	Punctiform	49,161	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
32	Punctiform	78,254	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
33	Punctiform	1,650,000	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
W	Telocentric	5,160,000	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Z ^d	Metacentric	82,310,000	n.a.	n.a.	n.a.	42,760,000	42,800,000	40,000

n.a. not available

^a Chromosomes with similar centromere locations with both sources of information are italicized. Categorization of chromosomes according to the relative length of their arms was done according to Krasikova et al. (2006, 2012); Zlotina et al. (2010, 2012); and Miller et al. (2014)

^b Chromosome 1 could be considered submetacentric or acrocentric in chicken breeds and lines (Bitgood et al. 1982) due to the inversion of a large region

^c Both annotations were juxtaposed, indicating that the centromere in chromosome 14 localized between positions 15,034,742 and 15,590,000

^d Chromosome displaying no N gap (> 100) in their centromere located with CENP-A peaks, taking into account centromere positions indicated in Table 1 and the location of N-tracts in Centromer-ESM.xlsx

was not flanked by peaks of enrichment in CENP-A and in which a strong CENP-A peak was localized elsewhere (e.g., chromosomes 5 and 27, Fig. 1d, e); (4) chromosomes in which the location of the centromere was not indicated in the galGal5 annotation but positioned using ChIP-seq data in the galGal5 model (e.g., chromosomes 28 and Z, Fig. 1f, g). A summary of centromere locations identified using ChIP-seq data is shown in Table 1, columns 6 to 8. It revealed that the centromere location matched both information sources for only 7 (1, 2, 4, 7, 8, 11, and 14) of the 16 chromosomes that were annotated with a centromere in galGal5. The centromere locations in chromosome models 3, 5, 6, 10, 13, 25, and 27 were found to be different from those indicated in the galGal5 annotation and did not support the presence of a sequenced centromere in chromosome models 9 and 25, as annotated in galGal5 (Table 1). The probable reason for this was that there was no CENP-A-enriched region on chromosome models such as 9 and 25, the pericentromeric region of which could not be assembled. However, centromeres, or centromeric sequences, could be positioned in chromosome models 12, 15, 16, 18, 19, 20, 21, 26, 28, and Z. For chromosome 16, which contains several types of repeats (genes and satellite DNA), previous studies have shown that this centromere is subtelomeric (Miller et al. 2014). The near-telomeric location of the CENP-A peak in chromosome model 16 was due to the absence of the p-arm that mainly consisted of AT-rich repeats.

Confirmation of centromere locations using data from karyological markers and from prior centromeric sequences

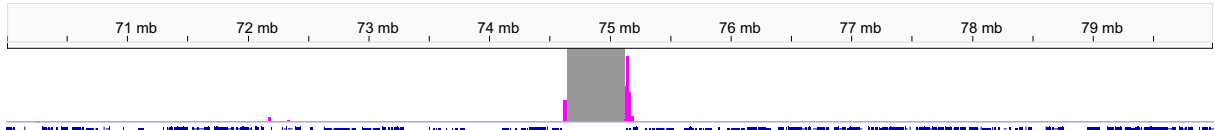
The sequence of marker probes flanking centromeres in chromosomes 1, 2, and 3 (Zlotina et al. 2012) were used to verify their location in these three chromosome models. We found that markers WAG43N11 and WAG53E23 were located at positions 65,721,251 and 76,397,565 in chromosome 1 and markers WAG21J8 and WAG18G1 at positions 50,277,689 and 53,029,789 in chromosome 2 and surrounded the centromere in each chromosome (Table 1). In chromosome 3, markers WAG35O13 and WAG44P17 were located at positions 2,133,568 and 5,508,192. They did not flank the centromere described in the galGal5 annotation but flanked the CENP-A peak (Fig. 1b). This confirmed that the largest CENP-A peak was the centromere in

chromosome 3, and those detected in all other chromosomes were very likely reliable. The CENP-A peaks flanking the chromosome 3 centromere described in the galGal5 annotation might correspond to the presence of a neocentromere. Neocentromeres are present in all eukaryote taxa and correspond to atypical centromeres spontaneously bound by CENP-A and able to form on unique sequence regions (Scott and Sullivan 2014). They are able to act as centromeres when the main centromere is lost by deletion (Shang et al. 2013).

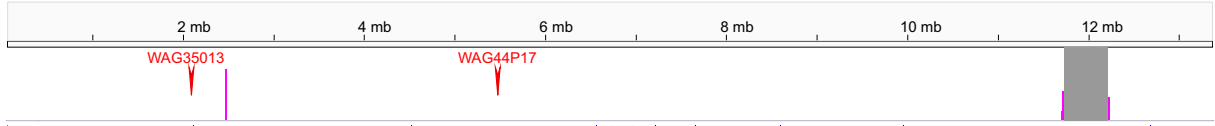
Finally, centromeric sequences (NCBI accessions AB556643 to AB556736) corresponding to major families of tandem repeats in the chicken genome, which were previously reconstructed from a ChIP-seq Illumina dataset obtained with a different antibody (Shang et al. 2010), were used to search chromosome sequences. Among the 7 centromeres located in the same region by the galGal5 annotation and the CENP-A peaks, sequence AB556722 was found to match with 6 repeats within the centromere of chromosome 1 (positions 74,615,536 to 75,136,859), AB556723 with 12 repeats within the centromere of chromosome 2 (positions 52,315,814 to 52,854,454), AB556725 with 15 repeats within the centromere of chromosome 4 (positions 18,841,161 to 18,852,575), AB556726 with 12 repeats within the centromere of chromosome 7 (positions 7,330,539 to 7,848,612), AB556727 with 10 repeats within the centromere of chromosome 8 (positions 10,498,281 to 11,007,717), and AB556728 with 7 repeats within the centromere of chromosome 11 (positions 3,305,308 to 3,323,207). In chromosome 8, AB556727 repeats were found interspersed between positions 11,308,276 to 11,379,802 with 8 copies of AF124927 sequences that belong to another family of partially inverted tandem repeats (PIR; Wang et al. 2002). This confirmed the specific centromeric origin of these sequences and their accuracy for locating centromeres.

We found that sequence AB556724 matched with six repeats between positions 2,464,414 to 2,476,273 in chromosome 3, AB556729 with two repeats between positions 3,032,478 to 3,061,773 in chromosome 5, AB556655.1 with one repeat between positions 30,655,666 to 30,656,292 in chromosome 6, and AB556731 with five repeats between positions 42,746,469 to 42,775,985 in chromosome Z. In these four macrochromosomes, the matches were located within the CENP-A peak, which is outside the

a. chromosome 1



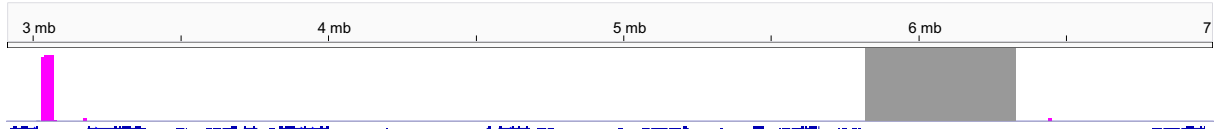
b. chromosome 3



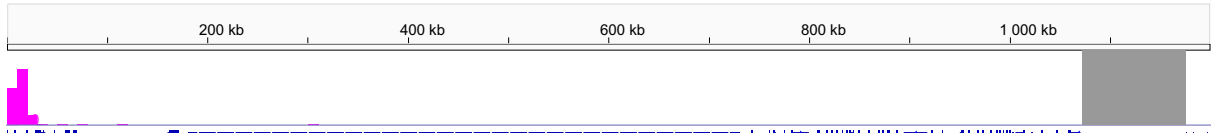
c. chromosome 4



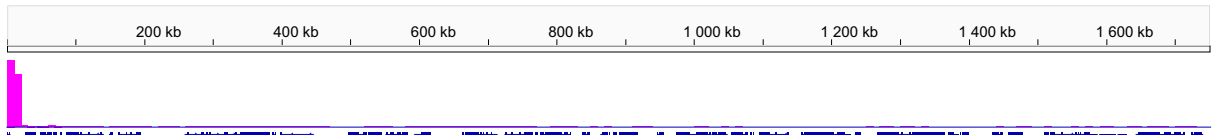
d. chromosome 5



e. chromosome 27



f. chromosome 28



g. chromosome Z



Fig. 1 Graphic representation, using the IGV, of inner regions of chromosome models 1 (a), 3 (b), 4 (c), 5 (d), 27 (e), and 28 (f), and of the complete model of chromosome Z (g). Centromeres corresponding to N-tracts in the official annotation of galGal5 are indicated with gray boxes. Peaks of CENP-A enrichment are indicated by pink bars. Markers (WAG35013 and WAG44P17;

Zlotina et al. 2012) used for fluorescent hybridization to locate the centromere in chromosome 3 are indicated with their names in red. Below each graphic, gene-containing regions are indicated in blue. Above each graphic, the scale of the region is indicated. In a, c, e, and g, small CENP-A peaks are shown reflecting the putative presence of neocentromeres (Shang et al. 2013)

centromeres described in the galGal5 annotation. For macrochromosome 9, no information besides the galGal5 annotation confirmed the location of a centromere in its sequence. Matches with sequences of centromeric origin were found within regions containing CENP-A peaks of microchromosomes 10, 11, 12, 16,

18, 20, 21, 27, and 28 (Table 2). No match was found between the X51431 sequence, which has been described as a 41 or 42 bp tandemly repeated sequence monomer, and centromeres of microchromosomes and macrochromosomes 7 and 8, where 41 and 42 bp repeat arrays were expected (Matzke et al. 1990).

Table 2 Features of complete or partial matches between centromeric sequences and some microchromosomes

Chromosome	Sequence Acc N°	Number of hits	Regions containing hits
10 ^d	AB556622.1	2	1,629,560 to 1,657,058
∞ ^d	AB556731.1 ^a	2	1,628,173 to 1,657,395
11	AB556728.1	7	3,305,308 to 3,323,207
12 ^d	AB556730.1 ^b	2	1,087,220 to 1,093,474
16 ^d	AB556462.1 and AB556516.1 ^c	6	5451 to 10,590
18 ^d	AB556716.1	1	8910 to 9073
20 ^d	AB556682.1	2	5,681,836 to 5,682,763
21 ^d	AB556456.1	> 20	6,834,279 to 6,860,061
∞ ^d	AB556671.1	> 20	6,838,826 to 6,859,100
∞ ^d	AB556439.1	> 20	6,838,465 to 6,852,276
27 ^d	AB556730.1 ^b	14	447 to 68,967
28 ^d	AB556462.1 and AB556516.1 ^c	> 20	4,968,416 to 4,973,573

^a Sequence displaying matches in centromeres of chromosomes 10 and Z

^b Sequence displaying matches in centromeres of chromosomes 12 and 27

^c Sequence displaying similarities and matching with some centromeric sequences of chromosomes 16 and 28

^d Chromosome displaying no N gap (> 100 bp) in their centromere located with CENP-A peaks; this takes into account centromere positions indicated in Table 1 and the location of N-tracts in Centromere-ESM.xls

Concluding remarks

The comparison of existing public datasets to the galGal5 annotation confirmed that CENP-A ChIP-seq datasets were a reliable tool to localize centromeres in the sequence of chromosome models, including chickens. In this study, we were able to localize centromere positions in 25 chromosomes (1 to 8, 10 to 22, 26 to 28, and Z). Contrary to popular belief, parts of these centromeres have successfully been sequenced and assembled in the galGal5 model and could be used for investigating synteny with other avian species. Our results also supported that there were at least three kinds of centromeres in the chicken genome: (i) centromeres consisting of chromosome-specific homologous tandem repetitive arrays that span over several hundred kilobases in chromosomes 1, 2, 4, 7, 8, 11, and 17; (ii) centromeres that do not contain tandem repetitive sequences and which span over regions of about 10–70 kb in chromosomes 3, 5, 6, 10, 12, 16, 18, 20, 21, 27, 28, and Z; and (iii) centromeres that contain several kinds of tandem repetitive sequences and span over regions of approximately 10–70 kb likely in chromosomes 13, 14, 15, 17, 19, 21, and 22. Although, the centromere status for all of these chromosomes is now established, these results should be taken cautiously, particularly for centromeres that do not contain tandem repetitive

sequences. Indeed, the lack of tandem repeats might result from an artifact, so-called muted gaps that can arise during the assembly process (Chaisson et al. 2015; Thomma et al. 2016). When tandem repeats are nearly identical or perfectly conserved in sequence, their assembly can transform the array into a unique copy. Such artifacts should be identified by performing copy number analyses using datasets of Illumina genomic resequencing (Abysov et al. 2011).

Our results have also highlighted two issues with the current galGal5 assembly. The first was the absence of centromeres in chromosome models 9, 23 to 25, 30 to 33, and W. It has commonly been assumed that this was due to sequencing and assembly difficulties and would likely be resolved in future genome models. The second was that the content of regions describing the centromeres in the galGal5 annotation was found to be inaccurate; these were the centromeres of chromosomes 3, 5, 6, 9, 10, 13, and 22. The reasons for these discrepancies may be due to two issues. First, they may have resulted from artifacts during the genome assembly step. Second, they were likely considered by the scientists in charge of the chicken genome project as large regions (~ 500,000 nucleotides) that were difficult to sequence and assemble. Such sequence characteristics are generally correlated with regions displaying an elevated GC content (Aird et al. 2011; Nakamura et al. 2011; Benjamini and Speed 2012;

Dabney and Meyer 2012; Oyola et al. 2012; Ross et al. 2013 Interestingly, highly GC-rich regions were found in the inner regions of some macrochromosome arms (Andreozzi et al. 2001, Federico et al. 2005, Costantini et al. 2007), but they are not present in the current sequence of macrochromosome models. Therefore, regions associated with GC-rich subtelomeric regions of macrochromosomes and the missing microchromosome models 34, 35, 36, 37, and 38 might be the hidden harbors of at least a part of the 1500 lost genes in the chicken genome that are located in GC-rich regions in other vertebrate genomes (Mello and Lovell 2018).

It is important to have reliable centromere annotations for a genome model such as the chicken, which is a reference in avian genomics and evolutionary studies. But it is also very important to understand the dynamic and the plasticity of avian genomes. Knowing the true location of centromeres should allow researchers to verify whether they are the seat of repeat expansions by retrotransposition during cellular differentiation, similar to those observed in mammalian genomes (Bersani et al. 2015; Tanne et al. 2015), and to verify whether there was a source of retrotransposition in a genome where all transposable elements were believed to be extinct (Guizard et al. 2016). Knowledge of centromere positions is also important for speciation and population research, for example when studying the effect they may have via the “centromere drive” and linked selection due to low meiotic recombination rates (Henikoff et al. 2001; Weissensteiner et al. 2017; Wolf and Ellegren 2017). It should therefore be important for the scientific community to re-examine the gold standard procedures for annotating centromeres and putative neocentromeres in chromosome models. For this, the existing literature in karyology and CENP-A ChIP-seq data might be used more systematically to verify their location.

Acknowledgements Peter Arensburger holds a senior researcher fellowship from the STUDIUM.

Author contribution YB and BP conceived the study and analyzed data. YB, PA and FG wrote the paper.

Funding information This work was funded by the Project Région Centre AviGeS, the C.N.R.S., the I.N.R.A., the Groupement de Recherche CNRS 2157, and the Ministère de l'Éducation Nationale, de la Recherche et de la Technologie.

References

- Abyzov A, Urban AE, Snyder M, Gerstien M (2011) CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Gen Res* 21:974–984. <https://doi.org/10.1101/gr.114876.110>
- Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol* 12:R18. <https://doi.org/10.1186/gb-2011-12-2-r18>
- Aldrup-Macdonald ME, Sullivan BA (2014) The past, present, and future of human centromere genomics. *Genes (Basel)* 5: 33–50
- Andreozzi L, Federico C, Motta S, Saccone S, Sazanov AL, Sazanov AA, Smirnov AF, Galkina SA, Lukina NA, Rodionov AV, Carels N, Bernardi G (2001) Compositional mapping of chicken chromosomes and identification of the gene-richest regions. *Chromosom Res* 9:521–532
- Benjamini Y, Speed TP (2012) Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res* 40(10):e72. <https://doi.org/10.1093/nar/gks001>
- Bersani F, Lee E, Kharchenko PV, Xu AW, Liu M, Xega K, MacKenzie OC, Brannigan BW, Wittner BS, Jung H, Ramaswamy S, Park PJ, Maheswaran S, Ting DT, Haber DA (2015) Pericentromeric satellite repeat expansions through RNA-derived DNA intermediates in cancer. *Proc Natl Acad Sci U S A* 112:15148–15153. <https://doi.org/10.1073/pnas.1518008112>
- Bitgood JJ, Shoffner RN, Otis JS, Wang N (1982) Recombinant inversion chromosomes in phenotypically normal chickens. *Science* 215:409–411
- Chaisson MJP, Wilson RK, Eichler EE (2015) Genetic variation and the de novo assembly of human genomes. *Nat Rev Genet* 16:627–640. <https://doi.org/10.1038/nrg3933>
- Copenhaver GP (2003) Using Arabidopsis to understand centromere function: progress and prospects. *Chromosom Res* 11: 255–262
- Costantini M, Di Filippo M, Auletta F, Bernardi G (2007) Isochore pattern and gene distribution in the chicken genome. *Gene* 400:9–15. <https://doi.org/10.1016/j.gene.2007.05.025>
- Dabney J, Meyer M (2012) Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *BioTechniques* 52:87–94. <https://doi.org/10.2144/000113809>
- Eden FC, Hendrick JP, Gottlieb SS (1978) Homology of single copy and repeated sequences in chicken, duck, Japanese quail, and ostrich DNA. *Biochemistry* 17:5113–5121
- Fechheimer NS (1990) Chromosomes of chickens. *Adv Vet Sci Comp Med* 34:169–207
- Federico C, Cantarella CD, Scavo C, Saccone S, Bed'Hom B, Bernardi G (2005) Avian genomes: different karyotypes but a similar distribution of the GC-richest chromosome regions at interphase. *Chromosom Res* 13:785–793. <https://doi.org/10.1007/s10577-005-1012-7>
- Guizard S, Piégu B, Arensburger P, Guillou F, Bigot Y (2016) Deep landscape update of dispersed and tandem repeats in the genome model of the red jungle fowl, *Gallus gallus*, using

- a series of de novo investigating tools. *BMC Genomics* 17: 659. <https://doi.org/10.1186/s12864-016-3015-5>
- Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUASt: quality assessment tool for genome assemblies. *Bioinformatics* 29:1072–1075. <https://doi.org/10.1093/bioinformatics/btt086>
- Henikoff S, Ahmad K, Malik HS (2001) The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* 293: 1098–1102. <https://doi.org/10.1126/science.1062939>
- Hori T, Kagawa N, Toyoda A, Fujiyama A, Misu S, Monma N, Makino F, Ikeo K, Fukagawa T (2017) Constitutive centromere-associated network controls centromere drift in vertebrate cells. *J Cell Biol* 216:101–113. <https://doi.org/10.1083/jcb.201605001>
- Hori T, Shang WH, Toyoda A, Misu S, Monma N, Ikeo K, Molina O, Vargiu G, Fujiyama A, Kimura H, Earnshaw WC, Fukagawa T (2014) Histone H4 Lys 20 monomethylation of the CENP-A nucleosome is essential for kinetochore assembly. *Dev Cell* 29:740–149. <https://doi.org/10.1016/j.devcel.2014.05.001>
- International Chicken Genome Sequencing Consortium (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432:695–716. <https://doi.org/10.1038/nature03154>
- Kapusta A, Suh A (2017) Evolution of bird genomes—a transposon’s-eye view. *Ann N Y Acad Sci* 1389:164–185. <https://doi.org/10.1111/nyas.13295>
- Kapusta A, Suh A, Feschotte C (2017) Dynamics of genome size evolution in birds and mammals. *Proc Natl Acad Sci U S A* 114:E1460–E1469. <https://doi.org/10.1073/pnas.1616702114>
- Khiste N, Ilie L (2015) LASER: large genome ASsembly Evaluator. *BMC Res Notes* 8:709. <https://doi.org/10.1186/s13104-015-1682-y>
- Khost DE, Eickbush DG, Larracuente AM (2017) Single-molecule sequencing resolves the detailed structure of complex satellite DNA loci in *Drosophila melanogaster*. *Genome Res* 27:709–721. <https://doi.org/10.1101/gr.213512.116>
- Krasikova A, Deryusheva S, Galkina S, Kurganova A, Evteev A, Gaginskaya E (2006) On the positions of centromeres in chicken lampbrush chromosomes. *Chromosom Res* 14: 777–789. <https://doi.org/10.1007/s10577-006-1085-y>
- Krasikova A, Fukagawa T, Zlotina A (2012) High-resolution mapping and transcriptional activity analysis of chicken centromere sequences on giant lampbrush chromosomes. *Chromosom Res* 20:995–1008. <https://doi.org/10.1007/s10577-012-9321-0>
- Kretschmer R, Ferguson-Smith MA, de Oliveira EHC (2018) Karyotype evolution in birds: from conventional staining to chromosome painting. *Genes (Basel)* 9. doi:<https://doi.org/10.3390/genes9040181>
- Langmead B, Salzberg SL (2012) Fast gapped-read alignment with bowtie 2. *Nat Methods* 9:357–359. <https://doi.org/10.1038/nmeth.1923>
- Matzke MA, Varga F, Berger H, Schemthaner J, Schweizer D, Mayr B, Matzke AJ (1990) A 41–42 bp tandemly repeated sequence isolated from nuclear envelopes of chicken erythrocytes is located predominantly on microchromosomes. *Chromosoma* 99:131–137
- Mello CV, Lovell PV (2018) Avian genomics lends insights into endocrine function in birds. *Gen Comp Endocrinol* 256:123–129. <https://doi.org/10.1016/j.ygcen.2017.05.023>
- Miller MM, Robinson CM, Abernathy J, Goto RM, Hamilton MK, Zhou H, Delany ME (2014) Mapping genes to chicken microchromosome 16 and discovery of olfactory and scavenger receptor genes near the major histocompatibility complex. *J Hered* 105:203–215. <https://doi.org/10.1093/jhered/est091>
- Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, Ishikawa S, Linak MC, Hirai A, Takahashi H, Altaf-Ul-Amin M, Ogasawara N, Kanaya S (2011) Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res* 39:e90. <https://doi.org/10.1093/nar/gkr344>
- Oyola SO, Otto TD, Gu Y, Maslen G, Manske M, Campino S, Turner DJ, Macinnis B, Kwiatkowski DP, Swerdlow HP, Quail MA (2012) Optimizing Illumina next-generation sequencing library preparation for extremely AT-biased genomes. *BMC Genomics* 13:1. <https://doi.org/10.1186/1471-2164-13-1>
- Peona V, Weissensteiner MH, Suh A (2018) How complete are ‘complete’ genome assemblies?—an avian perspective. *Mol Ecol Resour*. <https://doi.org/10.1111/1755-0998.12933>
- Plohl M, Meštrović N, Mravinac B (2014) Centromere identity from the DNA point of view. *Chromosoma* 123:313–325. <https://doi.org/10.1007/s00412-014-0462-0>
- Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, Nusbaum C, Jaffe DB (2013) Characterizing and measuring bias in sequence data. *Genome Biol* 14:R51. <https://doi.org/10.1186/gb-2013-14-5-r51>
- Scott KC, Sullivan BA (2014) Neocentromeres: a place for everything and everything in its place. *Trends Genet* 30:66–74. <https://doi.org/10.1016/j.tig.2013.11.003>
- Sedlazeck FJ, Lee H, Darby CA, Schatz MC (2018) Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat Rev Genet* 19:329–346. <https://doi.org/10.1038/s41576-018-0003-4>
- Seroussi E, Pitel F, Leroux S, Morisson M, Bormelöv S, Miyara S, Yosefi S, Cogburn LA, Burt DW, Anderson L, Friedman-Einat M (2017) Mapping of leptin and its syntenic genes to chicken chromosome 1p. *BMC Genet* 18:77. <https://doi.org/10.1186/s12863-017-0587-2>
- Shang WH, Hori T, Martins NM, Toyoda A, Misu S, Monma N, Hiratani I, Maeshima K, Ikeo K, Fujiyama A, Kimura H, Earnshaw WC, Fukagawa T (2013) Chromosome engineering allows the efficient isolation of vertebrate neocentromeres. *Dev Cell* 24:635–648. <https://doi.org/10.1016/j.devcel.2013.02.009>
- Shang WH, Hori T, Toyoda A, Kato J, Popendorf K, Sakakibara Y, Fujiyama A, Fukagawa T (2010) Chickens possess centromeres with both extended tandem repeats and short non-tandem-repetitive sequences. *Genome Res* 20:1219–1228. <https://doi.org/10.1101/gr.106245.110>
- Tanne A, Muniz LR, Puzio-Kuter A, Leonova KI, Gudkov AV, Ting DT, Monasson R, Cocco S, Levine AJ, Bhardwaj N, Greenbaum BD (2015) Distinguishing the immunostimulatory properties of noncoding RNAs expressed in cancer cells. *Proc Natl Acad Sci U S A* 112:15154–15159. <https://doi.org/10.1073/pnas.1517584112>

- Thomma BPHJ, Seidl MF, Shi-Kunne X, Cook DE, Bolton MD, van Kan JAL, Faino L (2016) Mind the gap; seven reasons to close fragmented genome assemblies. *Fungal Genet Biol* 90: 24–30. <https://doi.org/10.1016/j.fgb.2015.08.010>
- Thorvaldsdóttir H, Robinson JT, Mesirov JP (2013) Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* 14:178–192. <https://doi.org/10.1093/bib/bbs01>
- Wang X, Li J, Leung FC (2002) Partially inverted tandem repeat isolated from pericentric region of chicken chromosome 8. *Chromosom Res* 10:73–82
- Warren WC, Hillier LW, Tomlinson C, Minx P, Kremitzki M, Graves T, Markovic C, Bouk N, Pruitt KD, Thibaud-Nissen F et al (2017) A new chicken genome assembly provides insight into avian genome structure. *G3* 1:109–117. <https://doi.org/10.1534/g3.116.035923>
- Weissensteiner MH, Pang AWC, Bunikis I, Höijer I, Vinnere-Petterson O, Suh A, Wolf JBW (2017) Combination of short-read, long-read and optical mapping assemblies reveals large-scale tandem repeat arrays with population genetic implications. *Genome Res* 27:697–708. <https://doi.org/10.1101/gr.215095.116>
- Wolf JBW, Ellegren H (2017) Making sense of genomic islands of differentiation in light of speciation. *Nat Rev Genet* 18:87–100. <https://doi.org/10.1038/nrg.2016.133>
- Zhang G, Li C, Li Q, Li B, Larkin DM, Lee C, Storz JF, Antunes A, Greenwold MJ, Meredith RW et al (2014) Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* 346:1311–1320. <https://doi.org/10.1126/science.1251385>
- Zhang J, Li C, Zhou Q, Zhang G (2015) Improving the ostrich genome assembly using optical mapping data. *Gigascience* 4: 24. <https://doi.org/10.1186/s13742-015-0062-9>
- Zlotina A, Galkina S, Krasikova A, Crooijmans RP, Groenen MA, Gaginskaya E, Deryusheva S (2010) Precise centromere positioning on chicken chromosome 3. *Cytogenet Genome Res* 129:310–313. <https://doi.org/10.1159/000314923>
- Zlotina A, Galkina S, Krasikova A, Crooijmans RP, Groenen MA, Gaginskaya E, Deryusheva S (2012) Centromere positions in chicken and Japanese quail chromosomes: de novo centromere formation versus pericentric inversions. *Chromosom Res* 20:1017–1032. <https://doi.org/10.1007/s10577-012-9319-7>