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

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

[1]Department of Comparative Biomedical Sciences, Royal Veterinary College, Royal College Street, University of London, London, NW1 0TU, UK

[2]School of Biosciences, University of Kent, Canterbury, CT2 7NY, UK

[3]Institute of Biological, Environmental and Rural Sciences, Aberystwyth University, Aberystwyth, SY23 3DA, UK

[4]School of Human and Life Sciences, Canterbury Christ Church University, Canterbury, CT1 1QU, UK

*Joint first authors

**Joint last authors

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

Running title: Building chromosome-level assemblies for animals

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

ABSTRACT

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

INTRODUCTION

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

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

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

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

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

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

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

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

RESULTS

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

Construction of PCFs from fragmented assemblies

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

Verification of scaffolds essential for revealing species-specific chromosome architectures

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

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

Creation of a refined set of pigeon and falcon PCFs

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

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

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

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

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

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

Physical assignment of refined PCFs on the species' chromosomes

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

Pigeon chromosome assembly

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

Falcon chromosome assembly

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

Fate of CNEs in avian inter- and intrachromosomal EBRs

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

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

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

DISCUSSION

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

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

such as has recently been demonstrated in the B chromosomes of two deer species (Makunin et al. 2016)

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

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

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

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

METHODS

Avian genome assemblies, repeat masking and gene annotations

The chicken (ICGSC Gallus_gallus 4.0; Hillier 2004), zebra finch (WUGSC 3.2.4; Warren et al. 2010), and turkey (TGC Turkey_2.01; Dalloul et al. 2010) chromosome assemblies were downloaded from the UCSC Genome Browser (Kent et al. 2002). The collared flycatcher (FicAlb1.5; Ellegren et al. 2012) genome was obtained from NCBI. Scaffold-based (N50>2 Mb) assemblies of pigeon, falcon, and 16 additional avian genomes were provided by the Avian Phylogenomics Consortium (Zhang et al. 2014a). All sequences were repeat-masked using Window Masker (Morgulis et al. 2006) with *-sdust* option and Tandem Repeats Finder (Benson 1999). Chicken gene (version of 27/04/2014) and repetitive sequence (version of 11/06/2012) annotations were downloaded from the UCSC genome browser (Rosenbloom et al. 2015). Chicken genes with a single ortholog in the human genome were extracted from Ensembl Biomart (v.74; Kinsella et al. 2011).

Pairwise and multiple genome alignments, nucleotide evolutionary conservation scores and conserved elements

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

Reference-assisted chromosome assembly of pigeon and falcon genomes

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

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

PCR testing of adjacent SFs

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

BAC clone selection

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

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

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

Classification tree

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

Cell culture and chromosome preparation

Chromosome preparations were established from fibroblast cell lines generated from collagenase treatment of 5- to 7-day-old embryos or from skin biopsies. Cells were cultured at 40°C, and 5% CO_2 in Alpha MEM (Fisher), supplemented with 20% Fetal Bovine Serum (Gibco), 2% Pen-Strep (Sigma) and 1% L-Glutamine (Sigma). Chromosome suspension preparation followed standard protocols, briefly mitostatic treatment with colcemid at a final

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

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

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

Fluorescence *in-situ* hybridization (FISH)

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

EBR detection and CNE density analysis

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

Comparing CNE densities in EBRs and msHSBs

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

Densities of TEs in falcon intrachromosomal EBRs, fusions and fissions

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

DATA ACCESS

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

DISCLOSURE DECLARATION

Authors report no conflict of interests.

ACKNOWLEDGMENTS

The authors would like thank Dr. Yu-Mei Chang for the assistance with statistical analyses. This work was supported in part by the Biotechnology and Biological Sciences Research 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, Potamousis 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 FISHof 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[1]	Scaffold assembly	Default RACA	Adjusted RACA[1]
No. scaffolds (≥ 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[2])	15 (3.14[2])	NA	78 (13.64[2])	20 (3.50[2])

[1]RACA assembly after the use of adjusted coverage thresholds and post-processing of scaffolds verified by PCR.
[2]Percentage of all scaffolds included in the RACA assembly.

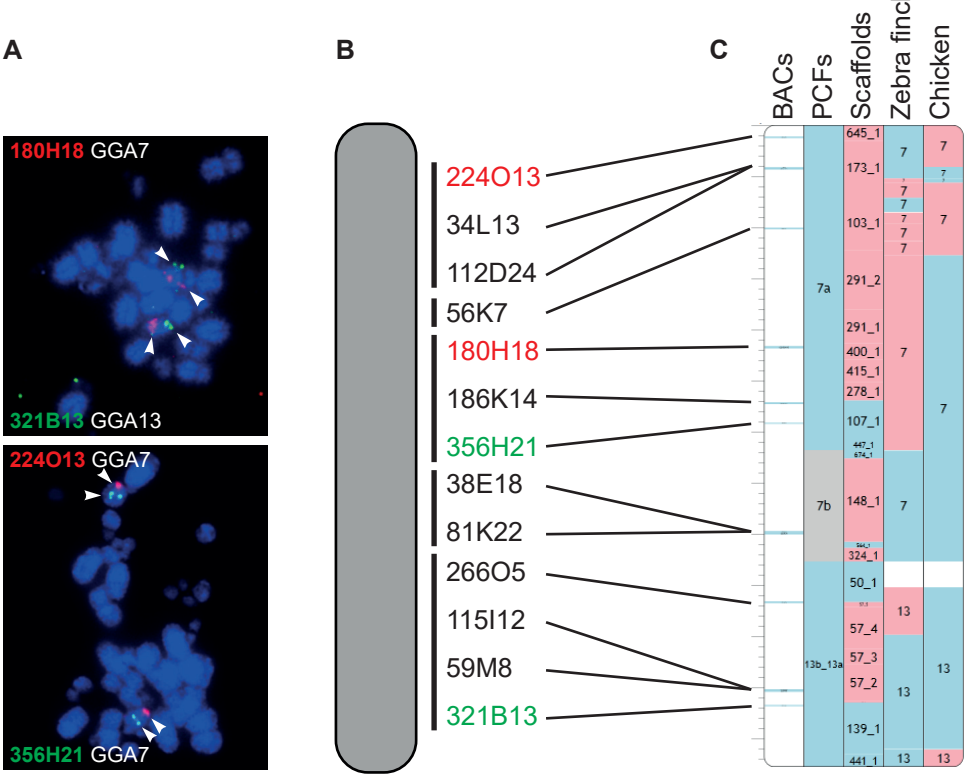
Table 2. Comparison of zoo-FISH success rate for random and selected set of BAC clones.

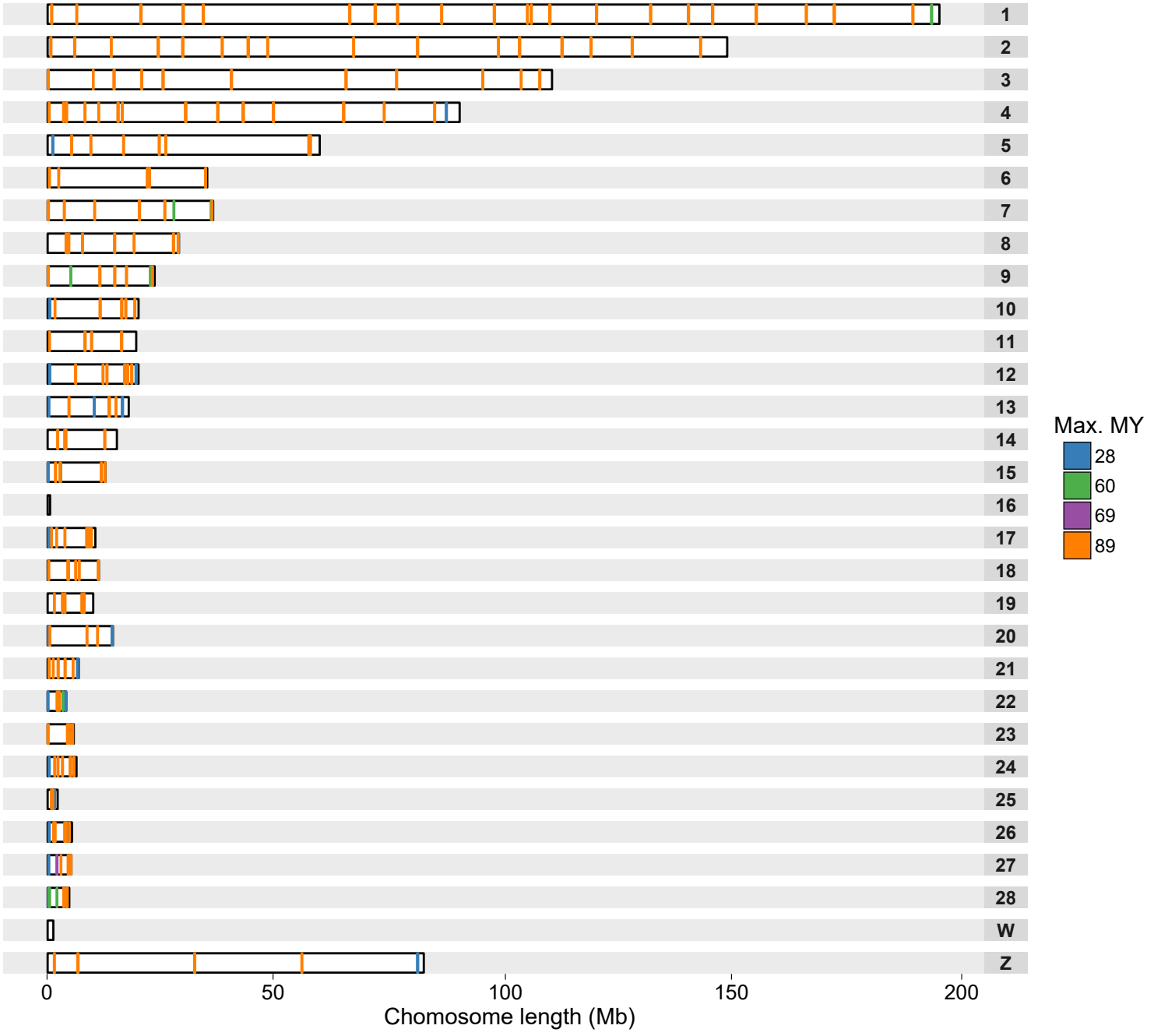
	Chicken BAC clones				Zebra finch BAC clones			
		Success rate (%)				Success rate (%)		
	Divergence time (MY)	Random set N = 53	Selected set N = 99	Ratio	Divergence time (MY)	Random set N = 48	Selected set N = 24	Ratio
Chicken	NA	NA	NA	NA	89	58.33	75.00	1.29
Turkey	28	88.68	100.00	1.13	89	54.17	83.33	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

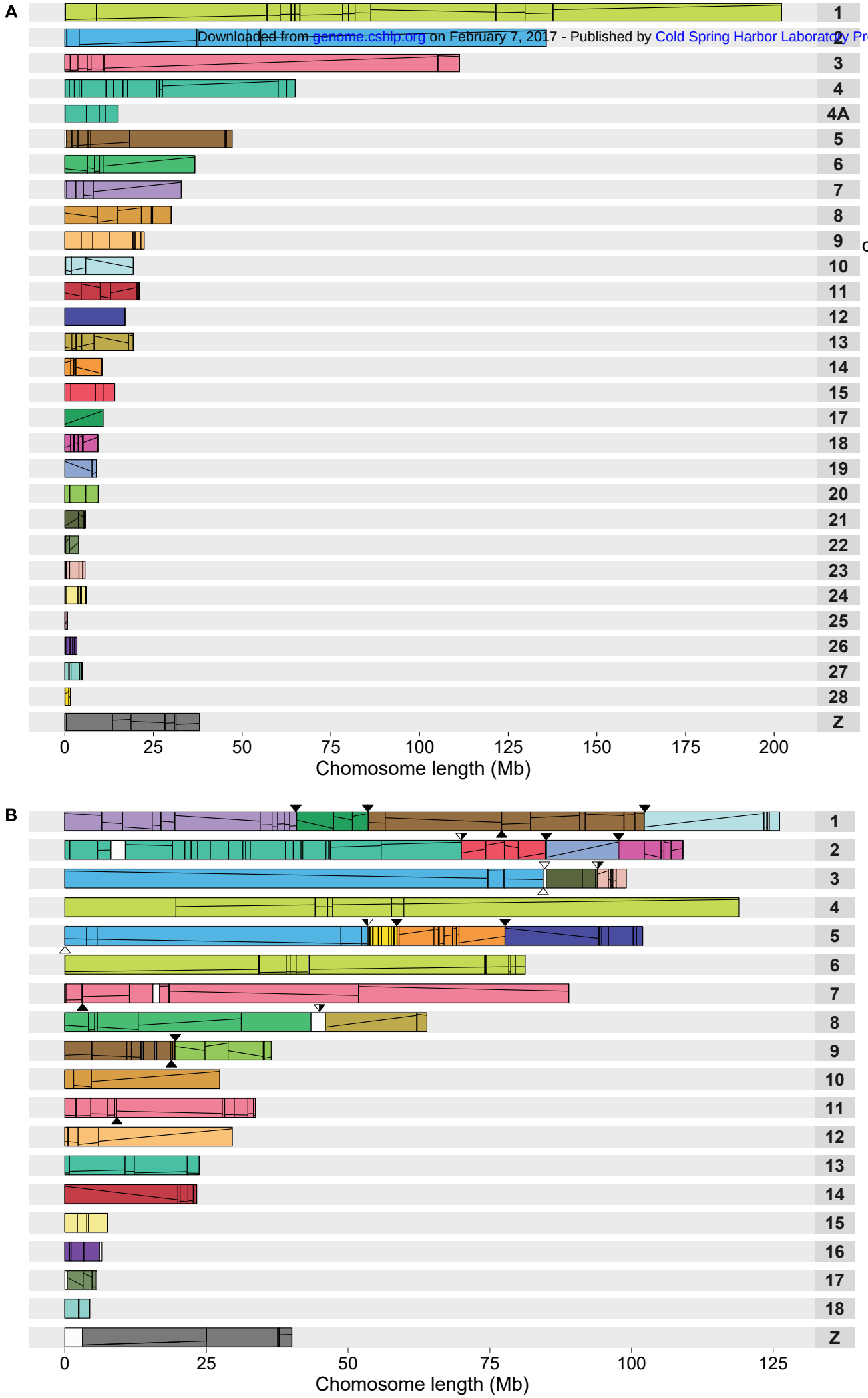
Divergence times are the average of the times reported on the ExaML TENT topology from Jarvis et al. 2014.

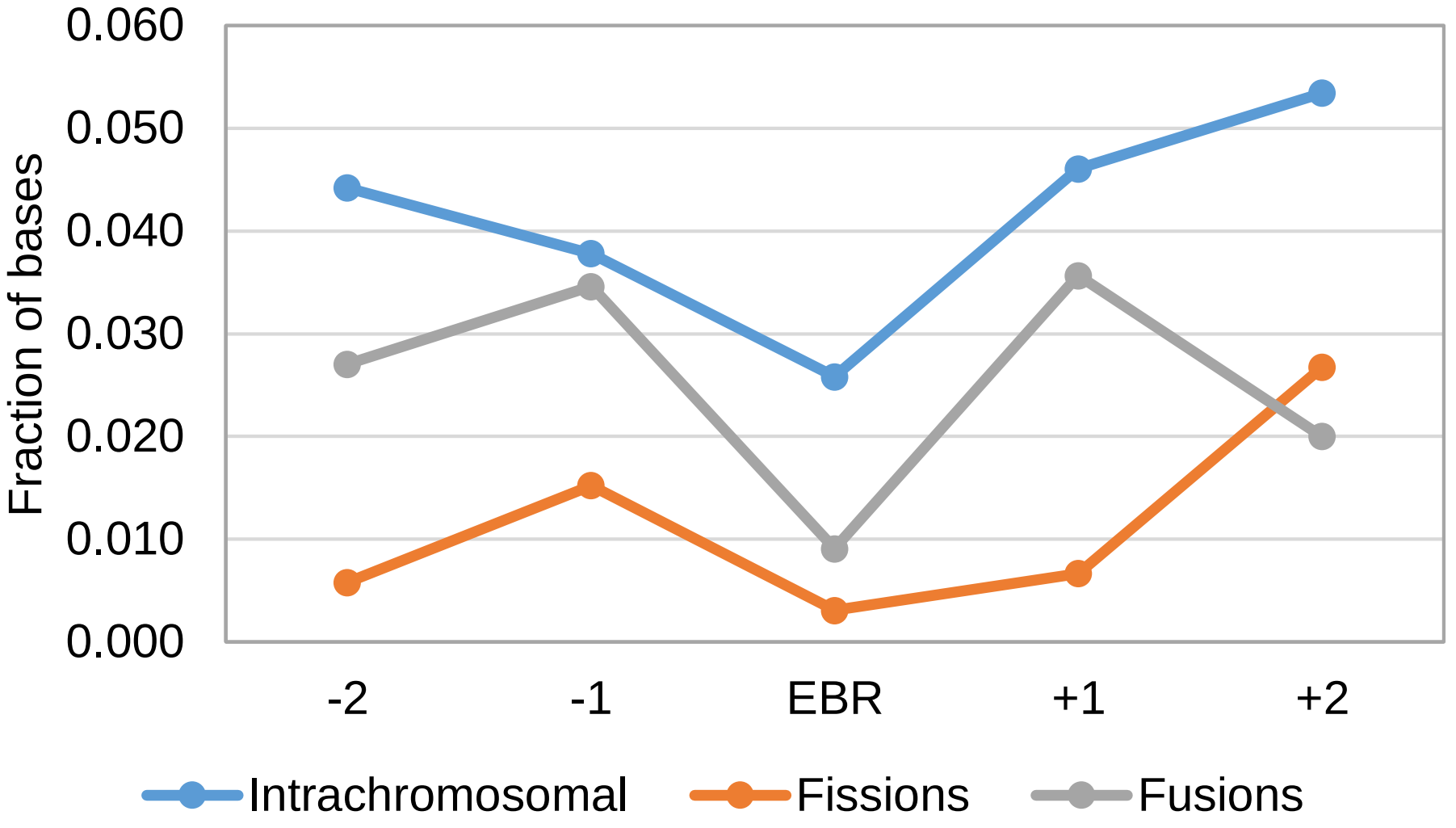
Table 3. Statistics for the chromosome assemblies of peregrine falcon and pigeon.

Statistics	Peregrine falcon	Pigeon
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
180H18 GGA7
321B13 GGA13
224O13 GGA7
356H21 GGA7
B
224O13
34L13
112D24
56K7
180H18
186K14
356H21
38E18
81K22
266O5
115I12
59M8
321B13
C
BACs
PCFs
Scaffolds
Zebra finch
Chicken
7a
7b
13b_13a
645_1
173_1
103_1
291_2
291_1
400_1
415_1
278_1
107_1
447_1
674_1
148_1
324_1
50_1
57_4
57_3
57_2
139_1
441_1


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
W
Z
Max. MY
28
60
69
89
0
50
100
150
200
Chomosome length (Mb)




0.060
0.050
0.040
0.030
0.020
0.010
0.000
Fraction of bases
-2
-1
EBR
+1
+2
Intrachromosomal
Fissions
Fusions

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