

# Kent Academic Repository

## Full text document (pdf)

### Citation for published version

Damas, J. and O'Connor, Rebecca and Farré, Marta and Lenis, V.P.E. and Martell, H.J. and Mandawala, A. and Fowler, K. and Joseph, S. and Swain, M.T. and Griffin, Darren K. and Larkin, D.M. (2016) Upgrading short read animal genome assemblies to chromosome level using comparative genomics and a universal probe set. *Genome Research*, 27 . pp. 875-884. ISSN 1088-9051.

### DOI

<https://doi.org/10.1101/gr.213660.116>

### Link to record in KAR

<http://kar.kent.ac.uk/60239/>

### Document Version

Author's Accepted Manuscript

#### Copyright & reuse

Content in the Kent Academic Repository is made available for research purposes. Unless otherwise stated all content is protected by copyright and in the absence of an open licence (eg Creative Commons), permissions for further reuse of content should be sought from the publisher, author or other copyright holder.

#### Versions of research

The version in the Kent Academic Repository may differ from the final published version.

Users are advised to check <http://kar.kent.ac.uk> for the status of the paper. **Users should always cite the published version of record.**

#### Enquiries

For any further enquiries regarding the licence status of this document, please contact:

[researchsupport@kent.ac.uk](mailto:researchsupport@kent.ac.uk)

If you believe this document infringes copyright then please contact the KAR admin team with the take-down information provided at <http://kar.kent.ac.uk/contact.html>

1 **Upgrading short read animal genome assemblies to chromosome level using**  
2 **comparative genomics and a universal probe set**

3

4 Joana Damas<sup>1\*</sup>, Rebecca O'Connor<sup>2\*</sup>, Marta Farré<sup>1</sup>, Vasileios Panagiotis E. Lenis<sup>3</sup>, Henry J.  
5 Martell<sup>2</sup>, Anjali Mandawala<sup>2,4</sup>, Katie Fowler<sup>4</sup>, Sunitha Joseph<sup>2</sup>, Martin T. Swain<sup>3</sup>, Darren K.  
6 Griffin<sup>2\*\*</sup>, Denis M. Larkin<sup>1\*\*</sup>

7

8 <sup>1</sup>Department of Comparative Biomedical Sciences, Royal Veterinary College, Royal College  
9 Street, University of London, London, NW1 0TU, UK

10 <sup>2</sup>School of Biosciences, University of Kent, Canterbury, CT2 7NY, UK

11 <sup>3</sup>Institute of Biological, Environmental and Rural Sciences, Aberystwyth University,  
12 Aberystwyth, SY23 3DA, UK

13 <sup>4</sup>School of Human and Life Sciences, Canterbury Christ Church University, Canterbury, CT1  
14 1QU, UK

15

16

17 \*Joint first authors

18 \*\*Joint last authors

19

20 <sup>§</sup>Corresponding author: Denis M. Larkin, Department of Comparative Biomedical Sciences,  
21 Royal Veterinary College, Royal College Street, NW1 0TU, London, UK. Phone number:  
22 +44-207-121-1906. E-mail: [dlarkin@rvc.ac.uk](mailto:dlarkin@rvc.ac.uk)

23

24 Running title: Building chromosome-level assemblies for animals

25

26 Keywords: *de novo* assembly, avian chromosome assembly, evolutionary breakpoint  
27 regions, peregrine falcon, pigeon, chicken

**1 ABSTRACT**

2 Most recent initiatives to sequence and assemble new species' genomes *de-novo* fail to  
3 achieve the ultimate endpoint to produce contigs, each representing one whole  
4 chromosome. Even the best-assembled genomes (using contemporary technologies) consist  
5 of sub-chromosomal sized scaffolds. To circumvent this problem, we developed a novel  
6 approach that combines computational algorithms to merge scaffolds into chromosomal  
7 fragments, PCR-based scaffold verification and physical mapping to chromosomes. Multi-  
8 genome-alignment-guided probe selection led to the development of a set of universal avian  
9 BAC clones that permit rapid anchoring of multiple scaffolds to chromosomes on all avian  
10 genomes. As proof of principle, we assembled genomes of the pigeon (*Columba livia*) and  
11 peregrine falcon (*Falco peregrinus*) to chromosome level comparable, in continuity, to avian  
12 reference genomes. Both species are of interest for breeding, cultural, food and/or  
13 environmental reasons. Pigeon has a typical avian karyotype ( $2n=80$ ) while falcon ( $2n=50$ ) is  
14 highly rearranged compared to the avian ancestor. Using chromosome breakpoint data, we  
15 established that avian interchromosomal breakpoints appear in the regions of low density of  
16 conserved non-coding elements (CNEs) and that the chromosomal fission sites are further  
17 limited to long CNE "deserts". This corresponds with fission being the rarest type of  
18 rearrangement in avian genome evolution. High-throughput multiple hybridization and rapid  
19 capture strategies using the current BAC set provide the basis for assembling numerous  
20 avian (and possibly other reptilian) species while the overall strategy for scaffold assembly  
21 and mapping provides the basis for an approach that (provided metaphases can be  
22 generated) could be applied to any animal genome.

23

## 1 INTRODUCTION

2 The ability to sequence complex animal genomes quickly and inexpensively has initiated  
3 numerous genome projects beyond those of agricultural/medical importance (e.g., Hu et al.  
4 2009; Groenen et al. 2012) and inspired ambitious undertakings to sequence thousands of  
5 species (Zhang et al. 2014a; Koepfli et al. 2015). *De novo* genome assembly efforts  
6 ultimately aim to create a series of contigs, each representing a single chromosome, from p-  
7 to q- terminus (“chromosome-level” assembly). Assembling genomes using next generation  
8 sequencing (NGS) technologies however typically relies on integration of the NGS data with  
9 a pre-existing chromosome-level reference assembly built with previous  
10 sequencing/mapping technologies (Larkin et al. 2012). Indeed, use of short read NGS data  
11 rarely produces assemblies at a similar level of integrity as those provided by traditional  
12 methodologies because of: a) an inability of NGS to generate long error-free contigs or  
13 scaffolds to cover chromosomes completely; and b) a paucity of inexpensive mapping  
14 technologies to upgrade NGS genomes to chromosome level. Even for projects with  
15 sufficient read-depths and long insert libraries, software algorithms at best, produce sub-  
16 chromosomal sized “scaffolds” requiring physical mapping to assemble chromosomes.  
17 Newer technologies such as optical mapping (Teague et al. 2010) including BioNano (Mak et  
18 al. 2016), Dovetail (Putnam et al. 2016), and PacBio long read sequencing (Rhoads and Au  
19 2015) provide a long-term solution to this problem. To date, however, such approaches  
20 suffer from multiple limitations: for instance, BioNano contigs do not extend across multiple  
21 DNA nick site regions, centromeres or large heterochromatin blocks while PacBio  
22 sequencing requires hundreds of micrograms of high molecular weight DNA which is often  
23 not easy to obtain.

24

25 Bioinformatic approaches, e.g., the Reference-Assisted Chromosome Assembly algorithm  
26 (RACA; Kim et al. 2013), were developed to approximate near chromosome-sized fragments  
27 for a *de novo* assembled NGS genome. RACA use requires a genome from the same clade  
28 (e.g., Order for mammals) of the target species being assembled to chromosomes (Kim et

1 al. 2013), sequencing of long-insert libraries and, at best, produces sub-chromosome sized  
2 predicted chromosome fragments (PCFs) that require further verification and subsequent  
3 chromosome assembly. It is worth mentioning that unlike RACA other reference-assisted  
4 assembly algorithms e.g., Ragout (Kolmogorov et al. 2014) or Chromosomer (Tamazian et  
5 al. 2016) do not use the target genome short- and long-range paired read data to verify  
6 synteny breaks in/between scaffolds, meaning that the target species-specific  
7 rearrangements could be missed from the reconstructed PCFs/pseudochromosomes making  
8 the reconstructed target chromosome structures more heavily biased to the reference  
9 genome(s) than when using RACA. RACA algorithm applied to the Tibetan antelope and  
10 blind mole rat genomes significantly improved continuities of these assemblies but they still  
11 contain more than one large PCF for most chromosomes (Kim et al. 2013; Fang et al. 2014).  
12 Therefore, a novel, integrative approach that would allow *de novo* assembled genomes to  
13 retain structures of the target species karyotypes is a necessity.

14

15 A dearth of chromosome-level assemblies for nearly all newly sequenced genomes limits  
16 their use for critical aspects of evolutionary and applied genomics. Chromosome-level  
17 assemblies are essential for species that are regularly bred (e.g., for food or conservation)  
18 because a known order of DNA markers facilitates establishment of phenotype-to-genotype  
19 associations for gene-assisted selection and breeding (Andersson and Georges 2004).  
20 While such assemblies are established for popular livestock species, they are not available  
21 for those species widely used in developing countries (e.g., camels, yaks, buffalo, ostrich,  
22 quail) or species bred for conservation reasons (e.g., falcons). Chromosome-level  
23 information is essential for addressing basic biological questions pertaining to overall  
24 genome (karyotype) evolution and speciation (Lewin et al. 2009). Karyotype differences  
25 between species arise from DNA aberrations in germ cells that were fixed throughout  
26 evolution. These are associated with repetitive sequences used for non-allelic homologous  
27 recombination (NAHR) in evolutionary breakpoint regions (EBRs) where ancestral  
28 chromosomes break and/or combine in descendant species genomes (Murphy et al. 2005).

1 An alternative theory however, suggests that proximity of DNA regions in chromatin is the  
2 main driver of rearrangements and repetitive sequences play a minor role (Branco and  
3 Pombo 2006). Regardless of the mechanism, comparisons of multiple animal genomes  
4 show that between EBRs are evolutionary stable homologous synteny blocks (HSBs). Our  
5 studies in mammals (Larkin et al. 2009) and birds (Farré et al. 2016) suggest that at least the  
6 largest HSBs are maintained non-randomly and are highly enriched for conserved non-  
7 coding elements (CNEs) many of which are gene regulatory sequences and miRNAs (Zhang  
8 et al. 2014b). We recently hypothesized that a higher fraction of elements under negative  
9 selection involved in gene regulation and chromosome structure in avian genomes (~7%)  
10 (Zhang et al. 2014b) compared to mammals (~4%) (Lindblad-Toh et al. 2011) could  
11 contribute to some avian-specific phenotypes and the evolutionary stability of most avian  
12 karyotypes (Farré et al. 2016). Whilst a high density of CNEs in avian multi-species  
13 (ms)HSBs supports this hypothesis (Farré et al. 2016) a more definitive answer might be  
14 obtained by examining the fate of CNEs in the “interchromosomal EBRs” (flanking  
15 interchromosomal rearrangements) of an avian genome with a highly rearranged karyotype.

16

17 In this study we focused on two avian genomes. The first, the peregrine falcon (*Falco*  
18 *peregrinus*) has an atypical karyotype ( $2n=50$ ) (Nishida et al. 2008). Falcon's ability to fly at  
19 speeds >300 km/h and its enhanced visual acuity make it the fastest predator on Earth  
20 (Tucker et al. 1998). A prolonged period of extinction risk due to persecution around the  
21 World War II and secondary poisoning from organochlorine pesticides (e.g., DDT) in the  
22 1950s-60s (Ferguson-Lees and Christie 2005) led to its placement on the CITES list of  
23 endangered species. The second avian genome that was focused on here, the pigeon  
24 (*Columba livia*) has a typical avian karyotype ( $2n=80$ ) similar to those of reference avian  
25 genomes: chicken, turkey and zebra finch. Pigeon is one of the earliest examples of  
26 domestication in birds (Driscoll et al. 2009) contemporarily used as food and in sporting  
27 circles (Price 2002). Pigeon breeds can vary significantly in appearance with color, pattern,  
28 head crest, body shape, feathers, tails, vocalization and flight display variations (Price 2002)

1 inspiring considerable interest in identifying the genetic basis for these variations (Stringham  
2 et al. 2012; Shapiro et al. 2013). For the above reasons, both species genomes were  
3 sequenced (Shapiro et al. 2013; Zhan et al. 2013), however their assemblies are highly  
4 fragmented and chromosome-level assemblies are thus essential.

5

6 The objective of this study was therefore to develop a novel, inexpensive, transferrable  
7 approach to upgrade fragmented genome assemblies (i.e. pigeon and falcon) to the  
8 chromosome level and to use them to address novel biological questions related to avian  
9 genome evolution. The method combines computational algorithms for ordering scaffolds  
10 into PCFs retaining local structures of the target genome chromosomes after verification of a  
11 limited number of scaffolds, and physical mapping of PCFs directly to chromosomes with a  
12 universal set of avian bacterial artificial chromosome (BAC) probes. Studying a highly  
13 rearranged genome (falcon) compared to the avian ancestor sheds light on why  
14 interchromosomal rearrangements are infrequent in bird evolution.

15

## 1 **RESULTS**

2 Our method involves: (1) the construction of PCFs for fragmented assemblies based on the  
3 comparative and sequence read data implemented in the RACA algorithm; (2) PCR and  
4 computational verification of a limited number of scaffolds that are essential for revealing  
5 species-specific chromosome structures; (3) creation of a refined set of PCFs using the  
6 verified scaffolds and adjusted adjacency thresholds in RACA; (4) the use of a panel of  
7 “universal” BAC clones to anchor PCFs to chromosomes in a high-throughput manner (Fig.  
8 1).

### 10 **Construction of PCFs from fragmented assemblies**

11 Predicted chromosome fragments were generated for fragmented falcon and pigeon whole-  
12 genome sequences using RACA (Kim et al. 2013). For falcon, the zebra finch chromosome  
13 assembly was used as reference (divergence 60 MYA) and the chicken genome as outgroup  
14 (divergence 89 MYA). We generated a total of 113 PCFs with N50 of 27.44 Mb (Table 1).  
15 For pigeon ( $\geq 69$  MY divergence from both the chicken and zebra finch), chicken was used  
16 as reference and zebra finch as outgroup because: a) fewer pigeon scaffolds were split in  
17 this configuration (Supplemental Table S1) and b) due to the high similarity of pigeon and  
18 chicken karyotypes (Derjushcheva et al. 2004). This resulted in 150 pigeon PCFs with N50 of  
19 34.54 Mb (Table 1). These initial PCF sets contained 72 (15.06%) and 78 (13.64%)  
20 scaffolds, for falcon and pigeon respectively, that were split by RACA due to insufficient read  
21 and/or comparative evidence to support their structures.

22

### 23 **Verification of scaffolds essential for revealing species-specific chromosome** 24 **architectures**

25 All scaffolds split by RACA contained structural differences between the target and reference  
26 chromosomes, suggesting their importance for revealing the architecture of target species  
27 chromosomes. The structures of these scaffolds were tested by PCR amplification across all  
28 the split regions defined to  $< 6$  kb in the target species scaffolds. Of these, 41 (83.67%) and



1 58 (84.06%) resulted in amplicons of expected length in pigeon and falcon genomic DNA,  
2 respectively (Supplemental Table S2). For the split regions with negative PCR results we  
3 tested an alternative (RACA-suggested) order of the flanking syntenic fragments (SFs). Out  
4 of these, amplicons were obtained for 2/4 in falcon and 7/7 in pigeon, confirming the  
5 chimeric nature of the original scaffolds properly detected in these cases (Supplemental  
6 Table S2). To estimate which of the remaining split regions (>6 kb; 36 in falcon and 40 in  
7 pigeon PCFs) were likely to be chimeric, we empirically identified two genome-wide  
8 minimum physical coverage (Meyerson et al. 2010) levels, one for falcon and one for pigeon,  
9 in the SFs joining regions for which (and higher) the PCR results were most consistent with  
10 RACA predictions. If the new thresholds were used in RACA without additional scaffold  
11 verification (e.g., by PCR) or mapping data, they would lead to splitting of nearly all scaffolds  
12 with large structural misassemblies in falcon and ~6% of them would still be present in  
13 pigeon PCFs. The number of scaffolds containing real structural differences with the  
14 reference chromosomes that would still be split by RACA was estimated as ~56% in the  
15 falcon and ~43% in pigeon PCFs (Supplemental Table S2). To reduce the number of the real  
16 structural differences split in the final PCF set, PCR verification of selected scaffolds and use  
17 of independent (cytogenetic) mapping have been introduced.

18

### 19 **Creation of a refined set of pigeon and falcon PCFs**

20 For new reconstructions the adjusted physical coverage thresholds were used. In addition,  
21 we kept intact those scaffolds confirmed by PCR, but split those shown to be chimeric and/or  
22 disagreeing with the cytogenetic map (see below) resulting in a total of 93 PCFs with N50  
23 25.82 Mb for falcon and 137 PCFs with N50 of 22.17 Mb for pigeon, covering 97.17% and  
24 95.86% of the original scaffold assemblies, respectively (Table 1). The falcon RACA  
25 assembly contained six PCFs homeologous to complete zebra finch chromosomes (TGU4A,  
26 9, 11, 14, 17 and 19) while five pigeon PCFs were homeologous to complete chicken  
27 chromosomes (GGA11, 13, 17, 22 and 25). Only 3.50% of the original scaffolds used by

1 RACA were split in pigeon and 3.14% in falcon final PCFs (Table 1). The accuracy for the  
2 PCF assembly was estimated as ~85% for falcon and ~89% for pigeon based on the ratio of  
3 the number of SFs to the number of scaffolds (Kim et al. 2013).

4  
5 **Construction of a panel of comparatively anchored BAC clones designed to hybridize**  
6 **in phylogenetically divergent avian species and link PCFs to chromosomes**

7 Initial experiments on cross-species BAC mapping using FISH on five avian species with  
8 divergence times between 28 and 89 MY revealed highly varying success rates (21-94%),  
9 with hybridizations more likely to succeed on species closely related to that of the BAC origin  
10 (Table 2). To minimize the effect of evolutionary distances between species on  
11 hybridizations, genomic features that were likely to influence hybridization success were  
12 measured in chicken, zebra finch and turkey BAC clones (Supplemental Tables S3, S4). The  
13 classification and regression tree approach (CART; Loh 2011) was applied to the 101  
14 randomly-selected BAC clones (Table 2). The obtained classification shows 87% agreement  
15 with FISH results (Supplemental Fig. S1). Correlating DNA features with actual cross-  
16 species FISH results led us to develop the following criteria for selection of chicken or zebra  
17 finch BAC clones very likely to hybridize on metaphase preparations of phylogenetically  
18 distant birds ( $\geq 69$  MY of divergence; where the hybridization success rate of random BAC  
19 clones was  $< 70\%$ ): the BAC had to have  $\geq 93\%$  DNA sequence alignable with other avian  
20 genomes and contain at least one conserved element (CE)  $\geq 300$  bp. Instead of a long CE,  
21 the BAC could contain only short repetitive elements ( $< 1290$  bp) and CEs of at least 3 bp  
22 long (Supplemental Fig. S1; Supplemental Table S4). The hybridization success rate with  
23 distant avian species for the set of newly selected clones obeying these criteria was high  
24 (71-94%; Table 2). The success rates for the selected chicken BAC clones only ranged from  
25 90% to 94%. From these chicken clones, 84% hybridized with chromosomes of all avian  
26 species in our set (Supplemental Fig. S2).

27 As a final result, we generated a panel of 121 BAC clones spread across the avian genome  
28 (GGA 1-28 +Z (except 16)) that successfully hybridized across all species attempted. The

1 collection was supplemented by a further 63 BACs that hybridized on the metaphases of at  
2 least one species that was considered phylogenetically distant (i.e.  $\geq 69$  MY; split between  
3 Columbea and the remaining Neoavian clades) and a further 33 that hybridized on at least  
4 one other species (Fig. 2; Supplemental Table S5).

5

## 6 **Physical assignment of refined PCFs on the species' chromosomes**

7 In order to place and order PCFs along chromosomes, BAC clones from the panel described  
8 above and assigned to PCFs based on alignment results were hybridized to falcon (177  
9 clones) and pigeon (151 clones) chromosomes (Table 3). The 57 PCFs cytogenetically  
10 anchored to the falcon chromosomes represented 1.03 Gb of its genome sequence (88% of  
11 the cumulative scaffold length). Of these, 888.67 Mb were oriented on the chromosomes  
12 (Table 3; Supplemental Table S6). The pigeon chromosome assembly consisted of 0.91 Gb  
13 in 60 pigeon PCFs representing 82% of the combined scaffold length. Of these 687.59 Mb  
14 were oriented (Table 3; Supplemental Table S7). Visualizations of both newly assembled  
15 genomes are available from the Evolution Highway comparative chromosome browser (see  
16 Supplemental Results) and our avian UCSC browser hub.

17

## 18 **Pigeon chromosome assembly**

19 No deviations from the standard avian karyotype ( $2n=80$ ) were detected for pigeon with each  
20 mapped chromosome having an appropriate single chicken and zebra finch homeologue.  
21 Compared to chicken, the only interchromosomal rearrangement identified was the ancestral  
22 configuration of GGA4 found as two separate chromosomes in pigeon and other birds  
23 (Derjushcheva et al. 2004; Hansmann et al. 2009; Modi et al. 2009) (Fig. 3A; Supplemental  
24 Fig. S4; <http://eh-demo.ncsa.uiuc.edu/birds>). Nonetheless, 70 intrachromosomal EBRs in the  
25 pigeon lineage were identified (Supplemental Table S8).

26

## 1 **Falcon chromosome assembly**

2 Homeology between the chicken and the falcon was identified for all mapped chromosomes  
3 with the exception of GGA16 and GGA25 (Fig. 3B; Supplemental Fig. S5; [http://eh-  
4 demo.ncsa.uiuc.edu/birds](http://eh-demo.ncsa.uiuc.edu/birds)). In total, 13 falcon-specific fusions and six fissions were detected  
5 (Supplemental Table S8). Each of the chicken largest macrochromosome homeologues  
6 (GGA1 to GGA5) were split across two falcon chromosomes. Both GGA6 and GGA7  
7 homeologues were found as single blocks fused with other chicken chromosome material  
8 within falcon chromosomes. Among the other chicken macrochromosomes, only GGA8 and  
9 GGA9 were represented as individual chromosomes. Of the 17 mapped chicken  
10 microchromosomes, 11 were fused with other chromosomes. A total of 69 intrachromosomal  
11 EBRs were detected in the falcon lineage (Supplemental Table S8; Supplemental Results).  
12 Consistent with our previous report (Farré et al. 2016) falcon intrachromosomal EBRs were  
13 found highly enriched for the LTR-ERV1 transposable elements (TEs; t-test p-value <0.05;  
14 Supplemental Table S9). Both fusion and fission EBRs were not significantly enriched for  
15 any type of TEs.

16

## 17 **Fate of CNEs in avian inter- and intrachromosomal EBRs**

18 The falcon chromosome assembly provided us with a set of 19 novel interchromosomal  
19 EBRs not previously found in published avian chromosome assemblies (Fig. 3B;  
20 Supplemental Table S8). To investigate the fate of CNEs in avian EBRs, we calculated  
21 densities of avian CNEs in the chicken chromosome regions corresponding to the chicken,  
22 falcon, pigeon, flycatcher and zebra finch intrachromosomal and interchromosomal EBRs  
23 defined to  $\leq 100$  kb in the chicken genome (Fig. 4; Supplemental Table S10). Avian EBRs  
24 had significantly lower fraction of CNEs than their two adjacent chromosome intervals of the  
25 same size each (up- and downstream (p-value =  $3.35e-07$ ; Supplemental Table S11)).  
26 Moreover, the interchromosomal EBRs (fusions and fissions) had on average  $\sim 12$  times  
27 lower density of CNEs than the intrachromosomal EBRs (p-value =  $2.40e-05$ ; Supplemental

1 Table S11). The lowest density of CNEs was observed in the fission breakpoints ( $p$ -value =  
2 0.04; Fig. 4, Supplemental Table S11).

3 To identify CNE densities and the distribution associated with avian EBRs at the genome-  
4 wide level, we counted CNE bases in 1 kb windows overlapping EBRs and avian msHSBs  
5 >1.5 Mb (Farré et al. 2016). The average density of CNEs in the EBR windows was lower  
6 (0.02) than in msHSBs (0.11). The density of CNEs in the fission EBRs was the lowest  
7 observed, zero CNE bases ('zero CNE windows'), while in the intrachromosomal EBRs the  
8 highest among the EBR regions (0.02; Supplemental Table S12). The genome-wide CNE  
9 density was 0.09, closer to the density observed in msHSBs. Of ~347 Mb of the chicken  
10 genome found in the 'zero CNE windows' 0.5% were associated with EBRs and 15% with  
11 msHSBs. To investigate if these intervals are distributed differently in the breakpoint and  
12 synteny regions we compared distances between the 'zero CNE windows' and the closest  
13 window with the average msHSB CNE density or higher in EBRs, msHSBs, and genome-  
14 wide. The median of the distances between these two types of windows was the lowest in  
15 the msHSBs (~4 kb), intermediate in the intrachromosomal (~19 kb) and fusion EBRs (~23  
16 kb), and highest in the fission EBRs (~35 kb) (Supplemental Table S13). All these values  
17 were significantly different from the genome-wide average distance of ~6 kb ( $p$ -values <2.2e-  
18 16) and also significantly different from each other ( $p$ -value  $\leq 0.004$ ; Supplemental Table  
19 S12; Supplemental Fig. S6).

20

## 1 DISCUSSION

2 In this study we present a novel integrative approach to upgrade sequenced animal  
3 genomes to the chromosome level. We have previously reported a limited success with the  
4 use of high-gene density and low-repeat content BAC clones for cross-species hybridization  
5 (Larkin et al. 2006; Romanov et al. 2011). However, the use of such probes for whole-  
6 genome chromosomal assembly has not hitherto been demonstrated. That is, in this study,  
7 we made use of the whole-genome sequences from multiple species and applied a  
8 systematic approach to design a panel of universally hybridizing BAC probes along the  
9 length of each chromosome. Using these probes as a basis, and in combination with  
10 comparative sequence analysis, targeted PCR and optimized high-throughput cross-species  
11 BAC hybridizations the approach herein presented thus represents a unique methodology to  
12 achieve chromosome-level reconstruction for scaffold-based *de-novo* assemblies that could  
13 be applied to any animal genome provided an actively growing population of cells can be  
14 obtained to generate metaphase preparations.

15

16 In this study we provide proof of principle for this new approach by generating such  
17 assemblies for two previously published, but highly fragmented, avian genomes. The  
18 resulting chromosome level assemblies contain >80% of the genomes (compared to current  
19 estimates of genome size) and, in continuity are comparable to those obtained by combining  
20 the traditional sequencing and mapping techniques (Deakin and Ezaz 2014) but require  
21 much less cost and resources. Given that it has been suggested that estimates of genome  
22 size based on cytology are inaccurate and usually overestimated (Kasai et al. 2012; Kasai et  
23 al. 2013) techniques such as flow cytometry should be used to estimate genome size more  
24 accurately (Kasai et al. 2012; Kasai et al. 2013). Flow cytometry will ultimately be able to  
25 determine the extent to which the genomes are actually covered by new procedures to  
26 upgrade their assemblies and will be invaluable in pointing out any remaining gaps to fill.  
27 Indeed, this approach could be augmented further by chromosome specific DNA sequencing

1 such as has recently been demonstrated in the B chromosomes of two deer species  
2 (Makunin et al. 2016)  
3  
4 Molecular and cytogenetic studies to date, suggest that the majority of avian genomes  
5 remain remarkably conserved in terms of chromosome number (in 60-70% of species  
6  $2n=80$ ) and that interchromosomal changes are relatively rare (Griffin et al. 2007; Schmid  
7 et al. 2015). Exceptions include representatives of *Psittaciformes* (parrots), *Sphenisciformes*  
8 (penguins) and *Falconiformes* (falcons). This study represents the first reconstruction of a  
9 highly rearranged avian karyotype (peregrine falcon). It demonstrates that fusion is the most  
10 common mechanism of interchromosomal change in this species, with some resulting  
11 chromosomes exhibiting as many as four fused ancestral chromosomes. There was no  
12 evidence of reciprocal translocations and all microchromosomes remained intact, even when  
13 fused to larger chromosomes. Recently we suggested possible mechanisms why avian  
14 genomes, with relatively rare exceptions, remain evolutionarily stable interchromosomally  
15 and why microchromosomes represent blocks of conserved synteny (Romanov et al. 2014;  
16 Farré et al. 2016). Absence of interchromosomal rearrangement (as seen in most birds)  
17 could either suggest an evolutionary advantage to retaining such a configuration or little  
18 opportunity for change. A smaller number of transposable elements in avian genomes  
19 compared to other animals would indicate that avian chromosomes indeed have fewer  
20 opportunities for chromosome merging using NAHR, explaining the presence of multiple  
21 microchromosomes. Our study provides an additional support for this hypothesis as in falcon  
22 lineage only intrachromosomal EBRs were significantly enriched in transposable elements,  
23 while interchromosomal EBRs (flanking both fusions and fissions) were not found  
24 significantly enriched. On the other hand, a strong enrichment for avian CNEs in the regions  
25 of interspecies synteny in birds and other reptiles suggests evolutionary advantage of  
26 maintaining established synteny (Farré et al. 2016), implying that fission events should be  
27 rare in avian evolution. In this study, we present the first analysis of a significant number of  
28 interchromosomal EBRs by analysis of the falcon genome, demonstrating that those rare

1 interchromosomal rearrangements that are fixed in the avian lineage-specific evolution did  
2 indeed appear in areas of a low density of CNEs. This applies to both fission and fusion  
3 events. Our results demonstrate moreover that, to be suitable for chromosomal fission, the  
4 sites of interchromosomal EBRs are restricted further as they need to be significantly more  
5 distant from the areas with high CNE density than the equivalent intervals found in the  
6 regions of multispecies synteny, other EBR types, or on average in the genome. This might  
7 also explain why falcon-specific fission breakpoints appear to be reused in other avian  
8 lineages as intrachromosomal EBRs. Study of intrachromosomal changes in pigeons,  
9 falcons (this study) and Passeriform species (Skinner and Griffin 2012; Romanov et al.  
10 2014) suggests that these events might have a less dramatic effect on *cis* gene regulation  
11 than interchromosomal events. Indeed, intrachromosomal EBRs appear in regions of  
12 significantly higher CNE density than interchromosomal EBRs. Why then, do species such  
13 as falcons and parrots undergo wholesale interchromosomal rearrangement (previously  
14 reported), but (according to this study) with fission restricted to a few events and fusion more  
15 common? Absence of positive selection for change in chromosome number (or lack of  
16 templates for NAHR) possibly explains why there was little fixation of any interchromosomal  
17 change among birds in general (Bush et al. 1977; Fontdevila et al. 1982; Burt et al. 1999;  
18 Burt 2002), however why this positive selection has been re-introduced (or barriers to it have  
19 been removed) in selected orders is still a matter of conjecture.

20

21 The design and use of a set of BAC probes intended to work equally well on a large number  
22 of diverged avian species created a resource for physical mapping that is transferrable to  
23 multiple species. In this regard, mammals are the greatest priority as they are the most  
24 studied phylogenetic Class of organisms in the scientific literature. Reasons for this include  
25 human interest (e.g. clinical studies), biomedical models (e.g. mouse, rat, rabbit, pig),  
26 companion animals (e.g. cat, dog) and agricultural mammals (pig, sheep, cattle etc.). Many  
27 are on the CITES threatened/endangered list, and, with impending global warming, tools for  
28 the study of ecology and conservation of these animals is a priority; many extinct species



1 also still attract considerable interest. Of the >5,000 extant species however, only ~20 have  
2 genomes assembled to chromosomes (with primates, rodents and artiodactyls  
3 disproportionally overrepresented) with more than ten of the 26 orders having no  
4 chromosome level assemblies at all. Recently a further >50 *de-novo* mammalian assemblies  
5 have been produced (more are inevitable); these however, at best, are collections of sub-  
6 chromosomal sized scaffolds. Moreover, several hundred are currently being assembled to  
7 scaffold level by individual projects or consortia such as Genome10K (Koepfli et al. 2015).  
8 Building a mammalian universal BAC set would be a greater challenge than in birds as  
9 mammalian genomes have more repetitive sequences and are about three times larger thus  
10 more BACs would be needed to achieve the same level of mapping resolution. On the other  
11 hand, the development of advanced mapping and sequencing techniques (e.g., Dovetail,  
12 BioNano or PacBio) will eventually provide an opportunity to replace RACA PCFs with longer  
13 and more complete sub-chromosomal sized superscaffolds or sequence contigs requiring  
14 fewer BACs to anchor them to chromosomes. The availability of large numbers of high-  
15 quality mammalian BAC clone libraries from many species makes our approach more  
16 applicable to mammals than to any other animal group. If we add the fact that our avian BAC  
17 set is showing good success rates on lizard and turtle chromosomes (unpublished results),  
18 building chromosomal assemblies for all vertebrate and ultimately all animal groups  
19 supported by universal collection of BACs is a realistic objective for the near future.

## 1 **METHODS**

### 2 **Avian genome assemblies, repeat masking and gene annotations**

3 The chicken (ICGSC Gallus\_gallus 4.0; Hillier 2004), zebra finch (WUGSC 3.2.4; Warren et  
4 al. 2010), and turkey (TGC Turkey\_2.01; Dalloul et al. 2010) chromosome assemblies were  
5 downloaded from the UCSC Genome Browser (Kent et al. 2002). The collared flycatcher  
6 (FicAlb1.5; Ellegren et al. 2012) genome was obtained from NCBI. Scaffold-based (N50>2  
7 Mb) assemblies of pigeon, falcon, and 16 additional avian genomes were provided by the  
8 Avian Phylogenomics Consortium (Zhang et al. 2014a). All sequences were repeat-masked  
9 using Window Masker (Morgulis et al. 2006) with *-sdust* option and Tandem Repeats Finder  
10 (Benson 1999). Chicken gene (version of 27/04/2014) and repetitive sequence (version of  
11 11/06/2012) annotations were downloaded from the UCSC genome browser (Rosenbloom et  
12 al. 2015). Chicken genes with a single ortholog in the human genome were extracted from  
13 Ensembl Biomart (v.74; Kinsella et al. 2011).

14

### 15 **Pairwise and multiple genome alignments, nucleotide evolutionary conservation** 16 **scores and conserved elements**

17 Pairwise alignments using chicken and zebra finch chromosome assemblies as references  
18 and all other assemblies as targets were generated with *LastZ* (v.1.02.00; Harris 2007) and  
19 converted into the UCSC “chains” and “nets” alignment formats with the Kent-library tools  
20 (Kent et al. 2003; Supplemental Methods). The evolutionary conservation scores and DNA  
21 conserved elements (CEs) for all chicken nucleotides assigned to chromosomes were  
22 estimated using PhastCons (Siepel et al. 2005) from the multiple alignments of 21 avian  
23 genomes (Supplemental Methods). Conserved non-coding elements obtained from the  
24 alignments of 48 avian genomes were used (Farré et al. 2016).

25

### 26 **Reference-assisted chromosome assembly of pigeon and falcon genomes**

27 Pigeon and falcon PCFs were generated using the Reference-Assisted Chromosome  
28 Assembly (RACA; Kim et al. 2013; Supplemental Methods) tool. We chose zebra finch

1 genome as reference and chicken as outgroup for falcon based on the phylogenetic  
2 distances between the species (Jarvis et al. 2014). For pigeon both chicken as reference  
3 and zebra finch as outgroup and the vice versa experiments were performed as pigeon is  
4 phylogenetically distant from chicken and zebra finch. Two rounds of RACA were done for  
5 both species. The initial run was performed using the following parameters:  
6 *WINDOWSIZE=10 RESOLUTION=150000 MIN\_INTRACOV\_PERC=5*. Prior to the second  
7 run of RACA we tested the scaffolds split during the initial RACA run using PCR  
8 amplification across the split intervals (see below) and adjusted the parameters accordingly  
9 (Supplemental Methods).

10

#### 11 **PCR testing of adjacent SFs**

12 Primers flanking split SF joints within scaffolds or RACA predicted adjacencies were  
13 designed using Primer3 software (v.2.3.6; Untergasser et al. 2012). To avoid  
14 misidentification of EBRs or chimeric joints we selected primers only within the sequences  
15 that had high quality alignments between the target and reference genomes and found in  
16 adjacent SFs. Due to alignment and SF detection settings some of the intervals between  
17 adjacent SFs could be >6 kb and primers could not be chosen for a reliable PCR  
18 amplification. In such cases we used CASSIS software (Baudet et al. 2010) and the  
19 underlying alignment results to narrow gaps between adjacent SFs where possible. Whole  
20 blood was collected aseptically from adult falcon and pigeon. DNA was isolated using  
21 DNeasy Blood and Tissue Kit (Qiagen) following standard protocols. PCR amplification was  
22 performed according to the protocol described in the Supplemental Methods.

23

#### 24 **BAC clone selection**

25 The chromosome coordinates of chicken (CHORI-261), turkey (CHORI-260) and zebra finch  
26 (TGMCB) BAC clones in the corresponding genomes were extracted from NCBI clone  
27 database (Schneider et al. 2013). We removed all discordantly placed BAC clones (based  
28 on BAC end sequence (BES) mappings) following the NCBI definition of concordant BAC

1 placement. Briefly, a BAC clone placement was considered concordant when the estimated  
2 BAC length in the corresponding avian genome is within [library average length  $\pm$   
3  $3 \times \text{standard deviation}$ ] and BAC BESs map to the opposite DNA strands in the genome  
4 assembly. Turkey and zebra finch BAC clone coordinates were translated into chicken  
5 chromosome coordinates using UCSC Genome Browser *LiftOver* tool (Kent et al. 2002) with  
6 the minimum ratio of remapped bases  $>0.1$ .

7

8 For each BAC clone mapped to the chicken chromosomes various genomic features  
9 selected to estimate the probability of clones to hybridize with metaphase chromosomes in  
10 distant avian species were calculated (Supplemental Table S3) using a custom Perl script or  
11 extracted from gene, repetitive sequence, conserved element and nucleotide conservation  
12 score files. The clones selected for mapping experiments were originally obtained from the  
13 BACPAC Resource Centre at the Children's Hospital Oakland Research Institute and the  
14 zebra finch TGMCBa library (Clemson University Genomics Institute).

15

### 16 **Classification tree**

17 The classification tree was created in R (v.3.2.3; Team 2015) using the classification and  
18 regression tree (CART) algorithm included in the rpart package (v.4.1-10; Therneau et al.  
19 2015). We introduced an adjusted weight matrix setting: the cost of returning a false positive  
20 was twice as high as the cost of a false negative. The tree was visualized with rattle package  
21 (v.4.1.0; Williams 2011).

22

### 23 **Cell culture and chromosome preparation**

24 Chromosome preparations were established from fibroblast cell lines generated from  
25 collagenase treatment of 5- to 7-day-old embryos or from skin biopsies. Cells were cultured  
26 at 40°C, and 5% CO<sub>2</sub> in Alpha MEM (Fisher), supplemented with 20% Fetal Bovine Serum  
27 (Gibco), 2% Pen-Strep (Sigma) and 1% L-Glutamine (Sigma). Chromosome suspension  
28 preparation followed standard protocols, briefly mitostatic treatment with colcemid at a final

1 concentration of 5.0 µg/ml for 1 h at 40°C was followed by hypotonic treatment with 75mM  
2 KCl for 15 min at 37°C and fixation with 3:1 methanol:acetic acid.

3

#### 4 **Preparation of BAC clones for fluorescence *in-situ* hybridization (FISH)**

5 BAC clone DNA was isolated using the Qiagen Miniprep Kit (Qiagen) prior to amplification  
6 and direct labelling by nick translation. Probes were labeled with Texas Red-12-dUTP  
7 (Invitrogen) and FITC-Fluorescein-12-UTP (Roche) prior to purification using the Qiagen  
8 Nucleotide Removal Kit (Qiagen).

9

#### 10 **Fluorescence *in-situ* hybridization (FISH)**

11 Metaphase preparations were fixed to slides and dehydrated through an ethanol series (2  
12 min each in 2xSSC, 70%, 85% and 100% ethanol at room temperature). Probes were diluted  
13 in a formamide buffer (Cytocell) with Chicken Hybloc (Insight Biotech) and applied to the  
14 metaphase preparations on a 37°C hotplate before sealing with rubber cement. Probe and  
15 target DNA were simultaneously denatured on a 75°C hotplate prior to hybridization in a  
16 humidified chamber at 37°C for 72 h. Slides were washed post-hybridization for 30 sec in  
17 2xSSC/ 0.05% Tween 20 at room temperature, then counterstained using VECTASHIELD  
18 anti-fade medium with DAPI (Vector Labs). Images were captured using an Olympus BX61  
19 epifluorescence microscope with cooled CCD camera and SmartCapture (Digital Scientific  
20 UK) system. In selected experiments, we used multiple hybridization strategies, making use  
21 of the Cytocell Octochrome (8 chamber) and Multiprobe (24 chamber) devices. Briefly,  
22 labeled probes were air dried on to the device. Probes were, re-hybridized in standard  
23 buffer, applied to the glass slide (which was sub-divided to correspond to the hybridization  
24 chambers) and FISH continued as above.

25

## 1 **EBR detection and CNE density analysis**

2 The multiple alignments of the chicken, zebra finch, flycatcher, pigeon and falcon  
3 chromosome sequences were obtained using progressiveCactus (Paten et al. 2011) with  
4 default parameters. Pairwise synteny blocks were defined using the maf2synteny tool  
5 (Kolmogorov et al. 2014) at 100, 300 and 500 kb resolution. Using chicken as reference  
6 genome, EBRs were detected and classified using the *ad hoc* statistical approach described  
7 previously (Farré et al. 2016). All well-defined (or flanking oriented PCFs) fusion and fission  
8 points were identified from pairwise alignments with the chicken genome. Only the EBRs  
9  $\leq 100$  kb were used for the CNE analysis. EBRs smaller than 1 kb were extended  $\pm 1$  kb. For  
10 each EBR, we defined two windows upstream (+1 and +2) and two downstream (-1 and -2)  
11 of the same size as the EBR. We calculated the fraction of bases within CNEs in each EBR  
12 site, upstream and downstream windows. Differences in CNE densities were tested for  
13 significance using the Kruskal-Wallis test followed by Mann-Whitney U test.

14

## 15 **Comparing CNE densities in EBRs and msHSBs**

16 Chicken chromosomes (excluding GGA16, W and Z) were divided into 1 kb non-overlapping  
17 intervals. Only windows with  $>50\%$  of their bases with chicken sequence data available were  
18 used in this analysis. All intervals were assigned either to msHSBs  $>1.5$  Mb (Farré et al.  
19 2016), avian EBRs flanking: fusions, fissions, intrachromosomal EBR, and the intervals  
20 found in the rest of the chicken genome. We estimated the average CNE density for each  
21 window type and also the distance, in number of 1 kb windows, between each window with  
22 the lowest CNE density (0 bp) and the nearest window with the average msHSB CNE  
23 density or higher. CNE densities were obtained using bedtools (v.2.20-1; Quinlan and Hall  
24 2010). Differences in distances between the two window types in msHSBs and EBRs were  
25 tested for significance using the Kruskal-Wallis test followed by Mann-Whitney U test.

26

1 **Densities of TEs in falcon intrachromosomal EBRs, fusions and fissions**

2 The TEs scaffold coordinates reported on Shapiro et al. 2013 were translated to falcon  
3 chromosome coordinates using a custom Perl script. The densities of TEs (>100 bp on  
4 average in the EBR- or non-EBR containing non-overlapping 10 Kbp genome intervals) were  
5 compared for the falcon-lineage specific interchromosomal EBRs, EBRs flanking fusion and  
6 fission events and the rest of the genome as previously described (Elsik et al. 2009; Larkin  
7 et al. 2009; Groenen et al. 2012; Farré et al. 2016).

8

9 **DATA ACCESS**

10 The falcon a pigeon chromosome assemblies are deposited at DDBJ/ENA/GenBank under  
11 the accessions MLQY00000000 and MLQZ00000000, respectively. Visualizations of falcon  
12 and pigeon genome assemblies are available from the Evolution Highway comparative  
13 chromosome browser: <http://eh-demo.ncsa.uiuc.edu/birds>; and our UCSC browser hub:  
14 <http://sftp.rvc.ac.uk/rvcpaper/birdsHUB/hub.txt>.

15

16 **DISCLOSURE DECLARATION**

17 Authors report no conflict of interests.

18

19 **ACKNOWLEDGMENTS**

20 The authors would like thank Dr. Yu-Mei Chang for the assistance with statistical analyses.  
21 This work was supported in part by the Biotechnology and Biological Sciences Research  
22 Council [BB/K008226/1 and BB/J010170/1 to D.M.L, and BB/K008161/1 to D.K.G].

## REFERENCES

- Andersson L, Georges M. 2004. Domestic-animal genomics: deciphering the genetics of complex traits. *Nat Rev Genet* **5**: 202-212.
- Baudet C, Lemaitre C, Dias Z, Gautier C, Tannier E, Sagot MF. 2010. Cassis: detection of genomic rearrangement breakpoints. *Bioinformatics* **26**: 1897-1898.
- Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**: 573-580.
- Branco MR, Pombo A. 2006. Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS Biol* **4**: e138.
- Burt DW. 2002. Origin and evolution of avian microchromosomes. *Cytogenet Genome Res* **96**.
- Burt DW, Bruley C, Dunn IC, Jones CT, Ramage A, Law AS, Morrice DR, Paton IR, Smith J, Windsor D et al. 1999. The dynamics of chromosome evolution in birds and mammals. *Nature* **402**: 411-413.
- Bush GL, Case SM, Wilson AC, Patton JL. 1977. Rapid speciation and chromosomal evolution in mammals. *Proc Natl Acad Sci U S A* **74**.
- Dalloul RA, Long JA, Zimin AV, Aslam L, Beal K, Blomberg Le A, Bouffard P, Burt DW, Crasta O, Crooijmans RP et al. 2010. Multi-platform next-generation sequencing of the domestic turkey (*Meleagris gallopavo*): genome assembly and analysis. *PLoS Biol* **8**.
- Deakin JE, Ezaz T. 2014. Tracing the evolution of amniote chromosomes. *Chromosoma* **123**: 201-216.
- Derjusheva S, Kurganova A, Habermann F, Gaginskaya E. 2004. High chromosome conservation detected by comparative chromosome painting in chicken, pigeon and passerine birds. *Chromosome Res* **12**: 715-723.
- Driscoll CA, Macdonald DW, O'Brien SJ. 2009. From wild animals to domestic pets, an evolutionary view of domestication. *Proc Natl Acad Sci U S A* **106**: 9971-9978.



- Elsik CG, Tellam RL, Worley KC. 2009 The genome sequence of a taurine cattle: a window to ruminant biology and evolution. *Science* **324**: 522-528.
- Ellegren H, Smeds L, Burri R, Olason PI, Backstrom N, Kawakami T, Kunstner A, Makinen H, Nadachowska-Brzyska K, Qvarnstrom A et al. 2012. The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature* **491**: 756-760.
- Fang X, Nevo E, Han L, Levanon EY, Zhao J, Avivi A, Larkin D, Jiang X, Feranchuk S, Zhu Y et al. 2014. Genome-wide adaptive complexes to underground stresses in blind mole rats *Spalax*. *Nat Commun* **5**: 3966.
- Farré M, Narayan J, Slavov GT, Damas J, Auvil L, Li C, Jarvis ED, Burt DW, Griffin DK, Larkin DM. 2016. Novel insights into chromosome evolution in birds, archosaurs, and reptiles. *Genome Biol Evol* **8**: 2442-2451.
- Ferguson-Lees J, Christie DA. 2005. *Raptors of the world*. Princeton University Press, Princeton, N.J.
- Fontdevila A, Ruiz A, Ocaña J, Alonso G. 1982. Evolutionary history of *Drosophila buzzatii*. II. How much has chromosomal polymorphism changed in colonization? *Evolution* **36**.
- Griffin DK, Robertson LBW, Tempest HG, Skinner BM. 2007. The evolution of the avian genome as revealed by comparative molecular cytogenetics. *Cytogenet Genome Res* **117**: 64-77.
- Groenen MAM, Archibald AL, Uenishi H, Tuggle CK, Takeuchi Y, Rothschild MF, Rogel-Gaillard C, Park C, Milan D, Megens H-J et al. 2012. Analyses of pig genomes provide insight into porcine demography and evolution. *Nature* **491**: 393-398.
- Hansmann T, Nanda I, Volobouev V, Yang F, Schartl M, Haaf T, Schmid M. 2009. Cross-species chromosome painting corroborates microchromosome fusion during karyotype evolution of birds. *Cytogenet Genome Res* **126**: 281-304.
- Harris RS. 2007. Improved pairwise alignment of genomic DNA. Vol Ph.D. The Pennsylvania State University.

- Hillier L. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**: 695-716.
- Hu X, Gao Y, Feng C, Liu Q, Wang X, Du Z, Wang Q, Li N. 2009. Advanced technologies for genomic analysis in farm animals and its application for QTL mapping. *Genetica* **136**: 371-386.
- Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SYW, Faircloth BC, Nabholz B, Howard JT et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* **346**: 1320-1331.
- Kasai F, O'Brien PC, Ferguson-Smith MA. 2013. Afrotheria genome; overestimation of genome size and distinct chromosome GC content revealed by flow karyotyping. *Genomics* **102**: 468-471.
- Kasai F, O'Brien PC, Ferguson-Smith MA. 2012. Reassessment of genome size in turtle and crocodile based on chromosome measurement by flow karyotyping: close similarity to chicken. *Biol Lett* **8**: 631-635.
- Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D. 2003. Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proc Natl Acad Sci U S A* **100**: 11484-11489.
- Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Res* **12**: 996-1006.
- Kim J, Larkin DM, Cai Q, Asan, Zhang Y, Ge R-L, Auvil L, Capitanu B, Zhang G, Lewin HA et al. 2013. Reference-assisted chromosome assembly. *Proc Natl Acad Sci U S A* **110**: 1785-1790.
- Kinsella RJ, Kahari A, Haider S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D, Derwent P, Kerhornou A et al. 2011. Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database* **2011**: bar030.
- Koepfli K-P, Benedict Paten, Scientists tGKCo, O'Brien SJ. 2015. The Genome 10K Project: A Way Forward. *Ann Rev Anim Biosci* **3**: 57-111.

- Kolmogorov M, Raney B, Paten B, Pham S. 2014. Ragout-a reference-assisted assembly tool for bacterial genomes. *Bioinformatics* **30**: i302-309.
- Larkin DM, Daetwyler HD, Hernandez AG, Wright CL, Hetrick LA, Boucek L, Bachman SL, Band MR, Akraiko TV, Cohen-Zinder M et al. 2012. Whole-genome resequencing of two elite sires for the detection of haplotypes under selection in dairy cattle. *Proc Natl Acad Sci U S A* **109**: 7693-7698.
- Larkin DM, Prokhorovich MA, Astakhova NM, Zhdanova NS. 2006. Comparative mapping of mink chromosome8p: in situ hybridization of seven cattle BAC clones. *Anim Genet* **37**: 429-430.
- Larkin DM, Pape G, Donthu R, Auvil L, Welge M, Lewin HA. 2009. Breakpoint regions and homologous synteny blocks in chromosomes have different evolutionary histories. *Genome Res* **19**: 770-777.
- Lewin HA, Larkin DM, Pontius J, O'Brien SJ. 2009. Every genome sequence needs a good map. *Genome Res* **19**: 1925-1928.
- Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E et al. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**: 476-482.
- Loh W-Y. 2011. Classification and regression trees. *WIREs Data Mining Knowl Discov* **1**: 14-23.
- Mak ACY, Lai YYY, Lam ET, Kwok T-P, Leung AKY, Poon A, Mostovoy Y, Hastie AR, Stedman W, Anantharaman T et al. 2016. Genome-wide structural variation detection by genome mapping on nanochannel arrays. *Genetics* **202**: 351-362.
- Makunin AI, Kichigin IG, Larkin DM, O'Brien PC, Ferguson-Smith MA, Yang F, Proskuryakova AA, Vorobieva NV, Chernyaeva EN, O'Brien SJ et al. 2016. *BMC Genomics* **17**: 618.
- Meyerson M, Gabriel S, Getz G. 2010. Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet* **11**: 685-696.

- Modi WS, Romanov M, Green ED, Ryder O. 2009. Molecular cytogenetics of the California condor: evolutionary and conservation implications. *Cytogenet Genome Res* **127**: 26-32.
- Morgulis A, Gertz EM, Schaffer AA, Agarwala R. 2006. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J Comput Biol* **13**: 1028-1040.
- Murphy WJ, Larkin DM, Everts-van der Wind A, Bourque G, Tesler G, Auvil L, Beever JE, Chowdhary BP, Galibert F, Gatzke L et al. 2005. Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science* **309**: 613-617.
- Nishida C, Ishijima J, Kosaka A, Tanabe H, Habermann FA, Griffin DK, Matsuda Y. 2008. Characterization of chromosome structures of Falconinae (Falconidae, Falconiformes, Aves) by chromosome painting and delineation of chromosome rearrangements during their differentiation. *Chromosome Res* **16**: 171-181.
- Paten B, Earl D, Nguyen N, Diekhans M, Zerbino D, Haussler D. 2011. Cactus: Algorithms for genome multiple sequence alignment. *Genome Res* **21**: 1512-1528.
- Price TD. 2002. Domesticated birds as a model for the genetics of speciation by sexual selection. *Genetica* **116**: 311-327.
- Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, Troll CJ, Fields A, Hartley PD, Sugnet CW et al. 2016. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res* doi:10.1101/gr.193474.115.
- Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841-842.
- Rhoads A, Au KF. 2015. PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics* **13**: 278-289.
- Romanov MN, Dodgson JB, Gonser RA, Tuttle EM. 2011. Comparative BAC-based mapping in the white-throated sparrow, a novel behavioral genomics model, using interspecies overgo hybridization. *BMC Res Notes* **4**: 211.

- Romanov MN, Farré M, Lithgow PE, Fowler KE, Skinner BM, O'Connor R, Fonseka G, Backström N, Matsuda Y, Nishida C et al. 2014. Reconstruction of gross avian genome structure, organization and evolution suggests that the chicken lineage most closely resembles the dinosaur avian ancestor. *BMC Genomics* **15**: 1-18.
- Rosenbloom KR, Armstrong J, Barber GP, Casper J, Clawson H, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M et al. 2015. The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res* **43**: D670-681.
- Schmid M, Smith J, Burt DW, Aken BL, Antin PB, Archibald AL, Ashwell C, Blackshear PJ, Boschiero C, Brown CT et al. 2015. Third report on chicken genes and chromosomes 2015. *Cytogenet Genome Res* **145**: 78-179.
- Schneider VA, Chen HC, Clausen C, Meric PA, Zhou Z, Bouk N, Husain N, Maglott DR, Church DM. 2013. Clone DB: an integrated NCBI resource for clone-associated data. *Nucleic Acids Res* **41**: D1070-1078.
- Shapiro MD, Kronenberg Z, Li C, Domyan ET, Pan H, Campbell M, Tan H, Huff CD, Hu H, Vickrey AI et al. 2013. Genomic diversity and evolution of the head crest in the rock pigeon. *Science* **339**: 1063-1067.
- Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**: 1034-2050.
- Skinner BM, Griffin DK. 2012. Intrachromosomal rearrangements in avian genome evolution: evidence for regions prone to breakpoints. *Heredity* **108**: 37-41.
- Stringham SA, Mulroy EE, Xing J, Record D, Guernsey MW, Aldenhoven JT, Osborne EJ, Shapiro MD. 2012. Divergence, convergence, and the ancestry of feral populations in the domestic rock pigeon. *Curr Biol* **22**: 302-308.
- Tamazian G, Dobrynin P, Krasheninnikova K, Komissarov A, Koepfli KP, O'Brien SJ. 2016. Chromosomer: a reference-based genome arrangement tool for producing draft chromosome sequences. *GigaScience* **5**: 38.

- Teague B, Waterman MS, Goldstein S, Potamouisis K, Zhou S, Reslewic S, Sarkar D, Valouev A, Churas C, Kidd JM et al. 2010. High-resolution human genome structure by single-molecule analysis. *Proc Natl Acad Sci U S A* **107**: 10848-10853.
- Team RC. 2015. R: a language and environment for statistical computing (R foundation for statistical computing, Vienna, 2012). <http://www.R-project.org>.
- Therneau T, Atkinson B, Ripley B. 2015. rpart: recursive partitioning and regression trees. R package version 4.1–10.
- Tucker V, Cade T, Tucker A. 1998. Diving speeds and angles of a gyrfalcon (*Falco rusticolus*). *J Exp Biol* **201**: 2061-2070.
- Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. 2012. Primer3-new capabilities and interfaces. *Nucleic Acids Res* **40**: e115.
- Warren WC, Clayton DF, Ellegren H, Arnold AP, Hillier LW, Kunstner A, Searle S, White S, Vilella AJ, Fairley S et al. 2010. The genome of a songbird. *Nature* **464**: 757-762.
- Williams G. 2011. *Data mining with Rattle and R: The art of excavating data for knowledge discovery*. Springer Science & Business Media.
- Zhan X, Pan S, Wang J, Dixon A, He J, Muller MG, Ni P, Hu L, Liu Y, Hou H et al. 2013. Peregrine and saker falcon genome sequences provide insights into evolution of a predatory lifestyle. *Nat Genet* **45**: 563-566.
- Zhang G, Li B, Li C, Gilbert MT, Jarvis ED, Wang J, Avian Genome C. 2014a. Comparative genomic data of the Avian Phylogenomics Project. *GigaScience* **3**: 26.
- Zhang G, Li C, Li Q, Li B, Larkin DM, Lee C, Storz JF, Antunes A, Greenwold MJ, Meredith RW et al. 2014b. Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* **346**: 1311-1320.

## FIGURE LEGENDS

**Figure 1.** Methodology for the placement of the PCFs on chromosomes. (A) dual-color FISH of universal BAC clones, (B) cytogenetic map of the falcon chromosome 8 (FPE8) with indication of the relative positions of the BAC clones along the chromosome, and (C) assembled chromosome containing PCFs 7a, 7b and 13b\_13a. Blue blocks indicate positive (+) orientation of tracks compared to the falcon chromosome, red blocks indicate negative (-) orientation and grey blocks show unknown (?) orientation.

**Figure 2.** Distribution of universal BAC clones along chicken chromosomes. Each rectangle represents a chicken chromosome and the lines inside the location of each BAC clone. BAC clones are colored accordingly to the maximum phylogenetic distance of the species they successfully hybridized. The distribution of spacing between all these BAC clones is shown on the Supplemental Fig. S3.

**Figure 3.** Ideogram of pigeon (A) and peregrine falcon (B) chromosomes. Numbered rectangles represent chromosomes and colored blocks inside represent regions of homeology with chicken chromosomes. Lines within colored blocks represent block orientation. Pigeon chromosomes 1-9 and Z were numbered according to Hansmann et al., 2009 and the remaining chromosomes according to their chicken homeologues. Falcon chromosomes 1-13 and Z were numbered accordingly to Nishida et al. 2008. The remaining chromosomes were numbered by decreasing combined length of the placed PCFs. Triangles above the falcon chromosomes point to the positions of falcon-specific fusions and below chromosomes demarcate the positions of fissions. Black filling within the triangles point to the EBR boundaries used in the CNE analysis.

**Figure 4.** Average fraction of bases within conserved non-coding elements (CNEs) in avian EBRs and two flanking regions upstream (-) and downstream (+).

**TABLES****Table 1.** Scaffold-based RACA assemblies for peregrine falcon and pigeon.

Statistics	Peregrine falcon			Pigeon		
	Scaffold assembly	Default RACA	Adjusted RACA <sup>1</sup>	Scaffold assembly	Default RACA	Adjusted RACA <sup>1</sup>
No. scaffolds ( $\geq 10$ kb)	723	478	478	1,081	572	572
No. PCFs	NA	113	93	NA	150	137
Total length (Gb)	1.17	1.14	1.14	1.10	1.07	1.07
N50 (Mb)	3.94	27.44	25.82	3.15	34.54	22.17
Fraction of scaffold assembly (%)	NA	97.17	97.17	NA	95.86	95.86
No. scaffolds split by RACA	NA	72 (15.06 <sup>2</sup> )	15 (3.14 <sup>2</sup> )	NA	78 (13.64 <sup>2</sup> )	20 (3.50 <sup>2</sup> )

<sup>1</sup>RACA assembly after the use of adjusted coverage thresholds and post-processing of scaffolds verified by PCR.

<sup>2</sup>Percentage of all scaffolds included in the RACA assembly.



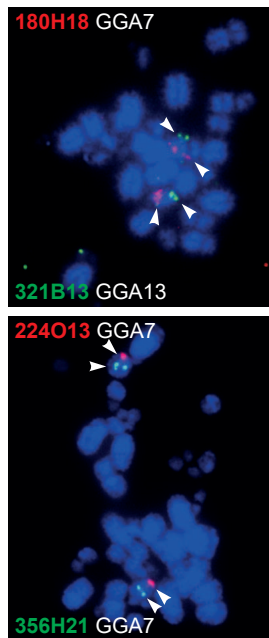
**Table 2.** Comparison of zoo-FISH success rate for random and selected set of BAC clones.

	Chicken BAC clones				Zebra finch BAC clones			
	Divergence time (MY)	Success rate (%)			Divergence time (MY)	Success rate (%)		
		Random set N = 53	Selected set N = 99	Ratio		Random set N = 48	Selected set N = 24	Ratio
Chicken	NA	NA	NA	NA	89	58.33	75.00	1.29
Turkey	28	88.68	100.00	1.13	89	54.17	83.33	1.54
Pigeon	89	26.42	91.92	3.48	69	68.75	70.83	1.03
Peregrine falcon	89	47.17	93.94	1.99	60	93.75	91.67	0.98
Zebra finch	89	20.75	90.91	4.38	NA	NA	NA	NA

Divergence times are the average of the times reported on the ExaML TENT topology from Jarvis et al. 2014.

**Table 3.** Statistics for the chromosome assemblies of peregrine falcon and pigeon.

<b>Statistics</b>	<b>Peregrine falcon</b>	<b>Pigeon</b>
No. informative BAC clones	177	151
No. PCFs placed on chromosomes	57	60
Combined length (Gb)	1.03	0.91
PCF assembly coverage (%)	90.03	85.23
Scaffold assembly coverage (%)	87.55	81.70
No. oriented PCFs	32	26
Combined length (Mb)	888.67	687.59

**A****B****224O13**

34L13

112D24

56K7

**180H18**

186K14

**356H21**

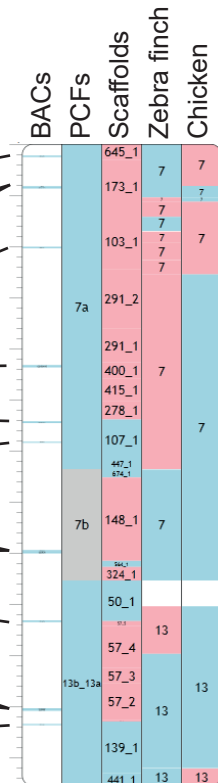
38E18

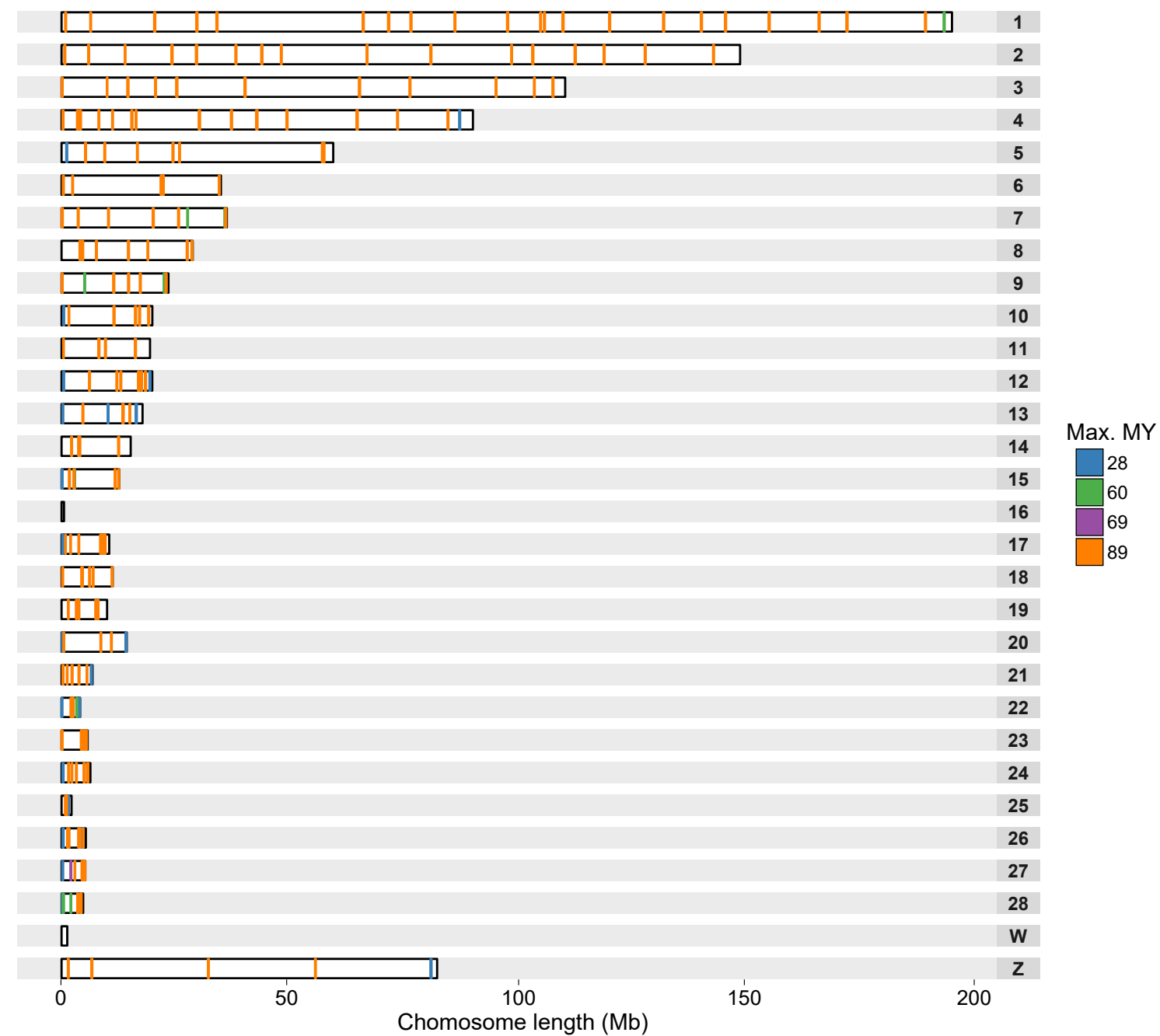
81K22

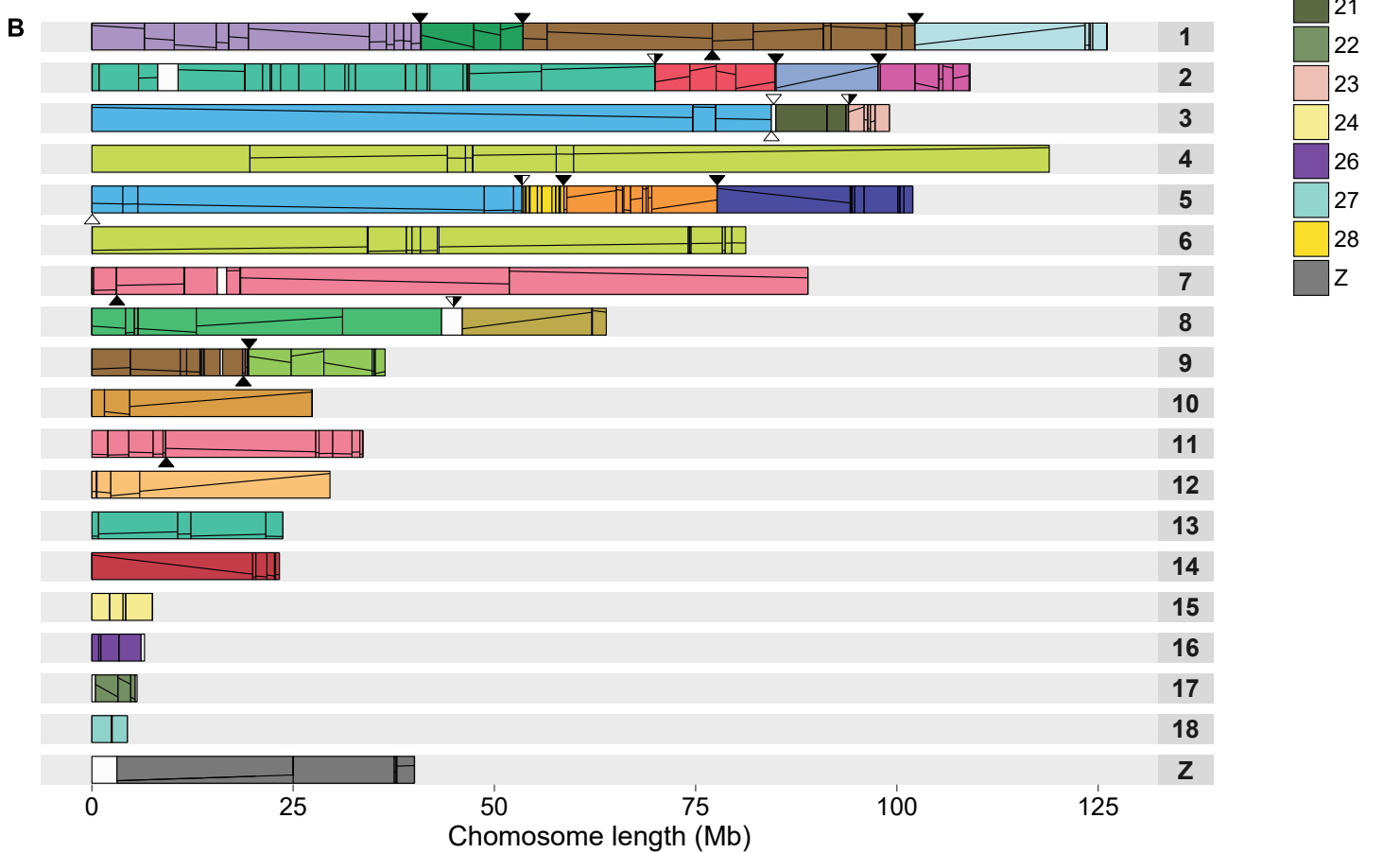
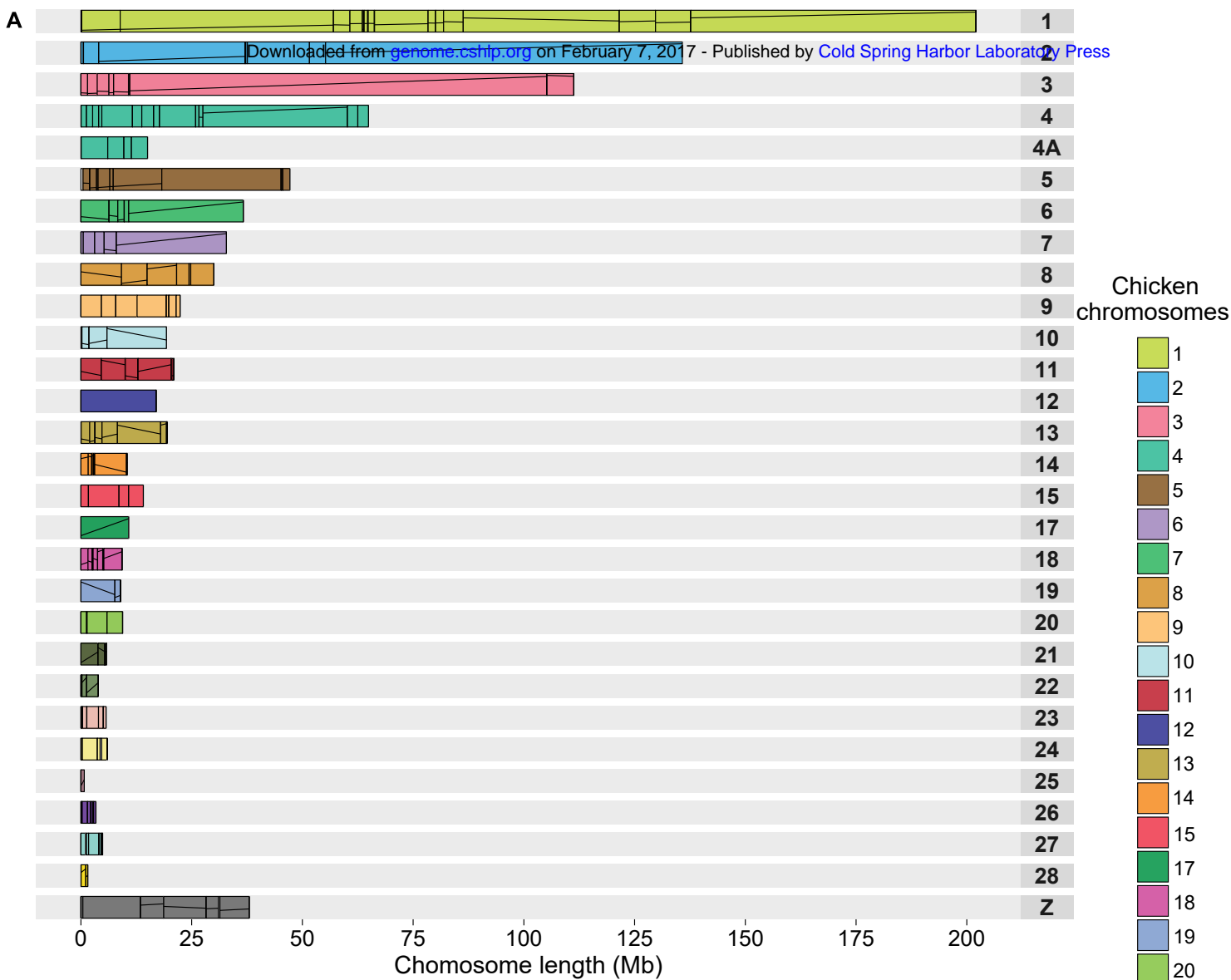
266O5

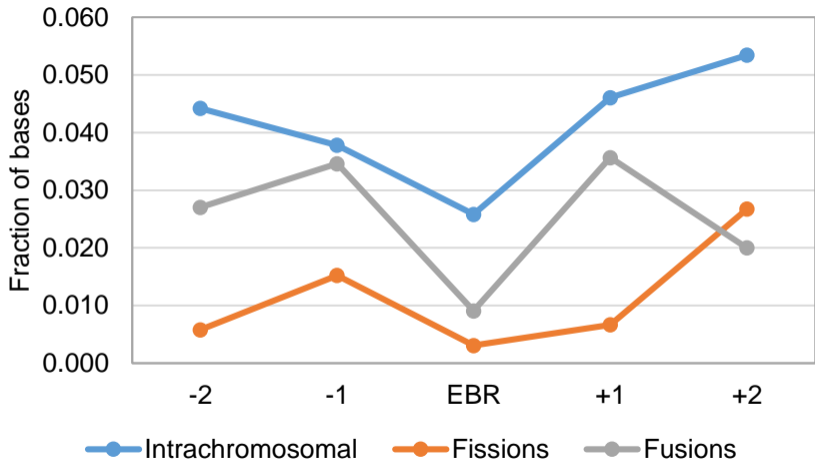
115I12

59M8

**321B13****C**









## Upgrading short read animal genome assemblies to chromosome level using comparative genomics and a universal probe set

Joana Damas, Rebecca O'Connor, Marta Farré, et al.

*Genome Res.* published online November 30, 2016  
Access the most recent version at doi:[10.1101/gr.213660.116](https://doi.org/10.1101/gr.213660.116)

---

<b>P&lt;P</b>	Published online November 30, 2016 in advance of the print journal.
<b>Accepted Manuscript</b>	Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.
<b>Open Access</b>	Freely available online through the <i>Genome Research</i> Open Access option.
<b>Creative Commons License</b>	This manuscript is Open Access. This article, published in <i>Genome Research</i> , is available under a Creative Commons License (Attribution 4.0 International license), as described at <a href="http://creativecommons.org/licenses/by/4.0/">http://creativecommons.org/licenses/by/4.0/</a> .
<b>Email Alerting Service</b>	Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or <a href="#">click here</a> .

---

---

Advance online articles have been peer reviewed and accepted for publication but have not yet appeared in the paper journal (edited, typeset versions may be posted when available prior to final publication). Advance online articles are citable and establish publication priority; they are indexed by PubMed from initial publication. Citations to Advance online articles must include the digital object identifier (DOIs) and date of initial publication.

---

To subscribe to *Genome Research* go to:  
<http://genome.cshlp.org/subscriptions>

---