Avian genomics:
from improvement of sequenced genomes to ancestral karyotypes and novel patterns of chromosomal evolution

Joana Daniela Mendes Damas

The Royal Veterinary College

University of London

A thesis submitted to the University of London in accordance with the requirements for the degree of Doctor of Philosophy in the Department of Comparative Biomedical Sciences, The Royal Veterinary College.

Declaration

This thesis represents the work of the author except where acknowledged. The views expressed in this thesis are those of the author and not necessarily those of the University.

I certify that:

1. The thesis being submitted for examination is my own account of my own research;
2. My research has been conducted ethically;
3. The data and results presented are the genuine data and results actually obtained by me during the conduct of the research;
4. Where I have drawn on the work, ideas and results of others this has been appropriately acknowledged in the thesis;
5. Where any collaboration has taken place with other researchers, I have clearly stated in the thesis my own personal share in the investigation;
6. The greater portion of the work described in the thesis has been undertaken subsequent to my registration for the higher degree for which I am submitting for examination;
7. The thesis submitted is within the required word limit as specified by the RVC.

_____

_____

Joana Damas

Abstract

Genomes assembled to chromosome-level are critical for the study of various aspects of evolutionary and applied genomics. Regrettably, genome assemblies produced with next-generation sequencing (NGS) techniques comprise thousands of short DNA fragments (scaffolds) instead of one scaffold per chromosome, limiting their use to the analysis of sequence variation within scaffolds. Traditionally, upgrading NGS assemblies to chromosome-level implies integration with independent data from traditional genome mapping projects. However, due to their significant time and cost requirements, these data do not exist for most newly sequenced NGS genomes. To overcome these problems, we developed a novel, inexpensive and transferable approach to upgrade fragmented genome assemblies to chromosome-level. Our model for the development of this method were avian genomes. Birds are important organisms for agricultural, cultural and environmental reasons. Nonetheless, due to the lack of chromosome assemblies for this Class, avian chromosome evolution is a relatively understudied topic. The recent release of 45 new avian genomes opened the door for a large-scale study of avian chromosome evolution; however, most of these genomes were only assembled to the scaffold-level. By application of our newly developed genome assembly methodology, we upgraded the genome assemblies of 18 of these genomes, out of which two to the chromosome-level (rock pigeon and peregrine falcon). These chromosome assemblies allowed the detection of intra- and interchromosomal rearrangements not previously described for avian genomes. The first large-scale analysis of interchromosomal evolutionary breakpoint regions (EBRs) revealed that avian interchromosomal EBRs locate in regions of low density of DNA conserved non-coding elements (CNEs) and that chromosomal fission sites are further limited to long CNE “deserts”. Additionally, the reconstruction of avian ancestral karyotypes provided valuable insights into the mechanisms behind the generation of chromosome rearrangements and their fixation in the avian lineage.

Table of contents

Declaration **3**
Abstract **5**
Table of contents **6**
List of tables **8**
List of figures **10**
List of appendices **12**
Acknowledgements **14**
Abbreviations **15**
1 General introduction **19**
1.1 Genomics, the science of genomes **21**
1.2 Genome evolution **22**
1.2.1 Genome size and composition 22
1.2.2 Genome organization 24
1.2.3 Origin of genome rearrangements 30
1.2.4 Implications of genome rearrangements 32
1.3 Obtaining genome sequences **35**
1.3.1 Sanger sequencing 35
1.3.2 Short-read next-generation sequencing 36
1.3.3 Long-read sequencing 42
1.3.4 Genome assembly 48
1.4 Ancestral karyotype reconstructions **58**
1.5 Birds **61**
1.5.1 Genome size 61
1.5.2 Karyotype structure 62
1.5.3 Phylogeny 64
1.5.4 Molecular evolution 67
1.5.5 Adaptive phenotypes 68
1.5.6 Chromosome evolution 71
1.6 Project aims **74**
2 Constructing avian predicted chromosome fragments **77**
2.1 Background **79**
2.2 Material and methods **84**
2.3 Results **90**

2.4 Discussion 115
3 Constructing avian chromosome-level assemblies 119
3.1 Background 121
3.2 Material and methods 123
3.3 Results 129
3.4 Discussion 143
4 Reconstructing avian ancestral karyotypes 147
4.1 Background 149
4.2 Material and methods 152
4.3 Results 157
4.4 Discussion 170
5 General discussion 173
5.1 Upgrading fragmented genome assemblies 175
5.2 Are avian genomes stable? 176
5.3 Avian genome evolution: the role of repetitive sequences 177
5.4 Avian genome evolution: advantage in maintaining synteny 178
5.5 Avian interchromosomal stability 179
5.6 Future directions 182
6 Bibliography 185
7 Appendices 207

List of tables

Table 1-1: Comparison of sequencing technologies. 47
Table 2-1: Chromosome number and scaffold assemblies' statistics (scaffolds ≥ 10 Kbp) for the 18 selected avian species. 89
Table 2-2: Reference selection for species diverged more than 67 MYA from both chicken and zebra finch. 92
Table 2-3: Statistics of the PCF reconstructions using RACA avian default parameters.....95
Table 2-4: Statistics for the scaffold split regions tested by PCR. 99
Table 2-5: Statistics of the refined PCF assemblies obtained for the 18 avian species.....102
Table 2-6: Number of intrachromosomal and fusion flanking lineage-specific EBRs detected from the PCF assemblies, and chicken and zebra finch genome assemblies.....108
Table 2-7: Significant comparisons of CNE densities for avian lineage-specific EBRs and their four adjacent intervals (±2) of the same size. 109
Table 2-8: Significant comparisons for the number of CNE bases in 1 Kbp windows overlapping avian EBRs, msHSBs, and genome-wide.....109
Table 2-9: Number of budgerigar, downy woodpecker and peregrine falcon EBRs used in the TE analysis.....110
Table 2-10: Comparison and ratios of the number of bases from populated TEs in 10 Kbp intervals overlapping peregrine falcon EBRs and the rest of the genome.....111
Table 2-11: Comparison and ratios of the number of bases from populated TEs in 10 Kbp intervals overlapping downy woodpecker EBRs and the rest of the genome.....111
Table 2-12: Comparison and ratios of the number of bases from populated transposable elements in 10 Kbp intervals overlapping budgerigar EBRs and the rest of the genome.....112
Table 3-1: DNA sequence feature information returned from the BAC clone analysis pipeline. 128
Table 3-2: Comparison of zoo-FISH success rate for random and selected set of BAC clones.....130

Table 3-3: Avian BAC clones and expected FISH success rates for the phylogenetically distant species (divergence time ≥69 MY). 130
Table 3-4: Statistics for the chromosome assemblies of the peregrine falcon and rock pigeon. 133
Table 3-5: Peregrine falcon and rock pigeon lineage-specific EBRs. 136
Table 3-6: Differences in TE densities in 10 Kbp intervals overlapping peregrine falcon EBRs and the rest of the peregrine falcon genome. 138
Table 3-7: Differences in TE densities in 10 Kbp intervals overlapping rock pigeon EBRs and the rest of the rock pigeon genome 139
Table 3-8: Significant differences in CNE densities in avian lineage-specific EBRs and their four adjacent intervals (±2) of the same size. 139
Table 3-9: Statistics for CNE density in 1 Kbp windows for avian EBRs, msHSBs, and genome-wide. 140
Table 3-10: Significant differences for distances in number of 1 Kbp windows between zero and high CNE density windows. 140
Table 4-1: Statistics for genome assemblies of descendant and outgroup species. 156
Table 4-2: Statistics for the Neognathae ancestor reconstructions at different resolutions of SF detection. 161
Table 4-3: Statistics of the reconstructed ancestors (100 Kbp resolution). 161
Table 4-4: Number of EBRs and EBR rates for the reconstructed ancestral genomes. 163
Table 4-5: EBR distribution and fraction within genes, CNEs, and TEs for each Avian ancestor chromosome 165

List of figures

Figure 1-1: Genome sizes of different organisms. 24
Figure 1-2: Cytogenetic techniques in the study of evolution and disease..... 26
Figure 1-3: Types of chromosomal rearrangements. 28
Figure 1-4: Karyotypes of (A) Indian muntjac and (B) red vizcacha rat, (C) chicken and (D) Japanese four-striped rat snake 30
Figure 1-5: Non-allelic homologous recombination (NAHR). 32
Figure 1-6: Dye-terminator sequencing methodology. 36
Figure 1-7: Illumina sequencing approach. 38
Figure 1-8: Single-molecule real-time sequencing approaches: (A) Pacific Biosciences and (B) Oxford Nanopore..... 44
Figure 1-9: De novo whole-genome sequencing assembly. 48
Figure 1-10: Hi-C methodology. 54
Figure 1-11: Overview of the RACA algorithm 57
Figure 1-12: Chicken karyotype 63
Figure 1-13: Genome-scale avian phylogeny. 66
Figure 1-14: Comparison of avian phylogenies reported by (Jarvis et al., 2014) and (Prum et al., 2015) 67
Figure 2-1: Overview of RACA algorithm 83
Figure 2-2: Sequencing and physical coverage. 86
Figure 2-3: Cladogram presenting the selected reference for each RACA reconstruction. 91
Figure 2-4: PCR verification strategy. 96
Figure 2-5: Physical coverage threshold establishment. 97
Figure 2-6: Continuity comparison for scaffold and PCF genome assemblies in chicken chromosome 3. 101
Figure 2-7: Comparison between numbers of split scaffolds in Pekin duck PCF and RH map assemblies. 104
Figure 2-8: Comparison between numbers of split scaffolds in rock pigeon PCF and Dovetail assemblies, and PCR verification results. 105
Figure 2-9: Comparison between RACA PCFs and super-scaffolds. 106
Figure 2-10: Comparison between super-scaffolds and RACA PCFs. 106
Figure 2-11: Average fraction of bases within conserved CNEs in avian EBRs and two flanking regions of the same size upstream (-) and downstream (+). 108

Figure 2-12: Anna's hummingbird PCF 17 representations on (A) Evolution Highway comparative chromosome browser and (B) UCSC Genome Browser hub. 114
Figure 3-1: Methodology for the placement of the PCFs on chromosomes. .. 129
Figure 3-2: Classification tree used to predict the non-successful (0) or successful hybridization (1) of a BAC clone on at least one phylogenetically distant species (divergence time ≥69 MY). 131
Figure 3-3: Distribution of universal BAC clones along chicken chromosomes.. 132
Figure 3-4: BAC clone tracks on UCSC genome browser. 133
Figure 3-5: Ideogram of rock pigeon (A) and peregrine falcon (B) chromosomes. 135
Figure 3-6: Average fraction of bases within CNEs in avian EBRs and two flanking regions upstream (-) and downstream (+). 137
Figure 3-7: GO terms enriched on lineage-specific EBRs. 141
Figure 4-1: Avian ancestor chromosomes 3, 17 and 26 representation on Evolution Highway comparative chromosome browser. 160
Figure 4-2: Phylogenetic tree of descendant species and reconstructed ancestors. 162
Figure 4-3: Correlation between the fraction of bases within CNEs, TEs and genes, and EBRs rates (EBRs per Mbp, observed-expected number of EBRs, and average EBR distance) for Avian ancestor chromosomes. 167
Figure 4-4: GO terms enriched on Avian ancestor chromosomes 17 and 22 169
Figure 5-1: Factors contributing to the generation of chromosomal aberrations in germ cells and their fixation in the avian and mammalian lineages. 182

List of appendices

Supplemental Table 1: Statistics of the Pekin duck original and RACA (2 rounds) genome assemblies. 209
Supplemental Table 2: Statistics of emperor penguin original and RACA (2 rounds) genome assemblies. 209
Supplemental Table 3: Statistics of Anna's hummingbird original and RACA (2 rounds) genome assemblies. 210
Supplemental Table 4: Statistics of chimney swift original and RACA (2 rounds) genome assemblies. 210
Supplemental Table 5: Statistics of killdeer original and RACA (2 rounds) genome assemblies. 211
Supplemental Table 6: Statistics of rock pigeon original and RACA (2 rounds) genome assemblies. 211
Supplemental Table 7: Statistics of American crow original and RACA (2 rounds) genome assemblies. 212
Supplemental Table 8: Statistics of common cuckoo original and RACA (2 rounds) genome assemblies. 212
Supplemental Table 9: Statistics of little egret original and RACA (2 rounds) genome assemblies. 213
Supplemental Table 10: Statistics of peregrine falcon original and RACA (2 rounds) genome assemblies. 213
Supplemental Table 11: Statistics of medium ground finch original and RACA (2 rounds) genome assemblies. 214
Supplemental Table 12: Statistics of the golden-collared manakin original and RACA (2 rounds) genome assemblies. 214
Supplemental Table 13: Statistics of the budgerigar original and RACA (2 rounds) genome assemblies. 215
Supplemental Table 14: Statistics of the crested ibis original and RACA (2 rounds) genome assemblies. 215
Supplemental Table 15: Statistics of the hoatzin original and RACA (2 rounds) genome assemblies. 216
Supplemental Table 16: Statistics of the downy woodpecker original and RACA (2 rounds) genome assemblies. 216

Supplemental Table 17: Statistics of the Adélie penguin original and RACA (2 rounds) genome assemblies. 217
Supplemental Table 18: Statistics of the ostrich original and RACA (2 rounds) genome assemblies. 217
Supplemental Table 19: Statistics for the super-scaffold and PCFs adjacencies comparisons..... 218
Supplemental Table 20: Statistics for the PCFs and super-scaffold adjacencies comparisons..... 219
Supplemental Table 21: Chicken genome intervals corresponding to lineage-specific intrachromosomal EBRs identified in PCFs. 220
Supplemental Table 22: Chicken genome intervals corresponding to lineage-specific interchromosomal EBRs identified in PCFs. 232
Supplemental Table 23: Chicken genome intervals corresponding to lineage-specific intrachromosomal EBRs identified in chromosome assemblies. 233
Supplemental Table 24: Chicken genome intervals corresponding to lineage-specific interchromosomal EBRs identified in chromosome assemblies. 239

Acknowledgements

About three years ago, I moved to this country not knowing how this experience would change my life both professionally and personally. It would not have been possible to perform this work and write this thesis without the help and support of the generous people around me. This thesis is dedicated to all of you!

First I would like to thank my supervisor Dr Denis Larkin. Thank you for allowing me to perform this work, for all you have taught me, and all your guidance and support during this journey. Without you, this thesis would not have been possible. *Спасибо*!

My special thanks to Dr Marta Farré. Thank you for making me feel welcome since the first day, for your advice, encouragement and for bearing with me in my "freeze" moments. Especially, thank you for your friendship. *Gràcies*!

To the present and former members of the Animal Genome and Evolution Lab and all our collaborators, in particular to Dr Laura Buggiotti, and Professor Darren Griffin and Dr Rebecca O'Connor, a very big thank you! Your helpful discussions, support and kindness, have made these three years some of the most rewarding in my life.

A very special thanks to all my family and friends back in Portugal, and all around the world. *Alguém disse que "Amizade verdadeira é aquela que o tempo não apaga, a distância não destrói, e acima de tudo o coração não esquece". Estes três anos mostraram-me que certamente nós fazemos parte da amostra usada para testar esta teoria.*

Lastly but definitely not the least, my deepest thank goes to my wonderful dad, mum and sister. *Obrigada por me deixarem voar e por me apoiarem em todas as decisões. Não há palavras que possam expressar a minha gratidão por fazerem parte da minha vida! Muito obrigada por tudo!*

Abbreviations

bp – Base Pair

BAC – Bacterial Artificial Chromosome

BES – BAC End Sequence

CAR – Contiguous Ancestral Region

CART – Classification and Regression Tree

CCS – Circular Consensus sequence

CE – Conserved Element

CGH - Comparative Genomic Hybridization

CNE – Conserved Non-Coding Element

cM – Centimorgan

CRT – Cyclic Reversible Termination

dNTP – Deoxynucleotide Triphosphate

ddNTP – Dideoxynucleotide Triphosphate

DSB – Double Strand Break

EBR – Evolutionary Breakpoint Region

EH – Evolution Highway Comparative Chromosome Browser

FISH – Fluorescence *in situ* Hybridization

FPE – Peregrine falcon chromosome

Gbp – Gigabase pair

GGA – Chicken chromosome

GL – Genetic Linkage

GO – Gene Ontology

HSB – Homologous Synteny Block

InDel – Insertion/Deletion Polymorphism

Kbp – Kilobase Pair

LCR – Low Copy Repeat

LINE – Long Interspersed Nuclear Element

LTR – Long Terminal Repeat

Mbp – Megabase pair

MY – Million Years

MYA – Million Years Ago

NAHR – Non-Allelic Homologous Recombination

NGS – Next Generation Sequencing

NHEJ – Non-Homologous End Joining

PCF – Predicted Chromosome Fragment

PCR – Polymerase Chain Reaction

RACF – Reconstructed Ancestral Chromosome Fragment

RH – Radiation Hybrid

SBS – Sequencing by Synthesis

SD – Segmental Duplication

SF – Syntenic Fragment

SINE – Short Interspersed Nuclear Element

SMRT – Single Molecule Real Time

SNA – Single Nucleotide Addition

Tbp – Terabase pair

TE – Transposable Element

TENT – Total Evidence Nucleotide Tree

TR – Tandem Repeat

ZMW – Zero-Mode Waveguide

“DNA is a biological code elegantly composed of only four letters: A, C, G and T. From this simplicity comes all the complexity of life.”

Dawn Field and Neil Davies
Biocode: the new age of genomics, 2015

1 General introduction

1.1 Genomics, the science of genomes

The term genome was introduced with reference to an organism's complete set of genes and chromosomes (Hieter and Boguski, 1997). The genome is often described as an 'information repository', carrying instructions used for the growth, development, and function of an organism, and which transmission through generations is the principal medium of inheritance of organismal traits (Goldman and Landweber, 2016). This information is contained in deoxyribonucleic acid (DNA) molecules and encrypted in the form of a four-nucleotide (adenine, thymine, cytosine and guanine) code.

In most prokaryotic cells, the genome is organised in one or more, usually circular, DNA molecules within the cell. In contrast, DNA in the nuclei of eukaryotic cells is distributed among multiple linear structures, the chromosomes. Because cellular DNA lengths are up to hundred thousand times the cell's diameter, it is essential for DNA to be condensed. Histones are the proteins responsible for organising and compacting chromosomal DNA, forming the chromatin. DNA inside interphase nuclei is organised in heterochromatic regions (highly condensed) that are believed to include transcriptionally inactive genes and euchromatic regions (lightly condensed) where most transcribed regions of DNA are found (Lodish, 2016). DNA condensation has then implications on cell architecture, gene expression, and also prevents incorrect segregation during cell division. In fact, during cell division, the chromatin is highly compacted, and chromosomes can be easily visualised by light microscopy.

Contrary to what was initially believed, genomes are much more than just the collection of an organism's genes. Indeed, non-coding regions of the genomes were found to play important roles in gene regulation and transcription. For instance, outside the 1.5% of its coding regions, 11% of the human genome was found associated with motifs in transcription factor-binding regions or high-resolution DNase footprints that are indicative of direct contact by regulatory proteins (Kellis et al., 2014). The acknowledgement of this fact led to the transition from genetics (i.e. the study of genes) to genomics (i.e. the study of genomes).

The concept of genomics was introduced 30 years ago by Thomas Roderick as *"(…) an activity, a new way of thinking about biology. (…) It encompassed*

sequencing, mapping, and new technologies. It also had the comparative aspect of genomes of various species, their evolution, and how they are related to each other." (Kuska, 1998).

The study of the genome structure – size, composition, chromosome numbers and shapes (karyotypes) – and function, both using cytogenetic techniques and genome sequences, presents us with the opportunity to unravel the mechanisms that produced the great variety of genomes that exist today. Furthermore, the elucidation of the structure of genomes and the function of the elements they encode will provide the foundation for many areas of the biological sciences to develop on (Collins et al., 2003).

1.2 Genome evolution

As mentioned above, the comparison of genome structures gives valuable insights regarding how genomes evolved. The structure of a genome can be assessed from multiple perspectives: its size, composition (e.g., repetitive and non-repetitive DNA content, coding and non-coding DNA content) and even the structure of karyotypes (e.g., chromosome number and appearance). Still, whatever aspect of a genome one is interested in, our understanding is enhanced through comparisons to closely and distantly related genomes (Miller et al., 2004). Using comparative methods, we can gain knowledge of which, how and when genomic changes occurred during evolution.

1.2.1 Genome size and composition

Genome size is characteristic of a species and at the same time distinguishes different taxa. Because it is more straightforward to characterise than the genome composition, structure, or organisation, the haploid DNA content of an organism's cell (C-value; given in picograms (pg)) was usually one of the first genomic features to be measured, using, for instance, flow cytometry. This value could then be approximated to genome size as one pg ≈ one gigabase pair (Gbp). Nowadays, genome size can also be estimated directly from sequencing data using k-mer frequency-based approaches (Hozza et al., 2015, Ekblom and Wolf, 2014). The significance of genome size research goes from the biological and evolutionary significance of genome size variation (Gregory, 2005, Jeffery and

Gregory, 2014) to setting the context for the analysis of genome composition, being a major consideration when choosing targets for complete genome sequencing projects, and being an important factor for the estimation of the amount of data needed for a specific sequencing project (Ekblom and Wolf, 2014).

We now have access to genome size estimates for more than 15,000 species of animals, plants, and fungi (Gregory et al., 2007). These data show that genome size varies enormously among taxonomical groups, as can be seen in Figure 1-1. We can find the smallest genomes in viruses, encoding less than ten genes, and the largest in eukaryotes, that have billions of base pairs of DNA encoding tens of thousands of genes (Pevsner, 2015). Prokaryote (bacteria and archaea) genomes are mainly comprised of protein-coding sequences (85-95%) and the variation in their size was found strongly associated with gene content (Lynch, 2007, Gregory, 2005). Moreover, bacteria with larger genome sizes present more complex behaviours and phenotypes, such as participating in multicellular interactions or differentiation processes (Pevsner, 2015), which led to the hypothesis that genome size was directly correlated with organismal complexity. This pattern is not found in eukaryotes (C-value paradox). In fact, protists can present larger genomes than many animals (Figure 1-1), and the number of genes in a genome does not scale with its size. The C-value paradox was partially explained by the observation that eukaryote genomes were filled with a large amount of non-coding DNA sequences and that these were heavily involved in genome size variation within this taxon (Pevsner, 2015, Elliott and Gregory, 2015, Gregory, 2005). However, important questions regarding the genomic fractions, types and functions of these non-coding elements emerged at the same time (Elliott and Gregory, 2015). Some valuable insights were already acquired through the comparison of genome sequences. For instance, some of the smallest genomes amongst mammals are found in bats, which together with birds are the only true volant vertebrates. In fact, bat genomes present less repetitive ribosomal DNA genes, heterochromatin, and microsatellites than those of other mammals (Van den Bussche et al., 1995, Zhang and Edwards, 2012). In addition, avian genomes possess shorter introns than mammals and non-avian reptiles (Zhang and Edwards, 2012) and have fewer repetitive DNA sequences (Zhang et al., 2014b). For these organisms (bats and birds), it is argued that there are

mechanisms at work to restrict genome size and that these might be related to the high metabolic demand of powered flight (Hughes and Friedman, 2008, Gregory, 2002b). Nonetheless, our understanding of the forces shaping genome evolution is still not complete and questions such as “What types and proportions of non-coding DNA sequences are found in different eukaryotic genomes? How do these elements accumulate and/or are lost over evolutionary timescales? What effects or functions, do these elements have? Why do some genomes remain (or become) compact, while others reach large sizes?” are still targets of active investigation (Elliott and Gregory, 2015).

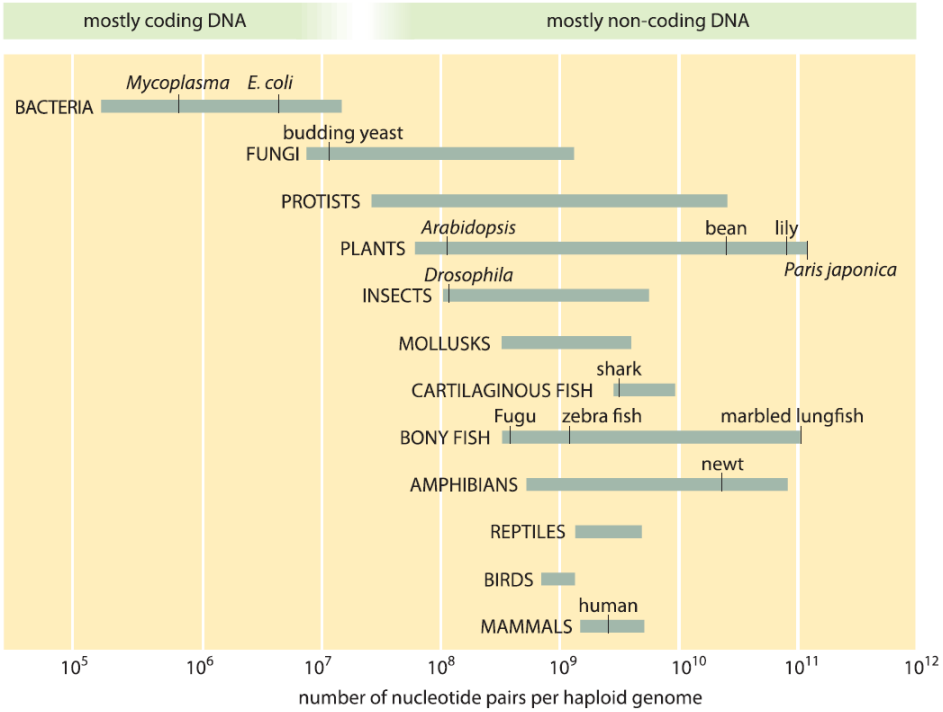

Figure 1-1: Genome sizes of different organisms (Milo and Phillips, 2016).

1.2.2 Genome organization

As previously referred, in eukaryote cells the DNA is packaged into chromosomes. The number and appearance of chromosomes in the cell nucleus are called the karyotype, which is uniform among the members of a species, with some exceptions. Because chromosomes are dynamic across evolutionary timescales, generations, populations, and even within individual lifetimes, karyotypes are not completely immutable (Pevsner, 2015). In that way, the study and comparison of karyotypes allow the detection of changes in number and shapes of chromosomes that can be associated with disease and/or give insights into the evolution of genomes.

1.2.2.1 Methods to detect karyotipical changes

Karyotyping, i.e. chromosome analysis, started with the staining of metaphase chromosomes, and subsequent microscopic observation. The differential staining levels of chromatin produce a series of landmarks along chromosomes, in the form of light and dark stripes, which allow the distinction of adjacent genomic segments and the recognition of individual chromosomes by their unique banding patterns. Moreover, distinct chromosome treatments allow the visualisation of different aspects of chromosomes. For example, the commonly used G-banding differentially stain regions of heterochromatin (dark bands) and euchromatin (light bands), while T-banding depicts the telomeric regions (Schreck and Disteche, 2001). These banding techniques can be used to detect breakpoint regions and chromosomes involved in translocations, as well as, deletions, insertions and inversions within individual chromosomes (Figure 1-2A).

Fluorescence *in situ* hybridization (FISH) allows a higher resolution analysis of karyotypes than chromosome banding. This technique provided powerful new tools and transformed cytogenetics into a molecular science. FISH uses specifically selected DNA probes that will hybridise to a region of interest on the chromosome. Nowadays, the probes are often obtained from bacterial artificial chromosomes (BACs), which are engineered DNA molecules used to clone a DNA sequence of interest in bacterial cells. The BAC clones are fluorescently labelled, and their location can be depicted through the observation with a fluorescence microscope (Speicher and Carter, 2005).

Chromosome painting, one of FISH variations, uses probe libraries obtained from a whole chromosome or a sub-chromosomal region. This approach is extremely valuable for the study of genome evolution as it allows the detection of homologous regions between chromosomes of different species (Figure 1-2B) and can also be used to detect karyotipical aberrations associated with disease phenotypes within individuals of the same species (Figure 1-2C) (Sharma and Sharma, 2001, Speicher and Carter, 2005).

In addition, comparative genomic hybridization (CGH) can be used to detect small quantitative DNA differences between samples (Weiss et al., 1999). This technique is based on competitive FISH and can be used to detect variation within

and between species. It involves the labelling, with different fluorescent dyes, of DNA from the two samples to be compared. These DNAs are denatured and hybridised together to a normal metaphase spread of chromosomes in a 1:1 ratio. Differences of fluorescent signals from one of the two compared samples indicate gain (increased signal) or loss (decreased signal) of genetic material in that particular genomic region (Weiss et al., 1999). Advances in this technology led to the development of microarrays (or array-CGH) in which the two compared samples are hybridised to a target such as BAC DNA printed on a glass slide. This technique increases the resolution of CGH. However, it is still limited to the detection of unbalanced chromosomal rearrangements (i.e. deletions or insertions), as balanced rearrangements (e.g., reciprocal translocation) do not affect the overall amount of DNA.

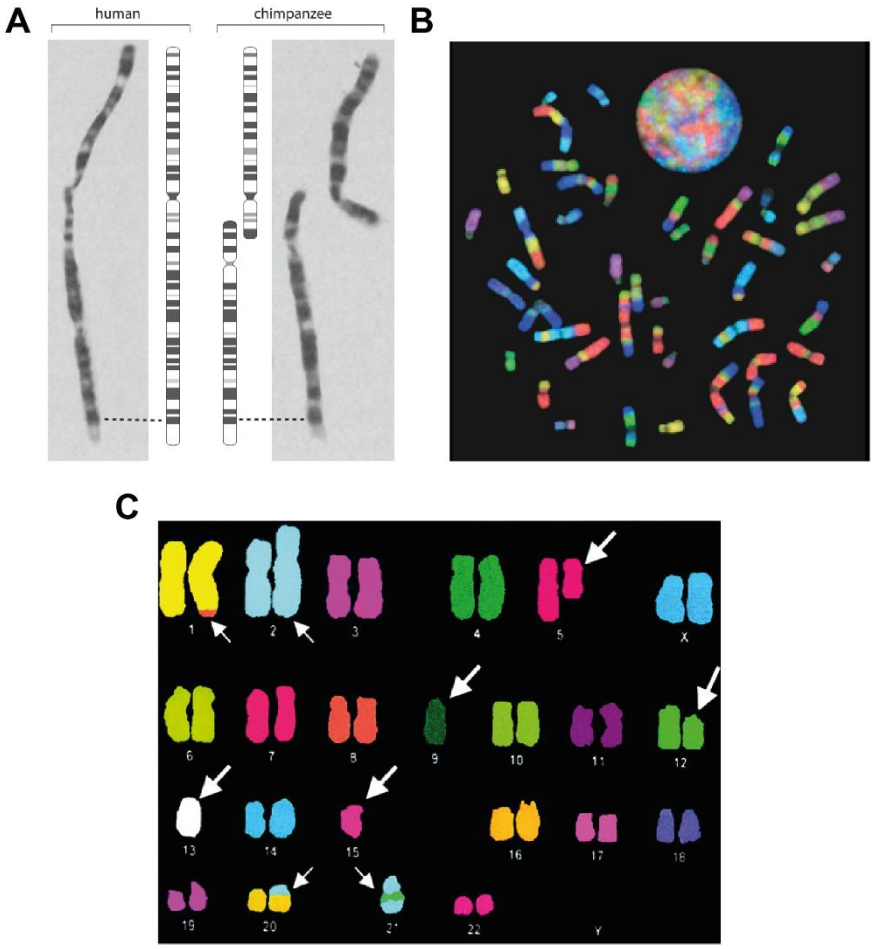

Figure 1-2: Cytogenetic techniques in the study of evolution and disease. (A) Chromosome banding revealing the homology of human chromosome 2 and chimpanzee chromosomes 2A and 2B (Yunis and Prakash, 1982). (B) Chromosome painting of gibbon probes on human chromosomes (Ferguson-Smith and Trifonov, 2007). (C) Chromosome painting depicting the karyotype of a patient with acute leukaemia; arrows show chromosomal abnormalities (Leroux et al., 2002).

1.2.2.2 Karyotype diversity in evolution

As mentioned above, chromosomes are dynamic entities that can change, among others, across evolutionary timescales (Pevsner, 2015). When karyotype alterations occur in germ cells, they have the potential to be fixed during evolution and result in distinct karyotype organisation between species. In this way, the comparison of species karyotypes using cytogenetic methodologies provided the first insights into genome evolution (Murphy et al., 2005). These comparisons reveal valuable information that is useful to reconstruct the evolutionary relationship between species, to understand the evolutionary forces that shape genomes, and the mechanisms that lead to the appearance of novel phenotypes during evolution.

1.2.2.2.1 Chromosome number

Chromosome numbers between species can vary due to a duplication of an entire or partial chromosome set, or be caused by chromosomal fusions or fissions (Figure 1-3). The duplication of entire chromosome sets, called polyploidization, can result from the fusion of two gametes from which one contained a non-reduced set of chromosomes or from spontaneous genome duplication within a non-dividing cell. In evolution, polyploidization is believed to have two main consequences. One of them is the creation of genetic diversity, due to the duplication of gene copies that can increase the number of members of a gene family, evolve to accrue new or slightly varied functions, or become pseudogenes. The other is the creation of reproductive isolation that may lead to speciation due to a greater complexity of chromosome pairing and segregation interactions (Ohno, 1970, Madlung, 2013).

During evolution, genome duplication events (paleopolyploidy) occurred in multiple eukaryote lineages (Chen et al., 2007, Madlung, 2013). In plants, polyploidy is very common, and the increase in ploidy was shown to be a frequent companion of speciation of angiosperm and ferns (Wood et al., 2009). Polyploid plants often present outstanding characteristics as the production of large flowers (as seen in dahlias that are octoploid, 2*n*=8x=64) or fruits, for instance, the banana and apple are triploid (2*n*=3x=33 and 2*n*=3x=51, respectively), and the strawberry is octoploid (2*n*=8x=56) (Pevsner, 2015). In animals, polyploidy is less frequent. Most animals contain two sets of chromosomes per somatic cell (i.e.

are diploid) with few exceptions. For example, male bees are monoploids (*n*=16) (Fernandes et al., 2013) and the red vizcacha rat (*Tympanoctomys barrerae*) is believed to be tetraploid (2*n*=4x=102) (Gallardo et al., 2006). It is however hypothesised that two rounds of whole-genome duplication occurred in early vertebrate evolution, leading to the genome complexity and size observed in extant species (Ohno, 1970).

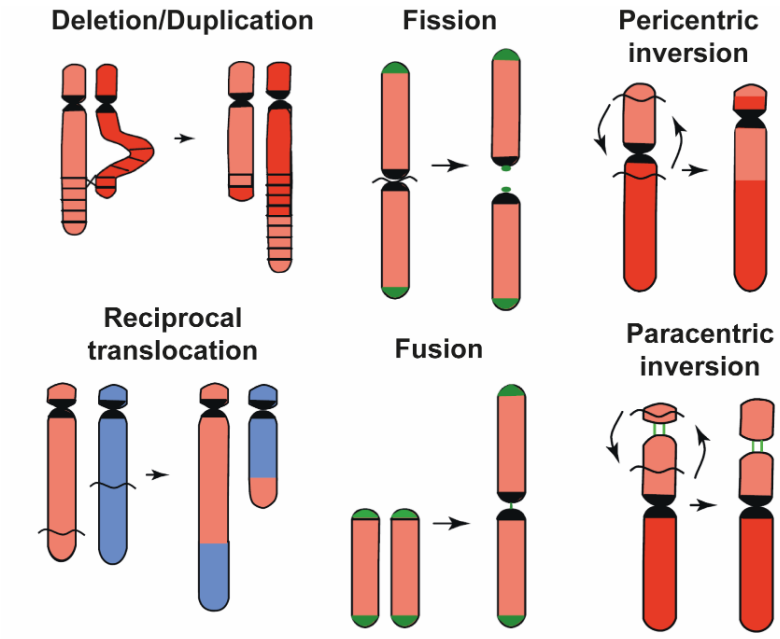

Figure 1-3: Types of chromosomal rearrangements. Adapted from (Schubert and Lysak, 2011).

Inter-species chromosome number variation can also result from the fusion or fission of chromosomes. Chromosomal fusions are the combination of two non-homologous chromosomes to form a new one (Figure 1-3). For instance, human (*Homo sapiens*) chromosome 2 is known to have resulted from the fusion of two ancestral great apes chromosomes, chromosomes 2A and 2B in chimpanzee (*Pan troglodytes*; Figure 1-2A; IJdo et al., 1991). Conversely, chromosomal fission involves the breakage of a chromosome producing two new ones (Figure 1-3). For example, human chromosomes 3 and 21 derive from the fission of a larger ancestral chromosome (Muzny et al., 2006). The occurrence of chromosomal fusion and fission in animal evolution resulted in a high haploid number variation that ranges from one in Jack jumper ants (*Myrmecia pilosula*; Crosland and Crozier, 1986) to 134 in Agrodiaetus butterfly (*Agrodiaetus shahrami*; Lukhtanov et al., 2005). In mammals, Indian muntjac (*Muntiacus*

muntjak) possesses the lowest haploid number of chromosomes (*n*=3; Wurster and Benirschke, 1970) and the aquatic rat (*Anotomys leander*) one of the highest (*n*=46; Schmid et al., 1988). Birds, however, show a very stable haploid number with more than 60% of their species presenting *n*=40. These differences indicate that distinct phylogenetic lineages seem to evolve at different rates and in distinct ways, and raises questions regarding the mechanisms that are shaping genome evolution in different taxa.

1.2.2.2.2 Chromosome structure

Karyotipical diversity is not limited to changes in chromosome numbers. Chromosome rearrangements can also involve individual chromosomes or occur between chromosomes while still maintaining haploid numbers (Hall and Quinlan, 2012, Liu et al., 2012). These rearrangements are duplications, deletions, inversions, and reciprocal translocations (Figure 1-3). Among multiple other cases, examples can be seen in *Drosophila* species that show multiple inversions (Guillén and Ruiz, 2012) and gorilla (*Gorilla gorilla*) where a reciprocal translocation shaped chromosomes 4 and 19 (Yunis and Prakash, 1982).

Multiple chromosomal rearrangements moulded the organisation and structure of extant genomes during evolution. In fact, distinct phylogenetic groups present strikingly different chromosome numbers, shapes and sizes, as can be noted in Figure 1-4. The variation within and between phylogenetic clades suggests different rates of karyotipical evolution. In this way, the comparative analysis of chromosome structures is a powerful tool for the establishment of relationships between species, and for a better understanding of the process of evolution and its effects on the biology of species.

1.2.2.3 Karyotype diversity and human disease

Besides providing insights into the evolution of species, karyotipical changes are frequently associated with disease phenotypes. Changes in chromosome numbers are often related to genomic instability and can be present in diseases such as cancer. Tumour cells can present changes in ploidy and an abnormal number of chromosomes within a set (aneuploidy). For instance, human epithelial tumours can show near-triploid or near-tetraploid cells (Mitelman et al., 2017). Moreover, Turner's syndrome (monosomy of the sex chromosomes; Sybert and

McCauley, 2004) and Down's syndrome (trisomy of chromosome 21; Patterson, 2009) are also related to abnormal chromosome numbers. Chromosomal inversions, especially those where breakpoints locate in regions of euchromatin (open chromatin, though to contain active genes) are also found associated with disease phenotypes. One such case is a pericentric inversion on human chromosome 8 that resulted in dysmorphic phenotypes and neurodevelopmental impairment (Ananthapur et al., 2012).

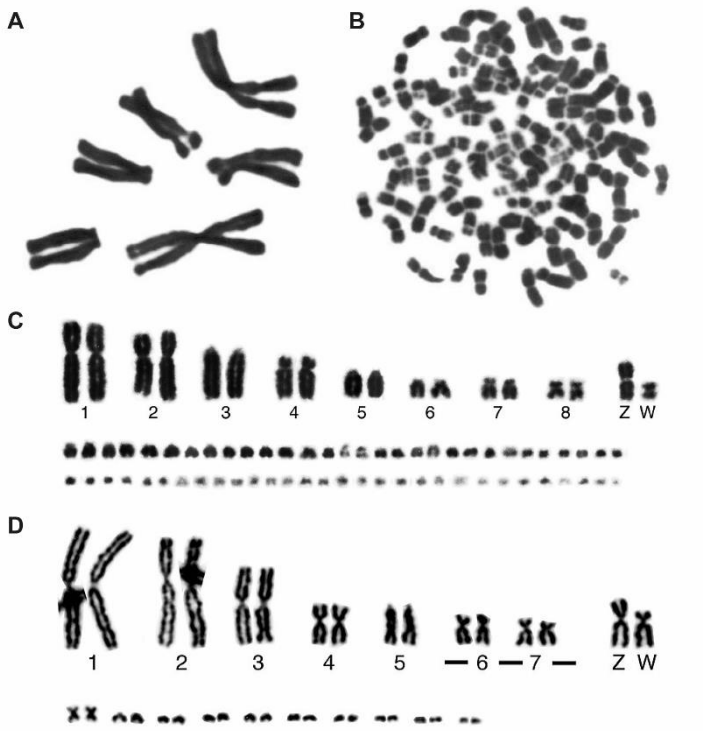

Figure 1-4: Karyotypes of (A) Indian muntjac and (B) red vizcacha rat (Graphodatsky et al., 2011), (C) chicken and (D) Japanese four-striped rat snake *(Uno et al., 2012)*.

1.2.3 Origin of genome rearrangements

Chromosomal rearrangements usually arise from the erroneous repair of double-strand breaks (DSBs). These can be caused by exogenous agents (e.g., radiation or chemical agents), occur during DNA replication when the DNA polymerase ensemble encounters obstacles (e.g., DNA lesions or unusual DNA structures), or be generated during cellular processes such as meiosis (Raynard et al., 2008, Capilla Pérez, 2015). The ill repair of these breaks can happen due to direct joining of incorrect DSBs (Branco and Pombo, 2006) or recombination of non-allelic homologous sequences (Schubert and Lysak, 2011). Repetitive DNA

sequences are often used as a template for non-allelic homologous recombination (NAHR) and because of that are considered one of the largest contributors to genome evolution in eukaryotes (Lynch, 2007, Gregory, 2005, Wessler, 2006). Furthermore, segmental duplications (SDs), transposable elements (TEs) and tandem repeats (TRs) were already described to play important roles in genome reshuffling.

SDs are blocks of DNA, often in the range of hundreds of kilobase pairs (Kbp) in size, spread across a genome (Sharp et al., 2005). Different copies of an SD usually present more than 90% identity and are known to predispose recurrent rearrangements by NAHR (Shaw et al., 2002, Bailey and Eichler, 2006). The relative orientation of the SDs involved in NAHR can result in deletion, duplication, or inversion of the intervening sequences (Figure 1-5; Shaw et al., 2002, Sharp et al., 2005, Liu et al., 2012). SDs are involved in the occurrence of chromosomal rearrangements implicated in both human disorders (Stankiewicz and Lupski, 2002) and chromosomal evolution (Locke et al., 2003, Bailey et al., 2004, Carbone et al., 2014). Moreover, SDs were also shown to be associated with the birth of new genes, contributing to functional diversification and the expansion of gene families (Wilson et al., 2006, Newman et al., 2005). These gene families might be relevant for species adaptation, as is the case of genes related to innate immunity and digestion that are commonly found duplicated in mammals (Liu et al., 2009, Beckmann et al., 2007).

TEs consist of stretches of DNA that move around the genome of a cell, creating interspersed repeats throughout eukaryote genomes (Biemont and Vieira, 2006). These elements can promote recombination between homologous or unrelated copies of an element, leading to either a gain of DNA with each insertion of a new TE or a loss of DNA due to the deletion of all sequences between two TEs (Figure 1-5; Devos et al., 2002). TEs can also mediate interchromosomal recombination, leading to major chromosomal rearrangements and translocations (Oliver and Greene, 2009). Initially, TEs were believed to be non-functional DNA sequences. However, they were already shown to perform important regulatory, coding or structural roles in the genome, as for instance, providing binding sites for several transcription factors (Chuong et al., 2013, Sundaram et al., 2014).

In addition to TEs and SDs, TRs also have the potential to shape genomes through the promotion of genome instability (Gregory, 2005). TRs consist of clusters of repeated sequences with sizes ranging from one to hundreds of base pairs per repeat. TRs are primarily generated by replication slippage (Viguera et al., 2001), and, along with TEs, represent a major source of DNA variation and mutation (Armour, 2006). TRs can form non-canonical secondary DNA structures, such as hairpins, cruciform and triplex conformations, that are known to promote genome instability (Zhao et al., 2010, Du et al., 2014a). TRs association with chromosomal rearrangements is shown, for instance, by their high density in human and primate-specific chromosomal breakpoint regions (Farré et al., 2011, Ruiz-Herrera et al., 2006).

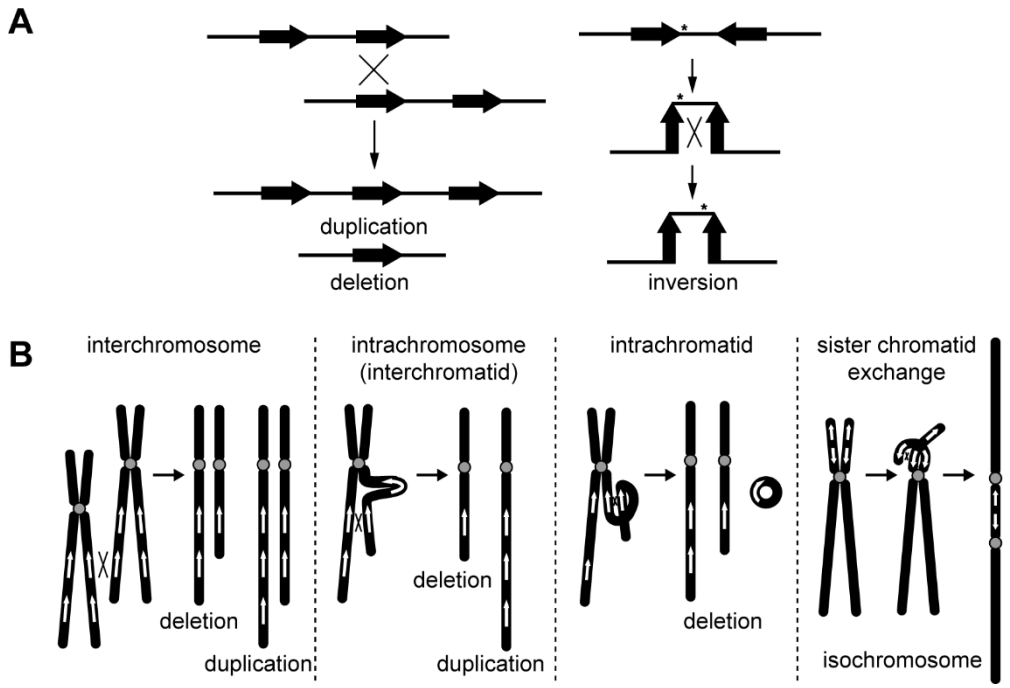

Figure 1-5: Non-allelic homologous recombination (NAHR). (A) Ectopic crossing-over between directly oriented repeats can result in deletion and duplication, while ectopic crossing-over between inversely oriented repeats can lead to inversions. (B) NAHR can produce deletion or duplication by interchromosomal crossover, interchromatid or intrachromatid crossover. NAHR between inverted repeats on sister chromatids can also lead to the formation of isochromosomes. Adapted from (Liu et al., 2012).

1.2.4 Implications of genome rearrangements

Chromosome rearrangements are believed to play a major role in speciation, and two main models have been proposed to explain this role. The hybrid dysfunction model hypothesises that speciation takes place when structural rearrangements

become fixed in a population (White, 1978). This model requires that hybrids for the chromosomal variant show reduced fertility (and underdominance), either caused by segregation problems or the generation of unbalanced gametes (Faria and Navarro, 2010, White, 1978). Natural selection will then promote mutations that reduce the probability of inter-crossing between populations carrying different chromosome arrangements, and thus their reproductive isolation. The main criticism made to this model, and its variants, relies on the underdominance of chromosomal rearrangements. Indeed, to be strong barriers to gene flow, chromosomal rearrangements need to be strongly underdominant, therefore it is unlikely that they become fixed in the population as they would be eliminated by natural selection (Rieseberg, 2001). The second model of chromosome speciation postulates that chromosome rearrangements contribute to speciation by the suppression of recombination (Rieseberg, 2001). Herein, rearrangements might not necessarily affect fertility, but they contribute to a reduction in gene flow by suppressing meiotic recombination within rearranged regions. This effect would allow the divergence of the genomic regions affected by the rearrangement, favouring the accumulation of genetic incompatibilities that would in the long term produce partial reproductive isolation (Faria and Navarro, 2010). This model is supported by evidence for the suppression of recombination in, for instance, the inversions of human and chimpanzee (Farré et al., 2013) and *Drosophila* (Noor et al., 2007).

Genomic rearrangements can also be associated with aberrant gene expression levels (Harewood and Fraser, 2014). Distinct rearrangements can result in balanced or unbalanced genomic alterations, which affect gene functionality in different ways. Balanced rearrangements (i.e. inversions and reciprocal translocations) result in alterations of nucleotide order without gain or loss of genetic material. These rearrangements might affect gene expression by the direct disruption of genes or regulatory pathways. On the other hand, unbalanced rearrangements (i.e. duplications, deletions and unbalanced translocations) can affect gene dosage through the gain or loss of genetic material. Chromosomal rearrangements have already been demonstrated to affect phenotype in multiple species. For instance, yeasts grown in stress-inducing environments, such as a glucose-limited setting, show many chromosomal rearrangements after few generations, and strains containing chromosomal rearrangements are more

resilient to starvation (Dunham et al., 2002, Coyle and Kroll, 2008). Chromosomal inversions were found associated with changes in size and developmental time in *Drosophila* (Hoffmann and Rieseberg, 2008), as well as, the adaptation of sticklebacks to freshwater (Jones et al., 2012b). In addition, pig lineage-specific evolutionary breakpoint regions (EBRs) were found enriched for genes related to the sensory perception of taste, which might explain why pigs can eat food that is unpalatable for humans (Groenen et al., 2012), and rhesus macaque EBRs are enriched for genes related to immune response (Ullastres et al., 2014), which were proposed to be involved in lineage-specific adaptation. Birds also show a similar pattern. For instance, budgerigar EBRs are enriched in genes involved in forebrain development that might relate to the different organisation of the "vocal brain nuclei" on this species when compared to other vocal-learning birds (Farré et al., 2016).

Besides giving insights into the effects of chromosomal changes, the comparison of chromosome structures also allows the detection of regions where sequence order was maintained during evolution (homologous synteny blocks; HSBs) and their genomic signatures. In fact, HSBs and EBRs were found to present distinct genomic signatures. For instance, eutherian multispecies (ms) HSBs were found enriched in genes related to organismal development, in particular of the central nervous system, bone and blood vessels (Larkin et al., 2009), and avian msHSBs were found enriched for conserved non-coding elements (CNEs), many of which are known to play important roles in gene regulation (Farré et al., 2016). These findings suggest that these conserved blocks might have been kept intact to avoid disruption of essential gene combinations and regulatory pathways.

The distinct genomic signatures found in EBRs and HSBs clearly show the importance of studying chromosomal evolution. Such studies result, not only, in a better understanding of the mechanisms that shaped the diversity of genomes seen today, but also unravel the origins of the phenotypical diversity presented by extant species. The finest level of information can be obtained by direct comparison of genome sequences. Regrettably, the availability of chromosome-level assemblies (one contig/scaffold representing a chromosome end-to-end) is limited in many taxonomical groups, which hinders the study of chromosome structures. Most currently used next generation sequencing methodologies

produce highly fragmented genome assemblies that lack sufficient robustness to study overall chromosome architecture. The development of new, fast and accurate methods to upgrade these genome assemblies to chromosome-level is in urgent need.

1.3 Obtaining genome sequences

Through whole-genome sequencing, researchers can obtain the most comprehensive catalogue of genomic information (Goodwin et al., 2016). In this regard, genome sequencing technologies, particularly next-generation sequencing, are transforming biology. With the assistance of both large (e.g., Bird 10K and Insect 5K) and small-scale sequencing projects, molecular sequence databases currently hold genome sequences for thousands of organisms. The power of comparative genomics is increasing, and we can now use DNA sequence analysis to learn how chromosomes were sculpted through evolution (Pevsner, 2015). However, obtaining a complete, chromosome-level assembled genome sequence is not an easy task. In this section, I will explore sequencing and genome assembly methodologies, giving particular attention to whole-genome sequencing approaches.

1.3.1 Sanger sequencing

Sanger and colleagues introduced the most commonly used ‘first-generation’ sequencing technique in 1977 (Sanger et al., 1977). Sanger sequencing relies on the selective incorporation, by a DNA polymerase, of chain-terminating dideoxynucleotides (ddNTPs) during *in vitro* DNA replication. Dye-terminator sequencing is one of the most recent variations of Sanger sequencing and involves the following (Figure 1-6): (A) a DNA template of interest is denatured to yield single-stranded DNA, and a universal oligonucleotide primer is added; (B) in the presence of DNA polymerase, the four deoxynucleotides (dNTPs) and fluorescently labelled ddNTPs, a second strand is synthesized. The addition of a ddNTP to an elongating strand inhibits further synthesis, preventing extension. This results in the production of extended fragments terminating at various positions. (C) The sequence is inferred through the detection of fluorescence emissions, corresponding to base calls, on fragments separated by capillary electrophoresis (Goodwin et al., 2016, Mardis, 2013). Sanger sequencing

produces reads with up to 1 Kbp of length and with less than 1% error rate per base call (Table 1-1). However, the high time and resources required by this technology clarified for the need of faster, higher throughput and cheaper sequencing technologies (Van Dijk et al., 2014).

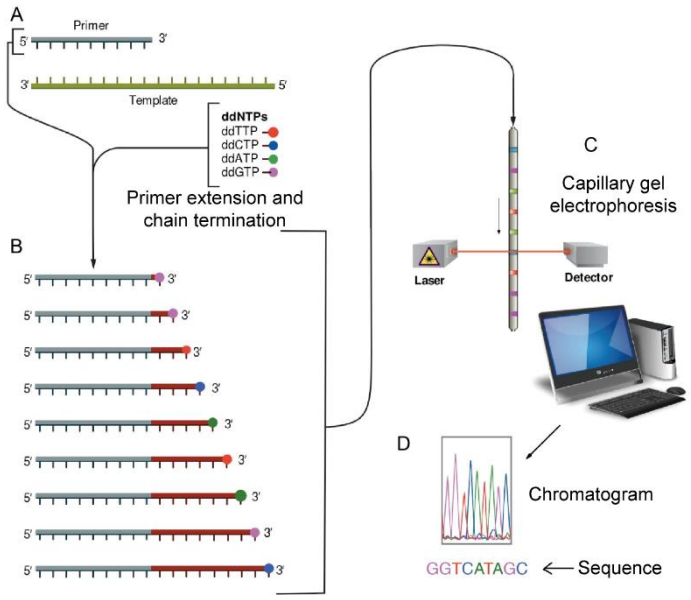

Figure 1-6: Dye-terminator sequencing methodology. (A) A primer, DNA polymerase, dNTPs and ddNTPs are added to a DNA template of interest. (B) Different sizes of extended fragments are produced. (C) Base calling is performed through the detection of fluorescence emissions. Adapted from (Estevezj).

1.3.2 Short-read next-generation sequencing

A collection of powerful new sequencing technologies, called 'next-generation' sequencing (NGS), emerged in recent years. These technologies possess significant advantages over Sanger sequencing: (1) instead of hundreds, thousands to billions of sequencing reactions are performed in parallel; and (2) the sequencing output detection is direct, without the requirement of electrophoresis (Van Dijk et al., 2014). NGS massive parallel output is the main contributor to its success. Together with the reasonable time frames and low cost they imply, NGS methodologies are now routinely used to perform whole-genome sequencing.

NGS methodologies are not, however, free of limitations. While still maintaining low error rates (<1%), most produce shorter reads than Sanger sequencing, usually within 35-600 base pairs (bp; Table 1-1), what complicates the genome

assembly process. Shorter sequencing reads complicate the assembly of repetitive genomic regions, specifically, if the repeat is longer than the reads generated. In these cases, genome assembly algorithms will collapse identical repeats resulting in reduced or lost genome complexity. The development of long insert (i.e. long distance) paired-end and mate pair read libraries reduced this effect, but could not eliminate it. Moreover, NGS generated data contain a mixture of random and systematic errors. Random errors can be minimised by increasing sequencing coverage (the number of times a base is sequenced) resulting in the production of a more accurate consensus sequence, but systematic errors are harder to resolve.

1.3.2.1 Sequencing by synthesis

Most NGS technologies share with Sanger sequencing the requirement of a DNA polymerase to replicate a template sequence. This polymerase-dependent sequencing is referred to as sequencing by synthesis (SBS), as each sequencing reaction synthesises a new DNA strand (Chen, 2014). SBS base call is performed by detection of a signal, such as a fluorophore or change in ionic concentration, at each nucleotide incorporation in an elongating strand (Goodwin et al., 2016). The most widely used technologies applying this approach are Illumina sequencing (Bentley et al., 2008) and Ion Torrent (Rothberg et al., 2011).

Illumina performs sequencing by cyclic reversible termination (CRT) with an approach that is very similar to that of Sanger sequencing (Bentley et al., 2008). Besides the advantages referred previously, and shared with all NGS approaches, Illumina differentiates from Sanger sequencing by using fluorescently labelled nucleotides from which the fluorophore can be removed at the end of each sequencing cycle. This allows a complete sequence to be obtained from the same original template, which was not possible with Sanger sequencing as the addition of a ddNTP would block the further elongation of a specific strand. Illumina high throughput is achieved by the parallel sequencing of millions of short reads. After DNA is fragmented and ligated to adapter sequences, each fragment is bound to a primer fixed on a solid surface, such as a flow cell (Figure 1-7). The free end can then bind with physically close primers, forming a bridge structure, and polymerase chain reaction (PCR) is used to create a second strand. This process is repeated several times resulting in the formation

of 100-200 million of clonal clusters. Then, the sequencing process starts. During each cycle, a mixture of all four individually labelled 3'-blocked dNTPs is added. After the incorporation of a single dNTP to each elongating strand, free dNTPs are washed out, and the surface is imaged to identify the incorporated nucleotide. Next, the fluorophore and blocking groups are removed, and a new cycle starts (Figure 1-7) (Goodwin et al., 2016).

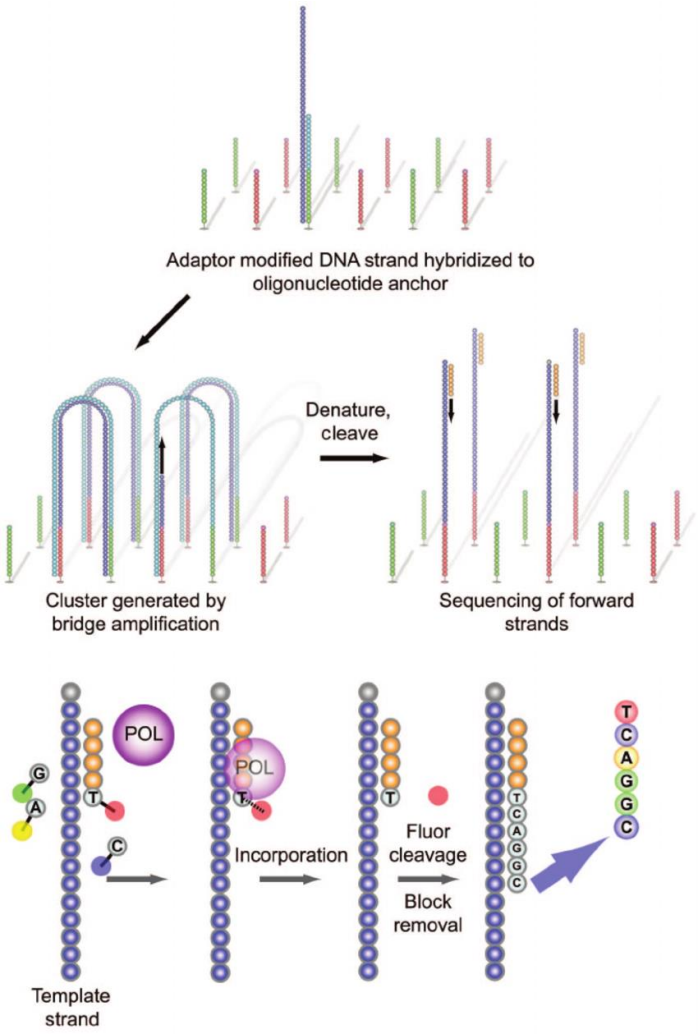

Figure 1-7: Illumina sequencing approach (Voelkerding et al., 2009).

The Illumina system can generate up to one terabase pair (Tbp) of DNA sequence data in a single run and produces reads up to 300 bp long with more than 99.9% accuracy (Table 1-1). Illumina platforms can perform single read, paired-end and mate pair sequencing (Bentley et al., 2008). Illumina data tend to have substitution errors with an increased frequency of them towards the end of

the sequencing reads. Some errors are also found associated with sequence motifs, such as GC-rich sequences, homopolymer stretches and around repeats. Fortunately, most of these errors are accompanied by lower quality scores and can be dealt with by trimming low-quality reads or read-ends (Laehnemann et al., 2016). Nonetheless, PCR biases are not reflected in changes in quality scores, and result in an underrepresentation of GC- and AT-rich sequences in the generated sequence data. PCR-free library preparation and improved PCR protocols reduced this effect, but they still do not remove it completely. This leads to a low sequencing coverage of low and high GC-rich genomic regions that might result in highly fragmented and incomplete genome assemblies (Chen et al., 2013b).

Despite its limitations, Illumina still demonstrates the lowest error rates of NGS technologies, which together with its flexibility makes it the most currently used sequencing technology. Its applications range from whole-genome sequencing to RNA, DNA methylation sequencing or chromatin immunoprecipitation followed by sequencing (Goodwin et al., 2016).

Another technology using CRT methodology is the Qiagen GeneReader. This platform is intended for clinical use, and relies on the targeted amplification of regions of interest, with a special focus on cancer gene panels. Amplification is followed by SBS, as with Illumina. In fact, this methodology shares Illumina's advantages and disadvantages, including the same error profile (Goodwin et al., 2016).

Ion Torrent uses a single nucleotide addition (SNA) approach to sequence DNA. Contrary to CRT, SNA does not use blocked dNTPs to prevent elongation, rather, each dNTP is added to a sequencing reaction at a time, and the absence of the next dNTP prevents elongation. Because of that, dNTPs do not need to be individually marked, and the signal received at each cycle is uniquely obtained from the dNTP added at that time. Ion Torrent is the first methodology not performing base call by imaging. Instead, the base call is performed by detection of pH change produced by the H^+ atoms released at each dNTP incorporation (Rothberg et al., 2011, Goodwin et al., 2016). Currently, Ion Torrent platform can produce reads up to 400 bp long that can sum up to 15 Gbp of raw data per run

(Table 1-1). These relatively long reads provide some advantages for applications focused on repetitive or complex DNA; however, they have error profiles that set some drawbacks. For this platform, insertion/deletion (InDel) errors are more frequent than substitutions, and the overall error rate is significantly higher than that of Illumina (1% for Ion Torrent versus 0.1% for Illumina) (Laehnemann et al., 2016). In fact, a considerable fraction of InDels is caused by homopolymers and usually results in the lack of coverage for homopolymers longer than 14 nucleotides. These errors are due to a decrease in base call accuracy with homopolymer length, as multiple base calls solely rely on the proportional increase of signal detected by the platform (Goodwin et al., 2016, Laehnemann et al., 2016). GC bias also has a noticeable effect in Ion Torrent generated data with read coverage significantly dropping in low and high GC content regions. Extremes of this limitation were shown by the failure of library preparation of GC-rich bacterial genomes, and the high error rate noticed for GC-poor bacterial genome sequencing (Laehnemann et al., 2016, Bragg et al., 2013). In addition, base call accuracy was also shown to decrease with consecutive cycles; however, these errors can be more easily dealt with by computational error corrections.

Roche's 454 pyrosequencing device also makes use the SNA methodology. However, the high cost and time associated with technology led Roche to discontinue this platform in 2016 (Margulies et al., 2005, Goodwin et al., 2016).

1.3.2.2 Sequencing by ligation

Contrary to the sequencing by synthesis methodologies, sequencing by ligation does not require the presence of a polymerase during the sequencing reaction. These approaches involve the hybridization and ligation of a labelled probe to a DNA strand (Tomkinson et al., 2006). Briefly, a fluorescently labelled (one-base- or two-base-encoded) probe hybridises to its complementary sequence adjacent to the primed template, and fluorescent imaging is performed to identify the ligated probes (Metzker, 2010).

The Applied Biosystems ABI SOLiD™ performs sequencing by ligation. It uses two-base-encoded probes, where each signal represents a dinucleotide (Valouev et al., 2008). In that way, reading a single colour does not specify a single base,

but rather corresponds to any of four possible dinucleotides. This approach allows a very high base-calling accuracy (~99.99%), as each base is read multiple times. As with Illumina, the SOLiD platform can generate both single-end and paired-end reads. However, maximum read length is only 75 bp (Table 1-1), which together with runtimes that take several days, strongly limits the use of this platform (Goodwin et al., 2016, Metzker, 2010).

Complete Genomics platform also performs sequencing by ligation. As with SOLiD, complete genomics chemistry produces relatively short reads (28-100 bp) with a very high accuracy (Drmanac et al., 2010, Peters et al., 2012). The limited use of this technology lead to the cancellation of the launch of a new platform, and older platforms are only used in mainland China (Goodwin et al., 2016).

1.3.2.3 Short-read NGS for *de novo* whole-genome sequencing

NGS technologies have revolutionised life sciences (Ellegren, 2014). The high parallelization achieved in NGS platforms led to a significant decrease in sequencing costs that, together with their reasonable time frames, makes NGS a widely-used tool for whole-genome sequencing. The release rate of genome sequences observed nowadays, especially for non-model organisms, and the resequencing of population samples facilitated by NGS methodologies, are empowering diverse scientific fields, such as comparative and population genomics (Ellegren, 2014). For instance, the comparative analyses of genomes are revealing important aspects of lineage-specific adaptations, such as the expansion of hypoxic stress related gene families and yak's (*Bos grunniens*) adaptation to high-altitude (Qiu et al., 2012); and, population genomics studies are unravelling unprecedented levels of variation in populations, as was shown in humans by the 1,000 Genomes Project (The Genomes Project Consortium, 2010). Functional genomics and medical genetics have also significantly benefited from NGS methodologies. NGS methodologies have increased the pace at which novel genetic variants are detected, being the best available tool to elucidate disease-causing mutations (Lohmann and Klein, 2014, Werner, 2010).

Nonetheless, NGS approaches are not free from some drawbacks. The amount of data generated by NGS technologies is one of the problems. The complexity

of the algorithms necessary for genome assembly and sequencing error corrections result in high computational costs, and the amount of data generated per sequencing run also requires massive storage capabilities. Indeed, storage costs were estimated to become more expensive than resequencing (Wuster, 2011), which was suggested as an alternative for easily obtainable samples. Additionally, the use of short sequence reads is problematic for the precise *de novo* assembly of complex genomic regions, such as long repeats and duplicated sequences (Alkan et al., 2011). As referred previously, during genome assembly, algorithms might collapse identical repeats, which leads to an underestimation or loss of genome complexity. The use of long insert size paired-end and mate pair read libraries ameliorated this bias, but it is still present in *de novo* genome assemblies. These limitations, together with the underrepresentation of AT- and GC-rich genomic regions, result in highly fragmented and incomplete genome assemblies. The fragmentation state of these genome assemblies directly influences gene annotation, as genes might be fragmented, missing or incorrectly annotated (Denton et al., 2014). Moreover, while useful for the analysis of small genomic variants, NGS genome assemblies lack sufficient contiguity to reveal larger structural variation and, without further assembly, to study chromosome architecture (Lee et al., 2016). In this way, NGS-generated genomes require the development of new approaches to increase the assembly accuracy of complex genomic regions and upgrade genome assemblies to chromosome-level.

1.3.3 Long-read sequencing

Long-read sequencing technologies were developed to assist the assembly of highly complex genomic regions. The produced reads, with lengths up to hundreds of Kbp, can span long repetitive elements, copy number or structural variants, eliminating ambiguity in the positions or size of these genomic elements. Moreover, the obtained reads can also span complete mRNA transcripts and so these technologies are also helpful for the study of transcriptomics. The main limitations of these technologies relate to their relatively high cost, compared with NGS methodologies, and the requirement of a large amount of high molecular weight DNA, which is often not easy to obtain. There are two main types of long-read technologies: single-molecule real-time sequencing and synthetic long-read technologies, which I will explore below.

1.3.3.1 Single-molecule real-time sequencing

Contrary to NGS short-read sequencing, single-molecule real-time (SMRT) sequencing approaches do not require the clonal amplification of a DNA template to generate a detectable signal, nor do they need the cyclic addition of dNTPs.

Pacific Bioscience (PacBio) platforms are currently the most widely used to perform SMRT sequencing (Eid et al., 2009). The PacBio methodology does not require the polymerase to move along a DNA template. Instead, the sequencing reaction is performed on flow cells containing many individual wells with a polymerase fixed to the bottom, and through which the DNA strand can progress (Figure 1-8A). This structure, called zero-mode waveguide (ZMW), allows the system to focus on a single-molecule. The incorporation of dNTPs is continuously visualised by a laser, and a camera system records the colour and duration of a light emission as the labelled dNTP pauses during incorporation at the bottom of the ZMW. During incorporation, the fluorophore is removed and diffuses away from the sensor before the next dNTP is incorporated. The PacBio platform uses a circular template, obtained through the ligation of two hairpin adapters to a double-stranded DNA template (Figure 1-8A). This allows the same template to be sequenced multiple times. However, templates longer than 3 Kbp are still difficult to sequence more than once. For shorter templates, these multiple passes are used to generate a circular consensus sequence (CCS) and reduce the high single-pass error rates associated with this methodology from ~13% to <0.1% (Goodwin et al., 2016, Eid et al., 2009). PacBio technology allows the production of reads that can average 5 Kbp, and are sometimes longer than 20 Kbp (Table 1-1) (Koren et al., 2013, Eid et al., 2009). Moreover, the GC content and homopolymer bias observed with NGS platforms are almost absent in PacBio reads. Nonetheless, these platforms have high error rate at single-pass and low throughput, which significantly increases its associated costs and limits their use. Even so, PacBio reads are extremely helpful to assist the assembly of genomes generated by NGS methodologies. Moreover, these platforms have already been used to sequence and assemble multiple bacterial, yeast, plant and animal genomes (Koren et al., 2013, Kim et al., 2014, Gordon et al., 2016).

More recently, Oxford Nanopore Technologies launched the first Nanopore sequencer, the MinION. Contrary to other technologies, Oxford Nanopore does

not detect the incorporation or hybridization of nucleotides to a template DNA strand. Instead, it reads the composition of a single-strand DNA molecule by measuring electrical changes while it passes through a protein pore (Clarke et al., 2009). The samples are prepared by adding a hairpin loop to one end of double-stranded DNA or complementary DNA fragments, linking both strands (Figure 1-8B). During analysis, DNA molecules are unwound by helicase enzymes into a continuous single-strand that is drawn through the protein pore to produce an electrical signal reflecting sequence composition (Figure 1-8B) (Deamer et al., 2016). Because the DNA molecules are hairpins, both strands can be analysed improving the base call accuracy. Each strand, however, is only sequenced once. The read lengths obtained with this methodology do not depend on the detection approach. Instead, they rely on the fragmentation method and can range from 6 to 48 Kbp (Jain et al., 2015). Accuracy, however, is just now surpassing 90% requiring further improvements, and contrary to PacBio, homopolymers tend to increase error rates (Table 1-1) (Deamer et al., 2016). Accuracy can be improved by resequencing the same native DNA fragment multiple times, which is only possible with this technology as Nanopore sequencing does not alter the DNA fragment it reads. In addition, the portability of Nanopore sequencers is a significant advantage. MinION, for instance, weighs mere 100 grammes and can be controlled by company designed software in a laptop computer. This portability proved extremely helpful during Ebola virus outbreak surveillance on site (Hoenen et al., 2016, Quick et al., 2016).

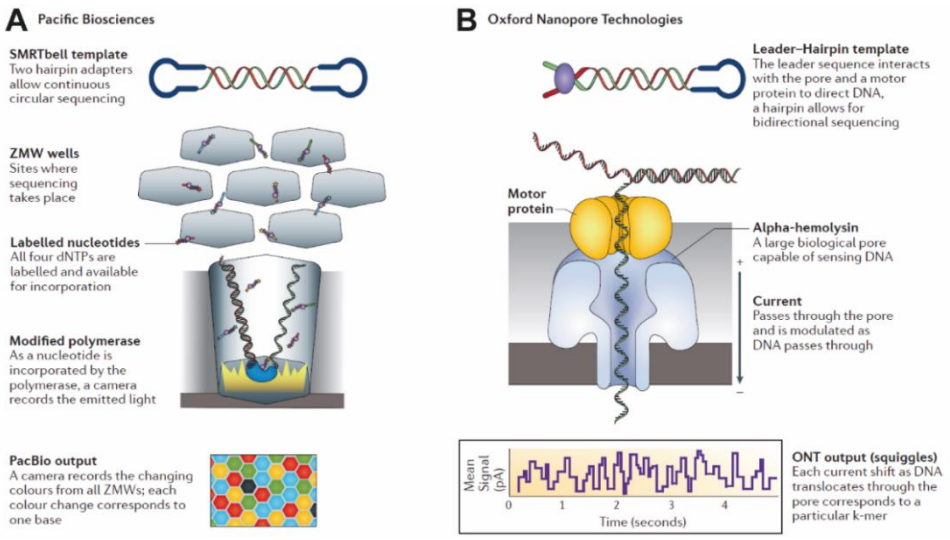

Figure 1-8: Single-molecule real-time sequencing approaches: (A) Pacific Biosciences and (B) Oxford Nanopore (Goodwin et al., 2016).

1.3.3.2 Synthetic long-read sequencing

Synthetic long-read sequencing approaches do not generate real long-reads. Instead, they rely on a system barcoding to associate fragments that are sequenced with short-read methodologies (Voskoboynik et al., 2013). These approaches partition large DNA fragments into either microtiter wells or an emulsion, such that only few hundred to thousand molecules exist in each partition. Then, within each partition, the molecules are fragmented, barcoded and sequenced using short-read NGS methodologies. The sequencing data obtained is split by barcode and reassembled with the knowledge that reads sharing a barcode result from the same original, large DNA template. This methodology allows for repetitive or complex genomic regions to be isolated and assembled locally, which prevents unresolvable areas in the assemblies, leading to lower genome fragmentation (Goodwin et al., 2016).

The Illumina synthetic long-read sequencing platform uses such an approach. This methodology allows the production of reads from 1.5 to 18.5 Kbp and has the same error profiles as Illumina short-read sequencing (Table 1-1). However, this system requires higher sequencing coverage than a typical NGS project, leading to an increase in the associated costs (McCoy et al., 2014).

10X Genomics also performs synthetic long-read sequencing. This system partitions fragments, up to 100 Kbp, into micelles called GEMs. Within a GEM, shorter barcode-identified DNA fragments (~350 bp) are amplified from the original larger fragment. After sequencing, the reads are aligned and linked, forming a series of anchored fragments across the span of the original fragment. Contrary to Illumina synthetic long-read system, this approach does not attempt to get full end-to-end coverage of a single DNA fragment from each GEM. Instead, the reads from a single GEM are dispersed across the original DNA fragment, and the cumulative coverage is derived from multiple GEMs (Goodwin et al., 2016, Zheng et al., 2016). 10X Genomics technology allows the use of as low as one nanogram of starting material, which can be advantageous for situations when DNA is difficult to obtain (Table 1-1). Once again, as it relies on short-read sequencing, the data generated presents the same error profiles and bias as that of NGS platforms.

1.3.3.3 Long-read sequencing for *de novo* whole-genome sequencing

Long-read sequencing technologies promise to generate longer reads enabling a more accurate and less fragmented genome assembly. Longer reads can assist *de novo* genome assembly projects, and be used to reveal complex long-range genomic structures (Goodwin et al., 2016). However, some limitations hinder the use of these technologies. For instance, synthetic long-read methodologies showed great potential towards making inexpensive and accurate *de novo* genome assembly a reality. Nevertheless, as they rely on short-read sequencing, they suffer from many limitations of underlying NGS methodologies, as is the case of GC-bias. Moreover, most synthetic read assembly processes require first the assembly of long fragments from short-reads, followed by the genome assembly with those long fragments. These strategies require vast amounts of short reads. For example, Illumina synthetic long-read approach usually requires 900-1,500X short-read coverage to assemble 30X coverage synthetic reads (Lee et al., 2016). Besides being sometimes completely unfeasible, this significantly increases sequencing costs and limits the utility of these approaches.

The use of ‘true’ long-reads alone for *de novo* genome assembly is strongly limited by both their high costs and the large amount of high molecular weight DNA that is required for their generation. For instance, PacBio and Nanopore sequencing cost around 1,000 US dollars per Gbp of generated data, while Illumina sequencing can be performed for just 10 US dollars per Gbp. These limitations, together with the high error rates at single-pass for these technologies result in a predominance of hybrid genome assembly approaches, where the continuity of ‘true’ long-reads is complemented by the accuracy of NGS short reads (Shi et al., 2016, Lok et al., 2017). Nonetheless, these hybrid approaches produce sequence fragments, so-called super-scaffolds, that represent at most chromosome arms, requiring further verification and ordering to obtain chromosome-level assemblies.

Table 1-1: Comparison of sequencing technologies. Adapted from (Goodwin et al., 2016).

Method	Maximum read length	Maximum accuracy	Maximum throughput per run	Advantages	Disadvantages
First generation					
Sanger sequencing	1 Kbp	99.90%	n.a.	Long single reads.	Very expensive and time-consuming.
Second generation					
Ion torrent	400 bp	99.00%	15 Gbp	Fast. Long reads.	Insertion/deletion and homopolymer errors.
Roche 454	1 Kbp	99.00%	700 Mbp	Long single reads.	Homopolymer errors.
Illumina	300 bp	99.90%	1 Tbp	High-throughput.	Short reads. GC bias.
SOLiD	75 bp	99.99%	320 Gbp	Very fast.	Very short reads.
Third generation					
PacBio	50 Kbp	99.90%[1]	7 Gbp	Very long reads.	Limited throughput. Large amount of high molecular weight DNA.
Oxford Nanopore	48 Kbp	92.00%[2]	4 Gbp	Very long reads.	Homopolymer errors. Large amount of high molecular weight DNA.
Illumina synthetic long-reads	200 Kbp[3]	99.90%	500 Gbp	No extra equipment.	Requires higher coverage.
10X Genomics	100 Kbp[3]	99.90%	500 Gbp	Small amount of DNA.	Inefficient DNA partitioning.

[1] Accuracy from circular consensus sequences;

[2] Single read accuracy;

[3] Synthetic lengths.

1.3.4 Genome assembly

The ultimate aim when assembling a genome *de novo* is to obtain a set of contigs each representing an entire chromosome end-to-end. *De novo* genome assembly up to the chromosome-level usually consists of three steps (Figure 1-9). (1) Sequencing reads are aligned, and overlapping reads are merged to form longer and gapless contiguous sequences, called contigs. Next, (2) the information from paired-ends and/or mate pair libraries is used to connect contigs and infer gap sizes between them, as reads from the same pair map on two separate contigs, forming scaffolds. Lastly, (3) information from mapping methodologies is integrated with scaffolds to order and orient them along the chromosomes (Figure 1-9) (Green, 2001).

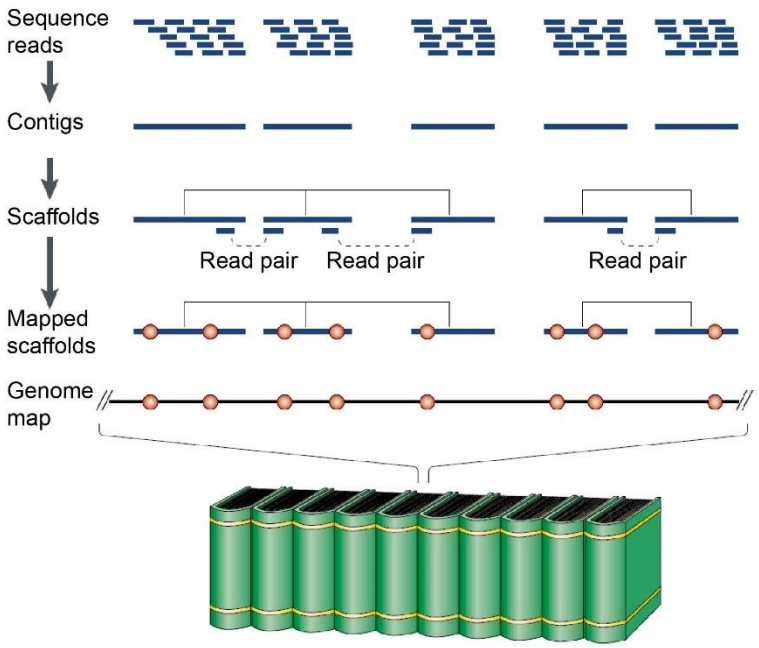

Figure 1-9: De novo whole-genome sequencing assembly. From (Green, 2001).

One of the major difficulties when performing genome assembly is the accurate assembly of complex genomic regions, for instance, those containing long repeats and duplicated sequences (Alkan et al., 2011). This limitation is particularly actual for NGS technologies where the short nature of the generated reads can result in *de novo* assemblies that are highly fragmented and lacking parts of the genome (Lee et al., 2016). Moreover, the dearth of genetic and physical maps for most *de novo* sequenced species, due to the costly and time-consuming nature of the mapping methodologies (discussed in the next section),

difficult the assembly process, and most genome assemblies publicly released nowadays are comprised of thousands of contigs or scaffolds. As referred, previously, this assembly fragmentation restricts their use for evolutionary and applied biology studies. For instance, genome annotation is directly influenced by the quality of genome assemblies, as genes might be missing and genes split between multiple scaffolds will only be annotated by using comparative approaches (Dessimoz et al., 2011). The fragmentation of the genome assemblies limits the extent of haplotype definitions and different haplotypes can even be assembled as distinct contigs or scaffolds. Highly fragmented genomes also complicate the establishment of long-range interactions in gene regulation and lack the robustness needed to study overall chromosome architecture and evolution.

These limitations could be minimised by using long and accurate sequencing reads obtained from PacBio or Nanopore sequencing, for example. These reads can span longer repetitive regions and other copy-number or structural variants, reducing ambiguities in assembling complex genomic elements (Schatz et al., 2010). Nonetheless, the high DNA requirements, low accuracy at single-pass and the inability of these technologies to generate super-scaffolds spanning chromosomes from end-to-end, still call for the development of novel approaches that would facilitate the generation of chromosome-level genome assemblies.

1.3.4.1 Mapping methodologies

Mapping technologies landmark a genome without sequencing every base. The recent advent of whole-genome sequencing is having a dramatic impact on the utility of genome maps and optimisation of map-generating methodologies are the target of constant developments. As mentioned in the previous section, genome maps are essential for *de novo* genome assembly as they assist the merging and ordering of scaffolds in obtaining chromosome-level genome assemblies. Moreover, they are also valuable tools for the detection of assembly errors. Herein, I will explore mapping methodologies focusing on their utility to assist genome assembly.

1.3.4.1.1 Genetic linkage maps

Genetic linkage (GL) maps, as the name implies, are based on the principle of genetic linkage. According to this principle, loci physically close on chromosomes tend to be inherited together, loci located far apart on the same chromosome tend to be inherited together less often, and loci found on different chromosomes are inherited independently. Initially, GL maps were based on the association of visual phenotypes, such as eye colour and wing size in *Drosophila* (Sturtevant, 1913). However, due to the limited number and complex nature of such phenotypes they were replaced by other genetic markers, such as restriction length fragment polymorphisms and single nucleotide polymorphisms.

The creation of GL maps starts with the establishment of a mapping population from which one can identify genetic differences between related individuals that can be used to determine recombination distances between loci. Next, maps are estimated through phenotype association or (more common nowadays) genotyping individual DNA samples for different markers, followed by the grouping of markers into linkage groups, and the ordering and spacing of markers within linkage groups (Fierst, 2015). The relative order and distance between markers are estimated accordingly to the rate of recombination between them, where two markers with a 50% recombination frequency are transmitted independently and assumed to be located further apart on the linkage group. The resulting GL maps are a set of linkage groups showing the positions of the markers relative to each other. Because recombination events are not randomly distributed along chromosomes, the genetic distance between markers is not proportional to their physical distance. Moreover, GL maps resolution depends on the number of crossovers that can be analysed and is usually low for long-lived organisms and those that are difficult to breed or grow under experimental conditions, as fewer meiosis can be studied. This usually results in fragmented GL maps, with more linkage groups than the number of chromosomes of the studied species. In addition, infrequent recombination between adjacent markers difficult their ordering and spacing, resulting in reduced accuracy at fine-scale resolution (Baxevanis and Ouellette, 2004). Nonetheless, the integration of GL maps with the high fine-scale accuracy of genome sequences can be used to produce high-quality *de novo* genome assemblies, through the reduction of genome assembly's fragmentation and correction of misassemblies. This was

already shown by the integration of this two data sources in the dog (*Canis lupus familiaris*; Wong et al., 2010) and collared flycatcher (*Ficedulla albicolis*; Ellegren et al., 2012), for example.

1.3.4.1.2 Radiation hybrid maps

The creation of radiation hybrid (RH) maps starts with the generation of radiation hybrid cells. DNA breaks are induced by the application of lethal doses of radiation to a donor cell line, which is then fused with a recipient cell line (usually hamster or mouse). An RH panel consists of a panel of independently fused radiation hybrid cells, each containing a separate collection of donor DNA fragments. Each radiation hybrid (or fusion cell) is tested for the presence of each marker of interest (Baxevanis and Ouellette, 2004). As with GL maps, during the creation of RH maps, it is assumed that physically close markers will present similar patterns of retention or loss (behaving as if they were linked), while physically distant markers will show different patterns. RH maps present three major benefits over GL maps. First, because the induced breaks are randomly distributed across the genome, break frequencies are roughly proportional to physical distances, and the constructed maps will be more accurate. Second, the markers do not need to be polymorphic within the species of interest, and any unique DNA sequence different from the recipient orthologue can be mapped. Third, RH maps do not require the establishment of a mapping population. The resolution of the RH map, however, directly depends on the size of the chromosomal fragments contained in the radiation hybrids, which in its turn is proportional to the amount of radiation the donor cell line was exposed. The generation of RH panels is very expensive and time-consuming which limits their creation, especially for non-model organisms. Nonetheless, the utility of RH maps for the increase in continuity and detection of genome assembly errors was clearly shown for, for instance, goat (*Capra hircus*; Du et al., 2014b) and rhesus macaque (*Macaca mulatta*; Karere et al., 2008).

1.3.4.1.3 Cytogenetic maps

While helpful, recombination-based maps do not correlate the molecular markers with the real structure where they are found, the chromosome (Griffiths, 2008). This can, however, be done through cytogenetic maps. Traditional cytogenetic

mapping relies on the hybridization of a radiolabelled or fluorescently labelled DNA probe containing a marker of interest (usually a BAC clone) to a chromosome preparation (Griffiths, 2008, Brown, 2007). The accuracy and resolution of cytogenetic maps are directly influenced by some subjective criteria as variability in technologies and methodologies, interpretation, reproducibility and the definition of banding boundaries (Chen and Chen, 2013). Because of that, traditional cytogenetic maps are helpful to grossly localise a marker of interest in a species chromosome but lack high resolution ordering. Recent developments in FISH comprise the use of naturally or mechanically stretched DNA molecules (fiber-FISH), which allows a significant increase in the map resolution, from ~1 Mbp on traditional metaphase chromosomes to ~1 Kbp on fiber-FISH (Ersfeld, 2004). Another limitation of FISH methodologies for the generation of cytogenetic maps is the low throughput of this technique. This relates to the difficult establishment of multiplex experiments. In addition, the success rate of FISH experiments is significantly reduced if target-specific probes are not available. In such cases, cross-species FISH (zoo-FISH) can be utilised, but the success rate of this approach is limited by the evolutionary distance and degree of sequence homology between the compared species (Graphodatsky et al., 2012). Nonetheless, this last limitation could potentially be tackled by the selection of probes according to DNA sequence features that are known to influence the success of hybridization, for instance, sequence homology, and GC and repetitive content.

1.3.4.1.4 Other physical maps

Physical maps (that include the cytogenetic and radiation hybrid maps discussed above) comprise maps that are either capable of directly measuring the distance between DNA markers or that use cloned DNA fragments to directly order these markers (Baxevanis and Ouellette, 2004).

Optical mapping visualises and measures the length of single DNA molecules, extended and digested with restriction enzymes, by high-resolution microscopy (Baxevanis and Ouellette, 2004). BioNano Genomics have significantly improved this mapping methodology on their Irys system. The Irys system detects, through direct visualisation, fluorescently marked restriction sites along a linearized DNA molecule. The collection of mapped DNA molecules is assembled into a larger

optical map that can span megabase pair (Mbp) (Cao et al., 2014). The Irys maps can then be compared to *in silico* scaffold restriction maps. The latter, generated by the identification of putative restriction sites in scaffolds, based on the recognition sequence of the restriction enzyme. The alignment of both maps allows the anchoring and ordering of scaffolds, creating a hybrid scaffold (Lee et al., 2016). The limitations of this methodology include the incomplete nicking of the DNA, which leads to unlabelled restriction sites, and the presence of physically close nick sites that might shear the DNA, limiting the total length of the generated hybrid scaffolds. Nonetheless, optical maps already successfully assisted the genome assemblies of many vertebrate species (Howe and Wood, 2015), for instance, goat (*Capra hircus*; Dong et al., 2013) and ostrich (*Struthio camelus*; Zhang et al., 2015).

Hi-C and its related protocols use proximity ligation and high-throughput sequencing to investigate the three-dimensional architecture of the genome within nuclei (Lieberman-Aiden et al., 2009). This methodology employs crosslinking of chromatin in intact cells; intra-molecular ligation of DNA fragments that were physically close at the time of crosslink; followed by high-throughput sequencing of the generated DNA junctions (Figure 1-10). A Hi-C map is then a genome-wide DNA-DNA contact probability map that allows the three-dimensional modelling of genomic conformation within a cell. The strong enrichment in Hi-C signal between genetically close loci (up to ~200 Mbp apart) allows, in theory, the scaffolding of entire chromosomes from fragmented draft assemblies. Still, Hi-C methodology suffers from various limitations. For instance, Hi-C requires intact cells, which might be difficult to obtain for many species. Published Hi-C experiments in human resulted in maps with ~1 Mbp resolution, obtained from ~10 million paired-end reads (Lieberman-Aiden et al., 2009), and an improvement by 10-fold of map resolution would represent a 100-fold increase in sequencing depth, leading to a rise in the associated costs. In addition, most Hi-C data describes only DNA-DNA contacts in the scale of tens or hundreds of Kbp, reducing map continuity, and the large-scale organisation of chromosomes in nuclei may provide confounding signals to assembly. These final limitations were overcome by the development of the Dovetail Genomics “Chicago” libraries (Putnam et al., 2016). This methodology relies on the generation of DNA crosslinks on *in vitro* reconstituted chromatin, obtained by adding histones to

input naked DNA (Putnam et al., 2016). The generated Chicago libraries share many characteristics with Hi-C data, as the relationship between read distance and read count that are helpful for genome assembly. At the same time, this approach tackles some of Hi-C limitations by reducing the number of momentary long-range or interchromosomal interactions, or the need for intact cells as starting material (Putnam et al., 2016). However, the large amount of DNA required for each experiment (~5 micrograms (µg)) limits the use of this methodology. Additionally, together with Hi-C and other NGS relying methodologies, the generated data suffers from GC-biases and difficulties on the assembly of repetitive sequences. Nevertheless, Hi-C was already used to perform *de novo* assembly of human, mouse, *Drosophila* (Burton et al., 2013), and threespine stickleback genomes (Peichel et al., 2016), and Chicago data were used for *de novo* assembly the American alligator (*Alligator mississipiensis*) genome (Putnam et al., 2016).

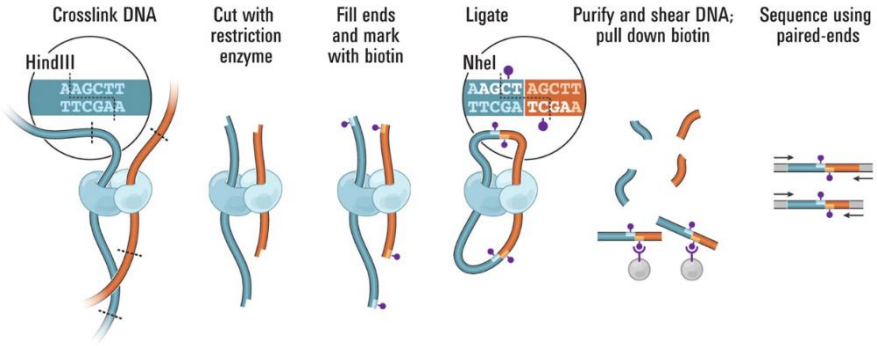

Figure 1-10: Hi-C methodology. Formaldehyde fixed cells are digested with a restriction enzyme. The resulting sticky ends are filled with nucleotides, one of which is biotinylated. Ligation is performed in conditions favouring intramolecular ligation events. DNA is purified, sheared and biotinylated junctions are isolated. Interacting fragments are identified by paired-end sequencing (van Berkum et al., 2010).

Genetic and physical maps are essential tools for *de novo* assembly of genomes. They assist the correction, merging and ordering of scaffolds to generate chromosome-level genome assemblies. Nonetheless, mapping methodologies are very expensive and time-consuming, especially for the generation of high-resolution maps. This results in a dearth of genome maps for most of the newly sequenced genomes and consequently a lack of chromosome-level assemblies.

In this way, the development of approaches to (1) reduce the resolution of maps required to assemble genomes to chromosome-level, and (2) quickly and inexpensively generate such maps are a necessity.

1.3.4.2 Computational approaches

Diverse algorithms were developed with the aim of further assemble scaffold-level genomes to chromosome-level, making use of available chromosome assemblies of related species. While many of the available algorithms are efficient for bacterial genomes, only a few can handle complex eukaryotic genomes. Examples of such software's are the reference-assisted chromosome assembly (RACA) (Kim et al., 2013), Ragout (Kolmogorov et al., 2014, Kolmogorov et al., 2016) and Chromosomer (Tamazian et al., 2016). Chromosomer relies on the alignment of the target scaffolds to a reference chromosome assembly. While this methodology is fast and efficient to predict chromosomes of eukaryotes, the use of just one reference and no additional data may result in the reconstruction of the reference chromosome structures instead of that of the target species. Ragout was first designed to work on prokaryote genomes, and only recently upgraded to deal with eukaryote genomes. This software can use one or multiple reference genomes, what will decrease the risk of reconstructing reference-like chromosomes. In its turn, RACA uses comparative information, as the two previously mentioned algorithms, but also sequencing data from the target species (paired-end and mate pair library reads) to order and orient scaffolds into predicted chromosome fragments (PCFs). The use of multiple data sources in this software, on one hand, reduces the number of introduced errors, and on the other increases the chance of detecting target-specific structural differences, in comparison with the reference genomes. RACA approach is summarised in Figure 1-11. Briefly, (A) RACA requires as inputs the pairwise alignments between the reference, target and outgroup genomes. Here, the reference genome pertains to a phylogenetically close species and the outgroup genome to a more distantly related species. (B) The algorithm will merge collinear alignments into syntenic fragments (SFs; regions of maintained sequence order) keeping the ones that exceed a user-set length threshold. (C) For each pair of SFs, an adjacency score is calculated combining information from the presence of such adjacency on the reference and outgroup genomes,

and the number of target species paired-end reads supporting the adjacency. (D) Once all adjacency scores are calculated, an SFs graph is constructed in which each head and tail of an SF are connected to another SF. Each connection is represented by a different weight, obtained from the adjacency score. (E) A chain of SFs is created by merging two adjacent SFs with the highest edge weight at each step. (F) Lastly, by using the order and orientation of SFs, RACA concatenates the scaffolds of the target genome creating PCFs.

The use of this algorithm proved useful on the increment of continuity of Tibetan antelope (*Pantholops hodgsonii*; Kim et al., 2013) and blind mole rat (*Spalax galili*; Fang et al., 2014) genomes. Still, RACA approach contains two main limitations. The obtained PCFs are often at a sub-chromosomal level, requiring further placement on the target species chromosomes, and, in some cases, RACA algorithm cannot properly distinguish chimeric scaffolds from those harbouring evolutionary breakpoint regions (EBRs). Such distinction is imperative as, chimeric scaffolds result from errors in the assembly process, connecting contigs that belong to different genomic regions, while EBRs represent target-species-specific structural differences that are critical for the accurate reconstruction of its genome structure. In this way, the efficient use of this algorithm that by itself does not add almost any extra costs to a whole genome sequencing project requires the development of an effective approach to verify the correctness of the PCFs, and further assemble PCFs to chromosome-level.

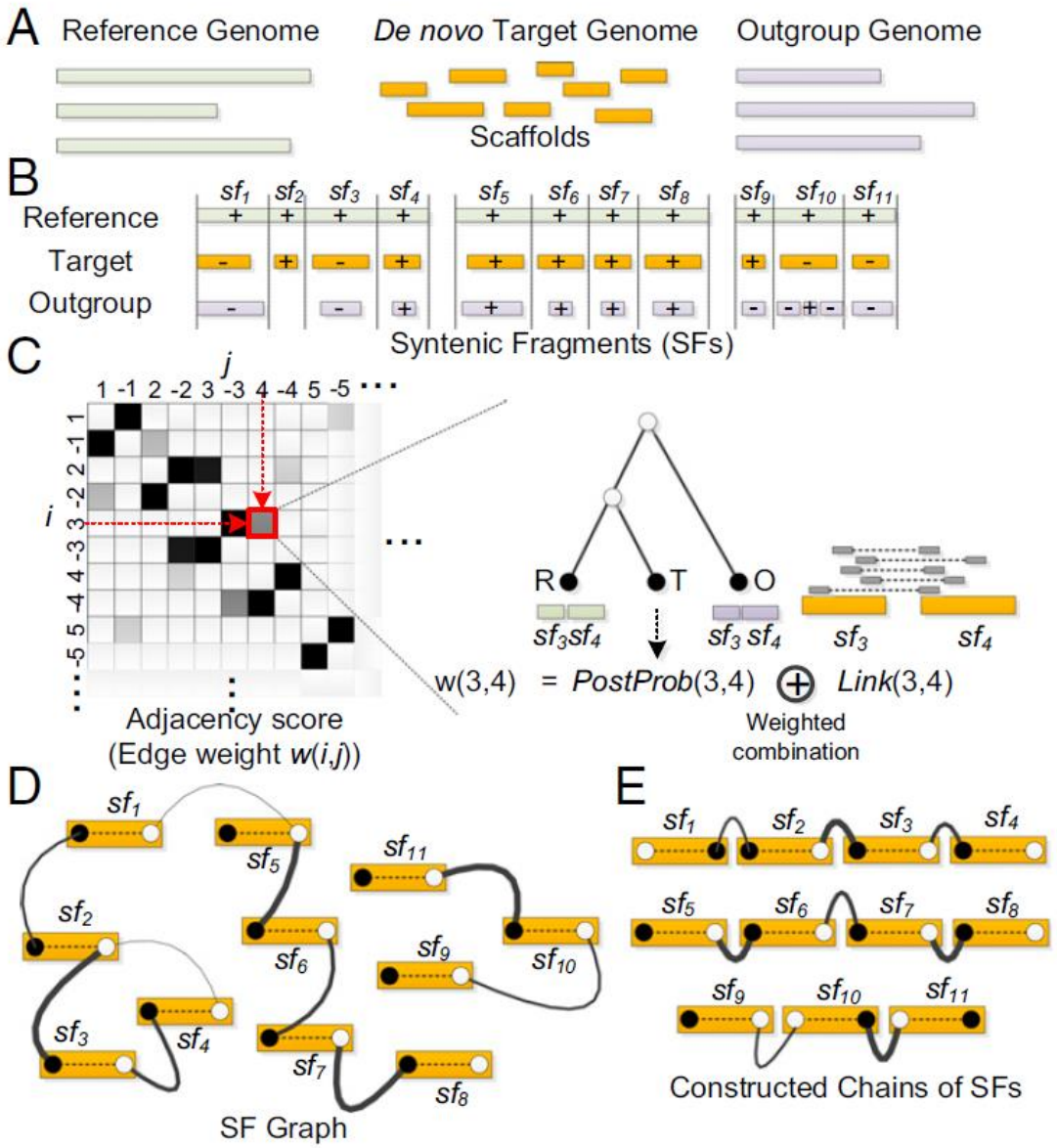

Figure 1-11: Overview of the RACA algorithm. (A) RACA takes a reference, a *de novo* sequenced target (in scaffolds), and one or more outgroup genomes as input data. (B) Syntenic fragments (SFs) delimited by vertical dashed lines are constructed by aligning reference and target genome sequences and merging collinear alignments. The outgroup may not always be aligned to SFs (e.g., sf2) and may contain rearrangements within a SF (e.g., sf10). Pluses and minuses represent the orientations of the target and outgroup DNA sequences on the reference genome, and three groups of SFs represent different reference chromosomes. (C) For each pair of SFs, an adjacency score (edge weight) that combines the posterior probability of the adjacency and the coverage of paired-end reads is calculated. (D) The SF graph is built by connecting SFs whose edge weight in C is higher than a certain threshold. Head (closed circle) and tail (open circle) vertices from the same SF are always connected with a maximum weight (dashed edge). (E) Constructed chains of SFs that are extracted by the RACA algorithm. From (Kim et al., 2013).

1.4 Ancestral karyotype reconstructions

As mentioned previously, changes in the structure of chromosomes are often related to the appearance of new phenotypes and can even lead to speciation. In this way, tracing the chromosome organisation of a common ancestor sheds light on the evolutionary process and the role of chromosomal rearrangements in phenotypic evolution and speciation (Deakin and Ezaz, 2014).

The first evolutionary relationships between the chromosomes of many organisms were determined by comparative cytogenetics (banding; e.g., IJdo et al., 1991), zoo-FISH (e.g., Richard et al., 2003, Griffin et al., 2007) and the comparison of gene maps (e.g., Murphy et al., 2003). In these studies, a genome organisation shared between all, or most of, the descendant species was believed to represent the ancestral state. Cytogenetics and genetic map based comparisons gave the first insights on the genome rearrangements that shaped extant genome structures. For instance, cytogenetic information was used to propose the ancestral chromosome structure of diverse phylogenetic groups, e.g. eutherian mammals (Richard et al., 2003), ruminants (Kulemzina et al., 2011), carnivores (Beklemisheva et al., 2016) and birds (Griffin et al., 2007). These reconstructions clearly show that genome structures in different taxa evolved in different ways and rates, and raised questions to which mechanisms drove the evolution of chromosomes in these lineages (Ruiz-Herrera et al., 2012). Cytogenetic-only ancestral reconstructions present two main limitations. The first relates to their lack of resolution, which results in undetected intrachromosomal rearrangements. The second is the limitation of evolutionary depth of the reconstructed ancestral karyotypes, which is restricted by the limits of detection of chromosome paintings (Deakin and Ezaz, 2014). Nowadays, the availability of genome sequences for most eukaryotic taxa not only expands the evolutionary depth of such reconstructions but also increases the resolution at which rearrangements are detected. The first attempts to reconstruct ancestral karyotypes from sequence data were performed for mammals, using chicken as an outgroup (Bourque et al., 2005). This reconstruction generated a Boreoeutherian ancestor karyotype with a haploid number of 21 and allowed the detection of a larger number of genome rearrangements between rodents and humans than what was previously known (Bourque et al., 2005). Moreover, a

combination of sequence and radiation hybrid map comparisons were used to propose the structure of the common ancestor of human, horse, cat, dog, pig, cow, rat and mouse (Murphy et al., 2005). Therein, Murphy and colleagues noted, for instance, that 20% of all detected EBRs were reused in different lineages, EBRs located in high gene density regions, and that EBRs were associated with the location of segmental duplications (Murphy et al., 2005). This work clearly shows the utility of ancestral karyotype reconstruction for the unravelling of the mechanisms that shaped extant genomes.

Diverse algorithms can perform the reconstruction of ancestral karyotypes based on genome sequence data. Two main approaches can be used to obtain such reconstructions: (a) the global parsimony methods that rely on the minimum number of genome rearrangements required to obtain the synteny of extant genomes, and (b) the local parsimony methods that infer the most parsimonious scenario for each individual adjacency (Ma et al., 2006, Jones et al., 2012a). Examples of ancestral chromosomes reconstruction software are InferCARs (Ma et al., 2006), ANGES (Jones et al., 2012a), AGORA (Muffato, 2010) and DESCHRAMBLER (Kim et al., 2017). InferCARs requires as inputs the pairwise alignments between the reference and all the other genomes used in the ancestral reconstruction. These alignments will be used to progressively construct “orthology blocks”, “conserved segments” and “contiguous ancestral regions” (CARs) (Ma et al., 2006). At its release, InferCARs was used to reconstruct the Boreoeutherian ancestral karyotype and gave important insights into the mechanisms governing genome rearrangements in mammals. This reconstruction based on human, mouse, rat and dog, and using chicken and opossum as outgroups resulted in 29 CARs, and was used to detect lineage-specific genome rearrangements and identify genomic features associated with EBRs. Within the latter, Ma and colleagues observed that 57 out of the 742 identified EBRs were reused in different lineages. Moreover, EBRs were found enriched for genes, repetitive DNA sequences, and human-specific EBRs were also significantly enriched for segmental duplications (Ma et al., 2006). Despite their utility, all the ancestral chromosome reconstructions obtained from the algorithms mentioned above were generated from chromosome-level genome assemblies, and their suitability to deal with fragmented assemblies was never proven (Kim et al., 2017). Because of that, and due to the limited availability of

chromosome-level assemblies for many phylogenetic groups, their use is also limited. This led Kim and colleagues to develop the DESCHRAMBLER algorithm. DESCHRAMBLER generates reconstructed ancestral chromosome fragments (RACFs) using SFs identified from whole-genome comparisons of both chromosome-level and scaffold-level genome assemblies (Kim et al., 2017). Contrary to other reconstruction software where orthologous chromosomal regions are required for all descendant species, DESCHRAMBLER can also utilise SFs where some species have deletions or missing data, which results in a better coverage of the ancestral genome (Kim et al., 2017). In addition, this algorithm can also accommodate a large number of descendant species. For example, at its release, DESCHRAMBLER was applied to the genomes of 21 species (14 chromosome-level and 7 scaffold-level) to reconstruct the chromosome structure of seven eutherian ancestors. In this study, Kim and colleagues compared the reconstructed RACFs with previous FISH-determined eutherian, boreoeutherian and simian ancestral karyotypes. They observed that the RACFs contained 12 out of 16 well-established interchromosomal rearrangements between the ancestral and human chromosomes that were previously detected by FISH (Kim et al., 2017). Moreover, the analysis of chromosome evolutionary rates was also highly consistent with previous studies (Kim et al., 2017).

The usefulness of the reconstruction of ancestral genomes structures for the study of chromosome evolution is apparent from the discoveries enumerated above. Such studies, not only allow the unravelling of the mechanisms that drive genome evolution but also the role of genome rearrangements in the phenotypical diversity of extant species. One of the main limitations for the performance of ancestral chromosome reconstructions is the high level of fragmentation of most available genome assemblies. Therefore, DESCHRAMBLER algorithm creates new opportunities, as it can handle both chromosome- and scaffold-level genome assemblies. Nonetheless, the fragmentation level of a genome directly influences the accurate identification of lineage-specific EBRs, as many could exist between scaffolds and be lost. In this way, highly accurate ancestral chromosome reconstructions also require accurate chromosome-level genome assemblies.

1.5 Birds

The Class Aves is the most species-rich class among tetrapod vertebrates, comprising more than 10,000 species (Zhang et al., 2014b). Birds are distributed worldwide, demonstrate a vast amount of physiological and morphological adaptations (Jetz et al., 2012) and are important model organisms for several scientific fields, such as developmental biology (Gilbert, 2000a), cancer research (Williams et al., 2014) and toxicology (Prauchner et al., 2013). Additionally, birds are also important from several agricultural (e.g., food), cultural (e.g., sports) and environmental (e.g., pest control and conservation) reasons. Despite their importance, in 2010 there were only available the annotated genomes of three avian species [chicken (*Gallus gallus*; Hillier, 2004), turkey (*Meleagris gallopavo*; Dalloul et al., 2010) and zebra finch (*Taeniopygia guttata*; Warren et al., 2010)]. At the end of 2014, the Avian Phylogenomics Consortium released the genome assemblies of 45 birds, aiming to tackle the lack of genomic information for this taxon and start investigating the links between genomic variation and phenotypic diversity (Zhang et al., 2014a, Zhang et al., 2014b). In the next sections, I will be exploring what is known about avian genomes and which questions remain unanswered.

1.5.1 Genome size

Birds possess the smallest genomes among amniotes averaging 1.35 Gbp, and ranging from 0.9 Gbp in the black-chinned hummingbird (*Archilochus alexandri*) to 2.1 Gbp in the ostrich (*Struthio camelus*; Scanes, 2014). The reduction of genome size in birds is hypothesised to reflect an adaptation to the high metabolic requirements of powered flight (Hughes and Friedman, 2008, Gregory, 2002b), which is supported by the fact that flightless birds have larger genomes than flying birds, and bats possess smaller genomes than their mammalian sister groups (Gregory, 2005). Nonetheless, comparative analyses also suggest that the evolution of compact genomes may have occurred before the emergence of flight (Tiersch and Wachtel, 1991, Organ et al., 2007). Regardless of when it occurred, the compactness of avian genomes is shown in extant species by a lower repetitive DNA content, by shorter genes and non-coding regions, and by the loss of gene family members. The avian genome content of TEs and other repetitive elements varies from 4 to 22%, very low values when compared to the 35 to 52%

observed in mammalian genomes (Lander et al., 2001, Zhang et al., 2014b). In fact, 47 out of the 48 avian genomes analysed by the Avian Phylogenomics Consortium had a TE content below 10%. The exception is downy woodpecker (*Picoides pubescens*) with 22% of its genome containing TEs, which resulted from an either species- or lineage-specific expansion of LINE-CR1 (long interspersed elements; chicken repeat I) transposons (Zhang et al., 2014b). Avian genomes were also shown to have shorter introns and intergenic regions. This compression, shared with bats, may result from a lower number of TEs within these regions, instigated by the fast gene regulation required for flight (Zhang et al., 2014b, Zhang et al., 2013). Likewise, bird genomes show an overall reduction in the number of gene family members when compared to other vertebrates, and the loss of these paralogs seems to correlate with segmental deletions associated with large-scale structural rearrangements (Hughes and Friedman, 2008, Lovell et al., 2014, Warren et al., 2016).

1.5.2 Karyotype structure

1.5.2.1 Bimodal karyotypes

Sizes of avian chromosomes vary from 200 Mbp to <10 Mbp (Ellegren, 2010). In fact, birds have bimodal karyotypes usually formed by up to ten pairs of macrochromosomes, comparable in size with mammalian chromosomes, and a varying number of small, almost indistinguishable, microchromosomes (Figure 1-12). Size is not the only difference between macro- and microchromosomes. Indeed, microchromosomes are GC-rich, gene-dense and demonstrate higher mutation rates than macrochromosomes. The presence of microchromosomes is not unique to birds, as they share this feature with turtles and lizards. Interestingly, crocodilians, the sister clade of birds do not have microchromosomes (Olmo, 2008). One possible explanation for this absence is the merging of micro- with macrochromosomes in crocodiles after the crocodiles-birds split (Ellegren, 2010).

1.5.2.2 Centromeres and telomeres

Avian chromosomes tend to be acrocentric (centromere locates near the chromosome end). Most chicken centromeres contain long (>100 Kbp) arrays of chromosome-specific simple repeats. However, chicken chromosome (GGA) 5,

GGA27 and GGAZ centromeres are remarkably short (~30 Kbp) and lack the usual repeat structure (Shang et al., 2010). Moreover, avian centromeres seem to be relatively labile, as centromere repositioning and the formation of new centromeres is frequently observed (Zlotina et al., 2012). As with other vertebrates, avian telomeres possess the canonical TTAGGG repeat structure, they constitute, however, unusually large repeat blocks that can go up to 4 Mbp in length (Delany et al., 2007, O'Hare and Delany, 2009).

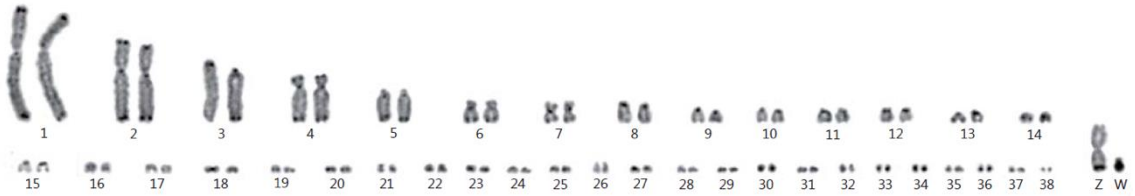

Figure 1-12: Chicken karyotype (Schmid et al., 2015).

1.5.2.3 Sex chromosomes

Unlike mammals, in the Class Aves females are the heterogametic sex. Females have a copy of chromosome Z and a copy of chromosome W, that is mostly non-recombining, and males have two copies of chromosome Z. This heterogamety evolved independently many times and can also be observed in butterflies, fishes, non-avian reptiles and amphibians (Matsubara et al., 2006). As with mammals, where the X chromosome is conserved across eutherian species, avian chromosome Z is also highly syntenic across most avian species (Nanda et al., 2008). Chromosome W, however, is minimally diverged from chromosome Z in ratites (e.g., ostrich and emu), and is smaller, gene-poor, and repeat-rich in all other birds (Marshall Graves and Shetty, 2001).

1.5.2.4 Chromosome numbers

Avian karyotypes show an incredible stability. In fact, more than 60% of all avian species present 74 to 86 chromosomes (Christidis, 1990, Griffin et al., 2007), and chromosomal paintings show that interchromosomal rearrangements were extremely rare during avian evolution. There are however exceptions to the typical avian 2*n*≈80, with the extremes of the spectrum being the stone curlew (*Burhinus oedicemus*) that has a diploid number of 40, and the Southern Go-away-bird (*Corythaixoides concolor*) with 2*n*=142 (Christidis, 1990, Griffin et al., 2007). Interestingly, the deviations from the typical avian chromosome number

are limited to few avian clades, such as penguins, birds of prey (i.e. falcons and eagles), and parrots (Griffin et al., 2007, Nishida et al., 2008, De Oliveira et al., 2005, Nanda et al., 2007). This karyotipical stability clearly distinguishes birds from other clades, for instance, mammals, and raises such questions as: (i) why interchromosomal rearrangements are so rare in birds or so common in other clades (e.g., mammals and lizards)? (ii) why different avian orders present different propensities to interchromosomal change? These questions are still unresolved, and its resolution directly depends on the availability of chromosome-level assemblies for avian species, especially those with highly rearranged karyotypes. In fact, to date, all chromosome-level assemblies for birds are from species with a typical avian karyotype. This limited the access to information that would give insights on the forces that are shaping genome evolution in different avian clades, and more generally in birds versus other animal Classes.

1.5.3 Phylogeny

One important aspect to consider when performing comparative analyses is the phylogenetic relationship of species, depicted in a phylogenetic tree. The resolution of such relationships is, however, not easy, as distinct data sets and analytical methods usually result in the production of different tree structures (Jarvis et al., 2014). This is further complicated for clades that underwent rapid radiations, as it is believed to be the case of Neoaves (Hackett et al., 2008) and placental mammals. For birds, the latest resolution of their phylogenetic tree was based on the genome sequence data released by the Avian Phylogenomics Consortium (Jarvis et al., 2014). Using whole-genome sequences for 48 avian species and multiple outgroups, Jarvis and colleagues generated a total evidence nucleotide tree (TENT) including information from introns, ultra-conserved elements and first and second codon positions. The obtained TENT (Figure 1-13) recognised three major groups within extant birds: the infraclass Palaeognathae (e.g., ostrich, emu, kiwi), and within Neognathae, the subclasses Galloanseres (e.g., chicken, duck) and Neoaves (e.g., parrots, pigeon, songbirds). It also supports the hypothesis of a “big bang” radiation of birds after the mass extinction event that led to the extinction of non-avian dinosaurs, approximately 66 million years ago (MYA). This tree supported some previous phylogenetic placements and contradicted others. For instance, birds of prey are now forming two different

clades: Falconiformes (falcons) and Accipitrimorphae (Eagles and New World vultures) (Jarvis et al., 2014). Moreover, this phylogenetic tree resolved some taxa relationships that were previously undetermined, as the grouping of cuckoos, turacos and bustards or the inclusion of mousebird within landbirds. Overall, this work proved the utility of genome-scale data to assist the resolution of difficult relationships in the tree of life (Jarvis et al., 2014). Nonetheless, although strong support was obtained for the resolution of early branches of the avian phylogeny, the same was not achieved for some deeper divergences after the Columbea and Passerea divergence, which shows that genome-scale alignments are not sufficient for a complete phylogenetic resolution (Jarvis et al., 2014). This TENT also proved helpful to the study of convergent traits. For example, it suggests that vocal learning might have evolved independently three times during avian evolution and that the common ancestor of core landbirds might have been an apex predator (Jarvis et al., 2014).

One of the criticisms made to Jarvis and colleagues approach is that the number of taxa sampled is too low, especially within Neoaves, which complicates the resolution of the relationships between members of this clade (Thomas, 2015). This led Prum and colleagues to generate a new tree recurring to the targeted sequencing of anchor regions (highly conserved in vertebrates) flanked by faster-evolving genome sequences. This approach was proven suited to solve rapid radiations (Lemmon et al., 2012) and was applied to 198 avian species and 2 crocodilian species (Prum et al., 2015). The major differences between this tree and the previous TENT are shown in Figure 1-14. Many of the conflicts between the two phylogenies are in early branching events that separate non-passerine taxa from each other. For instance, the clade including hummingbirds is shown to be the sister clade of Neoaves by Prum and colleagues rather than from grebes and flamingos as in (Jarvis et al., 2014). Additionally, a new monophyletic waterbirds clade is present in the phylogeny reported in (Prum et al., 2015).

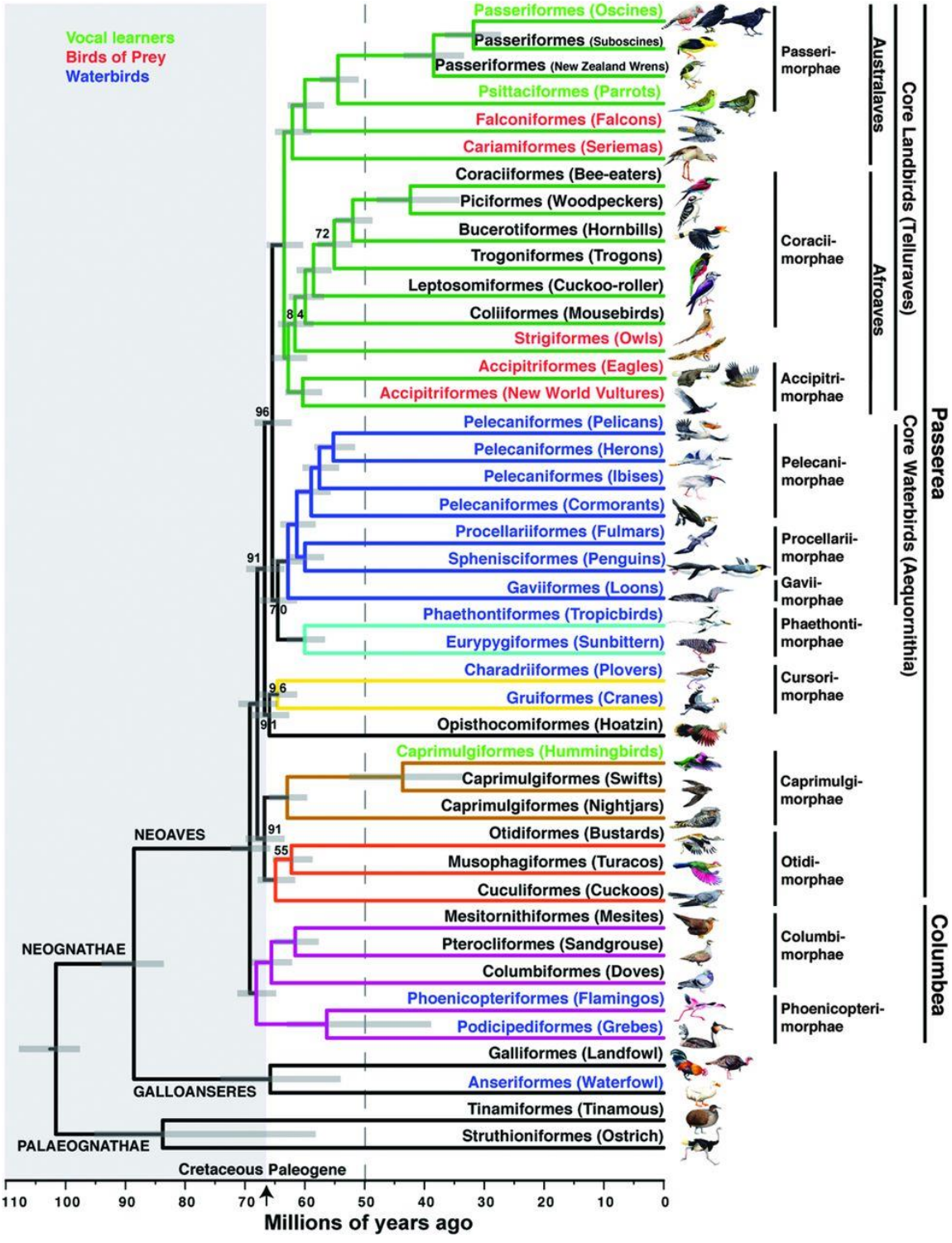

Figure 1-13: Genome-scale avian phylogeny as in (Jarvis et al., 2014).

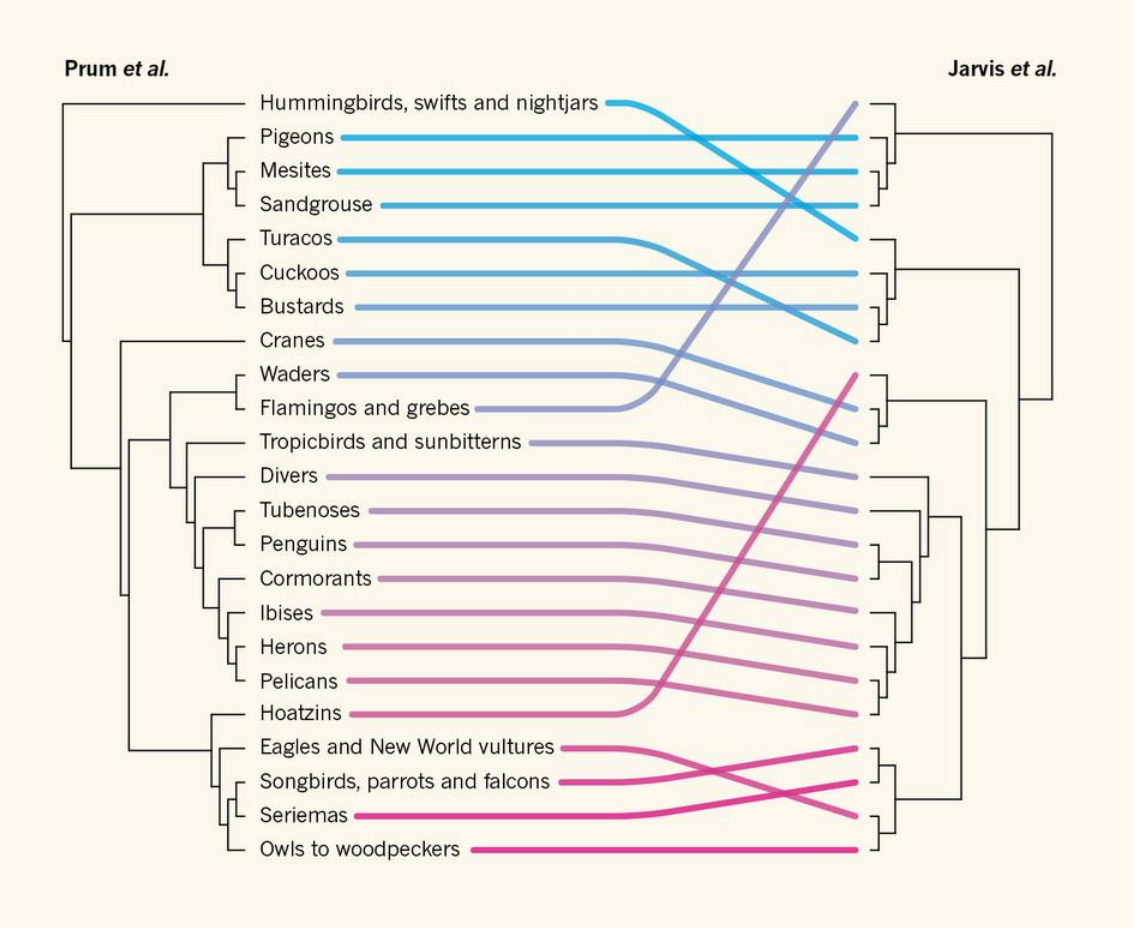

Figure 1-14: Comparison of avian phylogenies reported by (Jarvis et al., 2014) and (Prum et al., 2015) (Thomas, 2015).

1.5.4 Molecular evolution

Nucleotide substitution accumulation rates depend on several evolutionary factors as, for instance, the adaptive pressures that a species is subjected to and the effective population sizes (N_e) (Baer et al., 2007). The comparison of avian genomes started providing insights on how these factors influence avian molecular evolution. The overall substitution rate in birds (~1.9 x 10^{-3} substitutions per site per MY) is lower than in mammals (~2.7 x 10^{-3} substitutions per site per MY). However, mitochondrial DNA and nuclear DNA comparisons showed that the substitutions rates vary up to three-fold between bird lineages (Nabholz et al., 2009, Nabholz et al., 2011, Zhang et al., 2014b). Moreover, substitution rates were found positively correlated with the number of species within a phylogenetic order. For example, Passeriformes, the most speciose avian order, presents a neutral nucleotide substitution rate almost two times higher than that of Neoaves (Zhang et al., 2014b). In addition, landbirds also presented higher substitution rates than waterbirds, which agrees with the higher diversification rates of

landbirds (Zhang et al., 2014b). These observations suggested that the genome-wide variation in nucleotide substitution rates results from the avian radiation and the adaptive pressures the species were subjected to.

In agreement with the hypothesis of evolutionary stasis of avian genomes, orthologous genomic intervals in birds contain more constrained sequences than orthologous regions of mammalian genomes. In fact, ~7% of the avian genome (in comparison with ~4% in mammals; Lindblad-Toh et al., 2011) is found in sequences that are evolving slower than at the neutral substitution rate estimated for birds, and, from these, only less than 13% are associated with protein-coding genes (Zhang et al., 2014b). Thus, most of these slowly evolving sequences are CNEs, which may be regulatory elements or sites for transcription regulatory factors and contribute to the maintenance of synteny in avian genomes (Woolfe et al., 2004). The improvement of our understanding of the role of these sequences in avian genome evolution could be significantly improved through the analysis of their association with EBRs, which in its turn would benefit from the availability of a larger number of chromosome-level assemblies for avian species.

1.5.5 Adaptive phenotypes

Genomic variation (mutation or genomic rearrangement) is a primary source of phenotypic diversity (Chen et al., 2013a). Novel phenotypes can reach fixation during evolution. When a phenotype is advantageous in a specific environment and fixed through the action of natural selection, it is called adaptive. Herein, I will explore what is known regarding avian adaptive phenotypes and their distinct genomic signatures.

A major characteristic of birds is the ability to fly. This ability implies several major skeleton and bone modifications, such as the development of bones that are light and strong (Zhang et al., 2014b, Machado et al., 2016). In fact, 49 out of 89 ossification-related genes were found under positive selection in birds, a number twice as high as in mammals. Moreover, the highest ratio of non-synonymous to synonymous substitutions (d_N/d_S) within this set of genes was detected for the genes related to bone mineral density (*AHSG*) and bone homoeostasis (*P2RX7*) (Zhang et al., 2014b, Machado et al., 2016).

Even though feathers (composed of α- and β-keratin) are a classic example of exaptation, as they originated before the emergence of flight (Gould and Vrba, 1982), their further adaptation was influenced by species' lifestyles. Landbirds present two times more β-keratin genes than aquatic and semi-aquatic birds, and some domesticated species (zebra finch, chicken, pigeon, and budgerigar) have up to 8 times more β-keratin genes than other birds (Zhang et al., 2014b, Greenwold et al., 2014).

Another striking feature of modern birds is the lack of teeth. Fossil records show that several extinct avian lineages had teeth (O'Connor and Chiappe, 2011), suggesting that this phenotype evolved independently (Meredith et al., 2014). Indeed, multiple pseudogenes of enamel and dentin genes with many frame-shift or exon-deletion mutations were recovered from avian genomes. Most of these mutations differ between species, supporting the hypothesis of convergent changes within the Class. However, four enamel genes and one dentin-related gene shared the same deletions in all genomes suggesting that the common ancestor of all birds likely had no mineralized teeth (Meredith et al., 2014, Zhang et al., 2014b).

Genes related to beak morphology were also found to be under positive selection in both falcons (Zhan et al., 2013) and Darwin's finches (Rands et al., 2013, Lamichhaney et al., 2015, Chaves et al., 2016). These changes are believed to result from the adaptation to a predatory lifestyle and the need to expand the utilisation of food resources available in the Galápagos, respectively. Anna's hummingbird lineage-specific EBRs were also found associated with diet-related features. Specifically, these EBRs were found enriched for genes on the hexose metabolic process that might relate to their ability to digest sugar (Farré et al., 2016).

Birds possess not only the largest eyes (relative to body size) in the animal kingdom, but are also able to distinguish colours over a wider range of wavelengths than mammals, for instance (Scanes, 2014, Birkhead, 2012). In fact, for most vertebrate visual opsin genes, birds hold more copies than mammals, and the four classes of opsin genes found in most birds suggest that they are tetrachromatic (Zhang et al., 2014b). The avian visual acuity is however very

dependent on a species lifestyle. For instance, penguins were found to possess only three classes of opsin genes suggesting a trichromatic vision (Li et al., 2014), consistent with the loss of 1-2 cone pigments in aquatic mammals (Newman and Robinson, 2005). Kiwis (genus *Apteryx*) also present visual adaptations related to their nocturnality. Contrary to most nocturnal animals, which have large eyes in relation to their body size, kiwis have small eyes and reduced optic lobes in the brain (Martin et al., 2007). Moreover, the kiwi also has several opsin genes inactivated (Le Duc et al., 2015). In these birds, the loss of colour vision was accompanied by an improved sense of smell as shown by their increased repertoire of olfactory receptor genes (Le Duc et al., 2015).

As referred previously, vocal learning is believed to have evolved three times during avian evolution, and indeed, ~200 genes were found to contain signatures of convergent accelerated evolution in vocal learners. Interestingly, 73% of these genes are expressed in the songbird brain, and the vast majority in the song-learning nuclei (Whitney et al., 2014). Furthermore, budgerigar lineage-specific EBRs were found enriched for genes related to forebrain development, neuron differentiation and development, that might be linked to the different organisation of the "vocal brain nuclei" of this species when compared to other vocal-learning birds (Farré et al., 2016).

Some avian species were also found to possess segregating inversion polymorphisms (Itoh et al., 2011). These events can lead to the restriction of recombination, building genetic incompatibilities that might result in speciation. Such an example is a large pericentric inversion (~100 Mbp) found in white-throated sparrow (*Zonotrichia albicollis*) chromosome 2, which is associated with behavioural and plumage variations and believed to be an example of speciation in action (Thomas et al., 2008, Davis et al., 2011, Zinzow-Kramer et al., 2015).

The unravelling of specific effects of genomic variation on adaptive phenotypes is significantly limited by the high fragmentation level and incompleteness of most available avian genomes. As mentioned previously, genome annotation is directly influenced by the quality of the genome assemblies, as genes split between multiple scaffolds will not be annotated and others might be incorrectly annotated. Moreover, the lack of continuity of the genome assemblies will limit the extent of

haplotype definitions, as well as complicate the establishment of gene regulation links, especially long-range interactions. In this way, our understanding of how genomic variation shapes phenotypic diversity in birds would gain from the availability of more contiguous and accurate genome assemblies.

1.5.6 Chromosome evolution

Avian cross-species chromosome paintings showed a high degree of synteny conservation, even between distantly related avian lineages such as chicken and zebra finch that diverged ~90 MYA (Shetty et al., 1999, Derjusheva et al., 2004, Guttenbach et al., 2003, Romanov et al., 2014). This conservation is also noticeable with non-avian reptiles such as crocodiles and turtles, which diverged from birds ~230 MYA (Kasai et al., 2012, Pokorna et al., 2012). This fact made it relatively easy to predict the ancestral avian karyotype using cytogenetic data. The reconstructed Avian ancestral karyotype represented chicken chromosomes 1 to 9 plus Z with only one difference with chicken karyotype, where chromosome 4 results from the fusion of two ancestral chromosomes (Griffin et al., 2007). This organisation is maintained in most avian lineages studied so far. Passeriformes' karyotypes show only one difference to the proposed Avian ancestral karyotype were the ancestral chromosome 1 is split into two independent chromosomes. Galliformes (e.g., chicken, guinea fowl, quail, pheasant and turkey) also underwent only a few fusions or fissions, and most ancestral chromosomes were maintained intact (Griffin et al., 2007). The exceptions are birds of prey, where multiple rearrangements were identified (Griffin et al., 2007).

While cytogenetic techniques gave valuable insights into the gross structural stability of avian chromosomes, their lack of resolution could not assess if this constancy was also present within chromosomes. Indeed, the comparison of high-resolution gene maps and genome sequences started to reveal that, unlike interchromosomal events, intrachromosomal events (e.g., chromosomal inversions) were relatively frequent in avian evolution (Völker et al., 2010, Skinner and Griffin, 2012, Zhang et al., 2014b). These studies clearly demonstrated the importance of the availability of assembly data for the detection of the full collection of events that shaped genomes through evolution. The comparison of 21 avian genomes revealed that birds present an average rearrangement rate of ~1.25 EBRs/MY (Zhang et al., 2014b), which is significantly higher than the ~0.35

EBRs/MY of all mammals (Farré et al., 2011), and suggests that intrachromosomal rearrangements might be significant contributors to the phenotypic diversity presented by the members of this Class. Furthermore, rearrangement rate is highly variable between avian lineages. For instance, the origin of Neognathae was accompanied by an elevated rate of chromosomal rearrangements, ~2.87 EBRs/MY (Zhang et al., 2014b). Interestingly, vocal learning species show higher rearrangement rates than their non-vocal-learning relatives, and even higher than all the other non-vocal-learning species, which might relate with the larger radiations these clades experienced relative to the other bird groups (Zhang et al., 2014b). Nevertheless, because most analysed genomes were scaffold-level assemblies these rearrangement rates could be inaccurate, as EBRs located between scaffolds would not be included and assembly errors could be misinterpreted as EBRs.

Contrary to macrochromosomes, the difficult distinction, even by flow cytometry, between avian microchromosomes results in a lack of comparative data for these chromosomes and a lag in our understanding of their evolution. Still, the limited number of studies using chicken microchromosome paints on other birds showed that, in most cases, synteny is also conserved for these chromosomes, with most chicken microchromosomes maintained as a single chromosome in other species (Deakin and Ezaz, 2014). Exceptions to this rule are found for the species with an atypical chromosome number, where the reduction in the number of chromosomes is usually the result of microchromosomes tandem fusion or the fusion of micro- and macrochromosomes (Griffin et al., 2007, Nishida et al., 2008, De Oliveira et al., 2005, Nanda et al., 2007, Nie et al., 2009). The utilisation of BAC mapping for the establishment of homology between chromosomes could overcome the limitations of microchromosome paints probe generation. Moreover, if BACs are selected from regions of high sequence conservation, the same probe could be used to rapidly screen multiple species and help clarify the evolutionary history of microchromosomes.

Multiple hypotheses have been proposed to explain the stability of avian genomes. Some relate to their low repeat content, and how the lack of templates for NAHR might create fewer opportunities for avian genomes to change (Burt,

2002, Ellegren, 2010). Others suggest that there may be an advantage in maintaining synteny. Indeed, a combined mechanism may be acting. Avian msHSBs were found enriched for CNEs that are known to play important roles in gene regulation and be related to distinctive phenotypes (Farré et al., 2016). The disruption of these syntenic regions could then affect regulatory pathways, which might play a role in the maintenance of synteny. Additionally, avian EBRs were found associated with transposable elements that are usual NAHR templates (Farré et al., 2016). Indeed, avian genomes have ~4 times lower repeat content than mammal genomes, which could translate into fewer opportunities for avian genomes to change (merge). Why then some avian lineages are more prone to genomic rearrangement than others, both regarding inter- (e.g., birds of prey) or intrachromosomal rearrangements (e.g., vocal learning species), is still to uncover. Answering such questions is complicated by the lack of avian genomes assembled to chromosome-level, which prevents the study of chromosome evolution in this Class. Indeed, to date, the study of chromosome evolution in birds has been limited to karyotype comparisons or comparison of the chicken genome with mammalian genome sequences. Moreover, as previously mentioned, all avian chromosome-level genome assemblies belong to species with a karyotype close to the typical avian 2*n*=80. In this way, the unravelling of aspects of avian chromosome evolution would strongly benefit from the availability of new chromosome-level genome assemblies, in particular for species with highly rearranged karyotypes.

1.6 Project aims

The fragmented state of most newly sequenced genomes poses multiple challenges for the study of important aspects of applied biology and genome evolution. Among others, these limitations resulted in an understudy of avian genome biology. Having this in mind, this project aimed at the development of a novel approach to upgrade fragmented genome assemblies to chromosome-level, the generation of new avian chromosome assemblies and their use to answer some outstanding questions regarding patterns of chromosome evolution in birds.

This thesis first goal was then to upgrade avian NGS genome assemblies to (near) chromosome-scale and investigate the genomic signatures associated with avian chromosomal rearrangements. The achievement of this objective involved:

I. Testing the reliability of the reference-assisted chromosome assembly (RACA; Kim et al., 2013) to near chromosome-scale fragments from scaffold-level avian genome assemblies.
II. Developing a targeted-PCR-based methodology to evaluate a limited number of scaffolds important for accurate reconstruction of the target-species predicted chromosome fragments (PCFs) and using this approach to increase the accuracy of RACA-generated PCFs by minimising bias to the reference and outgroup genome structures.
I. Utilising the improved PCF genome assemblies to detect genomic signatures of avian evolutionary breakpoint regions (EBRs) flanking intra- and interchromosomal rearrangements.

The second objective of this thesis was the development of a novel, inexpensive and transferable approach to further assemble to the chromosome-level the PCF genome assemblies, and further study avian chromosome evolution. The achievement of this goal involved:

I. Designing an avian bacterial artificial chromosome (BAC) probe set efficiently hybridising with any avian species metaphase chromosomes to inexpensively produce sparse chromosomal physical maps for any avian

genome, permitting the assignment of scaffolds and PCFs to chromosomes.

II. Applying the novel genome assembly methodology to two avian genomes, the rock pigeon (a species with a typical avian karyotype) and the peregrine falcon (a species with an atypical avian karyotype).

III. Detecting the genomic signatures associated with newly identified avian EBRs flanking intra- and interchromosomal rearrangements and testing new hypotheses about the reasons behind the evolutionary stability of avian karyotypes in most avian species.

The third aim of this thesis was the study of the dynamics of chromosome evolution in birds using a combination of chromosome-level (previously existing and newly generated) and scaffold-level avian genome assemblies. The achievement of this goal implied:

I. Reconstructing the ancestral karyotypes for 14 avian clades, starting with the Avian ancestor and leading to the zebra finch lineage, including the Neognathae, Neoavian, landbirds and Passeriformes ancestors.

II. Inferring the evolutionary history of avian chromosomes through the detection of the type, number and time of occurrence of chromosomal rearrangements in avian lineages.

III. Testing hypotheses on the evolutionary stability and dynamics of different avian chromosomes related to their DNA feature content.

2 Constructing avian predicted chromosome fragments

2.1 Background

A scarcity of chromosome-level assemblies for most newly sequenced genomes impedes their use for critical aspects of evolutionary and applied genomics. For example, chromosome-level assemblies are crucial for studying genomics/genetics of species that are regularly bred (e.g., for food or conservation) as a known order of DNA markers allows the establishment of phenotype-to-genotype associations for gene-assisted selection and breeding (Andersson and Georges 2004). Such assemblies are established for common livestock species; however, they are not available for species used in developing countries or species bred for conservation reasons (e.g., camels and ostrich, and falcons, respectively). Chromosome-level information is also essential for addressing questions related to the overall genome (karyotype) evolution and speciation (Lewin et al. 2009). For instance, the study of evolutionary breakpoint regions (EBRs) revealed that EBRs, regions where synteny is disrupted between species due to evolutionary changes, are usually associated with repetitive sequences (e.g., transposable elements (TEs)) (Farré et al., 2011, Farré et al., 2016, Groenen et al., 2012) and influence genomic regions associated with genes related to an organism's response to external stimuli (Larkin et al., 2009). Disparately, homologous synteny blocks (HSBs), genomic regions where synteny and the order of homologous sequences are maintained intact during evolution between genomes of different species, are usually enriched for genes related to organismal development (Larkin et al., 2009) and conserved non-coding elements (CNEs) that are known to perform important roles in gene regulation (Farré et al., 2016).

The emergence of next-generation sequencing (NGS) methodologies made the sequencing of complex animal genomes a routine procedure. However, the genome assemblies produced using only NGS data are incomplete and highly fragmented. This limitation relates to the inability of NGS to generate long error-free contigs, and the lack of inexpensive mapping technologies to upgrade NGS genomes to chromosome-level. Generating traditional genetic and physical maps that could assist the genome assembly process is still more expensive than sequencing and assembling a genome up to scaffold-level, whereby they do not exist for most *de novo* sequenced species. Newer technologies, such as PacBio long read sequencing (Rhoads and Au, 2015) and BioNano optical mapping (Mak

et al., 2016) could provide a long-term solution to this problem. Nonetheless, these approaches suffer from various limitations that hinder their use. Besides their high costs, BioNano contigs, for example, cannot extend across multiple DNA nick regions, which decreases the continuity of the generated maps, and PacBio requires hundreds of micrograms of high molecular weight DNA that might be difficult to obtain in many cases. To provide and alternative solution, Kim and collaborators developed the reference-assisted chromosome assembly (RACA) algorithm (Kim et al., 2013), a computational approach to improve NGS genome assemblies' continuity. RACA usefulness was already shown for mammalian species, for instance, the Tibetan antelope (*Pantholops hodgsonii*; Kim et al., 2013) and the blind mole rat (*Spalax galili*; Fang et al., 2014). Besides significantly improving the genome assemblies' continuity, RACA also allowed for efficient detection of problematic genome assembly regions, such as putative chimeric scaffolds. RACA-enhanced assemblies have been used to identify lineage-specific intra- and interchromosomal rearrangements (Kim et al., 2013, Fang et al., 2014). Nevertheless, RACA approach has two main limitations. The obtained predicted chromosome fragments (PCFs) are still at a sub-chromosomal level, requiring further mapping and ordering on the target species chromosomes to obtain chromosome-level genome assemblies, and RACA algorithm does not always properly distinguish chimeric adjacencies within scaffolds that result from assembly errors from EBRs that depict lineage-specific structural differences. This might lead to the misinterpretation of chimeric adjacencies as EBRs or vice versa. In this way, this distinction is essential for an accurate reconstruction of a species genome.

RACA makes use of a combination of comparative information and target species sequencing data to verify, order and merge scaffolds, producing PCFs (Kim et al., 2013). The use of raw sequencing data to construct PCFs differentiates RACA from other available computational tools with the same purpose and reduces the reference genome bias during reconstruction. Moreover, RACA does not rely on comparative information from a single species. Instead, RACA uses the chromosome-level assemblies of a reference genome, phylogenetically close (e.g., same Order for mammals) and expected to have a similar genome structure to that of the target species, and one or more outgroup genomes that are phylogenetically farther from the target (Figure 2-1A). RACA starts by

constructing syntenic fragments (SFs; Figure 2-1B) merging collinear alignments from the pairwise alignments of the reference, the target and the outgroup genome assemblies. This information is obtained from the UCSC “net” nucleotide alignment files (Kent et al., 2003), which represent putative orthologous regions between two genomes. Next, for each pair of SFs, RACA calculates an adjacency score representing the likelihood of these fragments being adjacent in the target genome. This score combines the posterior probability of the adjacency being present in the target genome given its existence/absence on the reference and outgroup genomes, and the link probability based on the amount of target paired-end reads supporting the adjacency (Figure 2-1C). The following step consists of the creation of an SFs graph where each head and tail of an SF is connected to other SF head or tail. Each connection in weighted by the previously calculated adjacency score (Figure 2-1D). Chains of SFs are then constructed by merging two adjacent SFs with the highest edge score (Figure 2-1E). Finally, the order and orientation of each SF are used to concatenate the scaffolds of the target genome assembly and obtain PCFs.

Birds are important model organisms for multiple biological and medical fields (Gilbert, 2000b, Williams et al., 2014, Prauchner et al., 2013), and for many cultural, agricultural and environmental reasons. Moreover, the Class Aves comprises more than 10,000 species (Zhang et al., 2014b), being the most speciose Class among tetrapod vertebrates, that distribute worldwide and present numerous physiological and morphological adaptations (Jetz et al., 2012). Nevertheless, despite their importance, avian chromosomal evolution is an understudied topic. In fact, avian chromosomal evolution has been mostly restricted to karyotype comparisons or comparisons of the chicken genome with mammalian genome sequences. These comparisons showed that avian species do not show the same rate of interchromosomal changes as mammals or non-avian reptiles. Most birds have ~80 chromosomes and exceptions are limited to few avian lineages, such as falcons or penguins. This raises questions as: do the same mechanisms drive genome evolution in birds and mammals; and, why some avian lineages seem more prone to interchromosomal changes than others. The answer to these questions is restricted by the scarcity of chromosome-level genome assemblies for this Class. By 2010 only three avian

species had been sequenced: chicken (*Gallus gallus*; Hillier, 2004), turkey (*Meleagris gallopavo*; Dalloul et al., 2010) and zebra finch (*Taeniopygia guttata*; Warren et al., 2010). The release of an additional set of 45 avian genomes (Zhang et al., 2014a) showed potential to tackle this problem, however, the limitations of NGS methodologies and the lack of genetic or physical maps to assist genome assembly resulted in the generation of incomplete and highly fragmented scaffold-level assemblies.

Herein, we target the improvement of avian genome assemblies to allow a more detailed understanding of the forces driving avian genome evolution. We applied RACA algorithm to 18 avian scaffold-level genome assemblies. To minimise the number of structural errors in RACA-generated PCFs, we developed a targeted-PCR-based approach, which through the verification of a limited number of scaffolds facilitates the distinction of chimeric scaffolds from those harbouring target-specific structural differences. The results of the PCR-verification step were used to adjust RACA parameters and generate refined PCF assemblies. These refined PCF assemblies led to a significant improvement on the assemblies continuity, demonstrated by a reduction in the number of assemblies' fragments. Moreover, we proved that RACA is a reliable tool to upgrade avian NGS genomes by comparing the generated PCF assemblies with super-scaffold assemblies obtained from radiation hybrid maps, optical maps and Dovetail approaches. The generation of PCF assemblies allowed the detection of EBRs related with intrachromosomal events and flanking chromosomal fusions not previously reported for avian genomes. The analysis of the genomic patterns found on these EBRs provided important insights into the mechanisms governing avian genome evolution.

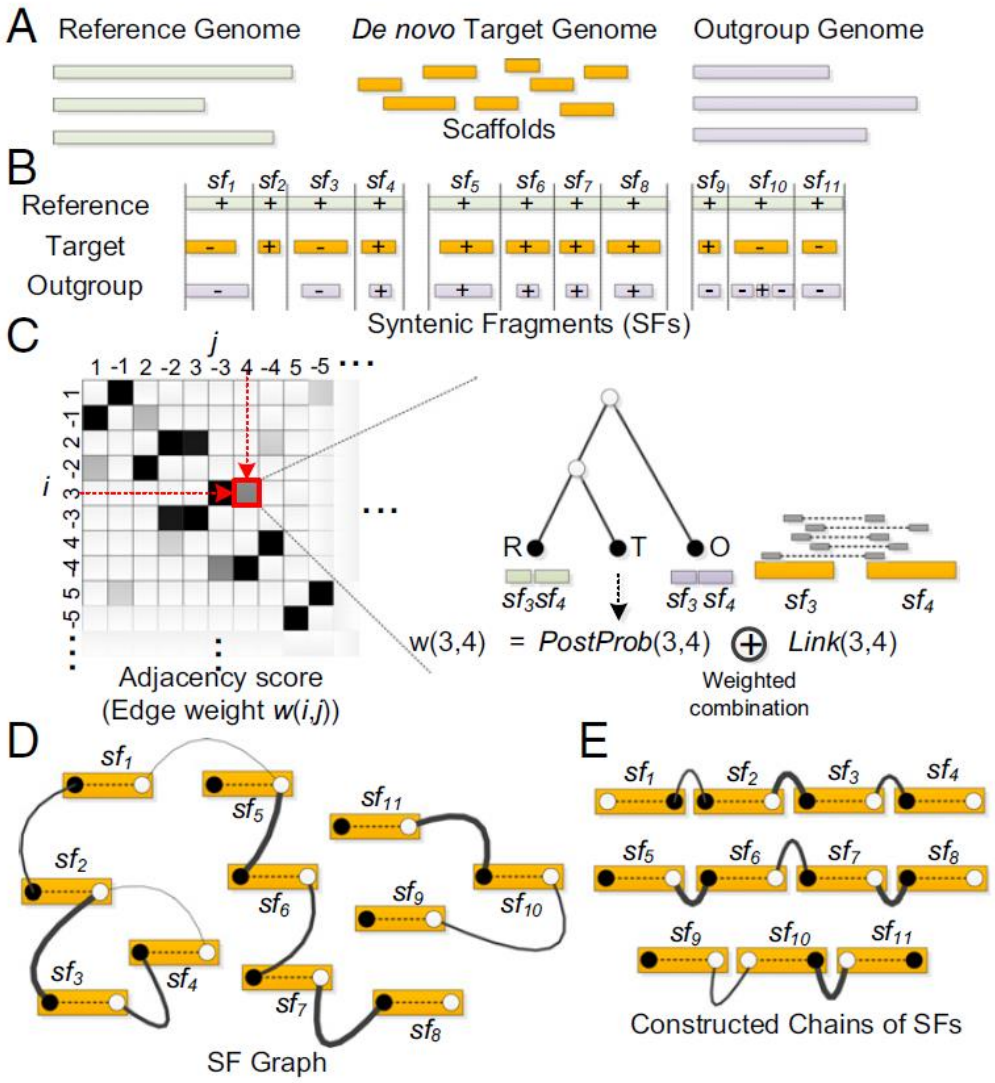

Figure 2-1: Overview of the RACA algorithm. (A) RACA takes a reference, a *de novo* sequenced target (in scaffolds), and one or more outgroup genomes as input data. (B) Syntenic fragments (SFs) delimited by vertical dashed lines are constructed by aligning reference and target genome sequences and merging collinear alignments. The outgroup may not always be aligned to SFs (e.g., sf2) and may contain rearrangements within a SF (e.g., sf10). Pluses and minuses represent the orientations of the target and outgroup DNA sequences on the reference genome, and three groups of SFs represent different reference chromosomes. (C) For each pair of SFs, an adjacency score (edge weight) that combines the posterior probability of the adjacency and the coverage of paired-end reads is calculated. (D) The SF graph is built by connecting SFs whose edge weight in C is higher than a certain threshold. Head (closed circle) and tail (open circle) vertices from the same SF are always connected with a maximum weight (dashed edge). (E) Constructed chains of SFs that are extracted by the RACA algorithm. From (Kim et al., 2013).

2.2 Material and methods

Avian genome assemblies, sequencing reads and annotations

The chicken (*Gallus gallus*; ICGSC Gallus_gallus 4.0; Hillier, 2004) and zebra finch (*Taeniopygia guttata*; WUGSC 3.2.4; Warren et al., 2010) genome assemblies were downloaded from the UCSC Genome Browser (Kent et al., 2002). Scaffold-based assemblies (N50 >2 Mbp; except Pekin duck with N50 of 1.26 Mbp), gene and repeat annotations for 18 avian genomes were downloaded from the GigaScience Database (Table 2-1; Sneddon et al., 2012, Zhang et al., 2014a). Sequence reads for the same 18 avian species were obtained from Dr Cai Li (BGI-Shenzhen). Scaffold assemblies were filtered for scaffolds <10 Kbp long using faFilter from Kent utilities (Kent et al., 2003). The ostrich (*Struthio camelus*) optical map assisted super-scaffold assemblies were obtained from the GigaScience Database (Sneddon et al., 2012). The Pekin duck (*Anas platyrhynchos*) radiation hybrid (RH) map was obtained from Dr Thomas Faraut (INRA, France), and Dovetail assisted rock pigeon super-scaffold assembly from Professor Thomas Gilbert (University of Copenhagen, Denmark). Divergence times and topologies were obtained from the total evidence nucleotide tree (TENT) reported in (Jarvis et al., 2014).

Pairwise alignments

Pairwise alignments using zebra finch chromosome assembly as reference and medium ground finch or emperor penguin scaffold assemblies as targets were generated using SatsumaSynteny (v.3.0; Grabherr et al., 2010) with "-m 1 -ni 10 -chain_only" parameters. Pairwise alignments using chicken or zebra finch chromosome assemblies as references and all other assemblies as targets were generated with LastZ (v.1.02.00; Harris, 2007) using the following parameters: *C=0 E=30 H=2000 K=3000 L=2200 O=400*. Both sets of pairwise alignments were converted into the UCSC “chain” and “net” alignment formats with axtChain (parameters: *-minScore=1000 -linearGap=medium -verbose=0*) followed by chainAntiRepeat, chainSort, chainPreNet, chainNet and netSyntenic, all with default parameters (Kent et al., 2003).

Read mapping

Sequence read qualities were assessed using the FastQC software (v0.10.1; Andrews, 2010). Mapping of each target genome mate pair and paired-end read libraries to their scaffold assemblies were performed with Bowtie2 (v2.0.1; Langmead and Salzberg, 2012) using the parameters suitable for each sequencing library, including read trimming based on FastQC results. For the libraries with read length ≤90 bp the following settings were used: *-N 1 -3 5 --no-discordant*. For libraries with reads >90 bp: *-N 1 -3 30 --no-discordant.* In addition, for mate pair libraries (>2 Kbp insert size) we used *–rf* option. Read pairs mapping to the same scaffold, but with mapping distance larger than (library insert size) ± (0.5 x insert size) were discarded.

Predicted chromosome fragments

The RACA algorithm (Kim et al., 2013) allows the adjustment of multiple parameters for a more accurate reconstruction of a species PCFs. One can adjust the minimum length of the SFs to be included in the reconstruction (RESOLUTION) and the size of the window used to estimate the physical coverage (WINDOWSIZE). Physical coverage differs from sequencing depth as it corresponds to the number of times a base is read (depth) or spanned by paired-end or mate pair reads (Figure 2-2; Meyerson et al., 2010). The disregard of adjacencies without support from sequencing data (IGNORE_ADJS_WO_READS) can also be set in RACA. In addition, RACA calculates the minimum physical coverage for each SF adjacency that could represent a potential evolutionary breakpoint or chimeric joint within a scaffold (Kim et al., 2013). A cut-off based on the distribution of these values (in percentage; MIN_INTRACOV_PERC) can also be set. This percentage is converted to the actual minimum physical coverage necessary for RACA to consider an SF adjacency within a scaffold as a putative breakpoint region or chimeric joint. If an SFs adjacency within scaffold is supported by physical coverage below the set threshold, it will be considered as a putative chimeric joint (and scaffold).

For the establishment of the optimal RACA parameters for avian chromosome reconstructions, RACA was tested for: (a) resolutions of SFs: 50, 80 and 150 Kbp (default); (b) window sizes of 10 bp, 50 bp, 100 bp, 1 Kbp, 10 Kbp (default), 20

Kbp, 50 Kbp, 100 Kbp and 150 Kbp; (c) ignoring, or not (default), adjacencies without paired-end read support; and (d) for physical coverage distribution cut-offs of 0%, 1E-13%, 1E-03%, 1% and 5% (default).

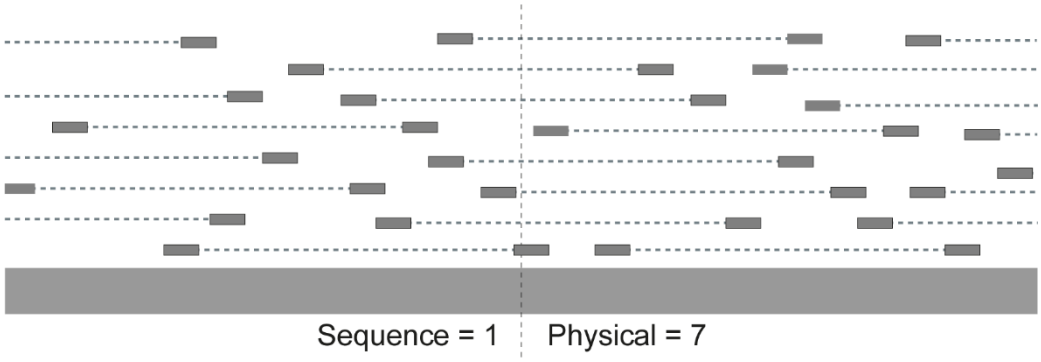

Figure 2-2: Sequencing and physical coverage. Sequence reads are represented as grey bars with dashed lines connecting paired-ends. Sequence coverage is the number of sequence reads that cover the site and physical coverage is the number of fragments (delimited by paired-end reads) that spans the site. Adapted from (Meyerson et al., 2010).

After the establishment of avian RACA default parameters, PCFs were generated for each of the 18 avian scaffold-level assemblies. The reference and outgroup species were selected from chicken and zebra finch genomes accordingly to divergence times, except when the target species diverged from both chicken and zebra finch at least 67 million years ago (MYA). For those species, we obtained RACA reconstructions using both chicken and zebra finch as references. Then, we chose as reference the genome causing the lowest number of split scaffolds in the resulting PCFs.

Two rounds of RACA were performed for each species. The initial run using the avian default parameters: *WINDOWSIZE=10 RESOLUTION=150000 MIN_INTRACOV_PERC=5*. Before the second run of RACA, we tested the structure of the putative chimeric scaffolds reported in the initial RACA round, using PCR amplification across the split adjacencies. Based on the PCR results we established thresholds for the minimum physical coverage across the SF joint intervals that allowed us to separate scaffolds that were likely to be chimeric from those that were likely to be real. These thresholds were used to update the MIN_INTRACOV_PERC parameter. Confirmed scaffolds were also included in the RACA “reliable adjacencies” input file. For species we did not have DNA to

test the scaffolds' structure, we used the physical coverage threshold from the tested species with the closest genome-wide physical coverage.

Verification of putative chimeric scaffolds

Verification of putative chimeric scaffolds was performed in two stages. First, we tested SF joints within the scaffolds split by RACA. Then, we tested RACA predicted adjacencies for those SFs joints with negative PCR results on the first stage. Primers were designed using Primer3 software (v.2.3.6; Untergasser et al., 2012). To minimise misclassification of EBRs as chimeric joints, and chimeric joints as EBRs, we selected primers only within the sequences that had high-quality alignments between the target and reference genomes found in adjacent SFs. Additionally, only those SFs joints with the total expected size of PCR product <6 Kbp were selected for testing. Due to alignment and SF detection settings, some of the intervals between adjacent SFs could be >6 Kbp and primers could not be chosen for a reliable PCR amplification based on RACA results. In such cases, we used CASSIS software (Baudet et al., 2010). CASSIS is designed to detect and report the narrowest genomic breakpoints region between two SFs. Thus, this software could in some cases be used to narrow gaps between adjacent SFs.

Whole blood was collected aseptically from adult peregrine falcon, rock pigeon and ostrich, and muscle tissue was obtained from adult emperor penguin. DNA was isolated using DNeasy Blood and Tissue Kit (Qiagen) following standard protocols. PCR amplification was performed in a volume of 10 microliters (µL) as follows: 5 µL of Taq PCR Master Mix (Qiagen; for expected PCR product lengths up to 2 Kbp) or DreamTaq Master Mix (Fermentas; for expected product lengths from 2 to 6 Kbp), 1 µL of each primer at 2 µM and ≈30 ng DNA. The PCR amplification was carried out in a T100 Thermal Cycler (BioRad) using the following profile: initial denaturation at 95°C for 3 min; 35 cycles for 30 secs at 94°C, 1 min/Kbp at 55-60°C and 1 min at 72°C; and final extension at 72°C for 10 min. DNA fragments were stained with SYBR Safe (Invitrogen), separated on a 1.5% (for expected PCR product lengths up to 2 Kbp) or 1% (for expected PCR product lengths from 2 to 6 Kbp) agarose gel and visualised in a ChemiDOC MP system (BioRad).

EBR detection and CNE density analysis

The pairwise genome alignments of the chicken, zebra finch and the 18 PCF assemblies were obtained using LastZ and converted into “chain” and “net” format as described above. Alignments were converted to multiple alignment format (MAF) using the netToAxt, axtSort and axtToMaf tools from Kent utilities (Kent et al., 2003). Pairwise synteny blocks were defined at 100, 300 and 500 Kbp resolution using the maf2synteny tool (Kolmogorov et al., 2016). Conserved non-coding elements (CNEs) obtained from the alignments of 48 avian genomes were obtained from (Farré et al., 2016). Using chicken as the reference genome, EBRs were detected and classified using the *ad hoc* statistical approach described in (Farré et al., 2016). Only the EBRs defined to ≤100 Kbp were utilised for the CNE analysis. EBRs <1 Kbp were extended ±1 Kbp. 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 Kruskal-Wallis test followed by Mann-Whitney U test.

Comparing CNE densities in EBRs and msHSBs

Chicken chromosomes (GGA; excluding GGA16, W and Z) were divided into 1 Kbp non-overlapping windows. 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 Mbp (Farré et al., 2016), avian EBRs flanking fusions and intrachromosomal EBRs, or the intervals found in the rest of the chicken genome. We estimated the average CNE density for each window type using bedtools (v.2.20-1; Quinlan and Hall, 2010). Differences between window types were tested for significance using the Kruskal-Wallis test followed by Mann-Whitney U test.

Densities of TEs in intrachromosomal and fusion-flanking EBRs

Intrachromosomal and chromosome fusion-flanking peregrine falcon, downy woodpecker and budgerigar lineage-specific EBRs were detected from the PCF assemblies. Only EBRs defined to ≤100 Kbp were used for this analysis. Scaffold-

based TE annotations for peregrine falcon, downy woodpecker and budgerigar were downloaded from GigaScience Database (Sneddon et al., 2012, Zhang et al., 2014a) and converted to PCF 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 between EBR types 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).

Table 2-1: Chromosome number and scaffold assemblies' statistics (scaffolds ≥ 10 Kbp) for the 18 selected avian species.

Species	Common name	Haploid number (n)	Sequencing depth (X)	No. scaffolds	N50 (Mbp)	Total length (Gbp)	% of original assembly
Anas platyrhynchos	Pekin duck	40	50	2,368	1.26	1.08	97.27
Aptenodytes forsteri	Emperor penguin	36	61	682	5.08	1.25	99.24
Calypte anna	Anna's hummingbird	37	110	887	4.30	1.05	94.20
Chaetura pelagica	Chimney swift	Unk.	103	1,172	3.88	1.10	98.01
Charadrius vociferus	Killdeer	38	100	1,598	3.68	1.21	98.60
Columba livia	Rock pigeon	40	63	1,081	3.15	1.10	98.66
Corvus brachyrhynchos	American crow	40	80	1,156	7.08	1.08	98.93
Cuculus canorus	Common cuckoo	36	100	900	2.99	1.15	98.77
Egretta garzetta	Little egret	33	74	1,195	3.11	1.20	98.93
Falco peregrinus	Peregrine falcon	25	105	723	3.94	1.17	99.47
Geospiza fortis	Medium ground finch	Unk.	115	1,168	5.28	1.04	96.99
Manacus vitellinus	Golden-collared manakin	Unk.	110	954	2.86	1.05	90.25
Melopsittacus undulatus	Budgerigar	29	160	1,138	11.41	1.08	96.97
Nipponia nippon	Crested ibis	34	105	1,479	5.35	1.20	96.72
Opisthocomus hoazin	Hoatzin	40	100	1,620	2.94	1.20	98.90
Picoides pubescens	Downy woodpecker	46	105	1,944	2.12	1.15	97.66
Pygoscelis adeliae	Adélie penguin	48	60	819	5.23	1.21	98.56
Struthio camelus	Ostrich	40	85	1,179	3.64	1.22	99.34

2.3 Results

Software selection for avian whole genome pairwise alignments

Comparative genomics aims to find genomic differences and similarities, and requires pairwise nucleotide alignments at maximum sensitivity, specificity and acceptable speed. This led to the development of multiple whole-genome alignment software. Examples of such algorithms are LastZ (Harris, 2007) and SatsumaSynteny (Grabherr et al., 2010). These two algorithms are well established in comparative genomics research and allow the generation of synteny maps between two species. Herein, we tested which of these two programs would provide a more complete pairwise whole-genome alignment of avian genomes. To do that we obtained the pairwise alignments between medium ground finch and emperor penguin scaffold-level assemblies and zebra finch chromosomes using both algorithms. The resulting alignments were used as input in RACA reconstructions. We observed that LastZ creates longer alignments than SatsumaSynteny, which also results in a higher genome recovery in RACA. Using LastZ alignments, RACA reconstructed 992 Mbp of medium ground finch genome while using SatsumaSynteny alignments 988 Mbp were recovered. For emperor penguin, using LastZ alignments, 1.22 Gbp of the genome was recovered against the 1.21 Gbp recovered using SatsumaSynteny alignments. The selection of the alignment software to use prior to RACA reconstructions was then made accordingly to the reference coverage of the resulting alignments, and the genome recovery rate after RACA reconstructions, and resulted in the selection of LastZ to create the pairwise alignments for all the species.

Reference genome selection

Reference and outgroup genomes for each RACA reconstruction were either the chicken or zebra finch genome assemblies, the best quality avian genome assemblies available to date. RACA was shown to be more accurate when the reference genome is closely related to the target genome, due to a higher genome coverage in the pairwise alignments (Kim et al., 2013) and, in many cases, a more similar chromosome structure. Because of that, the selection of the reference and outgroup genomes was made according to the divergence times of chicken and zebra finch from each of the target species, except when

both chicken and zebra finch diverged at least 67 MYA from the target, just before the Galloanseres split (Jarvis et al., 2014). For these latter cases (common cuckoo, hoatzin, ostrich, Anna's hummingbird, chimney swift, killdeer and rock pigeon) we obtained independent RACA reconstructions using both chicken and zebra finch as references. The selected final reference was the one inducing the lowest number of scaffold splits in the resulting PCFs, which indicates a more similar genome structure (Table 2-2). Overall, 12 reconstructions were performed using zebra finch genome as the reference (core landbirds, core waterbirds, killdeer and common cuckoo) and 6 using chicken genome as the reference (hoatzin, Anna's hummingbird, chimney swift, rock pigeon, Pekin duck and ostrich) (Figure 2-3).

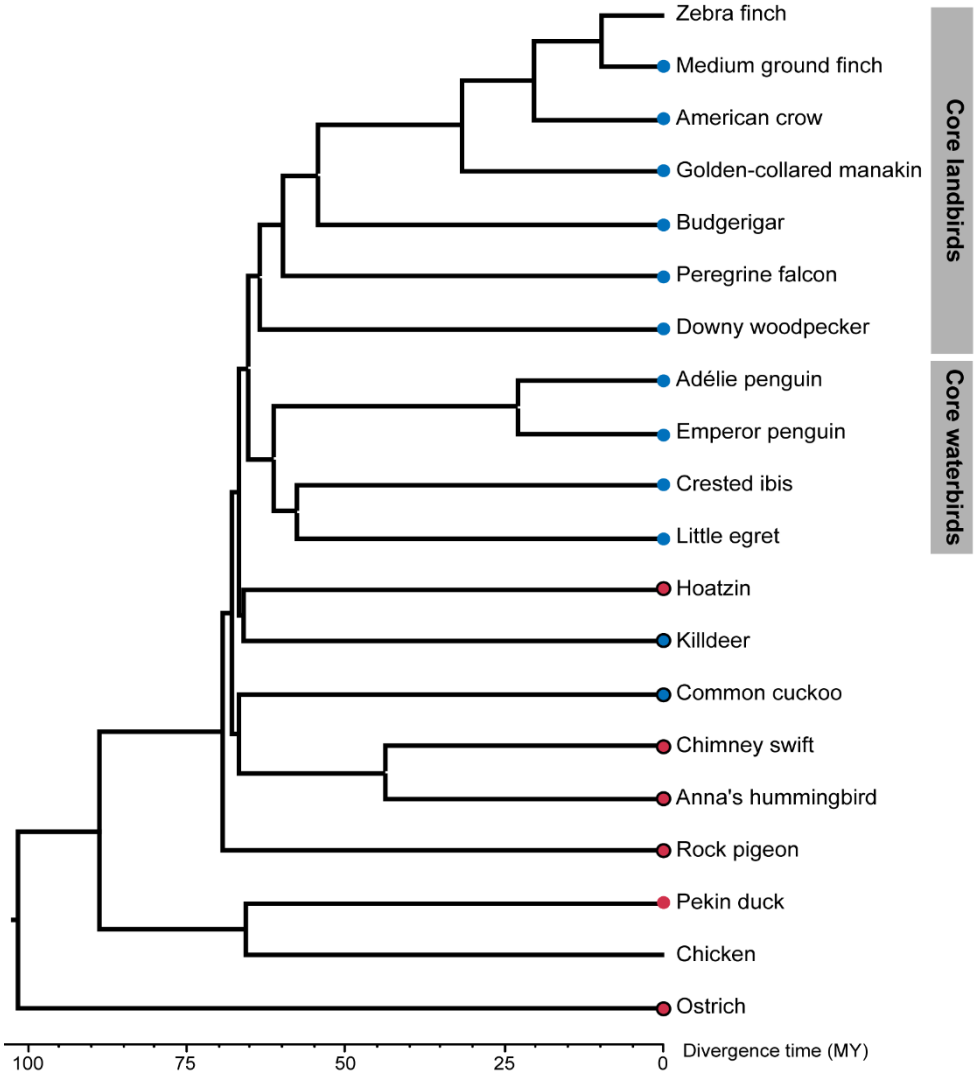

Figure 2-3: Cladogram presenting the selected reference for each RACA reconstruction. Red dots represent chicken as a reference and blue dots zebra finch as reference. Non-outlined dots represent selection by divergence time and outlined dots highlight species with selection based on the fraction of split scaffolds. Adapted from (Jarvis et al., 2014).

Table 2-2: Reference selection for species diverged more than 67 MYA from both chicken and zebra finch.

	Chicken			Zebra finch			
Species	**Divergence time (MY)**	**No. split scaffolds**	**No. unique split scaffolds**	**Divergence time (MY)**	**No. split scaffolds**	**No. unique split scaffolds**	**Selected reference**
Common cuckoo	89	54	26 (48%)	68	31	6 (19%)	Zebra finch
Hoatzin	89	12	1 (8%)	67	19	10 (53%)	Chicken
Ostrich	102	31	2 (6%)	102	30	3 (10%)	Chicken
Anna’s hummingbird	89	34	8 (24%)	68	30	8 (27%)	Chicken
Chimney swift	89	23	2 (9%)	68	23	4 (17%)	Chicken
Killdeer	89	8	2 (25%)	67	9	2 (22%)	Zebra finch
Rock pigeon	89	17	1 (6%)	69	22	8 (36%)	Chicken

RACA with default parameters for avian genomes

One of the first steps of this project was the establishment of the optimal RACA parameters for avian chromosomes reconstruction. The RACA parameters tested were SF resolution, window size, ignoring/not ignoring SF adjacencies without sequencing data support and minimum physical coverage threshold to connect SFs.

The SF resolution sets the minimum length of the SFs that will be included in the chromosome reconstructions. To establish this parameter, RACA was run with SF resolutions of 150, 80 and 50 Kbp with medium ground finch data. We verified that the lowest number of PCFs was obtained at the lowest resolution (150 Kbp) and using the higher resolutions allowed for the recovery of 0.5% of additional genome length. Forty-four PCFs comprising 992 Mbp were generated at 150 Kbp resolution, and 47 PCFs comprising 997 Mbp were obtained at 50 Kbp. These results showed that using higher resolutions does not significantly improve the obtained reconstructions. Moreover, the use of short SFs may introduce errors in the reconstructed PCFs caused by small misassemblies present in scaffolds. Considering all this, we established 150 Kbp as the default avian RACA resolution.

After running RACA for Pekin duck with default parameters, we noted a high number of split scaffolds (N=56). Aiming the reduction of this number, we checked if this was caused by an unoptimised parameter. All RACA parameters that can affect the number of split scaffolds were tested, namely window size, ignoring/not ignoring SF adjacencies without read support and the minimum physical coverage threshold.

The window size determines the length of the window used to calculate the physical coverage genome-wide. A precise determination of physical coverage is crucial for an accurate distinction of chimeric scaffolds from those harbouring lineage-specific structural differences. We compared Pekin duck RACA reconstructions using a window size of 10 bp, 50 bp, 100 bp, 1 Kbp (default), 10 Kbp, 20 Kbp, 50 Kbp, 100 Kbp and 150 Kbp. The lowest number of putative chimeric scaffolds (split scaffolds) was detected with window sizes of 10 Kbp or longer. However, the obtained number (N=50) was not much different from the

obtained with default parameters (N=56) or using a window size of 10 bp (N=53). In this way, to more accurately estimate the physical coverage across the genome we decided to use a more sensitive window size of 10 bp as default for avian RACA reconstructions.

RACA allows disregarding all the SFs adjacencies without paired-end read support. Using this option, the number of split scaffolds in Pekin duck PCFs decreased from 56 to 24, but the number of PCFs increased from 173 to 664. Because of that, we decided not to use this option in the subsequent reconstructions.

The physical coverage percentage cut-off sets the minimum physical coverage (i.e. the minimum number of spanning paired-end read pairs) used to maintain or break SF adjacencies during reconstruction. We analysed the differences in RACA results when this option was set to 0%, 1E-13%, 1E-3%, 1% and 5% (default). We verified that the number of putative chimeric scaffolds only differ by one from the default (N=56) to all the other experiments (N=55). Therefore, we decided to maintain 5% as the avian default. However, this parameter will be adjusted after the PCR verification of putative chimeric scaffolds, when we will establish the threshold that allows us to distinguish “true” chimeric from non-chimeric scaffolds.

PCF reconstruction using RACA default avian parameters

After the establishment of the RACA avian default parameters, we ran RACA for 18 avian species (scaffold assemblies N50 >1 Mbp). We noted a significant improvement on the genome assemblies’ continuity, shown by 90% average reduction in the number of assembly fragments and nine times higher N50 (Table 2-3, Supplemental Tables 1-18). The obtained PCFs cover ~96% of the scaffold-level assemblies. In addition, we also noted that on average 10% of the scaffolds used to construct the PCFs were split by RACA (Table 2-3, Supplemental Tables 1-18), which was higher than the 6% reported previously for NGS genome assemblies (Kim et al., 2013). Because of RACA inability to properly distinguish chimeric scaffolds from those containing target-specific structural variations, this high number of split scaffolds might negatively affect the accuracy of the reconstructions.

Table 2-3: Statistics of the PCF reconstructions using RACA avian default parameters.

Species	No. PCFs	Total length (Gbp)	Genome coverage (%)	N50 (Mbp)	No. used scaffolds	No. PCFs homeologous to reference chromosomes	% putatively chimeric scaffolds	No. lineage-specific EBRs	No. fragments reduction (%)	N50 increase (x)
Medium ground finch	46	0.99	95.19	55.14	373	21	5.63	8	96.06	9.44
Budgerigar	84	1.04	95.94	46.54	254	6	31.50	13	92.62	3.08
American crow	89	1.03	95.11	53.50	275	6	8.36	7	92.30	6.55
Emperor penguin	94	1.22	97.76	40.25	408	10	6.13	9	86.22	6.93
Golden-collared manakin	95	1.02	97.42	45.22	566	3	5.30	17	90.04	14.79
Little egret	100	1.15	95.91	37.58	601	6	5.66	9	91.63	11.09
Ostrich	100	1.17	95.90	37.95	588	7	10.03	13	91.52	9.42
Crested ibis	101	1.15	95.83	41.94	422	8	8.29	10	93.17	6.85
Anna's hummingbird	102	1.02	97.14	35.86	431	4	15.08	16	88.50	7.35
Adélie penguin	105	1.17	96.85	39.67	433	5	7.39	9	87.18	6.59
Killdeer	111	1.15	95.20	37.56	537	3	3.91	13	93.05	9.21
Peregrine falcon	113	1.14	97.60	27.44	478	8	15.06	0	84.37	5.97
Common cuckoo	114	1.12	97.82	35.04	581	8	12.39	10	87.33	10.72
Downy woodpecker	116	1.05	91.54	25.71	755	7	13.51	12	94.03	11.14
Chimney Swift	130	1.06	96.10	28.48	524	7	7.82	7	88.91	6.35
Rock pigeon	150	1.07	97.54	34.54	572	4	13.64	1	86.12	9.97
Hoatzin	158	1.14	95.32	35.40	680	2	3.82	8	90.25	11.04
Pekin duck	173	1.00	93.02	28.31	1,144	2	4.55	7	92.69	21.10

PCR verification of split adjacencies

As referred previously, the scaffolds split by RACA could be chimeric, resulting from assembly errors, or encompass regions of structural differences critical for the accurate reconstruction of the target species chromosomes. Aiming the reduction of misidentification of these events, we used PCR amplification as a method to distinguish which scaffolds were truly chimeric and which contained target-species-specific chromosome structural changes (Figure 2-4). We could test RACA split scaffolds for ostrich, peregrine falcon, rock pigeon and emperor penguin, as we had access to genomic DNA samples for these species. Due to PCR limitations, we only performed this verification when the split regions were defined to <6 Kbp in the target species scaffolds.

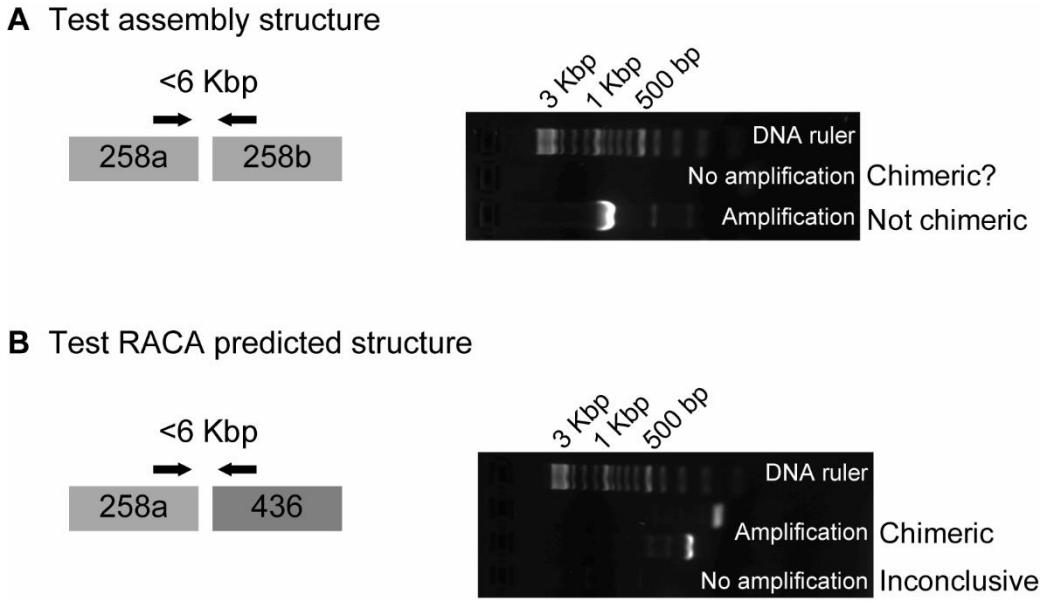

Figure 2-4: PCR verification strategy. For scaffolds split by RACA with split interval <6 Kbp we (A) verify the assembly reported structure by designing primers flanking the split adjacency. If there is amplification across the region, the assembly structure was considered correct and the scaffold was considered non-chimeric. If there is no amplification the scaffold was considered possibly chimeric and (B) the RACA predicted adjacency with another scaffold was tested. Primers flanking RACA predicted adjacency are designed. If there is amplification RACA predicted structure is considered correct and the scaffold is chimeric. If there is no amplification PCR testing in considered inconclusive.

We tested a total of 49, 69, 49 and 23 split regions for peregrine falcon, rock pigeon, ostrich and emperor penguin (Table 2-4). These represent all SF joints defined to <6 Kbp, corresponding to 57, 63, 71 and 82 percent of all RACA-split scaffolds, respectively. From these, the lowest percentage of confirmed SF

adjacencies was 43% in emperor penguin and the highest 84% in rock pigeon and peregrine falcon. For ostrich, 65% of the tested adjacencies were confirmed to be non-chimeric (Table 2-4). For the split SF adjacencies with negative PCR results, we tested the alternative (RACA-suggested) order of the flanking SFs when the split region was defined to <6 Kbp in chicken coordinates. We obtained amplicons for 2/4 in peregrine falcon, 7/7 in rock pigeon, 7/8 in ostrich and 5/10 in emperor penguin of the tested adjacencies, confirming the proper RACA identification of these scaffolds as chimeric (Table 2-4).

To estimate which of the remaining 5 to 40 non-tested split regions (Table 2-4) were likely to be chimeric, we identified the minimum physical coverage level in the SFs joining regions for which (and higher) the PCR results were most consistent with RACA predictions (Figure 2-5). These thresholds are 50X to emperor penguin, 85X for rock pigeon, 239X for ostrich and 583X for peregrine falcon (Table 2-4). Based on our estimates, when used without further PCR verification, these thresholds would lead to splitting of nearly all scaffolds with large structural misassemblies in peregrine falcon, ostrich and emperor penguin, and ~6% still maintained in the rock pigeon PCFs (Table 2-4). In addition, the fraction of scaffolds containing real structural differences that would be split was estimated as 56%, 43%, 52% and 35% in peregrine falcon, rock pigeon, ostrich and emperor penguin, respectively (Table 2-4).

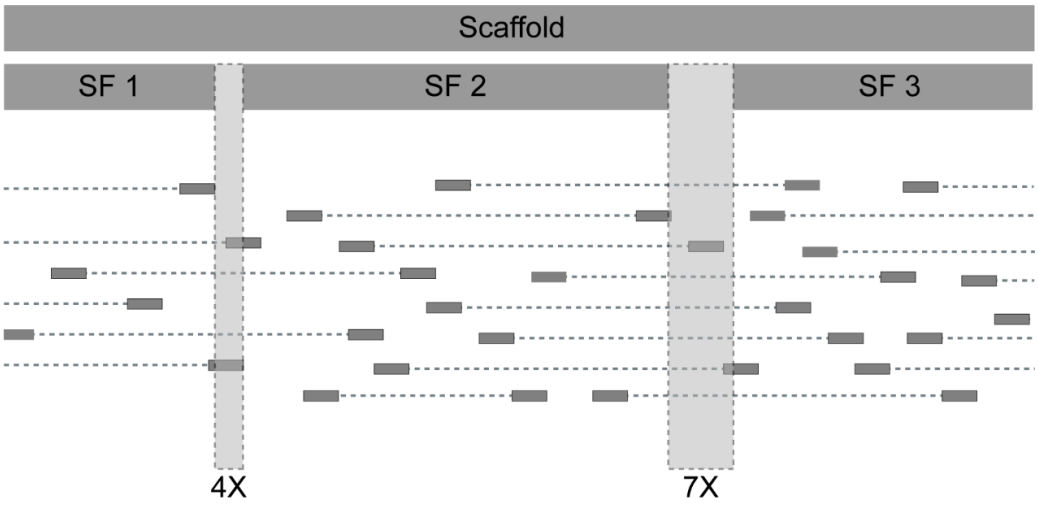

Figure 2-5: Physical coverage threshold establishment. Assuming that both SF1-2 and SF2-3 adjacencies were verified to be non-chimeric. The physical coverage threshold would be minimum physical coverage observed in these two cases (4X).

Aiming at the distinction of scaffolds more likely to be chimeric from scaffolds with real structural chromosomal differences for species that we were not able to test by PCR, we checked how reliable would be a threshold extrapolation from the tested species. Herein we used the mean genome-wide physical coverage to infer each threshold. We found that the extrapolated thresholds for emperor penguin and rock pigeon would be more stringent than those established by PCR (239X versus 50X, and 239X versus 85X, respectively). This would result in a stricter definition of chimeric scaffolds, leading to an increased number of split scaffolds and a reduction in the number of detected EBRs in the PCF assembly. We also noted the same when looking at the Pekin duck radiation hybrid (RH) map. The chimeric scaffolds detected therein have physical coverage lower than 15X in the split intervals and the inferred threshold for this species would be 50X. For peregrine falcon and ostrich, the pattern was the inverse. The PCR-established thresholds were more stringent than those inferred from the other tested species, which could lead to the misinterpretation of some chimeric scaffolds. Nonetheless, these inferred thresholds were still more stringent than the RACA default avian parameter.

Table 2-4: Statistics for the scaffold split regions tested by PCR.

Statistics	Peregrine falcon	Rock pigeon	Ostrich	Emperor penguin
Pair-end read physical coverage within tested scaffolds (X)	9-983	3-416	2-604	1-196
No. RACA-split SF adjacencies (default param.)	85	109	69	28
No. tested split SF adjacencies	49 (100%)	69 (100%)	49 (100%)	23 (100%)
No. amplified adjacencies (confirmed adjacencies)	41 (84%)	58 (84%)	32 (65%)	10 (43%)
No. non-amplified adjacencies	8 (16%)	11 (16%)	17 (35%)	13 (57%)
No. tested RACA-suggested SF adjacencies	4	7	8	10
No. amplified adjacencies (chimeric adjacencies)	2	7	7	5
No. non-amplified adjacencies	2	0	2	5
No. ambiguous SF adjacencies from tested adjacencies	6	4	10	8
Selected physical coverage threshold	583	85	239	50
No. tested split SF adjacencies below selected threshold	16 (100%)	7 (100%)	29 (100%)	20 (100%)
No. chimeric SF adjacencies	2 (13%)	3 (43%)	7 (24%)	5 (25%)
No. confirmed SF adjacencies	9 (56%)	3 (43%)	15 (52%)	7 (35%)
No. ambiguous SF adjacencies	5 (31%)	1 (14%)	7 (24%)	8 (40%)
No. tested split SF adjacencies above selected threshold	33 (100%)	62 (100%)	20 (100%)	3 (100%)
No. chimeric SF adjacencies	0 (0%)	4 (6%)	0 (0%)	0 (0%)
No. confirmed SF adjacencies	32 (97%)	55 (89%)	17 (85%)	3 (100%)
No. ambiguous SF adjacencies	1 (3%)	3 (5%)	3 (15%)	0 (0%)
No. non-tested SF split adjacencies	36	40	20	5

PCF reconstruction using RACA with adjusted parameters

The second round of RACA reconstructions was performed for each of the 18 selected genome assemblies. For peregrine falcon, rock pigeon, ostrich and emperor penguin, RACA parameters were adjusted based on the PCR-verified physical coverage threshold and by introducing the confirmed SFs adjacencies as “reliable adjacencies”. For the non-tested species, we utilised the threshold of the PCR-tested genome with the closest genome-wide physical coverage average.

The obtained RACA assemblies contain on average 117 PCFs (ranging from 86 in budgerigar to 175 in Pekin duck) comprising ~96% of their original scaffold genome assembly. The reduction in the number of assembly fragments is ~90% and the N50 increased ~7 times. The improvement of the assemblies’ continuity can be noted in Figure 2-6. Moreover, the fraction of split scaffolds in these assemblies varies between 2% and 8% (Table 2-5, Supplemental Tables 1-18), in agreement with the expected 6% reported for NGS genomes (Kim et al., 2013). The only exception was budgerigar for which 21% of the used scaffolds were split by RACA. Interestingly, more than one-quarter of the split scaffolds supports interchromosomal rearrangements that were not previously identified using chromosome paintings (Nanda et al., 2007), which might indicate multiple misassembled regions in the budgerigar genome.

For each species, 2 to 10 PCFs are completely homeologous to complete reference species chromosomes. Moreover, for all target species one of their PCFs was completely homeologous to the chicken or zebra finch chromosomes 17, which are also homeologous to one another, and for most of them (12 out of 18 species) another PCF was completely homeologous to chromosomes 25 (Supplemental Tables 1-18).

Most of the detected lineage-specific EBRs correspond to intrachromosomal rearrangements, however, for six species we were also able to detect chromosomal fusions. Those species were little egret, peregrine falcon, budgerigar, hoatzin, downy woodpecker and Adélie penguin (Supplemental Tables 1-18). Interestingly, some of the budgerigar fusions were already detected

using chromosome paintings. For instance, the fusion of chicken chromosomes 6 and 7, and 8 and 9 (Nanda et al., 2007).

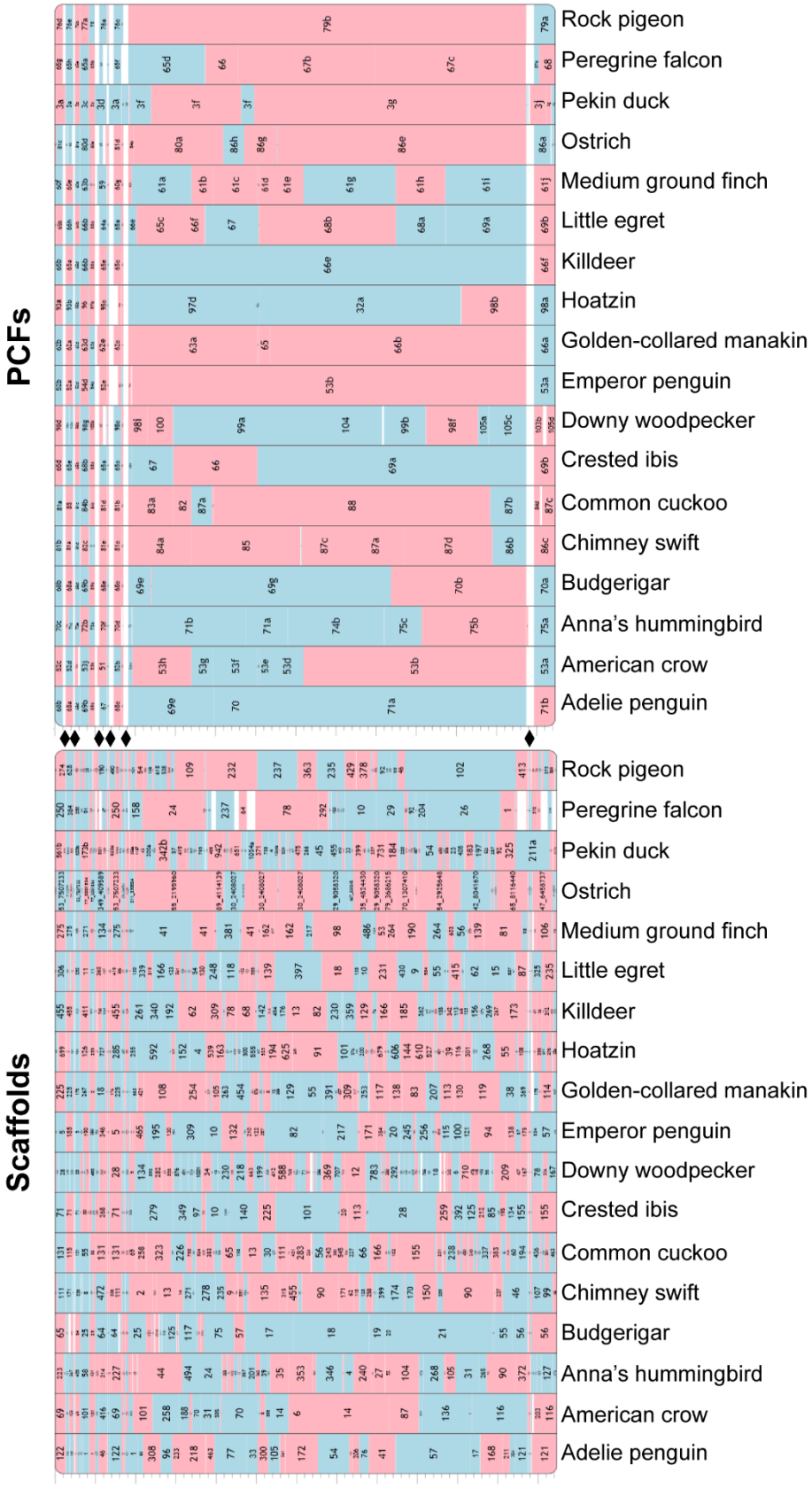

Figure 2-6: Continuity comparison for scaffold and PCF genome assemblies in chicken chromosome 3. Blue and pink blocks define syntenic fragments in the target genomes in "+" and "-" orientation, respectively compared to the chicken chromosome. Numbers within blocks represent scaffold or PCF identifiers. Black diamonds demarcate chicken specific breakpoint regions that are detectable from the scaffold and PCF assemblies.

Table 2-5: Statistics of the refined PCF assemblies obtained for the 18 avian species.

Species	Average physical coverage (X)	Physical coverage threshold (X)	No. PCFs	Total length (Gbp)	Genome coverage (%)	N50 (Mbp)	No. used scaffolds	No. PCFs homeologous to reference chromosomes	% putatively chimeric scaffolds	No. lineage-specific EBRs[1]
Budgerigar	246	50	86	1.04	95.94	38.57	254	8	21.26	81
American crow	544	85	88	1.03	95.11	53.50	275	6	6.18	16
Medium ground finch	509	85	91	0.99	95.19	36.73	373	6	3.49	17
Emperor penguin	375	50	94	1.22	97.76	40.25	408	10	6.13	10
Peregrine falcon	873	583	97	1.14	97.60	26.78	478	6	4.81	58
Golden-collared manakin	854	583	97	1.02	97.42	36.36	566	3	1.94	40
Killdeer	415	239	101	1.15	95.20	34.89	527	5	1.71	22
Little egret	529	85	103	1.15	95.91	30.49	601	7	2.83	26
Crested ibis	776	583	105	1.15	95.83	38.73	422	7	7.58	14
Adélie penguin	471	239	105	1.17	96.85	39.67	433	5	5.77	18
Anna's hummingbird	726	583	114	1.02	97.14	22.39	431	4	7.89	75
Chimney Swift	444	239	132	1.06	96.10	27.05	524	7	4.39	26
Rock pigeon	518	85	134	1.07	97.54	22.17	572	5	2.97	72
Ostrich	429	239	136	1.17	95.90	28.09	588	3	5.27	41
Common cuckoo	741	583	137	1.11	96.94	20.85	569	5	5.45	67
Downy woodpecker	527	85	144	1.05	91.54	14.74	755	2	2.12	127
Hoatzin	472	239	163	1.14	95.32	25.07	680	2	1.76	24
Pekin duck	248	50	175	1.00	93.02	21.32	1,144	2	2.88	26

[1] In comparison with reference and outgroup genomes.

RACA reliability for avian chromosome reconstruction

We assessed RACA reliability to reconstruct avian chromosomes by comparing the adjacencies it produces with existing super-scaffold assemblies of ostrich, Pekin duck and rock pigeon. These assemblies were obtained from the GigaScience Database (Sneddon et al., 2012), Dr Thomas Faraut (INRA, France), and Professor Thomas Gilbert (University of Copenhagen, Denmark), respectively. They were created with the assistance of physical maps what allowed us to compare the structure of RACA PCFs with the traditional and novel methods used for the creation of chromosome-level assemblies. Because different assembly methods work at different resolutions of contig/scaffold ordering and orienting, the following comparisons were performed at 150 Kbp SF resolution (resolution of RACA results) and only included those scaffolds found in both the compared assemblies.

Putatively chimeric scaffolds

We started with investigating the agreement in the number of split scaffolds between the super-scaffold and the PCF assemblies. These would be the number of scaffolds that were considered as chimeric by the RACA and an independent scaffolding method, allowing us to assess RACA reliability to detect regions of misassemblies on avian genomes. In this comparison, only split intervals located within ±5 Kbp from each other were considered as overlapping between the super-scaffold and PCF assemblies. We compared the number of scaffolds split by the Pekin duck RH map with the number of scaffolds split in the PCF assemblies (Figure 2-7). We observed that from the 22 scaffolds split by the RH map, 19 were also split in the RACA-default and RACA-adjusted PCF assemblies. We performed the same comparison for rock pigeon PCF assemblies, and Dovetail generated super-scaffolds. In these cases, we also incorporated information from our PCR verification step. We observed that 10 of the scaffolds split in the Dovetail super-scaffolds (N=53) were also split in the RACA-default PCFs (Figure 2-8). Two of these scaffolds were confirmed to be chimeric by PCR (Figure 2-8), five of them were verified to be non-chimeric, one had inconclusive PCR results and two could not be tested. When comparing with the RACA-adjusted PCF assembly, we observed that 7 scaffolds were split in both assemblies (Figure 2-8). From the three scaffolds that were split in the RACA-default PCFs and Dovetail assembly and were not split in the RACA-

adjusted PCFs, we observed that two were confirmed as non-chimeric by PCR and one had a physical coverage of the split interval above the established minimum physical coverage for rock pigeon. We also noted that the three scaffolds verified to be non-chimeric by PCR but still split on the adjusted PCFs had a physical coverage of the split joint below the established minimum physical coverage for rock pigeon.

Overall, we observed a reduction in the number of inconsistencies (uniquely split scaffolds) between the super-scaffold and PCF assemblies when adjusting RACA parameters and not splitting the PCR-supported scaffolds (20 fewer disagreements for Pekin duck and 34 fewer disagreements for rock pigeon). Moreover, for rock pigeon, we noted that the reduction in the number of inconsistencies was due to a lower number of split non-chimeric scaffolds in the RACA-adjusted PCF assembly. The putatively chimeric scaffold comparison could not be performed for ostrich, as its super-scaffold assembly does not report any split scaffold.

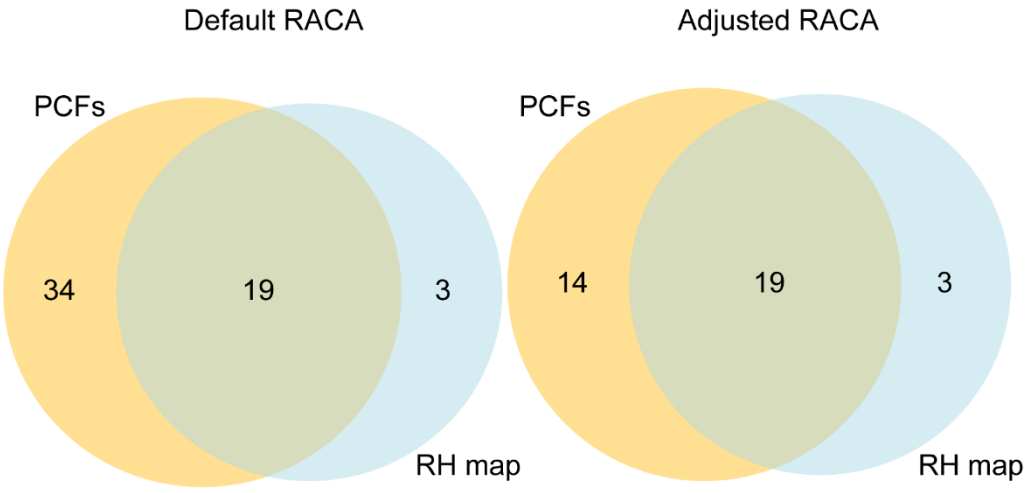

Figure 2-7: Comparison between numbers of split scaffolds in Pekin duck PCF and RH map assemblies.

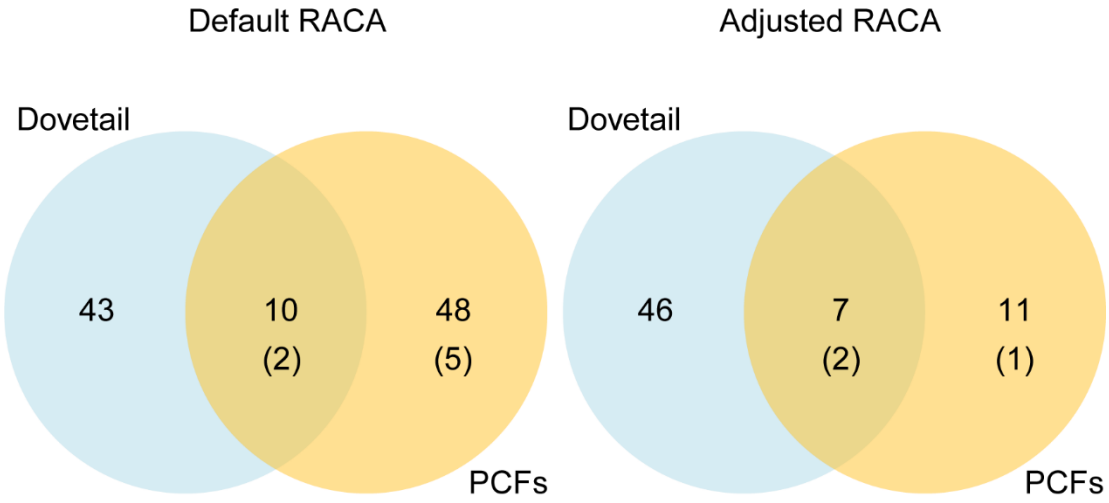

Figure 2-8: Comparison between numbers of split scaffolds in rock pigeon PCF and Dovetail assemblies, and PCR verification results. Number of scaffolds confirmed as chimeric by PCR are indicated in parenthesis.

Predicted adjacencies

In addition to the number of split scaffolds, we compared the level of consistency between RH map, Dovetail or optical map super-scaffold assemblies and RACA predicted adjacencies (and vice-versa) for Pekin duck, rock pigeon and ostrich. Each adjacency was classified as (a) *maintained*, if the adjacency is present in both the super-scaffold and PCF assemblies; (b) *missing*, if the adjacency is present in the super-scaffold assembly and would connect two PCFs or it is present in the PCF assembly and would connect two super-scaffolds; and, (c) *inconsistent*, if the adjacency is present in one of the assemblies and in the other the intervening scaffolds have different connections.

We observed that the average rate of agreement between the assembly methods was ~80% (Figure 2-9, Figure 2-10, Supplemental Table 19, Supplemental Table 20). We also noted that adjusting RACA parameters resulted in a 2% increase in the number of *maintained* adjacencies and that for the PCFs versus super-scaffolds comparison this increase is of 6% (Figure 2-9, Figure 2-10, Supplemental Table 19, Supplemental Table 20). In agreement with the previous observations, we observed an average 3% reduction of the fraction of *inconsistent* adjacencies, when adjusting RACA parameters (Figure 2-9, Figure 2-10, Supplemental Table 19, Supplemental Table 20). Additionally, we also noticed that most of the non-maintained adjacencies were *missing* adjacencies

that would connect two PCFs (12% on average; Figure 2-10, Supplemental Table 19) or two super-scaffolds (14% on average; Figure 2-9, Supplemental Table 20).

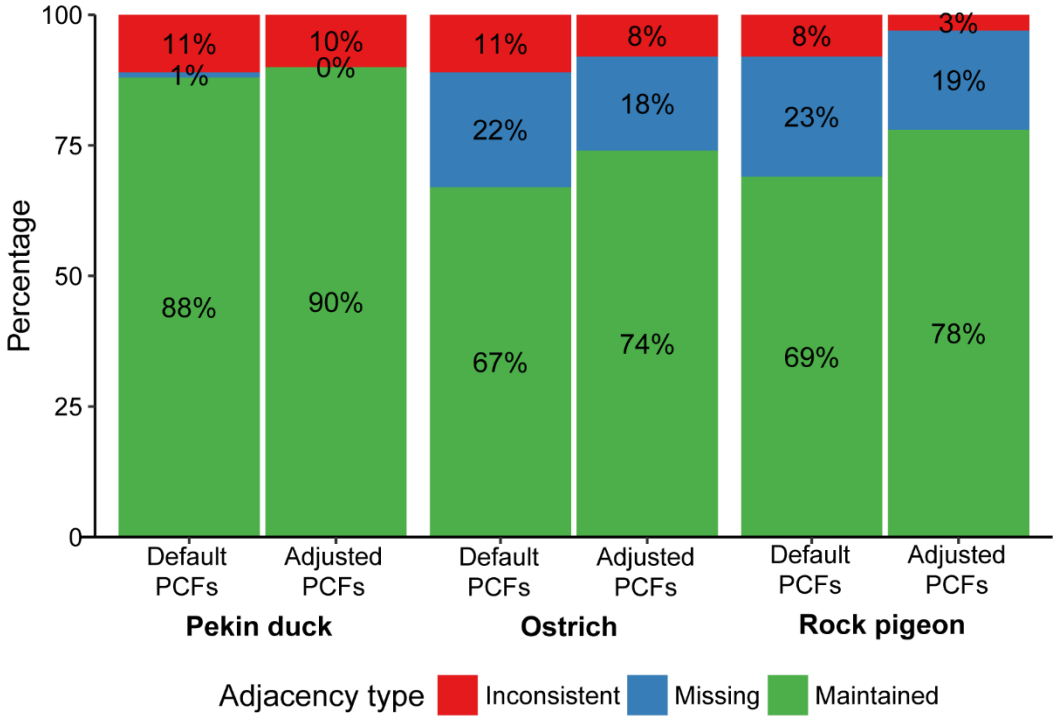

Figure 2-9: Comparison between RACA PCFs and super-scaffolds.

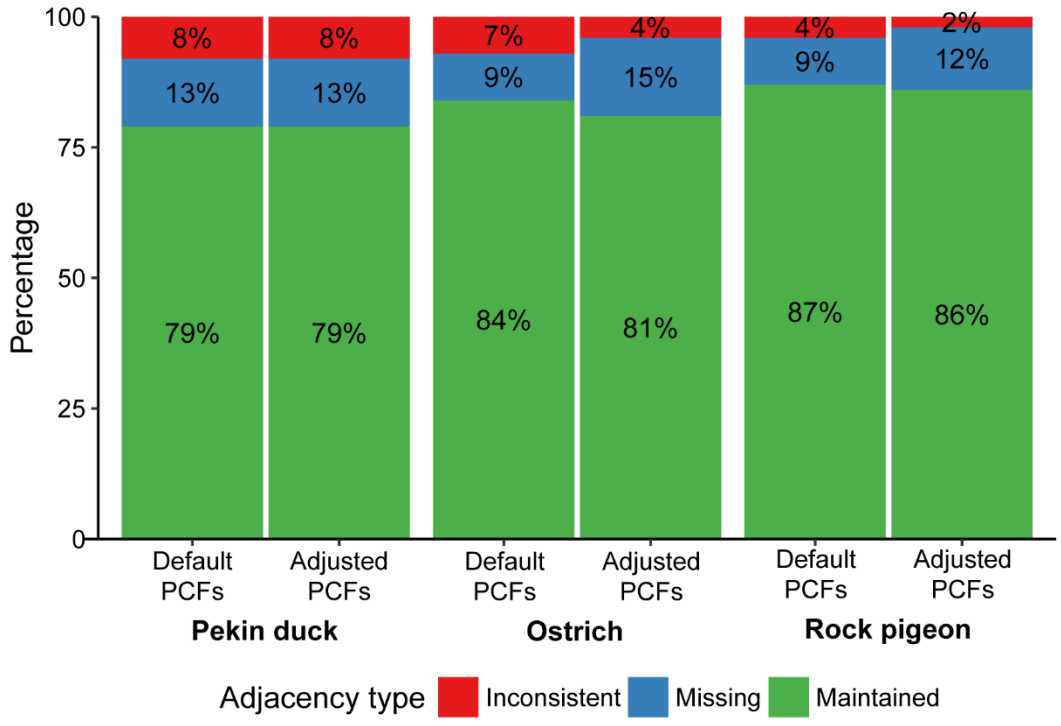

Figure 2-10: Comparison between super-scaffolds and RACA PCFs.

Conserved non-coding element densities in avian inter- and intrachromosomal EBRs

Supporting the hypothesis that the small number of interchromosomal rearrangements in avian genomes could result from an evolutionary advantage to retain synteny, it was recently found that avian HSBs are enriched for conserved non-coding elements (CNEs) (Farré et al., 2016). Many of these elements are known to play important roles in gene regulation. Thus, the disruption of these syntenic regions could have important effects on regulatory pathways. The PCF assemblies generated herein provided us with a set of new EBRs flanking chromosomal fusions not previously found in avian genome assemblies (Table 2-6). This allowed us to investigate the fate of CNEs in avian EBRs and, additionally, to evaluate if avian intra- and interchromosomal rearrangements present similar CNE signatures. We calculated densities of avian CNEs in the chicken chromosome regions corresponding to the lineage-specific EBRs defined to ≤100 Kbp in the chicken genome (Supplemental Table 21 and 22). Moreover, EBRs with a physical coverage below the established threshold for each species or with contradicting PCR or FISH data (Dr Rebecca O'Connor, personal communication) were removed from the analysis. A total of 31 chicken intervals flanking chromosomal fusions and 490 representing intrachromosomal rearrangements were included in this analysis (Supplemental Table 21 and 22). We observed that avian EBRs (intrachromosomal and fusions combined) had a significantly lower fraction of CNEs than their two adjacent chromosome intervals of the same size (up- and downstream; p-value <2.60E-08; Figure 2-11, Table 2-7). The same was observed when we considered intrachromosomal EBRs alone (p-value <2.10E-07; Figure 2-11, Table 2-7). We did not observe any significant CNE density differences between intrachromosomal and fusion EBRs.

To identify CNE densities and the distribution associated with avian EBRs at the genome-wide level, we counted CNE bases in 1 Kbp windows overlapping EBRs and avian msHSBs >1.5 Mbp (Farré et al. 2016). We noted that avian msHSBs contain a higher number of CNE bases (106.798) than the genome average (86.851) (p-value <2.00E-16; Table 2-8), in agreement with the CNE enrichment found in msHSB by Farré and colleagues (Farré et al., 2016). Moreover, the average density of CNE bases in the EBR windows was lower (24.196) than msHSBs and the genome average (p-value <2.00E-16; Table 2-8). Additionally,

we observed that fusion EBRs are associated with a lower number of CNE bases (10.416) than intrachromosomal EBRs (24.834) (p-value = 0.017; Table 2-8).

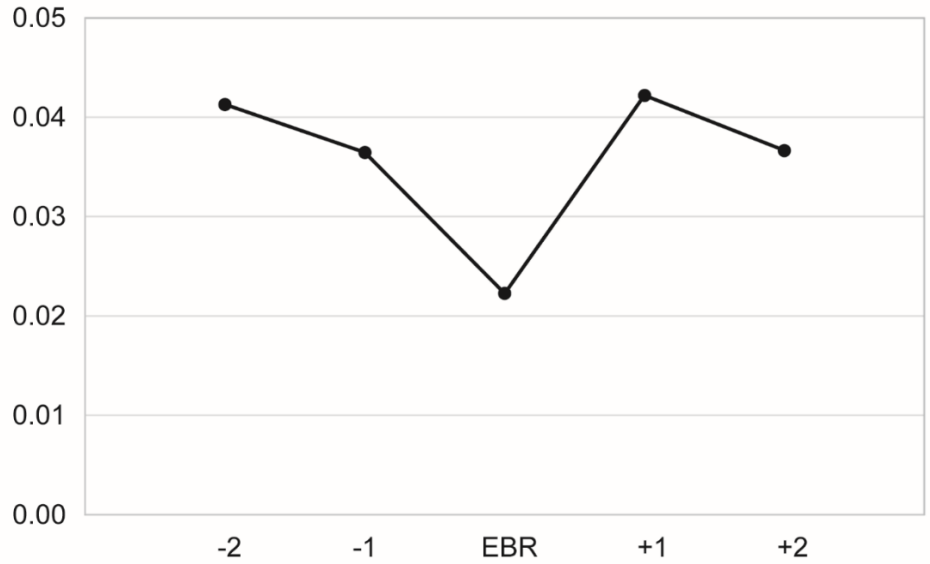

Figure 2-11: Average fraction of bases within CNEs in avian EBRs and two flanking regions of the same size upstream (-) and downstream (+).

Table 2-6: Number of intrachromosomal and fusion flanking lineage-specific EBRs detected from the PCF assemblies, and chicken and zebra finch chromosome assemblies.

Lineage	Intrachromosomal	Fusion
Adélie penguin	2	1
American crow	16	-
Anna's hummingbird	50	-
Budgerigar	43	15
Chimney swift	17	-
Common cuckoo	64	-
Crested ibis	6	-
Downy woodpecker	90	3
Emperor penguin	2	-
Golden-collared manakin	23	-
Hoatzin	25	-
Killdeer	14	-
Little egret	14	1
Medium ground finch	14	-
Ostrich	24	-
Pekin duck	26	-
Peregrine falcon	31	2
Rock pigeon	45	-
Chicken	11	1
Zebra finch	16	-

Table 2-7: Significant comparisons of CNE densities for avian lineage-specific EBRs and their four adjacent intervals (±2) of the same size.

Group1	Group2	Average fraction of CNE bases		p-value
		Group1	Group2	
Intra EBR	Intra (±1)	0.022	0.038	6.00E-11
Intra EBR	Intra (±2)	0.022	0.039	4.00E-13
Intra EBR	Intra (-2)	0.022	0.036	1.50E-07
Intra EBR	Intra (-1)	0.022	0.041	1.40E-08
Intra EBR	Intra (+1)	0.022	0.035	2.10E-07
Intra EBR	Intra (+2)	0.022	0.041	5.30E-12
EBR	(±1)	0.022	0.039	2.00E-12
EBR	(±2)	0.022	0.039	4.20E-14
EBR	(-2)	0.022	0.037	2.60E-08
EBR	(-1)	0.022	0.042	9.40E-10
EBR	(+1)	0.022	0.036	1.00E-08
EBR	(+2)	0.022	0.041	5.30E-13

Table 2-8: Significant comparisons for the number of CNE bases in 1 Kbp windows overlapping avian EBRs, msHSBs, and genome-wide.

Group 1	Group 2	Average no. CNE bases		p-value
		Group 1	Group 2	
EBR	msHSB	24.196	106.798	<2.00E-16
EBR	Genome	24.196	86.851	<2.00E-16
msHSB	Genome	106.798	86.851	<2.00E-16
Intra	msHSB	24.834	106.798	<2.00E-16
Fusion	msHSB	10.416	106.798	<2.00E-16
Intra	Genome	24.834	86.851	<2.00E-16
Fusion	Genome	10.416	86.851	<2.00E-16
Intra	Fusion	24.834	10.416	0.017

Transposable element densities in avian inter- and intrachromosomal EBRs

Another hypothesis trying to explain the reduced number of interchromosomal rearrangements fixed during avian evolution relates to the lower repetitive content of avian genomes. EBRs are usually found associated with repetitive sequences, especially transposable elements (TEs), which are typically used as templates for non-allelic homologous recombination (NAHR) (Groenen et al., 2012, Farré et

al., 2016). In this way, a smaller number of TEs in avian genomes compared to other animals might result in fewer opportunities to change. To identify TE densities associated with avian intra- and interchromosomal EBRs at the genome-wide level, we counted TE bases in 10 Kbp windows overlapping EBRs for peregrine falcon, downy woodpecker and budgerigar genomes. Only EBRs defined to ≤100 Kbp in these genomes and without contradicting PCR or FISH data (Dr Rebecca O'Connor, personal communication) were used in this analysis (Table 2-9).

Table 2-9: Number of budgerigar, downy woodpecker and peregrine falcon EBRs used in the TE analysis.

Lineage	Intrachromosomal	Fusion
Budgerigar	56	10
Downy woodpecker	113	2
Peregrine falcon	52	1

We observed that peregrine falcon EBRs (intrachromosomal and fusion combined) and intrachromosomal EBRs alone were enriched for LTR-ERVL (Table 2-10). This enrichment had already been reported by Farré and colleagues (Farré et al., 2016). Also in agreement with the previous report was the enrichment of downy woodpecker EBRs, and intrachromosomal EBRs alone, for LINE-CR1. Additionally, we also found an enrichment of LTR-ERVL on downy woodpecker EBRs that had not been previously reported (Table 2-11). For budgerigar EBRs, we did not find any TE enrichment (Table 2-12) which disagrees with (Farré et al., 2016) that detected enrichments of budgerigar lineage-specific EBRs in LTR-ERVL and LINE-CR1. The differences between the results reported here and those reported in (Farré et al., 2016) might be explained by the differences in the number of EBRs used in the analysis. Moreover, in the previous report, no filtering was performed for budgerigar EBRs, which might have resulted in an overrepresentation of chimeric scaffolds on that set.

Table 2-10: Comparison and ratios of the number of bases from populated TEs in 10 Kbp intervals overlapping peregrine falcon EBRs and the rest of the genome.

Repeat set	All EBRs			Intrachromosomal			Fusions		
	EBR	non-EBR	Ratio	EBR	non-EBR	Ratio	EBR	non-EBR	Ratio
LINE	305.174	280.330	1.089	300.478	280.334	1.072	516.500	280.346	1.842
LINE-CR1	304.163	274.858	1.107	299.444	274.862	1.089	516.500	274.877	1.879
LTR	289.587	110.461	2.622 *	285.156	110.468	2.581 *	489.000	110.596	4.421
LTR-ERVL	194.293	95.842	2.027 *	187.744	95.849	1.959 *	489.000	95.913	5.098

*****Statistically significant differences (FDR-corrected p-value <0.05). Only statistical significance for transposable elements covering ≥100 bp on average in the EBR and/or non-EBR 10 Kbp intervals in each individual comparison is reported.

Table 2-11: Comparison and ratios of the number of bases from populated TEs in 10 Kbp intervals overlapping downy woodpecker EBRs and the rest of the genome.

Repeat set	All EBRs			Intrachromosomal			Fusions		
	EBR	non-EBR	Ratio	EBR	non-EBR	Ratio	EBR	non-EBR	Ratio
LINE	1883.813	1611.434	1.169 *	1879.777	1611.450	1.166 *	2204.000	1611.990	1.367
LINE-CR1	1872.863	1601.272	1.170 *	1868.857	1601.287	1.167 *	2190.667	1601.827	1.368
LTR	60.685	55.737	1.089	57.088	55.744	1.024	346.000	55.740	6.207
LTR-ERVL	20.477	38.191	0.536 *	18.500	38.195	0.484 *	177.333	38.150	4.648
LTR-ERVK	6.556	3.127	2.097	4.668	3.131	1.491	156.333	3.130	49.947

*****Statistically significant differences (FDR-corrected p-value <0.05). Only statistical significance for transposable elements covering ≥100 bp on average in the EBR and/or non-EBR 10 Kbp intervals in each individual comparison is reported.

Table 2-12: Comparison and ratios of the number of bases from populated TEs in 10 Kbp intervals overlapping budgerigar EBRs and the rest of the genome.

Repeat set	All EBRs			Intrachromosomal			Fusions		
	EBR	non-EBR	Ratio	EBR	non-EBR	Ratio	EBR	non-EBR	Ratio
LINE	696.903	618.332	1.127	646.819	618.408	1.046	897.241	618.363	1.451
LINE-CR1	693.862	610.318	1.137	643.017	610.396	1.053	897.241	610.353	1.470
LTR	206.717	154.695	1.336	204.379	154.711	1.321	216.069	154.748	1.396
LTR-ERVL	176.428	136.122	1.296	168.698	136.142	1.239	207.345	136.158	1.523

Only transposable elements covering ≥100 bp on average in the EBR and/or non-EBR 10 Kbp intervals in each individual comparison are reported.

Data availability

The refined PCF sets are publicly available at the Evolution Highway (EH) comparative chromosome browser (http://eh-demo.ncsa.uiuc.edu/avian) and in our UCSC track hub BirdsHUB_test (http://sftp.rvc.ac.uk/rvcpaper/birdsHUB_test/hub.txt) (Figure 2-12). EH allows the representation of comparative information, for instance, regions of maintained synteny between species (i.e. HSBs) and areas where the synteny is broken (i.e. EBRs). It also shows the adjacency score calculated by RACA during the PCF reconstruction (Figure 2-12A). Each target species EH track was named using "RF" as a suffix (meaning RACA final), the first three letters of their scientific name (e.g., "Aptfor" for emperor penguin, *Aptenodytes forsteri*), followed by the used reference genome (e.g., "Taegut" for zebra finch, *Taeniopygia guttata*), and the established physical coverage threshold (e.g., 50X for emperor penguin). As an example, the emperor penguin EH track identifier is "RF:Aptfor:Taegut:50X". Our UCSC hub contains the comparative information used by RACA to construct the PCFs, and the translated gene and repeat annotations for each of the target species. These translations were performed from original scaffold coordinates to PCF coordinates using in-house Perl scripts. Only the genes and repeats fully covered in the PCFs were included (Figure 2-12B).

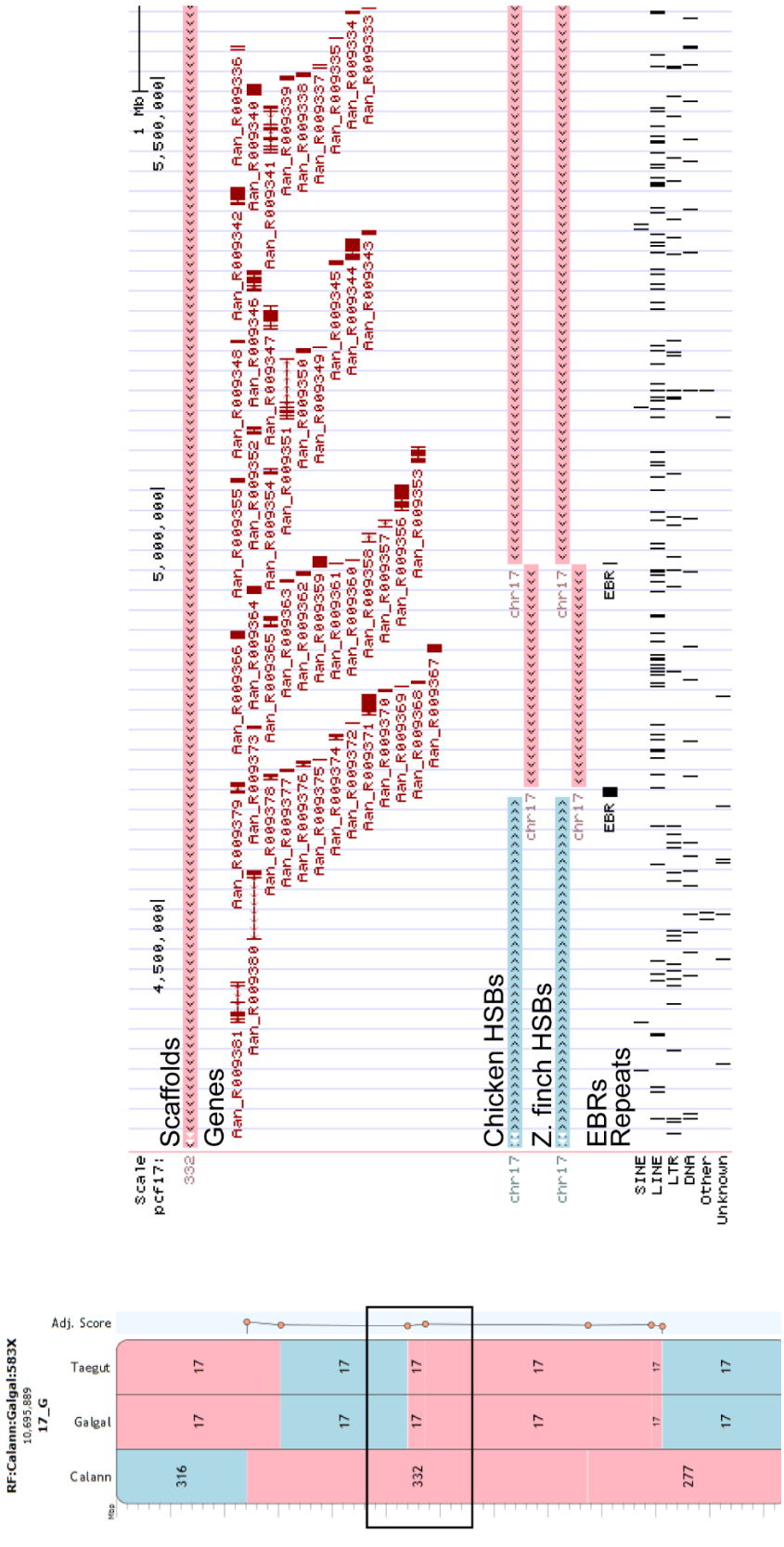

Figure 2-12: Anna's hummingbird PCF 17 representation on (A) Evolution Highway comparative chromosome browser and (B) UCSC Genome Browser hub. (A) Blue and pink blocks define syntenic fragments in the Anna's hummingbird scaffolds ("Calann" track), chicken ("Galgal" track) and zebra finch ("Taegut" track) genomes in "+" and "-" orientation, respectively compared to the PCF. Numbers within blocks represent scaffold or chromosome identifiers. "Adj. score" track shows adjacency scores calculated by RACA. Black rectangle delimits region represented on panel B.

2.4 Discussion

The existence of chromosome-level genome assemblies is of high importance for the study of many aspects of evolutionary biology. Due to the limitations of NGS and genome mapping methodologies, the rate of production of chromosome-level assemblies is very dissimilar from the frequency at which new genome sequences are released, limiting the utility of these generated sequences.

Herein, we report the first application of the RACA algorithm to avian genomes. By comparison of the RACA generated PCFs with pre-existing radiation hybrid (RH; Pekin duck), optical map (ostrich) and Dovetail (rock pigeon) super-scaffold assemblies we demonstrated that RACA could be highly consistent with the traditional chromosome assembly methods when proper parameters for an accurate distinction of chimeric and real joints in scaffolds are selected. This observation supports RACA reliability to reconstruct the chromosome structures of birds. Furthermore, RACA is also shown to be useful in detecting problematic regions on genome assemblies as seen through the high agreement between RH map and Dovetail, and RACA split scaffolds.

Even when using RACA avian default parameters, the obtained PCF assemblies show a considerable increase of genome continuity, as observed by the reduction in the number of assembly fragments and the increase of the assembly N50. Additionally, the establishment of the physical coverage threshold using PCR verification of syntenic fragment joints leads to the improvement of the constructed genome assemblies. This threshold proved critical as it directly influences the capacity for RACA to distinguish real and chimeric adjacencies in scaffolds. In fact, the fraction of scaffolds detected as putatively chimeric on the refined PCF assemblies agrees with the previously reported 6% for NGS genome assemblies (Kim et al., 2013), the only exception being budgerigar that presents a value more than three times higher. In addition, RACA PCFs also allow for the detection of lineage-specific EBRs that are known to have active roles in the formation of adaptive phenotypes. The EBRs detected from RACA PCFs can be related to intrachromosomal rearrangements or chromosomal fusions (some previously recognised using cytogenetic methodologies). In this way, PCF assemblies give an extra level of evidence that could be used to unravel the

forces driving avian genome evolution and the reasons behind the dissimilar frequencies of intra- and interchromosomal changes in avian genomes.

It was recently suggested that the small number of interchromosomal rearrangements in avian genomes could either relate to an evolutionary advantage to retain synteny or little opportunity for change. Indeed, the smaller number of TEs in avian genomes compared to other animals might indicate that avian chromosomes have fewer opportunities for chromosome merging using NAHR. Herein, we show support for this hypothesis as the peregrine falcon and downy woodpecker intrachromosomal EBRs were significantly enriched in TEs when compared with their immediately adjacent genomic regions. Conversely, no significant enrichment was found in EBRs flanking fusions, which might suggest that another mechanism is responsible for the generation of such events. Nonetheless, it is also possible that the small number of chromosomal fusions analysed could conceal the detection of statistical significance. In this way, the detection and verification of more interchromosomal EBRs are essential.

The strong enrichment for avian CNEs in the regions of interspecies synteny in birds and other reptiles, suggests an evolutionary advantage of maintaining the established synteny (Farré et al. 2016). Indeed, we observed that EBRs have a significantly lower density of CNEs than their same-size flanking windows. In a genome-wide analysis, regions flanking fusion events show a lower density of CNE bases than regions where intrachromosomal EBRs locate, demonstrating that those rare interchromosomal rearrangements appear in areas of a lowest density of CNEs. In fact, the study of intrachromosomal changes in Passeriformes (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.

Despite the many insights on avian chromosome evolution handed by the creation of chromosome-scale fragments using RACA, the detection of the full collection of chromosomal changes present in a species requires the generation of chromosome-level assemblies. Using RACA algorithm, one can significantly increase the continuity of an NGS generated genome. However, the generated PCFs are not error-free, and the number of produced PCFs is still higher than the haploid number of chromosomes of the target species. The generation of PCFs

is then an intermediate step in the construction of chromosome-level assemblies from NGS data, and it can significantly reduce time and costs associated with that task. Chromosome-level assemblies can be obtained through the mapping and ordering of chromosome-sized fragments along the target species chromosome, which would require only sparse physical maps generated for instance through cross-species BAC mapping.

3 Constructing avian chromosome-level assemblies

This chapter reports the results from:

Damas, J., O'Connor, R., Farré, M., Lenis, V. P. E., et al. (2017) 'Upgrading short-read animal genome assemblies to chromosome level using comparative genomics and a universal probe set', Genome Research, 27(5), pp. 875-884.

3.1 Background

In the previous chapter, I explored the utility of the reference-assisted chromosome assembly (RACA; Kim et al., 2013) algorithm to approximate near chromosome-sized predicted chromosome fragments (PCFs) for an avian *de novo* assembled NGS genome. Even though the use of this methodology significantly improves the continuity of NGS genome assemblies, using RACA, we still obtain a higher number of PCFs than the haploid number of the species' chromosomes, and PCFs might still contain misassemblies. This complicates the detection of the full collection of chromosomal changes that shaped a genome. Therefore, PCFs require further verification and mapping to chromosomes to obtain accurate chromosome-level assemblies, which could be achieved by integrating them with sparse physical maps.

Bacterial artificial chromosome (BAC) probes are widely used in fluorescence *in situ* hybridization (FISH) to pinpoint genomic alterations related to disease and/or evolution. Particularly, cross-species BAC clone mapping can be used to assist the reconstruction of the evolutionary relationship between species, through the creation of physical comparative maps. In this way, BAC probes could be a useful resource for the generation of the physical maps that would assist PCF verification and anchoring. Nonetheless, only a limited success of such probes for cross-species mapping was previously demonstrated when using gene-rich and low-repeat content BAC clones (Larkin et al., 2006, Romanov et al., 2011). Thus, the development of a selection approach that could increase the success rate of cross-species FISH (zoo-FISH) and allow a higher throughput of these experiments would be of great value for the scientific community.

As referred multiple times in the previous chapters, avian chromosome evolution has been a relatively understudied topic, mainly due to the dearth of chromosome assemblies for this clade. The scientific interest, together with a smaller genome size and the availability of BAC clones' libraries for representatives of this class, makes birds a perfect target for the development and testing of a novel genome assembly approach. In this work, we focused on two avian species: the rock pigeon (*Columba livia*) and the peregrine falcon (*Falco peregrinus*). The rock pigeon has a typical avian karyotype (2*n*=80) similar to those of reference avian genomes: chicken (*Gallus gallus*), turkey (*Meleagris gallopavo*) and zebra finch

(*Taeniopygia guttata*). Pigeon is one of the earliest examples of domestication in birds (Driscoll et al., 2009) and is currently used as food and in sporting circles (Price, 2002). Pigeon breeds display very diverse appearances (Price, 2002) inspiring interest in identifying the genetic basis for these variations (Stringham et al., 2012, Shapiro et al., 2013). The peregrine falcon has an atypical karyotype (2*n*=50) (Nishida et al., 2008). Peregrine 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) and it is also of great importance in sports circles (e.g., falconry). For the above reasons, both species genomes were sequenced (Shapiro et al., 2013, Zhan et al., 2013). However, their assemblies are highly fragmented, and hence chromosome-level assemblies are in need.

In this chapter, I report the development of an avian universal BAC clone panel that can be used to inexpensively verify and map near chromosome-scale fragments to species chromosomes. This approach can easily be applied to multiple species. As proof of principle for this approach, I applied this methodology to verify and order peregrine falcon and rock pigeon PCFs (reported in the previous chapter), which resulted in the generation of chromosome-level assemblies for these species. The study of the first chromosome assembly of a highly rearranged avian genome (peregrine falcon) provided novel insights on why interchromosomal rearrangements are infrequent in avian evolution.

3.2 Material and methods

Avian genome assemblies, repeat masking and gene annotations

The chicken (*Gallus gallus*; ICGSC Gallus_gallus 4.0; Hillier, 2004), zebra finch (*Taeniopygia guttata*; WUGSC 3.2.4; Warren et al., 2010), and turkey (*Meleagris gallopavo*; TGC Turkey_2.01; Dalloul et al., 2010) chromosome assemblies were downloaded from the UCSC Genome Browser (Kent et al., 2002). The collared flycatcher (*Ficedula albicollis*; FicAlb1.5; Ellegren et al., 2012) genome was obtained from NCBI. Peregrine falcon (*Falco peregrinus*) and rock pigeon (*Columba livia*) PCF assemblies were generated as described in chapter 2. 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).

Nucleotide evolutionary conservation scores and conserved elements

Nucleotide evolutionary conservation scores were obtained from Dr Vasilis Lenis (Aberystwyth University). Briefly, chicken chromosome as reference "net" alignments for 21 avian genomes were used to build multiple alignment files with MULTIZ (Blanchette et al., 2004). The evolutionary conservation scores and DNA conserved elements (CEs) for all chicken nucleotides assigned to chromosomes were estimated using PhastCons (Siepel et al., 2005) with the following parameters: *expected-length=45, target-coverage=0.3 and rho=0.2506*. Conserved non-coding elements (CNEs) obtained from the alignments of 48 avian genomes were used (Farré et al., 2016).

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 x 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 the UCSC Genome Browser *LiftOver* tool (Kent et al., 2002) with the minimum ratio of remapped bases set to >0.1.

For each BAC clone mapped to the chicken chromosomes, various genomic features selected to estimate the probability of clones to hybridise with metaphase chromosomes of distant avian species were calculated or extracted from the gene, repetitive sequence, conserved element and nucleotide conservation score files using a custom Perl script (Table 3-1). The chicken and turkey BAC clones selected for mapping experiments were initially obtained from the BACPAC Resource Centre at the Children's Hospital Oakland Research Institute (USA) and the zebra finch from the Clemson University Genomics Institute (USA).

Classification tree

A classification tree was created, in R (v.3.2.3; R CoreTeam, 2015) using the classification and regression tree (CART) algorithm included in the rpart package (v.4.1-10; Therneau et al., 2015), to detect genomic signatures associated with a higher chance of a BAC clone working on phylogenetically distant species. We introduced an adjusted weight matrix setting the cost of returning a false positive twice as high as the cost of a false negative. The tree was visualised with rattle package (v.4.1.0; Williams, 2011).

Fluorescence *in-situ* hybridization

FISH experiments were performed at Professor Darren Griffin's lab (University of Kent, UK) by Dr Rebecca O'Connor (95% of experiments) and myself (5%). Briefly, 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.

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

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 chambers) and Multiprobe (24 chambers) devices. Briefly, labelled probes were air dried onto the device. Probes were, re-hybridized in a standard buffer, applied to the glass slide (which was sub-divided to correspond to the hybridization chambers) and FISH continued as above.

Evolutionary breakpoint regions detection and sequence feature analysis

The multiple alignments of the chicken, zebra finch, flycatcher, rock pigeon and peregrine 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., 2016) at 100, 300 and 500 Kbp minimum resolution of syntenic fragments. Using chicken as the reference genome, evolutionary breakpoint regions (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 defined to ≤100 Kbp were used for the following analyses. EBRs smaller than 1 Kbp were extended ±1 Kbp. 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 from CNEs in each EBR site, the upstream and downstream windows. Differences in densities were tested for significance using the Kruskall-Wallis test followed by Mann-Whitney U test.

Comparing CNE densities in EBRs and multispecies homologous synteny blocks

Chicken chromosomes (GGA; excluding GGA16, W and Z) were divided into 1 Kbp 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 multispecies (ms) homologous synteny blocks (HSBs) >1.5 Mbp (Farré et al., 2016), avian EBRs flanking fusions, fissions, intrachromosomal rearrangements, or the intervals found in the rest of the chicken genome. We estimated the average CNE density for each window type and the distance, in the number of 1 Kbp 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. Differences with the genome average were tested for significance using a one-sample proportion test.

Densities of transposable elements in EBRs

The transposable elements (TEs) scaffold coordinates reported in (Zhan et al., 2013) and (Shapiro et al., 2013) were translated to the peregrine falcon and rock pigeon chromosome coordinates, respectively, 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 peregrine falcon lineage-specific intrachromosomal 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). The same method was applied for the comparison of rock pigeon intrachromosomal EBRs and the rest of the genome. Only windows with <50% of their bases within sequence gaps were used in this analysis.

Gene ontology enrichment analysis

The basic version of gene ontology (GO) annotations (version 3rd May 2016) was downloaded from the GO Consortium website (The Gene Ontology Consortium, 2015). Sequence coordinates and Ensembl identifiers for chicken genes were obtained from Ensembl Biomart (v.74; Kinsella et al., 2011). Chicken genes were used as the background list for the GO analysis. To evaluate gene functional enrichment in and near EBRs, we assigned genes from the background list to regions located within or ±300 Kbp from lineage-specific peregrine falcon and rock pigeon EBR boundaries. We used the GO::TermFinder Perl module (Boyle et al., 2004) to detect GO terms overrepresented in our gene sets. We considered as significantly enriched the terms with p-value <0.05 and false discovery rate (FDR) <5% in EBRs relative to all other regions on chicken chromosomes. In addition, to ensure that the enrichments observed were not due to gene family expansions we report GO terms for which the genes originated from at least four EBRs.

Table 3-1: DNA sequence feature information returned from the BAC clone analysis pipeline.

Group	Feature
General	BAC clone length
	Percentage of a BAC clone missed nucleotides (Ns)
DNA conservation	Percentage of a BAC clone sequence with conservation scores
	Average conservation score (only positions with score considered)
	Average conservation score (positions without score included as 0)
	Percentage of a BAC clone within conserved elements
	Minimum length of conserved elements
	Maximum length of conserved elements
	Average length of conserved elements
	Median of the length of conserved elements
	Percentage of a BAC clone within conserved elements of length ≥100nt
	Percentage of a BAC clone within conserved elements of length ≥200nt
	Percentage of a BAC clone within conserved elements of length ≥300nt
	Percentage of a BAC clone within conserved elements of length ≥400nt
	Percentage of a BAC clone within conserved elements of length ≥500nt
Gene content	Percentage of a BAC clone containing chicken genes
	Percentage of a BAC clone containing chicken exons
	Minimum length of chicken exons
	Maximum length of chicken exons
	Average length of chicken exons
	Median of the length of chicken exons
	Percentage of a BAC clone containing chicken exons of length ≥100nt
	Percentage of a BAC clone containing chicken exons of length ≥200nt
	Percentage of a BAC clone containing chicken exons of length ≥300nt
	Percentage of a BAC clone containing chicken exons of length ≥400nt
	Percentage of a BAC clone containing chicken exons of length ≥500nt
	Percentage of a BAC clone containing chicken-human orthologous genes
Repeat content	Percentage of a BAC clone containing repetitive elements
	Minimum length of repetitive elements
	Maximum length of repetitive elements
	Average length of repetitive elements
	Median of the length of repetitive elements
GC content	GC percentage of a BAC clone
	GC percentage of highly conserved bases (conservation score ≥0.5)
	Average GC percentage of conserved elements

3.3 Results

The novel genome chromosome-level assembly method developed within this project involves (Figure 3-1): (1) the generation of a refined set of PCFs for original fragmented (NGS) assemblies, as described in chapter 2, and (2) the use of a panel of “universal” BAC clones to anchor PCFs to chromosomes in a high-throughput manner, described herein.

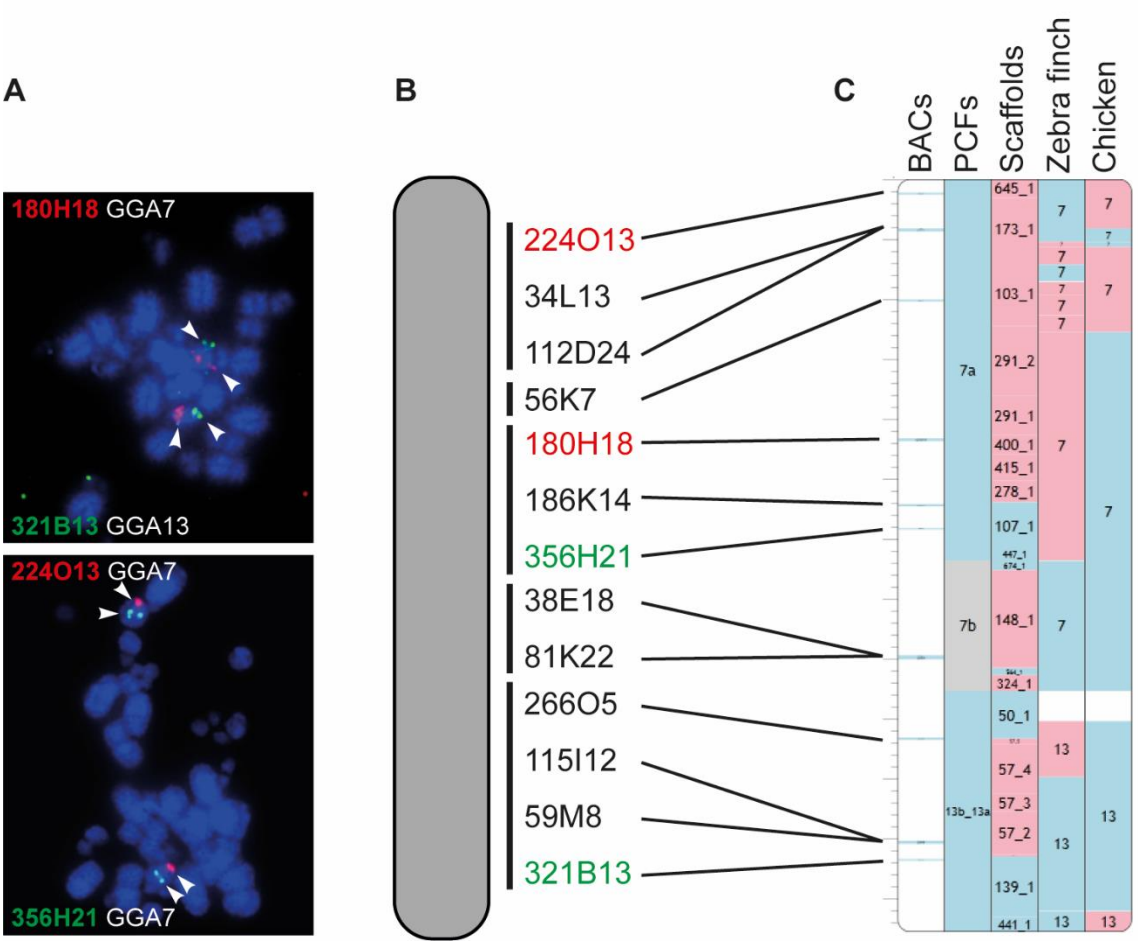

Figure 3-1: Methodology for the placement of the PCFs on chromosomes. (A) dual-color FISH of universal BAC clones, (B) cytogenetic map of the falcon chromosome 8 (FPE8) with indication of the relative positions of the universal BAC clones along the chromosome, and (C) assembled chromosome containing 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.

Construction of a panel of comparatively anchored BAC clones designed to hybridise 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 3-2). To minimise the effect of evolutionary distances between species on hybridizations, genomic features that were likely to influence hybridization success (e.g., repeat content, sequence conservation; Table 3-1) were measured in chicken, zebra finch and turkey BAC clones (Table 3-3).

Table 3-2: 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 3-3: Avian BAC clones and expected FISH success rates for the phylogenetically distant species (divergence time ≥69 MY).

Library	**Species**	**No. of analysed BAC clones ***	**Obey CART criteria**	**Expected success rate**
CHORI-260	Turkey	3,694	2,821	76.36
CHORI-261	Chicken	40,410	30,399	75.23
TGMCBA	Zebra finch	80,009	61,335	76.66
Total/Average		**124,113**	**94,555**	**76.18**

* No. BAC clones after filtering steps. Clones mapped to chicken unplaced regions or linkage groups, clones smaller than 50 Kbp or longer than 300 Kbp were not included in the analysis.

The classification and regression tree approach (CART; Loh, 2011) was applied to the 101 randomly-selected, but biased to subtelomeric regions, zebra finch and chicken BAC clones (Table 3-2). The inexistence of FISH preliminary results for turkey BAC clones led us to remove them from this analysis. The obtained

classification shows 87% agreement with FISH results (Figure 3-2). Correlating DNA features with actual cross-species FISH results led to the development of the following criteria for selection of chicken or zebra finch BAC clones that are very likely to hybridise on metaphase preparations of phylogenetically distant birds (≥69 MY of divergence, split between Columbea and the remaining Neoavian clades; where the hybridization success rate of random BAC clones was <70%). The BAC clone must have ≥93% DNA sequence alignable with other avian genomes and contain at least one conserved element (CE) ≥300 bp (Figure 3-2). Instead of a long CE, the BAC could contain only short repetitive elements (<1290 bp) and CEs at least four bp long (Figure 3-2).

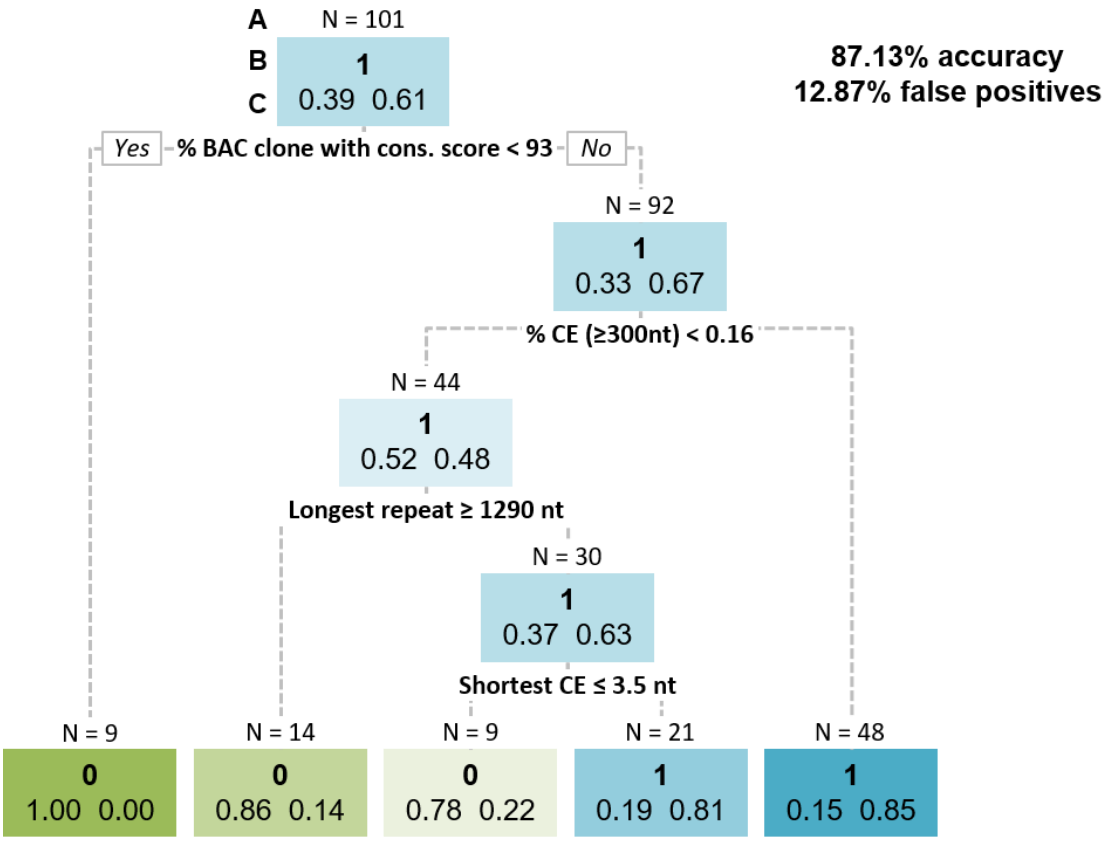

Figure 3-2: Classification tree used to predict the non-successful (0) or successful hybridization (1) of a BAC clone on at least one phylogenetically distant species (divergence time ≥69 MY). For each tree node (A) is the number of BAC clones at each node, (B) represents the classification with higher representation on the data, and (C) proportions of BAC clones classified as 0 (left) and 1 (right) on the data. At each intermediate node, a case goes to the right child node if the condition is satisfied. Accuracy depicts the fraction of BAC clones for which classification and FISH results agree. False positives depict the fraction of BAC clones classified as expected to hybridise but did not produce any specific FISH signal on species diverged ≥69 MY.

The hybridization success rate with distant avian species for the set of newly selected clones obeying these criteria was high (71-94%; Table 3-2). When considering only the chicken BAC clones, the success rates ranged from 90% to 94%. From these chicken clones, 84% hybridised with chromosomes of all avian species in our set. As a final result, we generated a panel of 121 BAC clones spread across the avian genome (chicken chromosome (GGA) 1-28 + Z; except 16) that successfully hybridised across all species attempted. This collection was supplemented by a further 63 BACs that hybridised on the metaphases of at least one species that was considered phylogenetically distant and a further 33 that hybridised on at least one other species (Figure 3-3). The universal panel of BAC clones is available at our UCSC track hub (Figure 3-4; http://sftp.rvc.ac.uk/rvcpaper/birdsHUB/hub.txt) under the “Chicken BAC clones” track on the chicken genome.

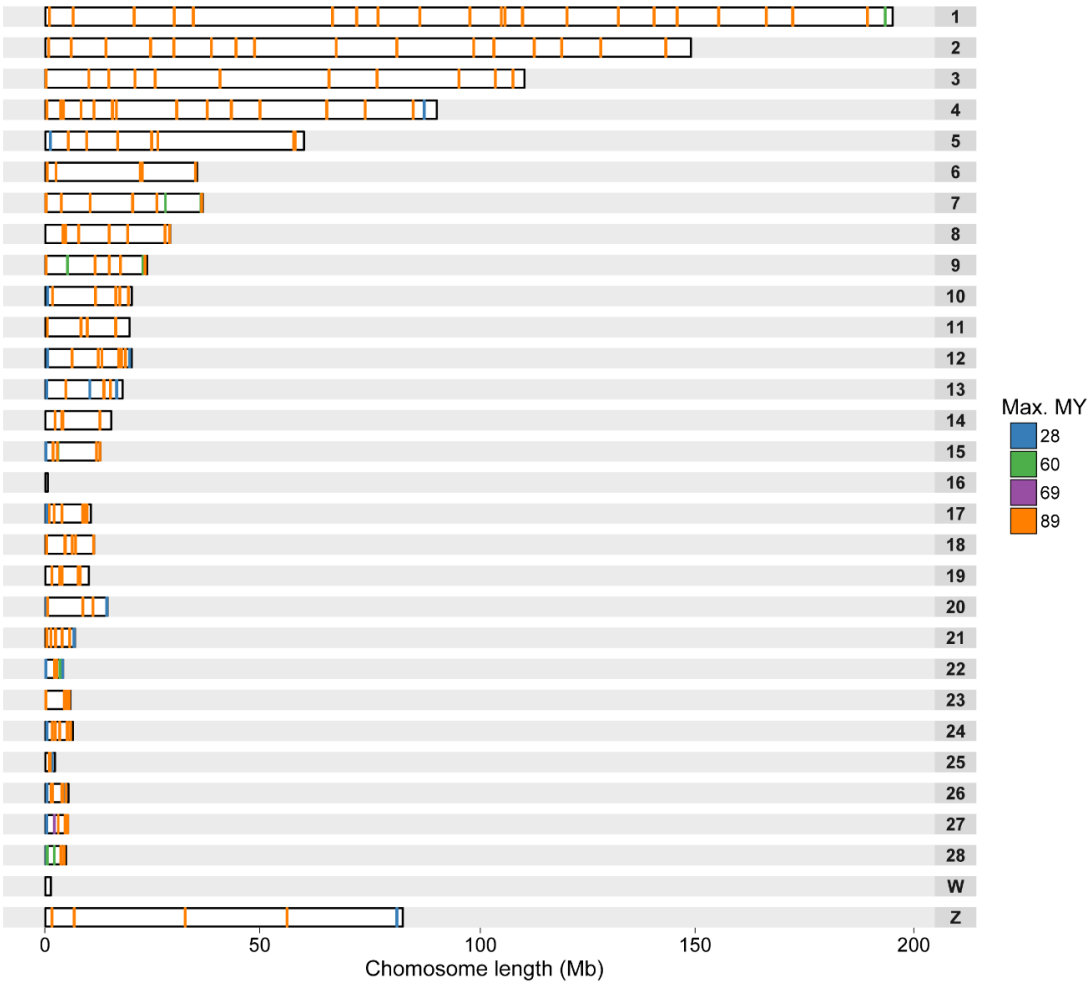

Figure 3-3: Distribution of universal BAC clones along chicken chromosomes. Each rectangle represents a chicken chromosome and the lines inside the location of each BAC clone. BAC clones are coloured accordingly to the maximum phylogenetic distance of the species they successfully hybridized.

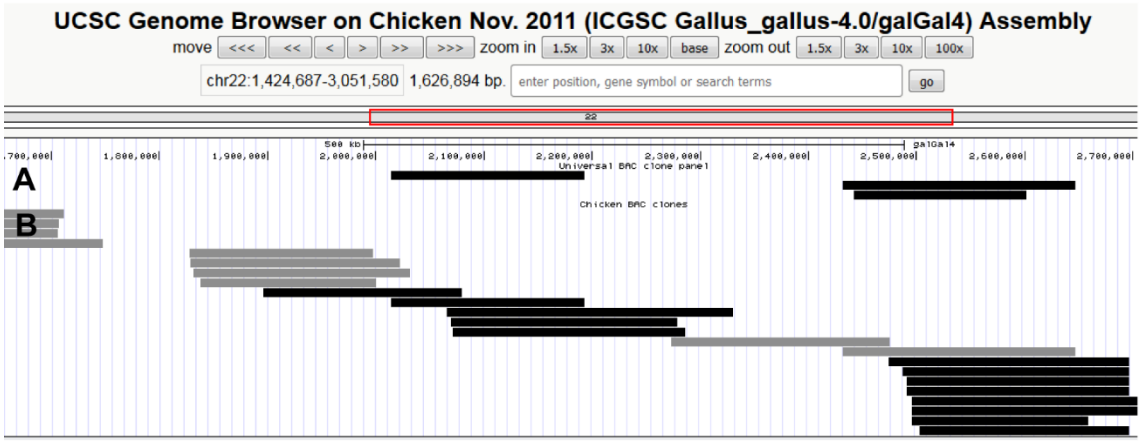

Figure 3-4: BAC clone tracks on UCSC genome browser. (A) BAC clones included on the avian universal BAC clone panel. (B) Full set of chicken selected BAC clones. Black and grey rectangles depict BAC clones predicted to hybridize on far related species with 93 and 85% probability, respectively.

Physical assignment of refined PCFs on the species' chromosomes

To place and order PCFs along chromosomes, BAC clones from the panel described above and assigned to PCFs based on alignment results, were hybridised to peregrine falcon (177 BAC clones) and rock pigeon (151 BAC clones) chromosomes (Table 3-4). The 57 PCFs cytogenetically anchored to the peregrine falcon chromosomes represented 1.03 Gbp of its genome sequence (88% of the cumulative scaffold length). Of these, 888.67 Mbp were oriented on the chromosomes (Table 3-4). The pigeon chromosome assembly consisted of 0.91 Gbp in 60 pigeon PCFs representing 82% of the combined scaffold length. Of these 687.59 Mbp were oriented (Table 3-4).

Table 3-4: Statistics for the chromosome assemblies of the peregrine falcon and rock pigeon.

Statistics	Peregrine falcon	Rock pigeon
No. informative BAC clones	177	151
No. PCFs placed on chromosomes	57	60
Combined length (Gbp)	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 (Mbp)	888.67	687.59

Rock pigeon chromosome assembly

No deviations from the standard avian karyotype (2*n*=80) were detected for rock 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) (Figure 3-5A). Nonetheless, 70 intrachromosomal EBRs in the pigeon lineage were identified (Table 3-5).

Peregrine falcon chromosome assembly

Homeology between the chicken and the peregrine falcon was identified for all mapped chromosomes, except GGA16 and GGA25 (Figure 3-5B). In total, 13 falcon-specific fusions and six fissions were detected (Table 3-5). Each of the chicken largest macrochromosome homeologues (GGA1 to GGA5) were split across two falcon chromosomes. The falcon GGA1 and GGA3 counterparts were represented as two entire chromosomes each (peregrine falcon chromosome (FPE) 4 and FPE6, FPE7 and FPE11, respectively). GGA2 was split across FPE3 and FPE5, both of which exhibited additional fusions of microchromosomes with GGA21 and GGA23 fused in FPE3 and GGA12, GGA14 and GGA28 fused in FPE5. Consistent with the rock pigeon assembly results (and most birds), GGA4 was found to be split across two falcon chromosomes (FPE2 and FPE13), the former of which exhibited three additional microchromosomal fusions (GGA15, GGA18 and GGA19). Both GGA6 and GGA7 homeologues were found as single blocks fused with other chicken chromosome material within falcon chromosomes FPE1 and FPE8, respectively. Among the other chicken macrochromosomes, only GGA8 and GGA9 were represented as individual chromosomes (FPE10 and FPE12, respectively). Of the 17 mapped chicken microchromosomes, 11 were fused with other chromosomes. Additionally, 69 intrachromosomal EBRs were detected in the falcon lineage (Table 3-5).

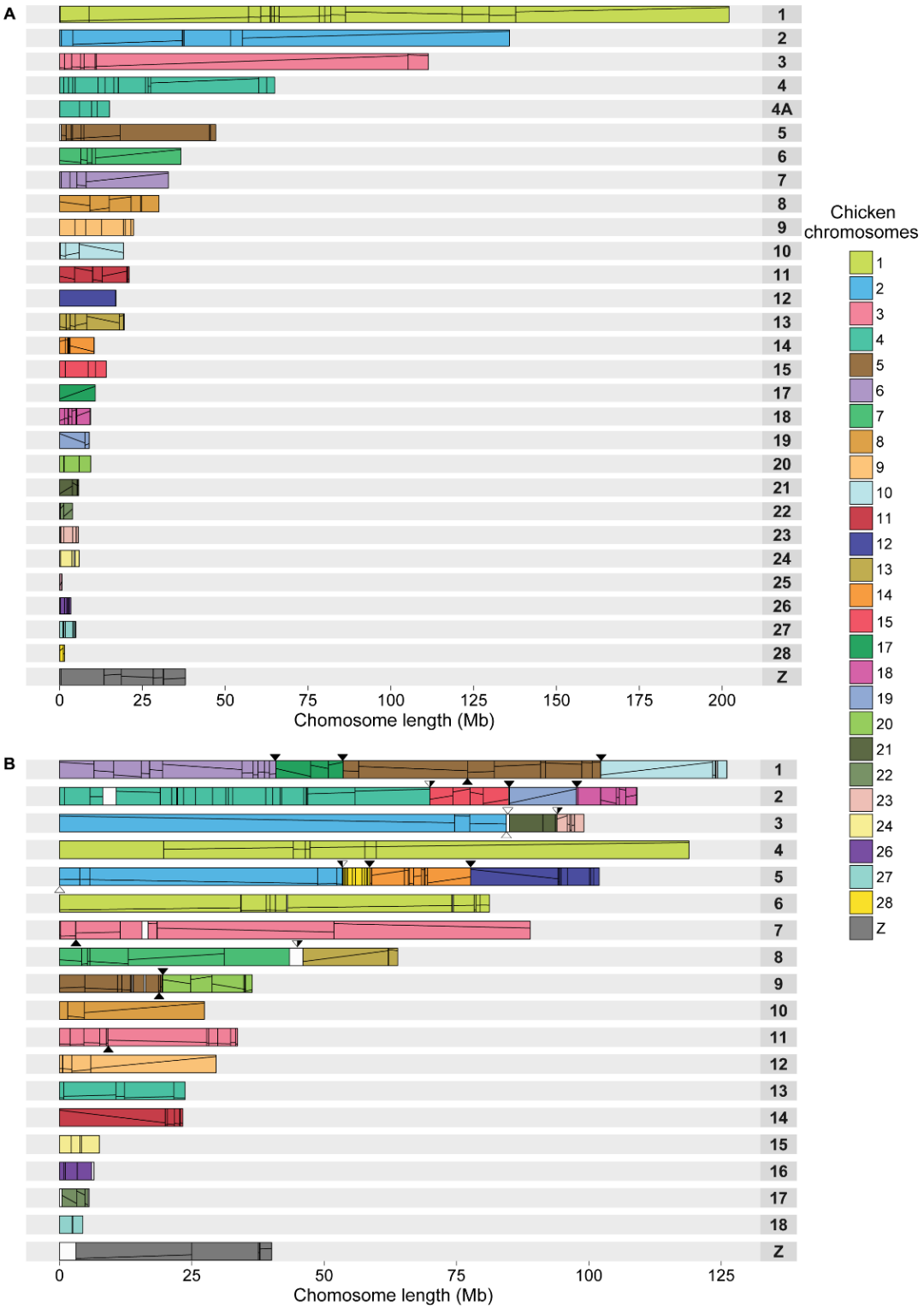

Figure 3-5: Ideogram of rock pigeon (A) and peregrine falcon (B) chromosomes. Numbered rectangles represent chromosomes and coloured blocks within show regions of homeology with chicken chromosomes. Lines within coloured blocks represent block orientation. Triangles above the falcon chromosomes point 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.

Table 3-5: Peregrine falcon and rock pigeon lineage-specific EBRs.

	Peregrine falcon	Pigeon
Fusions	13	0
Fissions	6	0
Intrachromosomal	69	70

Transposable element densities in peregrine falcon and rock pigeon EBRs

The peregrine falcon and rock pigeon chromosome assemblies provided us with a novel set of EBRs, especially flanking interchromosomal rearrangements, not previously found in published avian chromosome assemblies (Figure 3-5; Table 3-5). This allowed us to evaluate if the sequence features previously found associated with EBRs (Farré et al., 2016, Skinner and Griffin, 2012, Farré et al., 2011, Groenen et al., 2012, Larkin et al., 2009) show different patterns in avian inter- and intrachromosomal EBRs. We identified transposable element (TE) densities associated with avian EBRs at the genome-wide level. We counted the number of TE bases in 10 Kbp windows overlapping EBRs defined to ≤100 Kbp in the peregrine falcon and rock pigeon genomes. We observed that peregrine falcon intrachromosomal EBRs were significantly enriched for the LTR-ERV1 TEs (p-value <0.05; Table 3-6) in agreement with (Farré et al., 2016). For this species, neither fusion and fission EBRs were significantly enriched for any TE family. Rock pigeon intrachromosomal EBRs were also not found significantly enriched for any family of TEs (Table 3-7).

Conserved non-coding elements density in avian inter- and intrachromosomal EBRs

It is known that conservation of DNA sequences across distantly related species reflects functional constraints (Siepel et al., 2005). Furthermore, it is hypothesised that the significant enrichment of avian msHSBs for conserved non-coding elements (CNEs) might contribute to the maintenance of synteny in avian genomes (Farré et al., 2016). To test if this would result in EBRs being in regions of lower DNA sequence conservation, we calculated avian CNE densities in the chicken chromosome regions corresponding to the chicken, peregrine falcon, rock pigeon, collared flycatcher and zebra finch intra- and interchromosomal EBRs, and their adjacent regions (Figure 3-6, Supplemental Table 23 and 24). We observed that avian EBRs had a significantly lower fraction of CNEs than

their two adjacent chromosome intervals of the same size each (up- and downstream; p-value = 3.35e-07; Table 3-8). The interchromosomal EBRs (fusions and fissions) had on average ~12 times lower density of CNEs than the intrachromosomal EBRs (p-value = 2.40e-05; Table 3-8) and both fusion and fission EBRs separately presented significant differences in CNE density to intrachromosomal EBRs (p-value < 0.04; Table 3-8). The lowest density of CNEs was observed in the fission breakpoints (p-value = 0.04; Figure 3-6; Table 3-8).

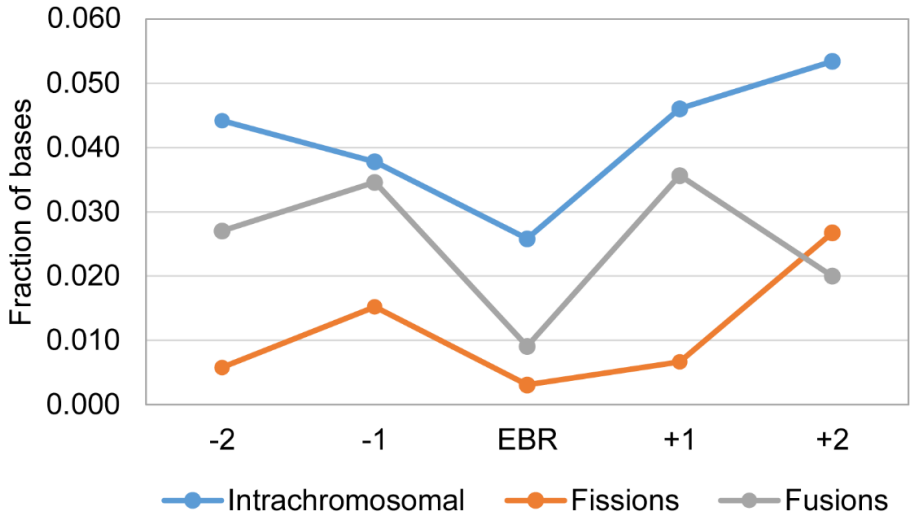

Figure 3-6: Average fraction of bases within CNEs in avian EBRs and two flanking regions upstream (-) and downstream (+).

To identify if avian EBRs are located in regions of lower CNE density genome-wide, we counted CNE bases in 1 Kbp windows overlapping EBRs and avian msHSBs >1.5 Mbp (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; Table 3-9). The genome-wide CNE density was 0.09, closer to the density observed in msHSBs. Of ~347 Mbp 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 Kbp), intermediate in the

intrachromosomal (~19 Kbp) and fusion EBRs (~23 Kbp), and highest in the fission EBRs (~35 Kbp) (Table 3-10). All these values were significantly different from the genome-wide average distance of ~6 Kbp (p-values <2.2e-16) and also significantly different from each other (p-value ≤0.004; Table 3-9; Table 3-10).

Table 3-6: Differences in TE densities in 10 Kbp intervals overlapping peregrine falcon EBRs and the rest of the peregrine falcon

Repeat set	All EBRs			Intrachromosomal			Fusions			Fissions		
	EBR	non-EBR	Ratio	EBR	non-EBR	Ratio	EBR	non-EBR	Ratio	EBR	non-EBR	Ratio
LINE	212.11	272.40	0.78	224.80	272.37	0.82	147.00	272.35	0.54	213.83	272.32	0.78
LINE-CR1	203.69	266.49	0.76	214.29	266.46	0.80	147.00	266.43	0.55	6.00	266.40	0.02
LTR	275.44	105.99	2.60 *	307.23	106.01	2.90 *	161.50	106.21	1.52	99.83	106.22	0.94
LTR-ERV1	107.15	7.22	14.83 *	118.58	7.24	16.38 *	78.18	7.34	10.64	0.00	7.36	0.00

*Statistically significant differences (FDR corrected p-values <0.05). Only statistical significance for transposable elements covering ≥100 bp on average in the EBR and/or non-EBR 10 Kbp intervals in each individual comparison is reported.

Table 3-7: Differences in TE densities in 10 Kbp intervals overlapping rock pigeon EBRs and the rest of the rock pigeon genome.

Repeat set	Intrachromosomal		
	EBR	**non-EBR**	**Ratio**
LINE	540.61	428.25	1.26
LINE-CR1	494.17	421.22	1.17

Table 3-8: Significant differences in CNE densities in avian lineage-specific EBRs and their four adjacent intervals (±2) of the same size.

Group 1	Group 2	Median Group 1	Median Group 2	Ratio	p-value
(± 2)	EBR	0.021	0.001	21.778	1.56e-09
Intra (± 2)	Intra EBR	0.022	0.002	11.204	1.02e-08
(+ 2)	EBR	0.021	0.001	21.978	3.61e-08
Intra (+ 2)	Intra EBR	0.023	0.002	11.681	1.64e-07
(± 1)	EBR	0.015	0.001	15.948	3.35e-07
(- 2)	EBR	0.019	0.001	20.426	5.58e-07
Intra (- 2)	Intra EBR	0.022	0.002	10.924	2.15e-06
Intra (± 1)	Intra EBR	0.017	0.002	8.812	3.06e-06
(+ 1)	EBR	0.015	0.001	15.948	5.92e-06
(- 1)	EBR	0.016	0.001	16.378	1.32e-05
Intra*	Inter[1]*	0.016	0.001	12.240	2.40e-05
Intra (+ 1)	Intra EBR	0.017	0.002	8.771	3.46e-05
Intra (- 1)	Intra EBR	0.018	0.002	8.933	6.71e-05
Intra*	Fusions*	0.016	0.001	13.662	0.0002
Intra (± 2)	Inter[1] (± 2)	0.022	0.003	8.475	0.0016
Intra (± 2)	Fusions (± 2)	0.022	0.002	11.911	0.0038
Intra (+ 2)	Inter (+ 2)	0.023	0.002	12.419	0.0233
Intra (+ 2)	Fusions (+ 2)	0.023	0.001	16.429	0.0270
Intra (- 2)	Inter[1] (- 2)	0.022	0.003	6.255	0.0299
Intra*	Fissions*	0.016	0.003	5.350	0.0400
Inter EBR	Intra EBR	0.000	0.002	0.000	0.0445
Intra (± 1)	Inter (± 1)	0.017	0.003	6.081	0.0460

[1] Fusions and fissions combined

* EBR and adjacent intervals combined

Table 3-9: Statistics for CNE density in 1 Kbp windows for avian EBRs, msHSBs, and genome-wide.

	Average no. CNEs	Average no. CNE bases	Average density of CNE bases	Fraction 'zero CNE windows' (%)
Genome	6.18	86.85	0.09	100
msHSB	7.54	106.81	0.11	15.44
Intra	1.90	23.72	0.02	0.41
Fusion	0.65	6.75	0.01	0.06
Fission	0.00	0.00	0.00	0.02
EBR*	1.71	21.25	0.02	0.50

*Fission, fusion and intrachromosomal EBRs combined.

Table 3-10: Significant differences for distances in number of 1 Kbp windows between zero and high CNE density windows.

Group 1	Group 2	Median Group 1	Median Group 2	Ratio	p-value
Fission	Genome	35	6	5.83	<2.20e-16
Fission	msHSB	35	4	8.75	<2.20e-16
Fusion	Genome	23	6	3.83	<2.20e-16
Fusion	msHSB	23	4	5.75	<2.20e-16
Intra	Genome	19	6	3.17	<2.20e-16
Intra	msHSB	19	4	4.75	<2.20e-16
msHSB	Genome	4	6	0.67	<2.20e-16
Fission	Intra	35	19	1.84	4.70e-14
Fusion	Intra	23	19	1.21	6.72e-05
Fission	Fusion	35	23	1.52	0.0038

Gene-functional enrichment in EBRs

Lineage-specific EBRs have been shown to be enriched for genes related to lineage-specific phenotypes (Groenen et al., 2012, Ullastres et al., 2014, Farré et al., 2016). To identify gene pathways associated with lineage-specific EBRs for the two newly assembled avian genomes, we measured gene ontology (GO) enrichment in the peregrine falcon and rock pigeon lineage-specific EBRs. Moreover, to evaluate if peregrine falcon intra- and interchromosomal EBRs have the same or different gene signatures we split them into two different sets. We observed that rock pigeon EBRs and peregrine falcon interchromosomal EBRs tend to reshuffle genes related with more general GO terms such as *cell*, *biological regulation* and *binding* (Figure 3-7). In their turn, peregrine falcon

intrachromosomal EBRs show an enrichment for genes related to *response to stimulus* and *response to bacterium* (Figure 3-7).

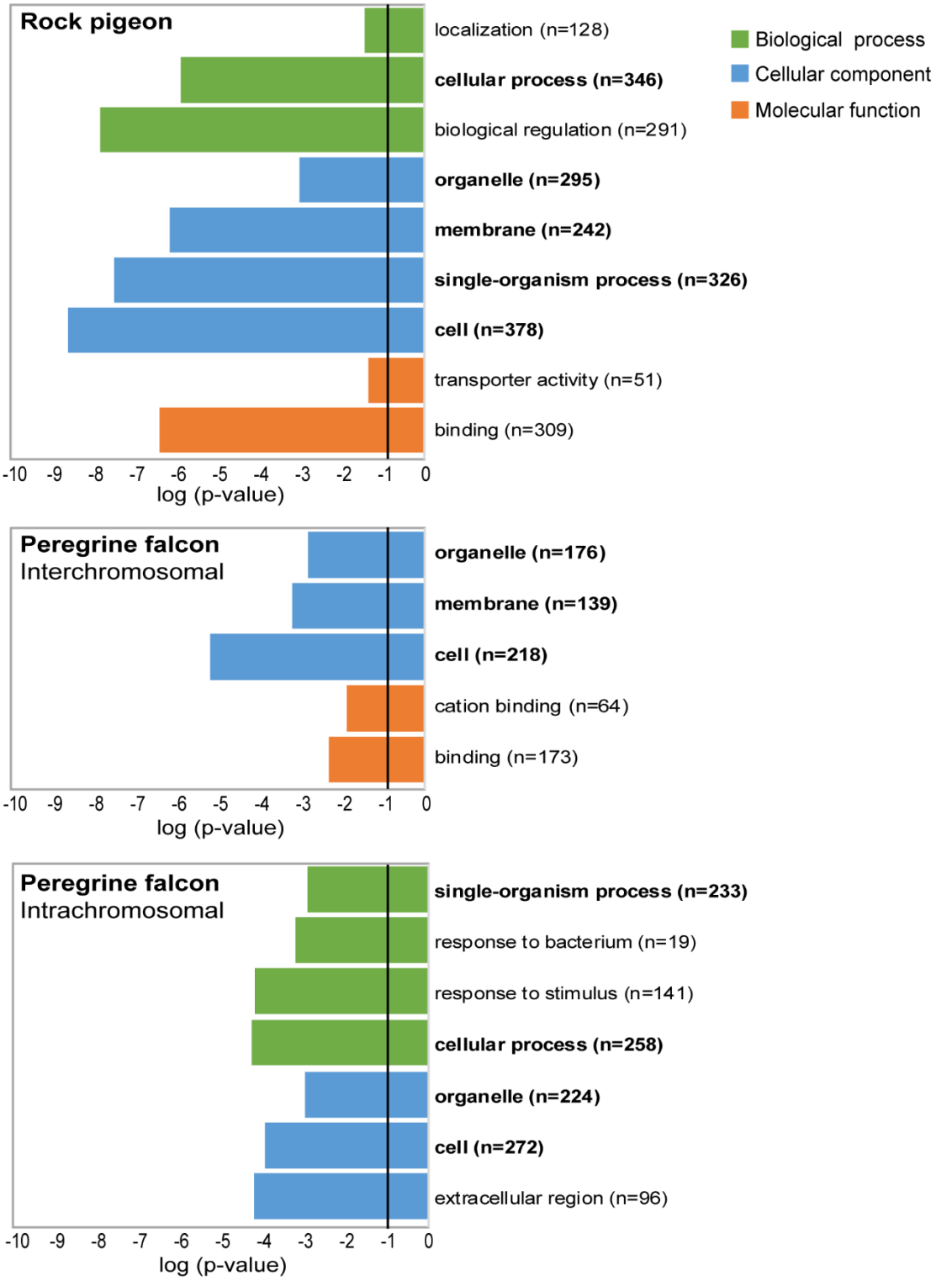

Figure 3-7: GO terms enriched on lineage-specific EBRs (p-value <0.05; FDR <5%). Number of genes in each term are given in parenthesis. Shared GO terms are depicted in bold. Black line depicts p-value threshold.

Data availability

Peregrine falcon and rock pigeon chromosome assemblies are available from DDBJ/ENA/GenBank under the accessions MLQY00000000 and MLQZ00000000, respectively. Comparative visualisations of both newly assembled genomes are available from the Evolution Highway comparative chromosome browser under the reference name “Peregrine:150K” and “Pigeon:150K” (http://eh-demo.ncsa.uiuc.edu/birds). Peregrine falcon chromosomes 1-13 and Z were named according to (Nishida et al., 2008), and chromosomes 14-18 were numbered by decreasing combined length of the placed PCFs. Rock pigeon chromosomes 1-9 and Z were named according to (Hansmann et al., 2009) with the remaining chromosome names assigned according to chicken homeologues. Unassigned PCFs were named after their reference chromosome homeologues with “un” added in front of their names to distinguish from chromosome assemblies.

Peregrine falcon and rock pigeon chromosome assemblies can also be visualised from our UCSC track hub (http://sftp.rvc.ac.uk/rvcpaper/birdsHUB/hub.txt). Each assembly contains four data tracks: (1) Assembly track, containing scaffolds from the original assemblies, PCFs obtained from running RACA (chapter 2) and BAC clones used to place PCFs on the chromosomes. (2) Genes track with translated gene annotation from the original scaffold assemblies. (3) Synteny track with the pairwise HSBs with chicken and zebra finch and lineage-specific EBRs. (4) Repeats track with translated repeat annotations from the original NGS scaffold assemblies.

3.4 Discussion

In this chapter, I present the development of the last step for a novel integrative approach to upgrade fragmented animal genomes to the chromosome-level. I report the design of a panel of avian universally hybridising BAC probes using a methodology that can easily be transferred to other species. The BAC selection approach can be applied to any clade provided cell lines and end-sequenced BAC clone libraries are available. This task could be, however, more challenging for genomes with a higher repetitive content, lower degree of sequence conservation, or requiring more probes to achieve the same level of mapping resolution. Furthermore, the combination of comparative sequence analysis, targeted PCR and optimised high-throughput cross-species hybridizations of universal BAC probes represents a unique methodology to achieve chromosome-level assemblies from scaffold-based *de novo* assemblies, which could be applied to any animal genome. This novel genome assembly methodology can also be easily adapted to accommodate other data types, such as super-scaffolds generated from more advanced mapping and sequencing techniques (e.g., Dovetail, BioNano or PacBio), which could be directly anchored to chromosomes. The chromosome assemblies generated for peregrine falcon and rock pigeon provide proof of principle for this approach. The resulting chromosome-level assemblies contain >80% of the original scaffold-level assemblies being, in continuity, comparable to those obtained by combining traditional sequencing and mapping techniques (Deakin and Ezaz, 2014).

Molecular and cytogenetic studies to date, suggest that most avian genomes remain remarkably conserved in terms of chromosome numbers (>60% of species with 2*n*=~80) and that interchromosomal changes were relatively rare during avian evolution (Griffin et al., 2007, Schmid et al., 2015). Exceptions include representatives of Psittaciformes (parrots), Sphenisciformes (penguins) and Falconiformes (falcons). Herein, I report the first chromosomal assembly of a highly rearranged avian karyotype (peregrine falcon). 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.

This new chromosome-level assembly allowed the analysis of the genomic signatures of avian interchromosomal EBRs and provided insights into the mechanisms driving avian genome evolution.

As discussed in chapter 2, the absence of interchromosomal rearrangements seen in most birds could either suggest an evolutionary advantage to retaining synteny or little opportunity for change (Farré et al., 2016, Romanov et al., 2014). On the one hand, repetitive DNA sequences, particularly TEs, are often used as a template for NAHR, after a double-strand break, resulting in the formation of a chromosomal rearrangement. Because of that, it was suggested that the low repetitive content of avian genomes could represent fewer opportunities for avian genomes to change due to the lack of templates for NAHR. On the other hand, CNEs were found strongly enriched in the regions of interspecies synteny in birds and other reptiles (Farré et al., 2016). Many CNEs are known to play important roles in gene regulation, and so the breakage of synteny could disrupt important regulatory pathways, suggesting an advantage in maintaining synteny. The results obtained during this work provide further support to both theories. Indeed, peregrine falcon lineage-specific intrachromosomal EBRs were enriched for TEs, suggesting NAHR as their generating mechanism. This was not observed for fusions nor fissions, which might suggest that a different mechanism is responsible for the merging of chromosomes and generation of fusions. The fact that EBRs locate in areas of low CNE density, and EBRs flanking chromosomal fissions present the lowest CNE density amongst EBR types and are restricted to “CNE deserts”, might explain why peregrine falcon lineage-specific fission EBRs appear to be reused in other avian lineages as intrachromosomal EBRs. Altogether, these results support the proposition that when disrupting CNE-dense regions a chromosome rearrangement has a higher probability of affecting gene regulation pathways and causing deleterious effects that would not be tolerated by selection. Moreover, the differences observed between EBR types agree with previous studies in Passeriform species that suggested that intrachromosomal rearrangements might have a less dramatic effect on *cis* gene regulation than interchromosomal events (Skinner and Griffin, 2012, Romanov et al., 2014). In fact, only neutral, nearly neutral or selectively advantageous rearrangements are likely to be fixed (Burt, 2001) and indeed, we found that peregrine falcon lineage-specific intrachromosomal EBRs were enriched for genes related to an

organisms' response to external stimuli. Interestingly, this pattern has been reported previously for mammals, such as pig (Groenen et al., 2012) and rhesus macaque (Ullastres et al., 2014), where they were suggested to give these species a selective advantage and to relate to lineage-specific phenotypes.

Despite the insights on the mechanisms driving avian genome evolution provided herein by the study of the first highly rearranged avian genome, there are still questions that lack answers. The reason behind why species such as falcons and parrots undergo large-scale interchromosomal rearrangement while other remain stable or, why are fissions restricted to a few events and fusions more common, remain a matter of future investigation. Moreover, the time and lineage of occurrence of genomic rearrangements during avian evolution also lack resolution. These questions can be addressed by tracing the chromosome organisation of avian ancestors throughout the avian phylogenetic tree. These reconstructions could shed light on the dynamics of the forces shaping avian genomes and the role of chromosomal rearrangements in phenotypic evolution and speciation.

4 Reconstructing avian ancestral karyotypes

4.1 Background

In the previous chapters, I demonstrated the utility of the comparison of extant species genomes to identify genome rearrangements, unravel the mechanisms driving genome evolution and give insights into the role of genome rearrangement in phenotypic diversity. Nonetheless, more information can be gained by tracing the chromosome organisation of common ancestors.

The first reconstructions of ancestral genome structures were based on low-resolution karyotype comparisons using chromosome paintings, or the comparison of gene maps. These studies offered the first insights onto the genome rearrangements that shaped extant genome structures. For instance, cytogenetic information was used to propose the chromosome structures of the ancestors of placental mammals (Richard et al., 2003), ruminants (Kulemzina et al., 2011), carnivores (Beklemisheva et al., 2016) and birds (Griffin et al., 2007). The differences in rearrangement rates detected from these reconstructions suggested different rates of evolution for distinct phylogenetic lineages (Ruiz-Herrera et al., 2012). For instance, contrary to mammals, birds have very stable chromosome numbers, and contrary to other amniotes, mammals and crocodiles do not have microchromosomes. These observations raised questions regarding the mechanisms that drove the evolution of these lineages. Despite their utility, the low resolution and limitations of the cytogenetic methodologies resulted in undetected intrachromosomal rearrangements and limited usefulness for reconstructing karyotypes of old ancestors (e.g., eutherian and amniote ancestors). These restrictions can be overcome by the comparison of genome sequences, which not only can expand the evolutionary depth of ancestral karyotype reconstructions but also increase the resolution at which genome rearrangements are identified.

The first attempts to reconstruct ancestral karyotypes from sequence data were performed for mammals (Bourque et al., 2005, Murphy et al., 2005). These reconstructions allowed the detection of a larger number of genome rearrangements than what was previously known, the observation that evolutionary breakpoint regions (EBRs) are often reused in different lineages, that EBRs locate in gene-dense regions, and that lineage-specific EBRs are associated with the locations of segmental duplications in mammals (Murphy et

al., 2005, Bourque et al., 2005). Overall, these findings showed the importance of genome sequence comparisons for the detection of the full catalogue of events that shaped extant genomes and led to the development of several algorithms to perform the reconstruction of ancestral karyotypes based on genome sequence data. Most of them, for example, InferCARs (Ma et al., 2006) or ANGES (Jones et al., 2012a), require chromosome-level genome assemblies as inputs and their suitability to deal with fragmented assemblies was never shown (Kim et al., 2017). In this way, while there have been many newly sequenced genomes released in the last few years, only a very limited number of them were assembled to chromosome-level and, because of that, were suitable for ancestral genome reconstruction (Kim et al., 2017). This resulted in a stagnation on the field and led Kim and colleagues to develop DESCHRAMBLER (Kim et al., 2017). DESCHRAMBLER algorithm produces reconstructed ancestral chromosome fragments (RACFs) using syntenic fragments (SFs) constructed from whole-genome alignments of both chromosome- and scaffold-level genome assemblies (Kim et al., 2017). The prediction of the order and orientation of the SFs in the target ancestor genome is established using a similar approach to that used by the reference-assisted chromosome assembly (RACA) algorithm (Kim et al., 2013). In a very simplistic way, the probability of adjacency for each pair of SFs in an ancestral chromosome is computed based on the SF adjacency on the extant (descendant and outgroup) species. Other advantages of DESCHRAMBLER are the ability to accommodate a large number of descendant species and the use of SFs where some species (except the reference) have deletions or missing data, which results in a more complete reconstruction of the ancestral karyotype (Kim et al., 2017). At its release, DESCHRAMBLER was applied to the genomes of 21 species (14 chromosome-level and 7 scaffold-level) to reconstruct the chromosome structure of seven eutherian ancestors. In this study, Kim and colleagues compared the reconstructed RACFs with previous FISH-determined eutherian, boreoeutherian and simian ancestral karyotypes and observed a high level of consistency between both methodologies to detect interchromosomal rearrangements (Kim et al., 2017). Additionally, the calculated rearrangement rates were also highly consistent with previous studies (Kim et al., 2017).

The limited availability of chromosome-level assemblies for birds restricted the study of chromosome evolution in this clade. Indeed, to date, the reconstruction of avian ancestral chromosome structures was mainly based on low-resolution karyotype comparisons (Griffin et al., 2007) and only recently the first sequence-based Avian ancestor genome structure was proposed (Romanov et al., 2014). Nonetheless, the low number of species utilised by Romanov and colleagues, limited the generated reconstructions to few Avian and Neognathae ancestor's chromosomes (Romanov et al., 2014).

Herein, making use of the DESCHRAMBLER algorithm (Kim et al., 2017) and a combination of chromosome- and scaffold-level genome assemblies currently available for birds, I report the first large-scale study of ancestral chromosome structure and evolution in this clade. The high number of genomes included in this study, and the extended sampling of the avian phylogenetic tree, allowed the reconstruction of the genome structure of 14 avian ancestors leading to the zebra finch, a representative of the most speciose avian clade (Passeriformes), which species also express a high phenotypical diversity. The analysis of these reconstructions provided important insights on the variability of rearrangement rates during avian evolution and allowed the detection of patterns related to the chromosome distribution of EBRs. Moreover, the inclusion of microchromosomes in our reconstructions allowed us to provide novel insights into the evolution of these avian chromosomes.

4.2 Material and methods

Avian and outgroup genome assemblies

The chicken (*Gallus gallus*; ICGSC Gallus_gallus 4.0; Hillier, 2004), zebra finch (*Taeniopygia guttata*; WUGSC 3.2.4; Warren et al., 2010), and turkey (*Meleagris gallopavo*; TGC Turkey_2.01; Dalloul et al., 2010) chromosome assemblies were downloaded from the UCSC Genome Browser (Kent et al., 2002). The collared flycatcher (*Ficedula albicollis*; FicAlb1.5; Ellegren et al., 2012) chromosome assembly was downloaded from NCBI. The Pekin duck (*Anas platyrhynchos*) chromosome assembly was obtained from Dr Thomas Faraut (INRA, France). Peregrine falcon (*Falco peregrinus*) and rock pigeon (*Columba livia*) chromosome assemblies generated during this project were used. The hooded crow (*Corvus cornix*; Hooded_crow_genome; Poelstra et al., 2014), canary (*Serinus canaria*; SCA1; Frankl-Vilches et al., 2015), Tibetan ground tit (*Pseudopodoces humilis*; PseHum1.0; Cai et al., 2013), golden eagle (*Aquila chrysaetos*; Aquila_chrysaetos-1.0.2; Doyle et al., 2014), bald eagle (*Haliaeetus leucocephalus*; Haliaeetus_leucocephalus-4.0) scaffold assemblies were obtained from NCBI. All remaining scaffold-based assemblies were downloaded from the GigaScience Database (Sneddon et al., 2012, Zhang et al., 2014a). Chromosome assemblies of outgroup genomes: anole lizard (*Anolis carolinensis*; AnoCar2.0; Alfoldi et al., 2011) and opossum (*Monodelphis domestica*; MonDom5; Mikkelsen et al., 2007), the scaffolds assemblies of the Chinese alligator (*Alligator sinensis*; ASM45574v1; Wan et al., 2013) and the painted turtle (*Chrysemys picta*; Chrysemys_picta_bellii-3.0.1; Shaffer et al., 2013) were obtained from NCBI. General assembly statistics for each genome used are presented in Table 4-1.

Pairwise alignments

Pairwise alignments using zebra finch chromosome assembly as reference and all other genomes as targets were generated with LastZ (v.1.02.00; Harris, 2007) using the following parameters: *C=0 E=30 H=2000 K=3000 L=2200 O=400*. The pairwise alignments were converted into the UCSC “chain” and “net” alignment formats with axtChain (parameters: *-minScore=1000 -verbose=0 -linearGap=loose* for anole lizard and opossum, and *-minScore=1000 -verbose=0 -linearGap=medium* for all other species) followed by chainAntiRepeat,

chainSort, chainPreNet, chainNet and netSyntenic, all with default parameters (Kent et al., 2003).

Reconstructed ancestral chromosome fragments

Reconstructed ancestral chromosome fragments (RACFs) were generated by Dr Jaebum Kim (Konkuk University, South Korea) using the DESCHARMBLER algorithm. First, DESCHRAMBLER was used to reconstruct RACFs of the Neognathae ancestor with a subset of species, as indicated in Table 4-1. This experiment was performed at 100, 300 and 500 Kbp SFs resolution. After the selection of the best SF resolution for avian ancestral chromosomes reconstruction (100 Kbp), DESCHRAMBLER was run with the full set of species to generate RACFs for all ancestors leading to zebra finch lineage, starting with the Avian ancestor. The full list of reconstructed ancestor RACFs is presented in Table 4-3.

Avian, Eufalconimorphae and Passeriformes ancestor chromosomes

The number of RACFs reconstructed by DESCHRAMBLER was higher than the number of Avian ancestor chromosomes previously proposed based on FISH experiments (Griffin et al., 2007). This fragmentation is mostly due to the predominance of scaffold-level assemblies for the descendant species, resulting in a reduction of adjacency support. To reduce the fragmentation of the Avian ancestor genome, we reorganised Avian ancestor RACFs by connecting RACFs which adjacency was supported by outgroup genomes or other, phylogenetically close and less fragmented, ancestors. Specifically, we first merged those Avian ancestor RACFs which adjacencies were supported (spanned) by an outgroup chromosome or scaffold. For the remaining adjacencies with no support from outgroup genomes, we merged Avian RACFs which adjacency was supported by other ancestor RACF, assuming that no rearrangement occurred between the Avian and the descendant ancestor in between RACFs. For each RACF adjacency, we used the support from the spanning RACF belonging to ancestors successively more distant to the Avian ancestor. That is, we used first the support from the Neognathae ancestor (the closest to the Avian ancestor) and successively more recent ancestors on the avian phylogenetic tree. The same approach was applied to Eufalconimorphae and Passeriformes ancestor RACFs.

Detection of EBRs and chromosome rearrangements

We detected EBRs relative to the avian ancestor in all other ancestors' RACFs, chicken and zebra finch chromosomes. Breakpoint rates (EBRs/MY) for each branch leading to zebra finch were calculated dividing the number of detected EBRs by the length of the branch (in MY). Differences in breakpoint rates compared to the average of all branches were tested as previously described (Kim et al., 2017). We used the genome rearrangements in man and mouse (GRIMM) webserver (Tesler, 2002) to predict the minimum number and the type of chromosomal rearrangements distinguishing the Avian ancestor chromosomes structure from those of the Eufalconimorphae and Passeriformes ancestors and the zebra finch.

EBR rates and DNA sequence feature associations on Avian ancestor chromosomes

We measured EBR density and distribution for the Avian ancestor chromosomes using the number of EBRs identified between the Avian ancestor and zebra finch. These measurements were obtained as the number of EBRs per Mbp, the average distance between EBRs and the difference between the expected and observed number of EBRs. The expected number of EBRs per chromosome was calculated by multiplying the length of the chromosome (in Mbp) by the genome-wide rate of EBRs. The latter calculated by dividing the total number of detected EBRs by the total length of the reconstructed Avian ancestor genome. Differences between chromosomes for each of the analysed features were tested as previously reported (Kim et al., 2017).

Avian conserved non-coding elements (CNEs) were obtained from (Farré et al., 2016). 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). We calculated the density of each of these features (CNEs, genes and transposable elements (TEs) for each Avian ancestor chromosome. The association between each sequence feature and chromosome-specific EBR density and distribution was tested using the Pearson's correlation coefficient.

Gene ontology enrichment analysis

The basic version of gene ontology (GO) annotations (version 3rd May 2016) was downloaded from the GO Consortium website (The Gene Ontology Consortium, 2015). Sequence coordinates and Ensembl identifiers for chicken genes we obtained from Ensembl Biomart (v.74; Kinsella et al., 2011). All chicken genes located in regions included in the Avian ancestor chromosomes were used as the background list. To evaluate gene functional enrichment in the Avian ancestor chromosomes that were maintained intact during avian evolution, we assigned genes from the background list to these chromosomes. We used the GO::TermFinder Perl module (Boyle et al., 2004) to detect GO terms overrepresented in our gene sets. We considered as significantly enriched the terms with p-value ≤0.05 and false discovery rate (FDR) <5%.

Table 4-1: Statistics for genome assemblies of descendant and outgroup species.

Species	Common name	Assembly type	No.[1]	N50 (Mbp)	Total length (Gbp)	On test recons.[2]
Taeniopygia guttata	Zebra finch	Chromosome	31	-	1.02	Yes
Geospiza fortis	Medium ground finch	Scaffold	1,168	5.28	1.04	Yes
Serinus canaria	Canary	Scaffold	887	25.15	1.05	No
Pseudopodoces humilis	Tibetan ground tit	Scaffold	661	16.34	1.04	No
Corvus brachyrhynchos	American crow	Scaffold	1,156	7.08	1.08	Yes
Corvus cornix	Hooded crow	Scaffold	366	16.36	1.05	No
Ficedula albicollis	Collared flycatcher	Chromosome	30	-	1.04	Yes
Manacus vitellinus	Golden-collared manakin	Scaffold	954	2.86	1.05	Yes
Melopsittacus undulatus	Budgerigar	Scaffold	1,138	11.41	1.08	Yes
Falco peregrinus	Peregrine falcon	Chromosome	19	-	1.03	No
Aquila chrysaetos	Golden eagle	Scaffold	470	9.23	1.19	No
Haliaeetus leucocephalus	Bald eagle	Scaffold	435	9.15	1.18	No
Picoides pubescens	Downy woodpecker	Scaffold	1,944	2.12	1.15	Yes
Pygoscelis adeliae	Adélie penguin	Scaffold	819	5.23	1.21	Yes
Aptenodytes forsteri	Emperor penguin	Scaffold	682	5.08	1.25	Yes
Nipponia nippon	Crested ibis	Scaffold	1,479	5.35	1.20	Yes
Egretta garzetta	Little egret	Scaffold	1,195	3.11	1.20	Yes
Opisthocomus hoazin	Hoatzin	Scaffold	1,620	2.94	1.20	Yes
Charadrius vociferus	Killdeer	Scaffold	1,598	3.68	1.21	Yes
Cuculus canorus	Common cuckoo	Scaffold	900	2.99	1.15	Yes
Chaetura pelagica	Chimney swift	Scaffold	1,172	3.88	1.10	Yes
Calypte anna	Anna's hummingbird	Scaffold	887	4.30	1.05	Yes
Columba livia	Rock pigeon	Chromosome	29	-	0.91	No
Gallus gallus	Chicken	Chromosome	30	-	1.00	Yes
Meleagris gallopavo	Turkey	Chromosome	32	-	1.04	Yes
Anas platyrhynchos	Pekin duck	Chromosome	29	-	0.94	Yes
Struthio camelus	Ostrich	Scaffold	1,179	3.64	1.22	Yes
Alligator sinensis	Chinese alligator	Scaffold	2,452	2.19	2.26	Yes
Chrysemys picta	Painted turtle	Scaffold	3,168	7.23	2.29	Yes
Anolis carolinensis	Anole lizard	Chromosome[3]	12	-	1.08	No
Monodelphis domestica	Opossum	Chromosome	9	-	3.50	No

[1] Number of scaffolds or chromosomes on the genome assembly.
[2] Species included on Neognathae test reconstructions.
[3] Chromosomes and linkage groups in anole lizard.

4.3 Results

Selection of reference and descendant genomes

Cytogenetic comparisons offered the first rough insight into the evolution of avian genomes (Griffin et al., 2007). Nonetheless, only the comparison of genome sequences provides access to the full collection of events that shaped genomes through evolution. Herein, we used the DESCHRAMBLER algorithm (Kim et al., 2017) to predict the chromosome structure of avian ancestors leading to the zebra finch.

We selected zebra finch as reference genome for the reconstruction of these ancestors because of: (a) the quality of the genome assembly for this species, which is one of the best avian genome assemblies currently available, and (b) the fact that DESCHRAMBLER requires the reference genome to be a descendant species of all reconstructed ancestors. Thus, using zebra finch as a reference genome allowed the reconstruction of a high number of avian ancestors and, subsequently, a thorough study of avian genome evolution. Moreover, zebra finch is a representative of Passeriformes, the avian clade with the highest number of extant species, and which species also exhibit a high phenotypical diversity.

The final set of descendant and outgroup species included 27 avian genomes (7 chromosome assemblies and 20 scaffold-level assemblies) and 4 outgroup genomes (2 chromosome- and 2 scaffold-level assemblies: 3 non-avian reptiles and one mammal; Table 4-1). The selection of these genomes was made according to their assembly continuity and alignment coverage of the reference (zebra finch) genome. Having a high coverage of the reference genome also assures a more thorough coverage of the ancestral genomes. In this way, all selected descendant species alignments cover >96% of zebra finch genome. Assembly continuity is also an important selection parameter because, (a) the use of highly fragmented descendant genomes will reduce the support to predicted ancestral adjacencies and result in fragmented ancestral reconstructions, and (b) increased genome continuity will increase the chances of detection of EBRs flanking genome rearrangements, as EBRs locating in between scaffolds in the extant species assemblies will be missed. Having this in mind, we selected from the set of Avian Phylogenomics Consortium sequenced

genomes only those with an N50 >2 Mbp. This allowed us to have a comprehensive sampling across the avian phylogenetic tree while maintaining a relatively high genome continuity. Peregrine falcon and rock pigeon scaffold-level assemblies were replaced by the chromosome assemblies generated within this project (Chapter 3). To further improve taxon sampling, we decided to complement the descendant genomes with other avian high-continuity genomes (chromosome-level assemblies and scaffold-level assemblies with N50 >9 Mbp) currently available (Table 4-1). These include the chromosome assemblies of the collared flycatcher and Pekin duck, and the scaffold level assemblies of canary, Tibetan ground tit, hooded crow, golden eagle and bald eagle. The final set of avian species included in our reconstructions represent 15 out of 37 avian orders and allowed the reconstruction of 14 ancestors starting with the Avian ancestor and leading to the zebra finch lineage.

Selection of resolution for syntenic fragment detection

We first performed the reconstruction of the Neognathae ancestor genome at three (100, 300 and 500 Kbp) different syntenic fragment (SF) resolutions, to set the minimum length of SF to be included in the reconstructions. We selected the Neognathae ancestor for this experiment as both ingroup and outgroup genome alignments cover >94% of the zebra finch genome. To be included in the RACFs an SF needs to be present in at least one of the outgroup genomes, in this way, by including only those outgroup genomes with alignments having a high coverage of the reference genome we could reduce the confounding effect of low outgroup coverage on the completeness of the RACFs. This experiment aimed the establishment of the optimal resolution for avian ancestral genome reconstructions. We verified that the number of RACFs obtained at 100 Kbp SF resolution was the lowest (N=62) and the reference genome coverage the highest (79%; Table 4-2). At 300 Kbp resolution, RACF covered ~46% of zebra finch genome and the number of reconstructed RACFs was much higher (N=80; Table 4-2). Lastly, at 500 Kbp resolution, there were 64 RACFs reconstructed that covered ~31% of the reference genome (Table 4-2). To minimise the fragmentation of the reconstructed ancestral genomes and, at the same time, maximise the coverage of the reference genome we decided to use the SF resolution of 100 Kbp in the final reconstructions.

Reconstructed ancestral chromosome fragments

After the selection of the optimal SF resolution for avian ancestral genome reconstructions (100 Kbp), we used DESCHRAMBLER to generate RACFs for 14 avian ancestors in the lineage leading to the zebra finch, starting with the Avian and passing through Neognathae, Neoavian, landbirds and Passeriformes ancestors (Table 4-3). The generated genome reconstructions range from 46 (Eufalconimorphae ancestor) to 89 (Passerea ancestor) RACFs and covered >77% of the reference genome (Table 4-3). We observed a lower number of reconstructed RACFs for those ancestors for which both sides of the bifurcation contained chromosome-level assemblies or scaffold-level assemblies with N50 > 9 Mbp (Ancestors 1, 2, 3, 4, 6, 7, 8, 12 and 13; Table 4-3). The only exception to this rule was the ancestor 6 (Psittacopasserae) presenting 65 RACFs. This result might relate to the presence of multiple potential misassemblies in the budgerigar genome as already mentioned in the previous chapters.

Avian ancestor chromosomes

The number of RACFs obtained for each ancestor was higher than the number of chromosomes previously proposed based on molecular cytogenetics (Griffin et al., 2007). For instance, Griffin and colleagues proposed an ancestral avian karyotype (chromosomes 1-10 and Z) containing only one interchromosomal difference to the chicken karyotype, where chicken chromosome 4 results from the fusion of ancestral chromosomes 4 and 10 (Griffin et al., 2007). Moreover, avian karyotypes are known to have a very low number of interchromosomal changes, with exceptions limited to few avian lineages, such as Falconiformes referred in the previous chapter. Considering this information and aiming at the reduction of the fragmentation of the Avian ancestor, we ordered Avian ancestor RACFs along chromosomes using both information from the outgroup genomes or other, more recent and less fragmented, ancestors. Using this approach, we generated an Avian ancestral karyotype comprising 27 chromosomes that are homeologous to zebra finch chromosomes 1-28, 4A and Z (except 16 and 25). We named the Avian ancestor chromosomes accordingly to their zebra finch homeologues. The comparative visualisation of all other ancestor RACFs against the Avian ancestor chromosomes is available at the Evolution Highway (EH) comparative chromosome browser under the reference genome name

"Avian:Ancestor" (http://eh-demo.ncsa.uiuc.edu/avian; subset shown in Figure 4-1).

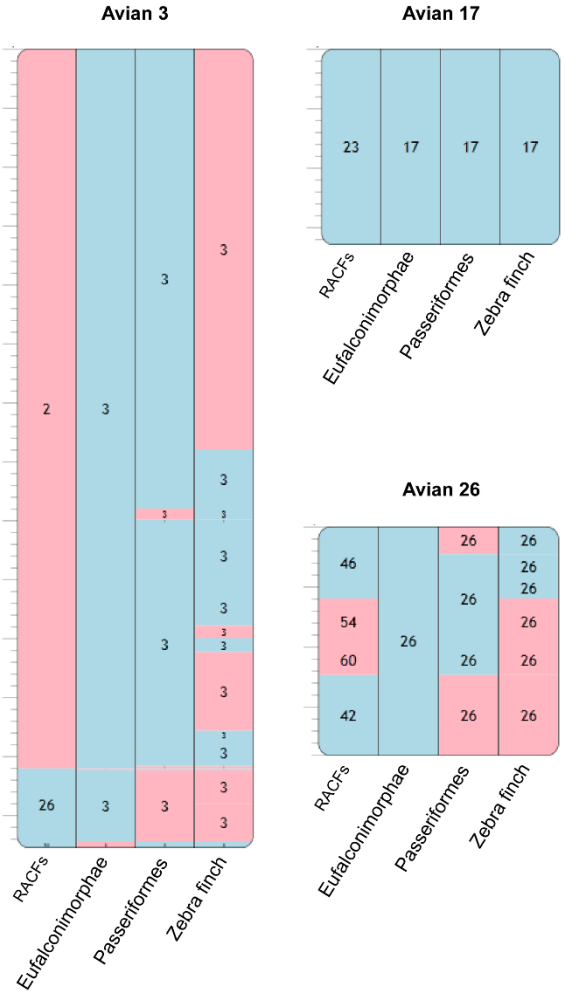

Figure 4-1: Avian ancestor chromosome 3, 17 and 26 representations on the Evolution Highway comparative chromosome browser. Blue and pink blocks define syntenic fragments, in "+" and "-" orientation, respectively. RACFs track represent DESCHRAMBLER-reconstructed Avian RACFs, with numbers within blocks depicting RACF identifiers. Eufalconimorphae, Passeriformes, and zebra finch tracks represent chromosomes and number within blocks depict chromosome identifiers.

Table 4-2: Statistics for the Neognathae ancestor reconstructions at different resolutions of SF detection.

Resolution (Kbp)	No. RACFs	Total length RACFs (Kbp)	Coverage (%) *	Longest RACF (Kbp)	Shortest RACF (Kbp)	No. SFs	Longest SF (Kbp)	Shortest SF (Kbp)
100	62	802,050.27	78.60	92,542.59	108.99	3,539	1,654.18	100.01
300	80	471,160.79	46.17	47,135.00	310.97	1,018	1,700.52	300.01
500	64	319,886.64	31.35	33,091.13	508.92	456	1,700.52	501.68

* Percentage of sequence coverage against the zebra finch genome (1,020,453,418 bp).

Table 4-3: Statistics of the reconstructed ancestors (100 Kbp resolution).

ID	Ancestor	No. RACFs	Total length RACFs (Kbp)	Coverage (%) *	Longest RACF (Kbp)	Shortest RACF (Kbp)	No. SFs	Longest SF (Kbp)	Shortest SF (Kbp)
14	Avian	79	790,916.41	77.51	90,786.44	100.10	3,488	1,554.45	100.04
13	Neognathae	54	806,245.38	79.01	92,678.53	108.57	3,443	1,554.45	100.04
12	Neoavian	56	831,511.32	81.48	96,145.84	111.70	3,440	1,554.45	100.04
11	Passerea	89	840,780.94	82.39	85,230.40	100.05	3,464	1,731.24	100.04
10	Telluraves & Aequornithia & Gruae	75	886,361.18	86.86	100,620.83	100.05	3,235	2,572.30	100.04
9	Telluraves & Aequornithia	68	916,263.77	89.79	104,425.83	100.05	2,979	2,572.30	100.00
8	Telluraves	53	957,749.67	93.86	109,264.85	109.58	2,424	3,779.46	100.00
7	Eufalconimorphae	46	981,131.47	96.15	112,969.25	109.58	1,775	5,889.26	100.00
6	Psittacopasserae	65	986,045.69	96.63	102,299.10	109.58	1,698	7,032.14	100.00
5	Passeriformes	64	996,905.97	97.69	154,087.81	100.05	1,435	8,844.43	100.02
4	Passeri	50	1,002,337.45	98.22	154,800.17	113.08	1,031	17,617.67	100.02
3	Passeroidea & Paroidea	54	1,005,719.08	98.56	117,574.56	113.08	981	17,617.67	100.02
2	Passeroidea	49	1,009,242.62	98.90	118,597.80	314.94	792	31,490.79	100.02
1	Estrilidade & Thraupidae & Fringilidae	51	1,011,702.12	99.14	155,663.35	287.41	689	31,490.79	100.31

* Percentage of sequence coverage against the zebra finch genome (1,020,453,418 bp).

Chromosome rearrangement rates during avian genome evolution

To estimate and compare rates of chromosomal rearrangement during avian evolution in the lineage leading to zebra finch, we calculated the number of EBRs for each branch of the phylogenetic tree (Figure 4-2). We detected 201 EBRs flanking rearrangements occurring during the ~100 MY of avian evolution (from the Avian ancestor to zebra finch). The average rearrangement rate was estimated as 2.01 EBRs/MY. The Neoavian to Passerea; Telluraves to Eufalconimorphae; Passeriformes to Passeri and zebra finch branches had rearrangement rates significantly higher than the average (> 3.5 EBRs/MY; FDR-corrected p-value ≤ 0.006; Table 4-4; Figure 4-2). The opposite trend was observed on the branches Neognathae to Neoavian; Passerea to Telluraves and Aequornithia; Passeri to Passeroidea and Paroidea, which had genome rearrangement rates significantly lower than the average (<1 EBR/MY; FDR-corrected p-value <0.03; Table 4-4; Figure 4-2).

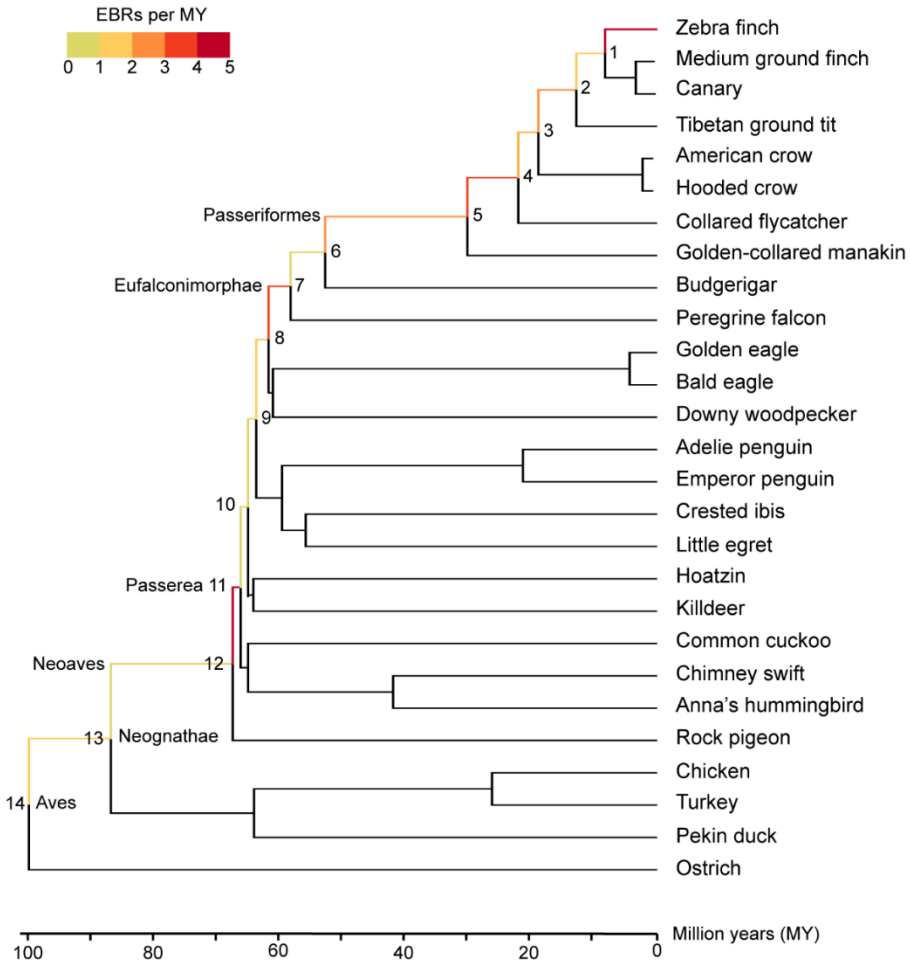

Figure 4-2: Phylogenetic tree of descendant species and reconstructed ancestors. Branch colour represent breakpoint rates in RACFs (EBRs/MY). Numbers on nodes represent ancestor ID as shown in Table 4-3.

Table 4-4: Number of EBRs and EBR rates for the reconstructed ancestral genomes.

Branch			Branch length (MY)	No. EBRs	EBRs per MY	FDR corrected p-value [1]
Avian	→	Neognathae	13.1	19	1.45	0.277
Neognathae	→	Neoavian	19.4	8	0.41	0.006
Neoavian	→	Passerea	1.2	6	5.00	1.61E-04
Passerea	→	Telluraves & Aequornithia & Gruae	1.2	1	0.83	0.031
Telluraves & Aequornithia & Gruae	→	Telluraves & Aequornithia	1.3	0	0.00	0.002
Telluraves & Aequornithia	→	Telluraves	2.0	2	1.00	0.057
Telluraves	→	Eufalconimorphae	3.5	14	3.71	0.005
Eufalconimorphae	→	Psittacopasserae	5.4	6	1.11	0.081
Psittacopasserae	→	Passeriformes	22.6	47	2.08	0.882
Passeriformes	→	Passeri	8.1	29	3.58	0.006
Passeri	→	Passeroidea & Paroidea	3.2	1	0.31	0.005
Passeroidea & Paroidea	→	Passeroidea	6.0	14	2.33	0.561
Passeroidea	→	Estrilidade & Thraupidae & Fringilidae	4.6	8	1.74	0.581
Estrilidade & Thraupidae & Fringilidae	→	Zebra finch	9.9	46	4.65	2.84E-04

[1] Compared to the average across all branches (2.01 EBRs/MY).

EBRs distribution in Avian ancestor chromosomes

It has been proposed that rearrangements in microchromosomes are rare and these chromosomes represent highly conserved blocks of synteny (Romanov et al., 2014). To test this hypothesis, we estimated the density of EBRs, detected between the Avian ancestor and the zebra finch, for each of the reconstructed Avian ancestor chromosomes.

We first tested if there was a difference in EBR density between chromosomes. We observed that microchromosomes 26, 27 and 28 had significantly higher EBR density than the average across all chromosomes (average ~0.48 EBR/Mbp; FDR-corrected p-value ≤1.73E-07; Table 4-5). Specifically, these chromosomes have EBR densities more than three times higher than the average (>1.4 EBR/Mbp; Table 4-5). Distinctively, Avian ancestor chromosomes 2, 3, 6, 8, 9 and 10 (all macrochromosomes) have EBR densities up to eight times lower than average (FDR-corrected p-value ≤0.05; Table 4-5). Also with an EBR density lower than average are Avian ancestor microchromosomes 17 and 22 that were maintained intact during the ~100 MY of avian evolution up to the zebra finch.

Next, we tested differences in EBR density by averaging the distance between EBRs and between the last/first EBR end/start of the chromosomes. We noted that Avian ancestor chromosomes 26 to 28 have average distances between EBRs significantly lower than the genome-wide average (FDR-corrected p-value ≤4.16E-05; Table 4-5), in agreement with their higher than average EBR density per Mbp. We also observed that the chromosomes with a lower EBR density per Mbp have a higher than average distance between EBRs (FDR corrected p-value ≤4.59E-06; Table 4-5).

Lastly, we tested if the EBRs were distributed uniformly across the genome, by calculating the difference between the number of observed and expected EBRs for each chromosome. We observed that all Avian ancestor chromosomes with an EBR density significantly lower than average also possess fewer EBRs than would be expected if the EBRs were distributed uniformly across the genome (FDR-corrected p-value <0.02; Table 4-5). We noted the same trend for the chromosomes with an EBR density higher than average, that is, these chromosomes contain a significantly higher number of EBRs than would be

expected from a uniform EBR distribution along the genome (FDR corrected p-value <0.03; Table 4-5).

Table 4-5: EBR distribution and fraction within genes, CNEs, and TEs for each Avian ancestor chromosome.

Avian chr.	Length (Mbp)	Chromosome fraction within			EBRs per Mbp	Average EBR distance (Mbp)	Obs – Exp no. EBRs
		Genes	TEs	CNEs			
1	151.05	0.51 *	0.04	0.11	0.20	4.72 *	- 7 *
2	126.80	0.46 *	0.10 *	0.11	0.18 *	6.04 *	- 8 *
3	89.44	0.51 *	0.05	0.11	0.14 *	5.59 *	- 9 *
4	56.83	0.46 *	0.03 *	0.09 *	0.28	3.34	2
4A	16.17	0.44 *	0.03 *	0.12	0.37	2.31	2
5	51.05	0.54 *	0.04	0.11	0.24	3.93	0
6	30.27	0.55	0.04 *	0.11	0.13 *	6.06 *	- 3 *
7	32.92	0.54 *	0.12 *	0.13 *	0.24	3.66	0
8	21.98	0.50 *	0.05	0.14 *	0.09 *	7.33 *	- 3 *
9	22.93	0.46 *	0.04 *	0.10	0.13 *	5.73 *	- 2 *
10	16.99	0.60	0.07	0.16 *	0.06 *	8.55 *	- 3 *
11	17.39	0.42 *	0.16 *	0.17 *	0.46	1.93 *	4 *
12	17.26	0.63	0.04 *	0.14 *	0.23	3.45	0
13	13.84	0.63	0.05	0.14 *	0.29	2.77	1
14	11.98	0.62	0.26 *	0.11	0.34	2.40	2
15	11.73	0.60	0.09 *	0.12	0.26	2.93	1
17	9.66	0.60	0.05	0.13 *	0.00 *	NA	- 2 *
18	8.79	0.64	0.04	0.11	0.57	1.46 *	3 *
19	8.90	0.66 *	0.04 *	0.12	0.22	2.97	0
20	10.91	0.59	0.05	0.14 *	0.28	3.64	1
21	4.15	0.82 *	0.04 *	0.11	0.72	1.04 *	2
22	2.03	0.64	0.03 *	0.10	0.00 *	NA	0
23	2.97	0.71 *	0.07	0.09	0.67	0.99 *	2
24	4.57	0.66 *	0.03 *	0.01 *	0.66	1.14 *	2
26	2.91	0.69 *	0.06	0.08 *	1.72 *	0.48 *	5 *
27	2.10	1.00 *	0.04 *	0.05 *	1.43 *	0.52 *	3 *
28	2.63	0.74 *	0.03 *	0.06 *	3.05 *	0.38 *	8 *
Z	42.67	0.67 *	0.06	0.05 *	0.61	1.64 *	16 *
Average	-	0.60	0.06	0.11	0.48	3.27	- 1

* Statistical significance (FDR-corrected p-value < 0.05) compared to the average across all chromosomes.

EBR distribution and DNA sequence features association

Diverse DNA sequence features were previously shown to associate with the presence of EBRs. Among them are CNEs, with EBRs locating in CNE-sparse regions (see Chapter 3), and TEs and genes that were found directly associated with EBRs (Murphy et al., 2005, Groenen et al., 2012, Ma et al., 2006, Farré et al., 2016). To detect if these DNA sequence features correlate with the distribution of EBRs in Avian ancestor chromosomes, we tested the association between avian CNEs, and chicken TEs and gene content, with the previously calculated EBR density measurements. We chose the chicken annotations for these comparisons as they are more exhaustive than those of the zebra finch.

We observed a moderate inverse correlation between the fraction of bases within CNEs for each chromosome, and both their EBR density (EBR/Mbp) and the difference between observed and expected number of EBRs (p-value ≤0.01; r = -0.62 and -0.47, respectively; Figure 4-3). The opposite trend was observed for the average distance between EBRs, which presented a direct association with the fraction of bases in CNEs on Avian ancestor chromosomes (p-value = 0.005; r = 0.53; Figure 4-3). We also noticed a direct association between the fraction of bases within genes and both EBR density (p-value = 0.05; r = 0.35; Figure 4-3) and the difference between observed and expected number of EBRs (p-value = 0.04; r = 0.38; Figure 4-3). The average distance between EBRs shows an inverse association (p-value = 0.005; r = -0.53; Figure 4-3) with the fraction of bases within genes on Avian ancestor chromosomes. No association was found between EBR distribution and the fraction of bases within chicken TEs (Figure 4-3).

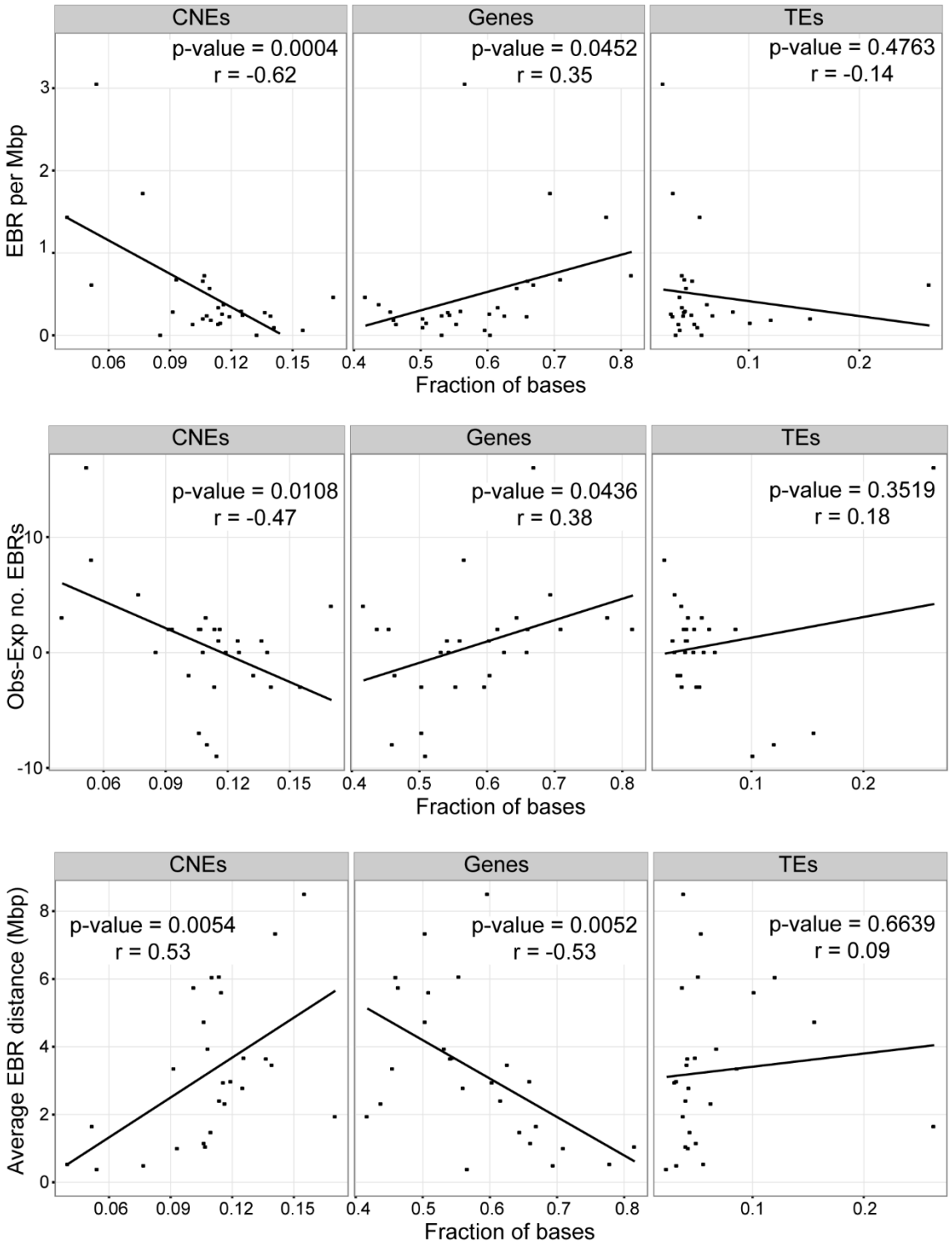

Figure 4-3: Correlation between the fraction of bases within CNEs, TEs and genes, and EBRs rates (EBRs per Mbp, observed-expected number of EBRs, and average EBR distance) for Avian ancestor chromosomes. The blue line shows linear correlation, and r and p-values show the Pearson correlation.

Types of chromosomal rearrangements

We used GRIMM webserver (Tesler, 2002) to detect types and number of chromosomal rearrangements that have occurred during the avian genome evolution, by comparing Avian, Eufalconimorphae and Passeriformes ancestors, and chicken and zebra finch genomes. The Eufalconimorphae ancestor was chosen for this comparison as its descendant species are mainly assembled to chromosomes or high continuity scaffold-level assemblies (6 out of 10; N50 > 9Mbp), resulting in the least fragmented ancestral genome (46 RACFs; Table 4-3). Passeriformes ancestor was also used in this comparison, as it is the ancestor of the most speciose avian clade, with species expressing a high phenotypical diversity. Prior to the GRIMM analysis, both Eufalconimorphae and Passeriformes ancestors' chromosomes were reconstructed from RACFs using the same approach as for the Avian ancestor described above.

Only one interchromosomal rearrangement was observed, which corresponds to the fission of the Avian ancestor chromosome 1 to form Eufalconimorphae ancestor chromosomes 1 and 1A. The remaining rearrangements were chromosomal inversions. We observed an increasing rate of chromosomal inversions (number of inversions per MY) for the ancestors phylogenetically closer to the zebra finch. From the Avian to the Eufalconimorphae ancestor the rate of inversions was calculated as 0.77 inversions/MY, from Eufalconimorphae to Passeriformes this rate increased to 1.64 inversions/MY and from Passeriformes to zebra finch we observed the highest rate of 2.58 inversions/MY. The high rearrangement rate for this last branch was previously proposed to relate with the larger radiations this clade experienced relative to the other bird groups (Zhang et al., 2014b).

The number of inversions detected between the five largest Avian ancestor chromosomes (1 to 5) and its zebra finch homeologues is highly consistent between our analysis (N=59) and that reported by Romanov and colleagues (N=54; Romanov et al., 2014). The same does not hold for the chicken. In this case, our reconstruction allows the detection of twice as many inversions as those reported in (Romanov et al., 2014; 43 versus 22). This inconsistency might relate to an overrepresentation of Galloanserae genomes in Romanov's work (3 out of 6 genomes), which might have led to a bias of the reconstructed

chromosomes towards Galloanserae genome structures. In fact, it is reported that Avian ancestor chromosomes 5 and 11 do not show any change relative to their chicken homeologues (Romanov et al., 2014). In our analysis, however, chicken chromosomes 5 and 11 contain 4 and 3 inversions, respectively, relative to the Avian ancestor chromosomes.

Gene ontology enrichment analysis in Avian ancestor chromosomes

In both RACF based and chromosome based analysis, we observed that Avian ancestor chromosomes 17 and 22 were maintained intact during the ~100MY evolution of avian genomes in the lineage leading to zebra finch. To identify if there are functional categories of genes associated with these two chromosomes we performed gene ontology (GO) enrichment analysis in these intact Avian ancestor chromosomes. We observed an enrichment for genes related to *cellular process*, *metabolic process* and *biological regulation* (274 out of 321 genes; Figure 4-4).

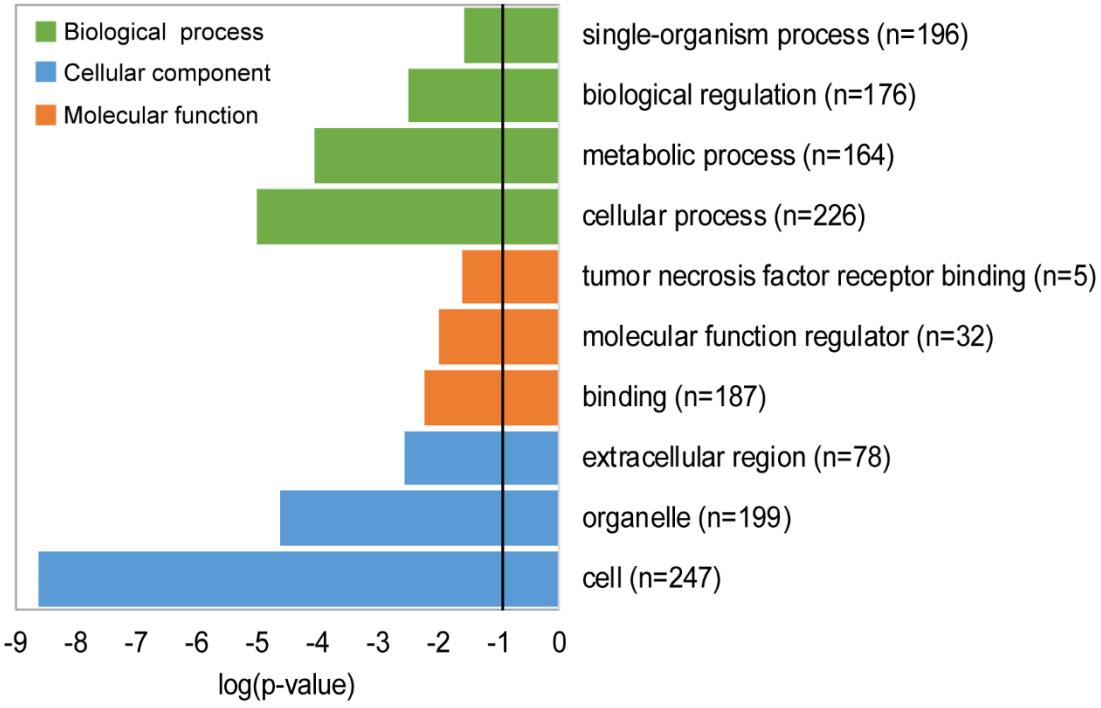

Figure 4-4: GO terms enriched on Avian ancestor chromosomes 17 and 22 (p-value <0.05; FDR <5%). Number of genes in each term are given in parenthesis. Black line depicts p-value threshold.

4.4 Discussion

Herein, using a combination of chromosome- and scaffold-level genome assemblies, we performed the reconstruction of the most likely chromosome structure of 14 avian ancestors leading to the zebra finch. The obtained RACFs showed a high coverage of the reference genome (zebra finch) and allowed the detection of structural differences, and their flanking EBRs, across the avian phylogenetic tree.

Contrary to previous studies, from which the proposed ancestral avian karyotypes included only macrochromosomes (Griffin et al., 2007, Romanov et al., 2014), the ancestor karyotypes generated herein include microchromosomes homeologous to zebra finch chromosomes 11 to 28 (except 16 and 25). This inclusion provided new important insights into the evolution of avian microchromosomes. Microchromosomes tend to have higher density of EBRs than macrochromosomes which might relate to their higher recombination rates (twice as high as in macrochromosomes and five times higher than in mammalian chromosomes) which, in its turn, could relate to the requirement of at least one chiasma point for their correct segregation during cell division (Burt, 2002, Rodionov et al., 1992). Moreover, microchromosomes are gene-rich, and it is also possible that the nature of transcription and the chromatin structure make these chromosomes more susceptible to breakage and non-allelic homologous recombination (Zody et al., 2006, Lemaitre et al., 2009). Nonetheless, we noted that two Avian ancestor microchromosomes (17 and 22) were maintained intact during the ~100 MY of avian evolution. Many genes located in these chromosomes have important roles in embryonic development and controlling cell cycle progression. For instance, *NKX2-6* is known to play a role in cardiovascular development in vertebrates (Tanaka et al., 1998) and *NODAL* plays a part in the left-right axis formation in the chicken embryo (Schier and Shen, 2000). In fact, a similar trend was previously observed in mammals where msHSBs were found enriched for genes related to the development of the central nervous system, bone and blood vessels (Larkin et al., 2009). The location of many genes essential for development in these intact chromosomes might explain why synteny was maintained during avian evolution, as changes in the

organisation of these genomic regions could disturb gene regulation and have deleterious effects.

EBRs were previously found associated (directly or inversely) with various DNA sequence features. Multiple studies showed that EBRs tend to locate in genomic regions rich in TEs, rich in genes, and sparse in CNEs (Chapter 3; Murphy et al., 2005, Groenen et al., 2012, Ma et al., 2006, Farré et al., 2016). In this work, we observed that the distribution of EBRs on the Avian ancestor chromosomes was inversely correlated with their density in CNEs, suggesting that CNE density might play a role in the fixation of chromosome rearrangements. Indeed, many CNEs are known to be regulatory elements (Woolfe et al., 2004), so EBRs located in regions with high CNE content would have a higher probability of disrupting regulatory pathways, reducing their chances of being fixed in evolution. In fact, Farré and colleagues showed that CNEs located near development-related genes, on avian msHSBs, might contain transcription factor binding sites, and the presence of these regulatory pathways might have contributed to the stability of these genomic regions during avian evolution (Farré et al., 2016).

Besides the variability in EBR densities observed between chromosomes, we also noted a high variability of rearrangement rates between phylogenetic clades, in agreement with what was reported in (Zhang et al., 2014b). We found only two disagreements between the differences from the average rearrangement rate reported herein and those reported by Zhang and colleagues (Zhang et al., 2014b). One relates to the Passeriformes to Passeri branch, which was previously reported to have a rearrangement rate lower than the average of all avian branches and we found it to be significantly higher. This inconsistency might be related to the higher number of chromosome assemblies and higher continuity of scaffold-level assemblies present in our descendant genomes set, which facilitated the detection of rearrangements unidentified in the previous work. In addition, the branch between the Avian and the Neognathae ancestors is also inconsistent between ours and Zhang's data. In this case, Zhang and colleagues reported this branch as presenting a rearrangement rate higher than the average while we observed a lower than average rearrangement rate. This disagreement is probably related to the fact that most of the established Avian ancestor RACF adjacencies were supported (spanned) by Neognathae ancestor

RACFs, which we assumed did not contain rearrangements, in this way, our data set might have an underrepresentation of Neognathae EBRs. Both these disagreements show the importance of having genomes assembled to chromosome-level. These genomes would allow the reconstruction of less fragmented ancestral genomes, reducing of even removing the requirement for manual ordering and merging of RACFs, and leading to the identification of the full collection of events that shaped extant genomes.

Despite the utility of the predicted ancestor genome structures to better understand avian chromosome evolution, they are not free of limitations. Due to the use of only one reference genome to define SFs, it is possible that some ancestral sequences that are not present in the reference genome, zebra finch in our case, were omitted from the reconstructions. Moreover, the predominance of scaffold-level assemblies in the descendant species result in fragmented predicted ancestral genomes. Indeed, we observed a lower number of reconstructed RACFs for those ancestors to which both sides of the bifurcation contained chromosome-level assemblies or scaffold-level assemblies with N50 >9 Mbp, which reinforces the importance of having high continuity genome assemblies to facilitate the study of chromosome evolution. Moreover, as mentioned above, this fragmentation may lead to an underrepresentation of the number of structural differences between the ancestors, as putative EBRs between RACFs will not be accounted, and can also result in the misclassification of the EBRs.

The reconstruction of ancestral karyotypes is essential for the detection of the full catalogue of events that shaped extant genomes, their time of occurrence and their implications on the biology of species. In this chapter, the reconstruction of the Avian ancestor karyotype offered valuable insights into the mechanisms behind the generation of chromosome rearrangements and their fixation in the avian lineage. Nonetheless, there are still many questions to answer. For instance, the mechanisms behind the different propensity of some avian lineages to generate or fix interchromosomal changes (e.g., birds of prey and penguins) are still a matter of speculation and require further investigation.

5 General discussion

5.1 Upgrading fragmented genome assemblies

The contributions this thesis has made to the identification of novel patterns of genome evolution in birds was only possible through the improvement of the existing avian genome assemblies. The genome assembly methodology developed herein is an important and inexpensive tool to upgrade fragmented (e.g., next-generation sequencing (NGS)) genome assemblies to the chromosome-level. Our approach tackles the inability of NGS methodologies to generate long error-free contigs or scaffolds, and the lack of inexpensive mapping technologies to upgrade NGS genomes to complete chromosome-level. Moreover, it allows the production of chromosome-level genome assemblies that are comparable, in continuity and reliability, to those generated using traditional approaches (e.g., with the assistance of radiation hybrid or optical maps) while having significantly lower cost and time requirements.

This methodology can be easily transferred to other species. The existence of chromosome-level assemblies for many vertebrate phylogenetic groups allows the approximation to near chromosome-scale of NGS genome assemblies, using the reference-assisted chromosome assembly (RACA) algorithm (Kim et al., 2013), provided these NGS genomes were generated using paired-end and mate pair read libraries. The generation of predicted chromosome fragments (PCFs) not only increases genome continuity but also serves to identify likely problematic fragments in the original genome assemblies (i.e. chimeric scaffolds). The BAC selection methodology can be applied to any phylogenetic clade provided BAC clone libraries are available and end-sequenced, and metaphase chromosomes could be obtained for FISH. This task could be, however, more challenging for other clades (e.g., mammals, amphibians or plants). For instance, a higher repetitive genome content and lower degree of genome conservation could reduce the number of selected probes, either because they would have limited specificity or hybridization success, respectively; the existence of multiple duplicated genomic regions within the target genomes (that in many cases present high genetic diversity), could limit the mapping accuracy; and, longer genomes would require a larger number of probes to achieve the same level of resolution obtained for avian genomes.

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 contigs or super-scaffolds that could be directly anchored to chromosomes. Nonetheless, these newer technologies still suffer from multiple limitations. They are still relatively expensive, require large amounts of high molecular weight DNA, which might be difficult to obtain for many species, and generate super-scaffolds that span, at most, chromosome arms. Therefore, the generated super-scaffold assemblies require further verification and mapping to chromosomes to produce genome assemblies at the chromosome-level. Our genome assembly approach can also tackle these limitations. On the one hand, it can be used to improve the genome assemblies for species to which re-sequencing with third-generation methodologies would not be possible, either because of DNA unavailability or cost limitations. On the other hand, it can simplify the verification and anchoring of super-scaffolds to the species chromosomes, reducing the requirements usually associated with this task.

5.2 Are avian genomes stable?

Avian genomes are the smallest among amniotes (Scanes, 2014). Their compact nature results from lower repetitive DNA content, shorter genes and non-coding regions, and the loss of paralogs (Zhang et al., 2014b). Another striking characteristic of avian genomes is the stability of the number of chromosomes in avian karyotypes. In fact, more than 60% of avian species karyotypes contain ~80 chromosomes (Christidis, 1990, Griffin et al., 2007) suggesting that interchromosomal rearrangements were extremely rare during avian evolution. Exceptions to this stability are limited to few avian lineages, such as penguins, birds of prey (i.e. falcons and eagles), and parrots (Griffin et al., 2007, Nishida et al., 2008, De Oliveira et al., 2005, Nanda et al., 2007). This karyotipical stability clearly distinguishes birds from other clades, such as mammals and lizards, where interchromosomal rearrangements are relatively common (Graphodatsky et al., 2011, Organ et al., 2008, Carvalho et al., 2015). Interestingly, unlike interchromosomal events, intrachromosomal rearrangements (e.g., chromosomal inversions) were frequent in avian evolution and are thought to be important contributors for the phenotypic diversity observed within this Class

(Völker et al., 2010, Skinner and Griffin, 2012, Zhang et al., 2014b). Indeed, the average rate of rearrangements during avian evolution detected in this study (~2 EBRs/MY from the Avian ancestor to the zebra finch; Chapter 4) is slightly higher than that observed in mammals (~1.5 EBRs/MY from the Eutherian ancestor to the human; Kim et al., 2017). These evolutionary rates, coupled with the smaller genome sizes in birds compared to mammals, show that despite their karyotypical stability, avian genomes have a higher density of genome rearrangements than, for instance, mammals.

5.3 Avian genome evolution: the role of repetitive sequences

Karyotype differences between species arise from DNA aberrations in germ cells that were fixed during evolution. These usually result from the erroneous repair of double-strand breaks (DSBs) either by the direct joining of incorrect DSBs or by recombination of non-allelic homologous sequences (Schubert and Lysak, 2011, Branco and Pombo, 2006). Repetitive DNA sequences, such as transposable elements (TEs), segmental duplications and tandem repeats, are often used as a template for non-allelic homologous recombination (NAHR) and because of that are considered one of the largest contributors to genome evolution in eukaryotes (Lynch, 2007, Gregory, 2005, Wessler, 2006). Indeed, EBRs are associated with repetitive sequences in many animal groups. Lineage-specific EBRs were previously found enriched for segmental duplications, tandem repeats and long terminal repeats (LTRs) in, for instance, mammals (Murphy et al., 2005, Groenen et al., 2012, Farré et al., 2011), yeasts (Chan and Kolodner, 2011) and *Drosophila* (Puerma et al., 2016). Also in birds lineage-specific EBRs are usually enriched in LTRs (Chapter 2 and 3; Farré et al., 2016, Skinner and Griffin, 2012), and songbird's (e.g., zebra finch) genomes show both an expansion of LTR elements (Kapusta and Suh, 2017, Zhang et al., 2014b) and higher rearrangement rates than other avian clades. Altogether, these data suggest that NAHR might have an important role generating chromosomal changes, particularly intrachromosomal rearrangements, in avian genomes.

The repeat-poor nature of avian genomes was previously hypothesised to contribute to the maintenance of independent chromosomes after breakage, due to a lack of templates for chromosome merging by NAHR (Burt, 2002). This fact could easily explain the occurrence of chromosomal fissions in the avian lineage.

Nonetheless, the reason behind the predominance of chromosomal fusions in the peregrine falcon, and other species with highly rearranged karyotypes (e.g., parrots and eagles; Nanda et al., 2007, De Oliveira et al., 2005) are still to unravel. It is worth mentioning that peregrine falcon and woodpecker EBRs flanking chromosomal fusions did show an increased density of TEs, but this enrichment did not reach statistical significance. This observation raises the question if the number of analysed fusion-flanking EBRs was just too low to detect the association of avian chromosome fusions and TEs. As a matter of fact, the frequency of chromosomal fusions, for instance, in mammals and amphibians is significantly higher than in birds (Voss et al., 2011, Uno et al., 2012), correlating with the higher repeat content of their genomes, which could endorse this association. The same pattern is not clearly followed within birds where the highly rearranged genomes of falcons, penguins and parrots, do not seem to associate with higher repetitive contents. However, NGS genome sequences usually show an underrepresentation of repetitive elements that might conceal this feature. Furthermore, it is also possible that, as seen in the highly rearranged gibbon genomes (Carbone et al., 2014), the generation of these chromosomal rearrangements is associated with segmental duplications or unidentified clade-specific TE families, or even result from repair mechanisms that do not require homologous templates, such as non-homologous end joining (Moore and Haber, 1996). Therefore, the detection and analysis at the sequence level of more EBRs flanking avian interchromosomal rearrangements, particularly chromosomal fusions, will be essential for the clarification of the mechanisms behind their generation.

5.4 Avian genome evolution: advantage in maintaining synteny

A recently proposed model to explain avian genome stability hypothesises that there is an advantage in maintaining synteny (Farré et al., 2016, Skinner and Griffin, 2012). Indeed, avian multispecies (ms) homologous synteny blocks (HSBs) were found significantly enriched for avian conserved non-coding elements (CNEs) (Farré et al., 2016), which are non-coding sequences evolving slower than at the neutral substitution rate. Many CNEs are known to be regulatory elements or sites for transcription regulatory factors (Woolfe et al., 2004) and, therefore, the disruption of CNE-dense regions could have a higher

chance of disturbing gene regulation pathways. Birds have a higher fraction of their genomes within CNEs (~7%; Zhang et al., 2014b) than, for instance, mammals (~4%; Lindblad-Toh et al., 2011), which together with their smaller genome sizes might contribute to an increased probability of genome rearrangements having significant functional implications, reducing their chances of fixation. In this way, one would expect that, as EBRs are regions where synteny is broken, they would locate in genomic regions were CNE density is lower. Indeed, the fact that the distribution of EBRs flanking intrachromosomal rearrangements in the Avian ancestor chromosomes inversely correlates with the fraction of the chromosome found within CNEs (Chapter 4), and genomic regions harbouring EBRs have lower CNE density than their immediately adjacent regions (Chapter 2 and 3), support the contribution of CNEs to the maintenance of synteny in avian genomes. Moreover, the higher fraction of CNEs in EBRs flanking intrachromosomal rearrangements, when compared with other EBR types (Chapter 3) agrees with the suggestion that intrachromosomal rearrangements might have a less dramatic effect on *cis* gene regulation than chromosomal fusions or fissions (Skinner and Griffin, 2012, Romanov et al., 2014). Additionally, the enrichment for development-related genes in the two chromosomes maintained intact from the Avian ancestor until the zebra finch, coupled with the association of housekeeping genes with long-range regulation (Mongin et al., 2009, Harmston and Lenhard, 2013) also supports the hypothesis that there is an advantage to maintain synteny in some chromosomes or chromosomal intervals. In summary, there seems to be an advantage in the maintenance of synteny in avian genomes as rearrangements disrupting regions with high density of regulatory elements would have a higher probability of disturbing regulatory pathways. Nonetheless, as not all CNEs are regulatory elements, this feature requires further verification, which could be achieved through the analysis of patterns of gene regulation (e.g., enhancers or transcription factor binding sites) and their association with avian EBRs and HSBs.

5.5 Avian interchromosomal stability

As with any other mutation, differences in chromosomal rearrangement rates, between lineages, can result from either change in their rate of mutation or their

rate of fixation (Burt, 2001). Indeed, these factors might help us to understand why avian genomes appear so stable at the interchromosomal level when compared to mammals and non-avian reptiles, for example.

Generation time and the repetitive content of a species genome are believed to be significant contributors to mutation rate variability (Figure 5-1). In birds, these two factors seem to have contrasting effects. The shorter generation times of birds, relatively to other animals, could lead to a higher chance of occurrence of genome rearrangements. In other words, for the same evolutionary period, the shorter generation times would mean a greater number of undergone meiosis and associated crossovers. As crossovers require the production of DSBs that could be misrepaired, chromosomes would then have a higher chance to rearrange. In fact, generation time was previously correlated with the differences in rearrangement rates in human and mouse (Burt et al., 1999), and could, at least partially, explain why the lineages leading to the short-generation-time songbirds possess some of the highest rearrangement rates on the avian phylogenetic tree. Interestingly, the recombination rates observed in birds (1.7-2.6 cM/Mbp; Pigozzi, 2016) are higher than those of eutherian mammals (0.2-1.8 cM/Mbp; Segura et al., 2013)), which could also imply higher mutation rates in avian genomes. Contrarily, as mentioned above, the low repetitive content of avian genomes could represent a lower opportunity to change due to a dearth of templates for NAHR (Burt, 2002) and could be counteracting the effects of generation time and crossover requirements on the increase of the rearrangement rates. Another factor that could influence the rate of rearrangements would be a different stringency of the meiosis checkpoint, proofreading and/or repair mechanisms responsible for the regulation of chromosome pairing and recombination. A lower stringency of these mechanisms could result in a higher frequency of rearrangements.

Independently of the rate of mutation observed in a lineage, after a chromosomal rearrangement occurs, it will only influence the evolution of species if it is fixed. The probability of fixation of a chromosome rearrangement can increase: (a) simply by chance (i.e. genetic drift) in small or inbreed populations, (b) if one of the chromosomal variants is transmitted at higher rate to the descendants (i.e. meiotic drive), or (c) if the novel chromosome variant has selectively

advantageous implications (Burt, 2001). Nonetheless, in large, randomly mating populations chromosome rearrangements have a higher probability of being fixed if they have neutral, nearly neutral or selectively advantageous implications (Burt, 2001). In this way, the compactness of avian genomes might be one of the main contributors to their stability (Figure 5-1). This compactness was previously hypothesised to relate to the high metabolic demands of powered flight (Gregory, 2002a, Hughes and Friedman, 2008). Indeed, this theory is supported both by the smaller genome sizes of bats, when compared to other mammals (Gregory, 2005), and the genome size variation within the Class Aves. Flightless birds (e.g., ostrich) present the largest avian genomes, and hummingbirds that have the highest metabolic rates among birds present the smallest (Wright et al., 2014). Nonetheless, having compact genomes with higher gene and regulatory elements density, together with a lower number of gene family members can also have some drawbacks. One of them would be a lower tolerance to rearrangement. That is, the packed nature of avian genomes might increase the chances of a chromosome rearrangement having significant functional implications that would not be tolerated by selection. Therefore, this lineage would have a reduced rate of fixation of novel chromosome variants. In fact, chromosomes with a higher density of CNEs tend to have a lower density of EBRs (Chapter 4), supporting the hypothesis that the disturbance of regulatory pathways may play a major role in the fixation of chromosome rearrangements in birds (Farré et al., 2016). According to this theory, we would expect that the fixation of interchromosomal rearrangements would be further restricted because, as mentioned above, these are believed to have more drastic effects on *cis* gene regulation (Skinner and Griffin, 2012, Romanov et al., 2014). Indeed, interchromosomal changes were rarely fixed in avian genomes, and are limited to few avian lineages (Chapter 3 and 4; Griffin et al., 2007). Additionally, we also observe that larger genomes, with longer non-coding regions, lower gene and CNE density, and higher repetitive content (e.g., those of mammals and non-avian reptiles) tend to have higher rates of interchromosomal rearrangements, endorsing this hypothesis. These larger genomes might also have a higher genomic plasticity. On the one hand, TEs and segmental duplications are known to play a role in the generation of chromosome rearrangements generating genetic diversity. On the other, the amplification of transposons and the

generation of segmental duplications were found associated with the birth of novel genes and regulatory sequences, and the increase of members of gene families. These could evolve to perform new or slightly varied functions, which could be relevant for species adaptation (Wilson et al., 2006, Newman et al., 2005, Piacentini et al., 2014, Canapa et al., 2015). Indeed, this might be one of the reasons why non-flying organisms present larger genomes.

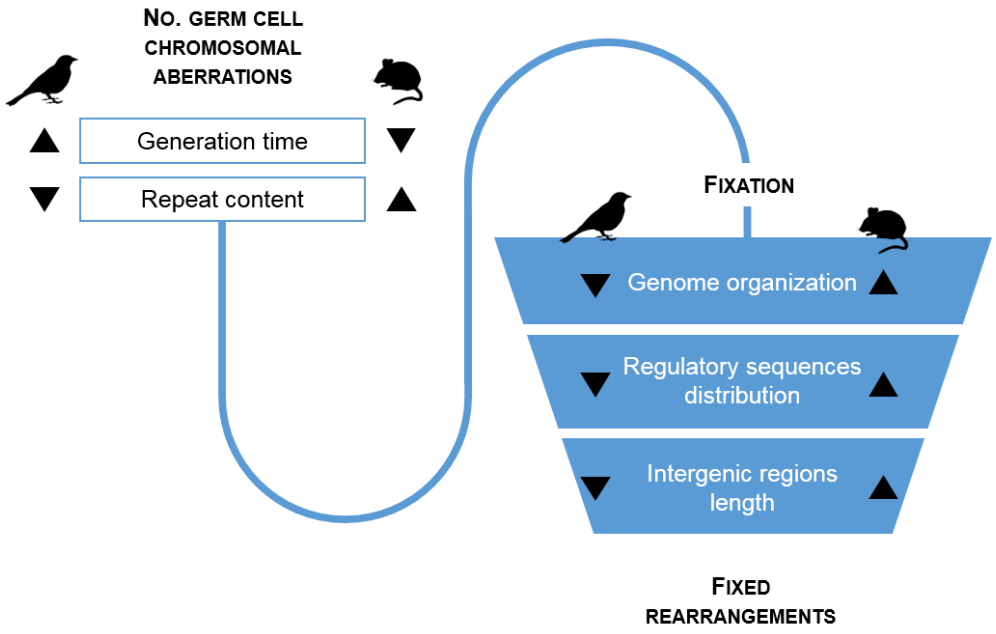

Figure 5-1: Factors contributing to the generation of chromosomal aberrations in germ cells and their fixation in the avian and mammalian lineages. ▲ and ▼depict a positive or negative contributions, respectively.

5.6 Future directions

The data reported in this thesis, together with previous works (Farré et al., 2016, Skinner and Griffin, 2012) suggest that the overall stability of avian genomes, and the low frequency of fixed interchromosomal changes, might relate to the higher chance of chromosomal rearrangements to have significant functional implications. Still, until we fully understand the forces that shaped avian genomes, there is a long way to go. For example, the reasons behind the higher propensity of some avian lineages to fix interchromosomal rearrangements (e.g., birds of prey and penguins) are still a matter of future investigations. The generation, herein, of the first chromosome-level assembly for an atypical avian karyotype provided the first insights into the genomic features differentiating intra- and interchromosomal changes, and the mechanisms that might be responsible for

their generation. However, the number of chromosomal fusions and fissions analysed so far was very low, limiting the power of the performed analysis. Therefore, the study of avian chromosome evolution would significantly benefit from the further availability of chromosome-level assemblies of species with atypical avian karyotypes. These chromosome assemblies will be a valuable resource for the detection of novel interchromosomal rearrangements fixed during avian evolution, as well as, the more accurate dating of those already identified. This data on their turn would be a useful resource for the study of the implications of chromosomal rearrangements on the biology of the species.

The availability of extra chromosome-level assemblies for birds would also be essential for a more accurate reconstruction of avian ancestral karyotypes. As we observed in Chapter 4, the continuity of the descendant and, to a lesser extent, outgroup genome assemblies, used for each reconstruction have a direct influence in the fragmentation of the predicted ancestral karyotypes. Therefore, the utilisation of less fragmented genome assemblies, ideally chromosome-level assemblies, would not only result in less fragmented reconstructed ancestral genomes but also increase the accuracy of the subsequent analyses. That is, one would have access to a more comprehensive catalogue of the events that shaped extant avian genomes and more accurately date the occurrence of such events on the evolutionary tree.

Additionally, the role of chromatin conformation on the generation of chromosome rearrangements in birds is also a matter for further investigation. Recently, Berthelot and colleagues showed that the location of EBRs, on five mammalian and three yeast genomes, could be easily explained as misrepaired breaks between open chromatin regions, which were brought into contact by the three-dimensional conformation of chromosomes in the nucleus (Berthelot et al., 2015). The direct correlation between EBR density and gene content in the Avian ancestor chromosomes (Chapter 4) could support this hypothesis. In fact, Avian ancestor microchromosomes have the highest EBR density, which could either be associated with their also higher recombination rates or with a higher breakage susceptibility related to the nature of transcription and chromatin conformation of these gene-rich chromosomes. Moreover, it is known that macro- and microchromosomes occupy distinct territories in the nucleus (Habermann et al.,

2001), with microchromosomes clustered in the centre and macrochromosomes locating near the nuclear periphery. In this way, chromatin conformation might also account for the predominance of fusions between microchromosomes on highly rearranged avian karyotypes (Nishida et al., 2008, Nanda et al., 2007, De Oliveira et al., 2005), since to be joined DSBs need to be physically close. Moreover, while the radial organisation of the chromosomes in the nucleus seems well established, the side-by-side arrangement of chromosomes is highly variable (Habermann et al., 2001), and could explain why different microchromosomes combinations appear fused in different avian lineages. The generation of chromatin conformation data for avian species will clarify its association with EBRs and give further insights into the mechanisms that govern avian genomes evolution.

6 Bibliography

Alfoldi, J., Di Palma, F., Grabherr, M., Williams, C., et al. (2011) 'The genome of the green anole lizard and a comparative analysis with birds and mammals', *Nature,* 477(7366), pp. 587-591.

Alkan, C., Sajjadian, S. and Eichler, E. E. (2011) 'Limitations of next-generation genome sequence assembly', *Nature Methods,* 8(1), pp. 61-65.

Ananthapur, V., Avvari, S., Madireddi, S., Nallari, P., et al. (2012) 'A Dysmorphic Child with a Pericentric Inversion of Chromosome 8', *Case Reports in Pediatrics,* 2012, pp. 3.

Andrews, S. (2010) *FastQC: a quality control tool for high throughput sequence data*. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Armour, J. A. (2006) 'Tandemly repeated DNA: why should anyone care?', *Mutation Research,* 598(1-2), pp. 6-14.

Baer, C. F., Miyamoto, M. M. and Denver, D. R. (2007) 'Mutation rate variation in multicellular eukaryotes: causes and consequences', *Nature Reviews Genetics,* 8(8), pp. 619-31.

Bailey, J. A., Baertsch, R., Kent, W. J., Haussler, D., et al. (2004) 'Hotspots of mammalian chromosomal evolution', *Genome Biology,* 5(4), pp. R23.

Bailey, J. A. and Eichler, E. E. (2006) 'Primate segmental duplications: crucibles of evolution, diversity and disease', *Nature Reviews Genetics,* 7(7), pp. 552-564.

Baudet, C., Lemaitre, C., Dias, Z., Gautier, C., et al. (2010) 'Cassis: detection of genomic rearrangement breakpoints', *Bioinformatics,* 26(15), pp. 1897-1898.

Baxevanis, A. D. and Ouellette, B. F. F. (2004) *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins.* Wiley.

Beckmann, J. S., Estivill, X. and Antonarakis, S. E. (2007) 'Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability', *Nature Reviews Genetics,* 8(8), pp. 639-646.

Behjati, S. and Tarpey, P. S. (2013) 'What is next generation sequencing?', Archives of Disease in Childhood. Education and Practice Edition, 98(6), pp. 236-238.

Beklemisheva, V. R., Perelman, P. L., Lemskaya, N. A., Kulemzina, A. I., et al. (2016) 'The Ancestral Carnivore Karyotype As Substantiated by Comparative Chromosome Painting of Three Pinnipeds, the Walrus, the Steller Sea Lion and the Baikal Seal (Pinnipedia, Carnivora)', *PLoS ONE,* 11(1), pp. e0147647.

Bentley, D. R. and Balasubramanian, S. and Swerdlow, H. P. and Smith, G. P., et al. (2008) 'Accurate Whole Human Genome Sequencing using Reversible Terminator Chemistry', *Nature,* 456(7218), pp. 53-59.

Berthelot, C., Muffato, M., Abecassis, J. and Roest Crollius, H. (2015) 'The 3D organization of chromatin explains evolutionary fragile genomic regions', *Cell Reports,* 10(11), pp. 1913-1924.

Biemont, C. and Vieira, C. (2006) 'Genetics: Junk DNA as an evolutionary force', *Nature,* 443(7111), pp. 521-524.

Birkhead, T. (2012) *Bird Sense: What It's Like to Be a Bird.* Bloomsbury Publishing.

Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., et al. (2004) 'Aligning multiple genomic sequences with the threaded blockset aligner', *Genome Research,* 14(4), pp. 708-715.

Bourque, G., Zdobnov, E. M., Bork, P., Pevzner, P. A., et al. (2005) 'Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages', *Genome Research,* 15(1), pp. 98-110.

Boyle, E. I., Weng, S., Gollub, J., Jin, H., et al. (2004) 'GO::TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes', *Bioinformatics,* 20(18), pp. 3710-3715.

Bragg, L. M., Stone, G., Butler, M. K., Hugenholtz, P., et al. (2013) 'Shining a Light on Dark Sequencing: Characterising Errors in Ion Torrent PGM Data', *PLoS Computational Biology,* 9(4), pp. e1003031.

Branco, M. R. and Pombo, A. (2006) 'Intermingling of Chromosome Territories in Interphase Suggests Role in Translocations and Transcription-Dependent Associations', *PLoS Biology,* 4(5), pp. e138.

Brown, T. A. (2007) *Genomes 3.* Garland Science Pub.

Burt, D. W. (2001) 'Chromosome Rearrangement in Evolution', *Encyclopedia of Life Sciences*: John Wiley & Sons, Ltd.

Burt, D. W. (2002) 'Origin and evolution of avian microchromosomes', *Cytogenetic and Genome Research,* 96, pp. 97-112.

Burt, D. W., Bruley, C., Dunn, I. C., Jones, C. T., et al. (1999) 'The dynamics of chromosome evolution in birds and mammals', *Nature,* 402, pp. 411-413.

Burton, J. N., Adey, A., Patwardhan, R. P., Qiu, R., et al. (2013) 'Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions', *Nature Biotechnology,* 31(12), pp. 1119-1125.

Cai, Q., Qian, X., Lang, Y., Luo, Y., et al. (2013) 'Genome sequence of ground tit Pseudopodoces humilis and its adaptation to high altitude', *Genome Biology,* 14(3), pp. R29.

Canapa, A., Barucca, M., Biscotti, M. A., Forconi, M., et al. (2015) 'Transposons, Genome Size, and Evolutionary Insights in Animals', *Cytogenetic and Genome Research,* 147(4), pp. 217-239.

Cao, H., Hastie, A. R., Cao, D., Lam, E. T., et al. (2014) 'Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology', *GigaScience,* 3(1), pp. 34.

Capilla Pérez, L. (2015) *On the role of chromosomal rearrangements in evolution: Reconstruction of genome reshuffling in rodents and analysis of Robertsonian fusions in a house mouse chromosomal polymorphism zone.* Ph.D., Universitat Autònoma de Barcelona [Online] Available at: http://hdl.handle.net/10803/314185.

Carbone, L., Harris, R. A., Gnerre, S., Veeramah, K. R., et al. (2014) 'Gibbon genome and the fast karyotype evolution of small apes', *Nature,* 513(7517), pp. 195-201.

Carvalho, N. D. M., Arias, F. J., da Silva, F. A., Schneider, C. H., et al. (2015) 'Cytogenetic analyses of five amazon lizard species of the subfamilies Teiinae and Tupinambinae and review of karyotyped diversity the family Teiidae', *Comparative Cytogenetics,* 9(4).

Chan, J. E. and Kolodner, R. D. (2011) 'A Genetic and Structural Study of Genome Rearrangements Mediated by High Copy Repeat Ty1 Elements', *PLoS Genetics,* 7(5), pp. e1002089.

Chaves, J. A., Cooper, E. A., Hendry, A. P., Podos, J., et al. (2016) 'Genomic variation at the tips of the adaptive radiation of Darwin's finches', *Molecular Ecology,* 25(21), pp. 5282-5295.

Chen, A. Y. and Chen, A. (2013) 'Fluorescence in situ hybridization', *Journal of Investigative Dermatology,* 133(5), pp. e8.

Chen, C.-Y. (2014) 'DNA polymerases drive DNA sequencing-by-synthesis technologies: both past and present', *Frontiers in Microbiology,* 5, pp. 305.

Chen, S., Krinsky, B. H. and Long, M. (2013a) 'New genes as drivers of phenotypic evolution', *Nature Reviews Genetics,* 14(9), pp. 645-660.

Chen, Y.-C., Liu, T., Yu, C.-H., Chiang, T.-Y., et al. (2013b) 'Effects of GC Bias in Next-Generation-Sequencing Data on De Novo Genome Assembly', *PLoS ONE,* 8(4), pp. e62856.

Chen, Z. J., Ha, M. and Soltis, D. (2007) 'Polyploidy: genome obesity and its consequences: Polyploidy workshop: Plant and Animal Genome XV Conference, San Diego, CA, USA, January 2007', *The New phytologist,* 174(4), pp. 717-720.

Christidis, L. (1990) 'Aves', in John, B., Kayano, H. & Levan, A. (eds.) *Animal Cytogenetics. Volume 4: Chordata 3 B*. Berlin: Gebrüder Borntraeger.

Chuong, E. B., Rumi, M. A., Soares, M. J. and Baker, J. C. (2013) 'Endogenous retroviruses function as species-specific enhancer elements in the placenta', *Nature Genetics,* 45(3), pp. 325-329.

Clarke, J., Wu, H.-C., Jayasinghe, L., Patel, A., et al. (2009) 'Continuous base identification for single-molecule nanopore DNA sequencing', *Nature Nanotechnology,* 4(4), pp. 265-270.

Collins, F. S., Green, E. D., Guttmacher, A. E. and Guyer, M. S. (2003) 'A vision for the future of genomics research', *Nature,* 422(6934), pp. 835-847.

Coyle, S. and Kroll, E. (2008) 'Starvation Induces Genomic Rearrangements and Starvation-Resilient Phenotypes in Yeast', *Molecular Biology and Evolution,* 25(2), pp. 310-318.

Crosland, M. W. J. and Crozier, R. H. (1986) 'Myrmecia pilosula an Ant with Only One Pair of Chromosomes', *Science,* 231(4743), pp. 1278.

Dalloul, R. A., Long, J. A., Zimin, A. V., Aslam, L., et al. (2010) 'Multi-platform next-generation sequencing of the domestic turkey (Meleagris gallopavo): genome assembly and analysis', *PLoS Biology,* 8(9).

Davis, J. K., Mittel, L. B., Lowman, J. J., Thomas, P. J., et al. (2011) 'Haplotype-based genomic sequencing of a chromosomal polymorphism in the white-throated sparrow (Zonotrichia albicollis)', *The Journal of Heredity,* 102(4), pp. 380-90.

De Oliveira, E. H., Habermann, F. A., Lacerda, O., Sbalqueiro, I. J., et al. (2005) 'Chromosome reshuffling in birds of prey: the karyotype of the world's largest eagle (Harpy eagle, Harpia harpyja) compared to that of the chicken (Gallus gallus)', *Chromosoma,* 114(5), pp. 338-343.

Deakin, J. E. and Ezaz, T. (2014) 'Tracing the evolution of amniote chromosomes', *Chromosoma,* 123(3), pp. 201-216.

Deamer, D., Akeson, M. and Branton, D. (2016) 'Three decades of nanopore sequencing', *Nature Biotechnology,* 34(5), pp. 518-524.

Delany, M. E., Gessaro, T. M., Rodrigue, K. L. and Daniels, L. M. (2007) 'Chromosomal mapping of chicken mega-telomere arrays to GGA9, 16, 28 and W using a cytogenomic approach', *Cytogenetic and Genome Research,* 117(1-4), pp. 54-63.

Denton, J. F., Lugo-Martinez, J., Tucker, A. E., Schrider, D. R., et al. (2014) 'Extensive error in the number of genes inferred from draft genome assemblies', *PLoS Computational Biology,* 10(12), pp. e1003998.

Derjusheva, S., Kurganova, A., Habermann, F. and Gaginskaya, E. (2004) 'High chromosome conservation detected by comparative chromosome painting in chicken, pigeon and passerine birds', *Chromosome Research,* 12(7), pp. 715-723.

Dessimoz, C., Zoller, S., Manousaki, T., Qiu, H., Meyer, A. and Kuraku S. (2011) 'Comparative genomics approach to detecting split-regions in a low-coverage genome: lesson from the chimaera Callorhinchus milii (Holocephali, Condrichthyes)', *Briefings in Bioinformatics*, 12(5), pp.474-484

Devos, K. M., Brown, J. K. M. and Bennetzen, J. L. (2002) 'Genome Size Reduction through Illegitimate Recombination Counteracts Genome Expansion in Arabidopsis', *Genome Research,* 12(7), pp. 1075-1079.

Dong, Y., Xie, M., Jiang, Y., Xiao, N., et al. (2013) 'Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (Capra hircus)', *Nature Biotechnology,* 31(2), pp. 135-141.

Doyle, J. M., Katzner, T. E., Bloom, P. H., Ji, Y., et al. (2014) 'The genome sequence of a widespread apex predator, the golden eagle (Aquila chrysaetos)', *PLoS One,* 9(4), pp. e95599.

Driscoll, C. A., Macdonald, D. W. and O'Brien, S. J. (2009) 'From wild animals to domestic pets, an evolutionary view of domestication', *Proceedings of the National Academy of Sciences of the United States of America,* 106(Suppl 1), pp. 9971-9978.

Drmanac, R., Sparks, A. B., Callow, M. J., Halpern, A. L., et al. (2010) 'Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays', *Science,* 327(5961), pp. 78-81.

Du, X., Gertz, E. M., Wojtowicz, D., Zhabinskaya, D., et al. (2014a) 'Potential non-B DNA regions in the human genome are associated with higher rates of nucleotide mutation and expression variation', *Nucleic Acids Research,* 42(20), pp. 12367-12379.

Du, X., Servin, B., Womack, J. E., Cao, J., et al. (2014b) 'An update of the goat genome assembly using dense radiation hybrid maps allows detailed analysis of evolutionary rearrangements in Bovidae', *BMC Genomics,* 15(1), pp. 625.

Dunham, M. J., Badrane, H., Ferea, T., Adams, J., et al. (2002) 'Characteristic genome rearrangements in experimental evolution of Saccharomyces cerevisiae', *Proceedings of the National Academy of Sciences of the United States of America,* 99(25), pp. 16144-16149.

Eid, J., Fehr, A., Gray, J., Luong, K., et al. (2009) 'Real-Time DNA Sequencing from Single Polymerase Molecules', *Science,* 323(5910), pp. 133-138.

Ekblom, R. and Wolf, J. B. W. (2014) 'A field guide to whole-genome sequencing, assembly and annotation', Evolutionary Applications, 7(9), pp. 1026-1042.

Ellegren, H. (2010) 'Evolutionary stasis: the stable chromosomes of birds', *Trends in Ecology & Evolution,* 25(5), pp. 283-291.

Ellegren, H., Smeds, L., Burri, R., Olason, P. I., et al. (2012) 'The genomic landscape of species divergence in Ficedula flycatchers', *Nature,* 491(7426), pp. 756-760.

Ellegren, H. (2014) 'Genome sequencing and population genomics in non-model organisms', Trends Ecol Evol, 29(1), pp. 51-63.

Elliott, T. A. and Gregory, T. R. (2015) 'What's in a genome? The C-value enigma and the evolution of eukaryotic genome content', *Philosophical*

Transaction of the Royal Society of London. Series B, Biological Sciences, 370(1678), pp. 20140331.

Elsik, C. G., Tellam, R. L., Worley, K. C., Gibbs, R. A., et al. (2009) 'The genome sequence of taurine cattle: a window to ruminant biology and evolution', *Science,* 324.

Ersfeld, K. (2004) 'Fiber-FISH: Fluorescence In Situ Hybridization on Stretched DNA', in Melville, S.E. (ed.) *Parasite Genomics Protocols*. Totowa, NJ: Humana Press, pp. 395-402.

Estevezj Sanger sequencing.

Fang, X., Nevo, E., Han, L., Levanon, E. Y., et al. (2014) 'Genome-wide adaptive complexes to underground stresses in blind mole rats Spalax', *Nature Communications,* 5, pp. 3966.

Faria, R. and Navarro, A. (2010) 'Chromosomal speciation revisited: rearranging theory with pieces of evidence', *Trends in Ecology & Evolution,* 25(11), pp. 660-669.

Farré, M., Bosch, M., López-Giráldez, F., Ponsà, M., et al. (2011) 'Assessing the Role of Tandem Repeats in Shaping the Genomic Architecture of Great Apes', *PLoS ONE,* 6(11), pp. e27239.

Farré, M., Micheletti, D. and Ruiz-Herrera, A. (2013) 'Recombination Rates and Genomic Shuffling in Human and Chimpanzee—A New Twist in the Chromosomal Speciation Theory', *Molecular Biology and Evolution,* 30(4), pp. 853-864.

Farré, M., Narayan, J., Slavov, G. T., Damas, J., et al. (2016) 'Novel insights into chromosome evolution in birds, archosaurs, and reptiles', *Genome Biology and Evolution*.

Ferguson-Lees, J. and Christie, D. A. (2005) *Raptors of the world. Princeton field guides* Princeton, N.J.: Princeton University Press.

Ferguson-Smith, M. A. and Trifonov, V. (2007) 'Mammalian karyotype evolution', *Nature Reviews Genetics,* 8(12), pp. 950-962.

Fernandes, A., Werneck, H. A., Pompolo, S. G. and Lopes, D. M. (2013) 'Evidence of separate karyotype evolutionary pathway in Euglossa orchid bees by cytogenetic analyses', *Anais da Academia Brasileira de Ciências,* 85, pp. 937-944.

Fierst, J. L. (2015) 'Using linkage maps to correct and scaffold de novo genome assemblies: Methods, challenges and computational tools', *Frontiers in Genetics,* 6.

Frankl-Vilches, C., Kuhl, H., Werber, M., Klages, S., et al. (2015) 'Using the canary genome to decipher the evolution of hormone-sensitive gene regulation in seasonal singing birds', *Genome Biology,* 16, pp. 19.

Gallardo, M. H., González, C. A. and Cebrián, I. (2006) 'Molecular cytogenetics and allotetraploidy in the red vizcacha rat, Tympanoctomys barrerae (Rodentia, Octodontidae)', *Genomics,* 88(2), pp. 214-221.

Gilbert, S. F. (2000a) *Developmental Biology.* Palgrave Macmillan.

Gilbert, S. F. (2000b) *Early development in birds.* Sunderland (MA): Sinauer Associates.

Goldman, A. D. and Landweber, L. F. (2016) 'What Is a Genome?', *PLOS Genetics,* 12(7), pp. e1006181.

Goodwin, S., McPherson, J. D. and McCombie, W. R. (2016) 'Coming of age: ten years of next-generation sequencing technologies', *Nature Reviews Genetics,* 17(6), pp. 333-351.

Gordon, D., Huddleston, J., Chaisson, M. J. P., Hill, C. M., et al. (2016) 'Long-read sequence assembly of the gorilla genome', *Science,* 352(6281).

Gould, S. J. and Vrba, E. S. (1982) 'Exaptation-A missing term in the science of form', Paleobiology, 8(1), pp. 4-15.

Grabherr, M. G., Russell, P., Meyer, M., Mauceli, E., et al. (2010) 'Genome-wide synteny through highly sensitive sequence alignment: Satsuma', *Bioinformatics,* 26(9), pp. 1145-51.

Graphodatsky, A., Ferguson-Smith, M. A. and Stanyon, R. (2012) 'A short introduction to cytogenetic studies in mammals with reference to the present volume', *Cytogenetic and Genome Research,* 137(2-4), pp. 83-96.

Graphodatsky, A. S., Trifonov, V. A. and Stanyon, R. (2011) 'The genome diversity and karyotype evolution of mammals', *Molecular Cytogenetics,* 4, pp. 22.

Green, E. D. (2001) 'Strategies for the systematic sequencing of complex genomes', *Nature Reviews Genetics,* 2(8), pp. 573-583.

Greenwold, M. J., Bao, W., Jarvis, E. D., Hu, H., et al. (2014) 'Dynamic evolution of the alpha (alpha) and beta (beta) keratins has accompanied integument diversification and the adaptation of birds into novel lifestyles', *BMC Evolutionary Biology,* 14, pp. 249.

Gregory, T. R. (2002a) 'A bird's-eye view of the c-value enigma: genome size, cell size, and metabolic rate in the class aves', *Evolution,* 56(1), pp. 121-130.

Gregory, T. R. (2002b) 'Genome size and developmental complexity', *Genetica,* 115(1), pp. 131-46.

Gregory, T. R. (2005) *The evolution of the genome.* Burlington, MA: Elsevier Academic.

Gregory, T. R., Nicol, J. A., Tamm, H., Kullman, B., et al. (2007) 'Eukaryotic genome size databases', *Nucleic Acids Research,* 35(Database issue), pp. D332-D338.

Griffin, D. K., Robertson, L. B., Tempest, H. G. and Skinner, B. M. (2007) 'The evolution of the avian genome as revealed by comparative molecular cytogenetics', *Cytogenetic and Genome Research,* 117.

Griffiths, A. J. F. (2008) *Introduction to Genetic Analysis.* W.H. Freeman and Company.

Groenen, M. A., Archibald, A. L., Uenishi, H., Tuggle, C. K., et al. (2012) 'Analyses of pig genomes provide insight into porcine demography and evolution', *Nature,* 491.

Guillén, Y. and Ruiz, A. (2012) 'Gene alterations at Drosophila inversion breakpoints provide prima facie evidence for natural selection as an explanation for rapid chromosomal evolution', *BMC Genomics,* 13(1), pp. 53.

Guttenbach, M., Nanda, I., Feichtinger, W., Masabanda, J. S., et al. (2003) 'Comparative chromosome painting of chicken autosomal paints 1–9 in nine different bird species', *Cytogenetic and Genome Research,* 103.

Habermann, F. A., Cremer, M., Walter, J., Kreth, G., et al. (2001) 'Arrangements of macro- and microchromosomes in chicken cells', *Chromosome Research,* 9(7), pp. 569-584.

Hackett, S. J., Kimball, R. T., Reddy, S., Bowie, R. C., et al. (2008) 'A phylogenomic study of birds reveals their evolutionary history', *Science,* 320(5884), pp. 1763-8.

Hall, I. M. and Quinlan, A. R. (2012) 'Detection and interpretation of genomic structural variation in mammals', *Methods in Molecular Biology,* 838, pp. 225-48.

Hansmann, T., Nanda, I., Volobouev, V., Yang, F., et al. (2009) 'Cross-species chromosome painting corroborates microchromosome fusion during karyotype evolution of birds', *Cytogenetic and Genome Research,* 126(3), pp. 281-304.

Harewood, L. and Fraser, P. (2014) 'The impact of chromosomal rearrangements on regulation of gene expression', *Human Molecular Genetics,* 23(R1), pp. R76-R82.

Harmston, N. and Lenhard, B. (2013) 'Chromatin and epigenetic features of long-range gene regulation', *Nucleic Acids Research,* 41(15), pp. 7185-7199.

Harris, R. S. (2007) *Improved pairwise alignment of genomic DNA.* Ph.D., The Pennsylvania State University

Hieter, P. and Boguski, M. (1997) 'Functional genomics: it's all how you read it', *Science,* 278(5338), pp. 601-2.

Hillier, L. (2004) 'Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution', *Nature,* 432(7018), pp. 695-716.

Hoenen, T., Groseth, A., Rosenke, K., Fischer, R. J., et al. (2016) 'Nanopore Sequencing as a Rapidly Deployable Ebola Outbreak Tool', *Emerging Infectious Diseases,* 22(2), pp. 331-4.

Hoffmann, A. A. and Rieseberg, L. H. (2008) 'Revisiting the Impact of Inversions in Evolution: From Population Genetic Markers to Drivers of Adaptive Shifts and Speciation?', *Annu Rev Ecol Evol Syst,* 39, pp. 21-42.

Howe, K. and Wood, J. M. (2015) 'Using optical mapping data for the improvement of vertebrate genome assemblies', *Gigascience,* 4, pp. 10.

Hozza, M., et al. (2015) 'How Big is that Genome? Estimating Genome Size and Coverage from k-mer Abundance Spectra', in Iliopoulos, C., Puglisi, S. & Yilmaz, E. (eds.) *String Processing and Information Retrieval: 22nd International Symposium, SPIRE 2015, London, UK, September 1-4, 2015, Proceedings*. Cham: Springer International Publishing, pp. 199-209.

Hughes, A. L. and Friedman, R. (2008) 'Genome size reduction in the chicken has involved massive loss of ancestral protein-coding genes', *Molecular Biology and Evolution,* 25(12), pp. 2681-8.

IJdo, J. W., Baldini, A., Ward, D. C., Reeders, S. T., et al. (1991) 'Origin of human chromosome 2: an ancestral telomere-telomere fusion', *Proceedings of the National Academy of Sciences,* 88(20), pp. 9051-9055.

Itoh, Y., Kampf, K., Balakrishnan, C. N. and Arnold, A. P. (2011) 'Karyotypic polymorphism of the zebra finch Z chromosome', *Chromosoma,* 120(3), pp. 255-264.

Jain, M., Fiddes, I. T., Miga, K. H., Olsen, H. E., et al. (2015) 'Improved data analysis for the MinION nanopore sequencer', *Nature Methods,* 12(4), pp. 351-356.

Jarvis, E. D. and Mirarab, S. and Aberer, A. J. and Li, B., et al. (2014) 'Whole-genome analyses resolve early branches in the tree of life of modern birds', *Science,* 346(6215), pp. 1320-1331.

Jeffery, N. W. and Gregory, T. R. (2014) 'Genome size estimates for crustaceans using Feulgen image analysis densitometry of ethanol-preserved tissues', *Cytometry Part A,* 85(10), pp. 862-868.

Jetz, W., Thomas, G. H., Joy, J. B., Hartmann, K., et al. (2012) 'The global diversity of birds in space and time', *Nature,* 491(7424), pp. 444-8.

Jones, B. R., Rajaraman, A., Tannier, E. and Chauve, C. (2012a) 'ANGES: reconstructing ANcestral GEnomeS maps', *Bioinformatics,* 28(18), pp. 2388-2390.

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., et al. (2012b) 'The genomic basis of adaptive evolution in threespine sticklebacks', *Nature,* 484(7392), pp. 55-61.

Kapusta, A. and Suh, A. (2017) 'Evolution of bird genomes—a transposon's-eye view', *Annals of the New York Academy of Sciences,* 1389(1), pp. 164-185.

Karere, G. M., Froenicke, L., Millon, L., Womack, J. E., et al. (2008) 'A high-resolution radiation hybrid map of rhesus macaque chromosome 5 identifies rearrangements in the genome assembly', *Genomics,* 92(4), pp. 210-218.

Kasai, F., O'Brien, P. C., Martin, S. and Ferguson-Smith, M. A. (2012) 'Extensive homology of chicken macrochromosomes in the karyotypes of Trachemys scripta elegans and Crocodylus niloticus revealed by chromosome painting despite long divergence times', *Cytogenetic and Genome Research,* 136(4), pp. 303-7.

Kellis, M., Wold, B., Snyder, M. P., Bernstein, B. E., et al. (2014) 'Defining functional DNA elements in the human genome', *Proceedings of the National Academy of Sciences of the United States of America,* 111(17), pp. 6131-6138.

Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W., et al. (2003) 'Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes', *Proceedings of the National Academy of Sciences,* 100(20), pp. 11484-11489.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., et al. (2002) 'The human genome browser at UCSC', *Genome Research,* 12(6), pp. 996-1006.

Kim, J., Farre, M., Auvil, L., Capitanu, B., et al. (2017) 'The Reconstruction and Evolutionary History of Eutherian Chromosomes', *Proceedings of the National Academy of Sciences*, In press.

Kim, J., Larkin, D. M., Cai, Q., Asan, et al. (2013) 'Reference-assisted chromosome assembly', *Proceedings of the National Academy of Sciences,* 110(5), pp. 1785-1790.

Kim, K. E., Peluso, P., Babayan, P., Yeadon, P. J., et al. (2014) 'Long-read, whole-genome shotgun sequence data for five model organisms', *Scientific Data,* 1, pp. 140045.

Kinsella, R. J., Kahari, A., Haider, S., Zamora, J., et al. (2011) 'Ensembl BioMarts: a hub for data retrieval across taxonomic space', *Database,* 2011, pp. bar030.

Kolmogorov, M., Armstrong, J., Raney, B. J., Streeter, I., et al. (2016) 'Chromosome assembly of large and complex genomes using multiple references', *bioRxiv*.

Kolmogorov, M., Raney, B., Paten, B. and Pham, S. (2014) 'Ragout—a reference-assisted assembly tool for bacterial genomes', *Bioinformatics,* 30(12), pp. i302-i309.

Koren, S., Harhay, G. P., Smith, T. P., Bono, J. L., et al. (2013) 'Reducing assembly complexity of microbial genomes with single-molecule sequencing', *Genome Biology,* 14(9), pp. R101.

Kulemzina, A. I., Yang, F., Trifonov, V. A., Ryder, O. A., et al. (2011) 'Chromosome painting in Tragulidae facilitates the reconstruction of Ruminantia ancestral karyotype', *Chromosome Research,* 19(4), pp. 531-9.

Kuska, B. (1998) 'Beer, Bethesda, and Biology: How “Genomics” Came Into Being', *Journal of the National Cancer Institute,* 90(2), pp. 93.

Laehnemann, D., Borkhardt, A. and McHardy, A. C. (2016) 'Denoising DNA deep sequencing data-high-throughput sequencing errors and their correction', *Briefings in Bioinformatics,* 17(1), pp. 154-79.

Lamichhaney, S., Berglund, J., Almen, M. S., Maqbool, K., et al. (2015) 'Evolution of Darwin's finches and their beaks revealed by genome sequencing', *Nature,* 518(7539), pp. 371-375.

Lander, E. S. and Linton, L. M. and Birren, B. and Nusbaum, C., et al. (2001) 'Initial sequencing and analysis of the human genome', *Nature,* 409(6822), pp. 860-921.

Langmead, B. and Salzberg, S. L. (2012) 'Fast gapped-read alignment with Bowtie 2', *Nature Methods,* 9(4), pp. 357-9.

Larkin, D. M., Astakhova, N. M., Prokhorovich, M. A., Lewin, H. A., et al. (2006) 'Comparative mapping of cattle chromosome 19: cytogenetic localization of 19 BAC clones', *Cytogenetic and Genome Research,* 112(3-4), pp. 235-40.

Larkin, D. M., Pape, G., Donthu, R., Auvil, L., et al. (2009) 'Breakpoint regions and homologous synteny blocks in chromosomes have different evolutionary histories', *Genome Research,* 19(5), pp. 770-7.

Le Duc, D., Renaud, G., Krishnan, A., Almén, M. S., et al. (2015) 'Kiwi genome provides insights into evolution of a nocturnal lifestyle', *Genome Biology,* 16(1), pp. 147.

Lee, H., Gurtowski, J., Yoo, S., Nattestad, M., et al. (2016) 'Third-generation sequencing and the future of genomics', *bioRxiv*.

Lemaitre, C., Zaghloul, L., Sagot, M.-F., Gautier, C., et al. (2009) 'Analysis of fine-scale mammalian evolutionary breakpoints provides new insight into their relation to genome organisation', *BMC Genomics,* 10(1), pp. 335.

Lemmon, A. R., Emme, S. A. and Lemmon, E. M. (2012) 'Anchored Hybrid Enrichment for Massively High-Throughput Phylogenomics', *Systematic Biology,* 61(5), pp. 727-744.

Leroux, D., Mugneret, F., Callanan, M., Radford-Weiss, I., et al. (2002) 'CD4+, CD56+ DC2 acute leukemia is characterized by recurrent clonal chromosomal changes affecting 6 major targets: a study of 21 cases by the Groupe Français de Cytogénétique Hématologique', *Blood,* 99(11), pp. 4154-4159.

Li, C., Zhang, Y., Li, J., Kong, L., et al. (2014) 'Two Antarctic penguin genomes reveal insights into their evolutionary history and molecular changes related to the Antarctic environment', *Gigascience,* 3(1), pp. 27.

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., et al. (2009) 'Comprehensive mapping of long-range interactions reveals folding principles of the human genome', *Science,* 326(5950), pp. 289-93.

Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M. F., et al. (2011) 'A high-resolution map of human evolutionary constraint using 29 mammals', *Nature,* 478(7370), pp. 476-482.

Liu, G. E., Ventura, M., Cellamare, A., Chen, L., et al. (2009) 'Analysis of recent segmental duplications in the bovine genome', *BMC Genomics,* 10(1), pp. 571.

Liu, P., Carvalho, C. M., Hastings, P. J. and Lupski, J. R. (2012) 'Mechanisms for recurrent and complex human genomic rearrangements', *Current Opinion in Genetics & Development,* 22(3), pp. 211-20.

Locke, D. P., Archidiacono, N., Misceo, D., Cardone, M. F., et al. (2003) 'Refinement of a chimpanzee pericentric inversion breakpoint to a segmental duplication cluster', *Genome Biology,* 4(8), pp. R50.

Lodish, H. F. (2016) *Molecular cell biology.* Eighth edition. edn. New York: W.H. Freeman-Macmillan Learning.

Loh, W.-Y. (2011) 'Classification and regression trees', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* 1(1), pp. 14-23.

Lohmann, K. and Klein, C. (2014) 'Next Generation Sequencing and the Future of Genetic Diagnosis', Neurotherapeutics, 11(4), pp. 699-707.

Lok, S., Paton, T. A., Wang, Z., Kaur, G., et al. (2017) 'De Novo Genome and Transcriptome Assembly of the Canadian Beaver (Castor canadensis)', *G3: Genes|Genomes|Genetics*.

Lovell, P. V., Wirthlin, M., Wilhelm, L., Minx, P., et al. (2014) 'Conserved syntenic clusters of protein coding genes are missing in birds', *Genome Biology,* 15(12), pp. 565.

Lukhtanov, V. A., Kandul, N. P., Plotkin, J. B., Dantchenko, A. V., et al. (2005) 'Reinforcement of pre-zygotic isolation and karyotype evolution in Agrodiaetus butterflies', *Nature,* 436(7049), pp. 385-389.

Lynch, M. (2007) *The origins of genome architecture.* Sunderland, Mass.: Sinauer Associates.

Ma, J., Zhang, L., Suh, B. B., Raney, B. J., et al. (2006) 'Reconstructing contiguous regions of an ancestral genome', *Genome Research,* 16(12), pp. 1557-65.

Machado, J. P., Johnson, W. E., Gilbert, M. T. P., Zhang, G., et al. (2016) 'Bone-associated gene evolution and the origin of flight in birds', *BMC Genomics,* 17, pp. 371.

Madlung, A. (2013) 'Polyploidy and its effect on evolutionary success: old questions revisited with new tools', *Heredity,* 110(2), pp. 99-104.

Mak, A. C. Y., Lai, Y. Y. Y., Lam, E. T., Kwok, T.-P., et al. (2016) 'Genome-Wide Structural Variation Detection by Genome Mapping on Nanochannel Arrays', *Genetics,* 202(1), pp. 351-362.

Mardis, E. R. (2013) 'Next-Generation Sequencing Platforms', *Annual Review of Analytical Chemistry,* 6(1), pp. 287-303.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., et al. (2005) 'Genome sequencing in microfabricated high-density picolitre reactors', *Nature,* 437(7057), pp. 376-380.

Marshall Graves, J. A. and Shetty, S. (2001) 'Sex from W to Z: Evolution of vertebrate sex chromosomes and sex determining genes', *Journal of Experimental Zoology,* 290(5), pp. 449-462.

Martin, G. R., Wilson, K.-J., Wild, J. M., Parsons, S., et al. (2007) 'Kiwi Forego Vision in the Guidance of Their Nocturnal Activities', *PLoS ONE,* 2(2), pp. e198.

Matsubara, K., Tarui, H., Toriba, M., Yamada, K., et al. (2006) 'Evidence for different origin of sex chromosomes in snakes, birds, and mammals and

step-wise differentiation of snake sex chromosomes', *Proceedings of the National Academy of Sciences,* 103(48), pp. 18190-18195.

McCoy, R. C., Taylor, R. W., Blauwkamp, T. A., Kelley, J. L., et al. (2014) 'Illumina TruSeq Synthetic Long-Reads Empower De Novo Assembly and Resolve Complex, Highly-Repetitive Transposable Elements', *PLoS ONE,* 9(9), pp. e106689.

Meredith, R. W., Zhang, G., Gilbert, M. T., Jarvis, E. D., et al. (2014) 'Evidence for a single loss of mineralized teeth in the common avian ancestor', *Science,* 346(6215), pp. 1254390.

Metzker, M. L. (2010) 'Sequencing technologies - the next generation', *Nature Reviews Genetics,* 11(1), pp. 31-46.

Meyerson, M., Gabriel, S. and Getz, G. (2010) 'Advances in understanding cancer genomes through second-generation sequencing', *Nature Reviews Genetics,* 11(10), pp. 685-696.

Mikkelsen, T. S., Wakefield, M. J., Aken, B., Amemiya, C. T., et al. (2007) 'Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences', *Nature,* 447(7141), pp. 167-177.

Miller, W., Makova, K. D., Nekrutenko, A. and Hardison, R. C. (2004) 'COMPARATIVE GENOMICS', *Annual Review of Genomics and Human Genetics,* 5(1), pp. 15-56.

Milo, R. and Phillips, R. (2016) *Cell biology by the numbers.* New York, NY: Garland Science.

Mitelman, F., Johansson, B. and Mertens, F. (2017) 'Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer'. Available at: http://cgap.nci.nih.gov/Chromosomes/Mitelman.

Modi, W. S., Romanov, M., Green, E. D. and Ryder, O. (2009) 'Molecular cytogenetics of the california condor: evolutionary and conservation implications', *Cytogenetic and Genome Research,* 127(1), pp. 26-32.

Mongin, E., Dewar, K. and Blanchette, M. (2009) 'Long-range regulation is a major driving force in maintaining genome integrity', *BMC Evolutionary Biology,* 9(1), pp. 203.

Moore, J. K. and Haber, J. E. (1996) 'Cell cycle and genetic requirements of two pathways of nonhomologous end-joining repair of double-strand breaks in Saccharomyces cerevisiae', *Molecular and Cellular Biology,* 16(5), pp. 2164-2173.

Muffato, M (2010) 'Reconstruction de génomes ancestraux chez le vertébrés', Ph.D., Université d'Evry Val d'Essonne

Murphy, W. J., et al. (2003) 'Reconstructing the genomic architecture of mammalian ancestors using multispecies comparative maps', Human Genomics, 1(1), pp. 30.

Murphy, W. J., Larkin, D. M., Everts-van der Wind, A., Bourque, G., et al. (2005) 'Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps', *Science,* 309(5734), pp. 613-7.

Muzny, D. M. and Scherer, S. E. and Kaul, R. and Wang, J., et al. (2006) 'The DNA sequence, annotation and analysis of human chromosome 3', *Nature,* 440(7088), pp. 1194-8.

Nabholz, B., Glemin, S. and Galtier, N. (2009) 'The erratic mitochondrial clock: variations of mutation rate, not population size, affect mtDNA diversity across birds and mammals', *BMC Evolutionary Biology,* 9, pp. 54.

Nabholz, B., Kunstner, A., Wang, R., Jarvis, E. D., et al. (2011) 'Dynamic evolution of base composition: causes and consequences in avian phylogenomics', *Molecular Biology and Evolution,* 28(8), pp. 2197-210.

Nanda, I., Karl, E., Griffin, D. K., Schartl, M., et al. (2007) 'Chromosome repatterning in three representative parrots (Psittaciformes) inferred from comparative chromosome painting', *Cytogenetic and Genome Research,* 117(1-4), pp. 43-53.

Nanda, I., Schlegelmilch, K., Haaf, T., Schartl, M., et al. (2008) 'Synteny conservation of the Z chromosome in 14 avian species (11 families) supports a role for Z dosage in avian sex determination', *Cytogenetic and Genome Research,* 122(2), pp. 150-6.

Newman, L. A. and Robinson, P. R. (2005) 'Cone visual pigments of aquatic mammals', *Visual Neuroscience,* 22(6), pp. 873-9.

Newman, T. L., Tuzun, E., Morrison, V. A., Hayden, K. E., et al. (2005) 'A genome-wide survey of structural variation between human and chimpanzee', *Genome Research,* 15(10), pp. 1344-1356.

Nie, W., O'Brien, P. C., Ng, B. L., Fu, B., et al. (2009) 'Avian comparative genomics: reciprocal chromosome painting between domestic chicken (Gallus gallus) and the stone curlew (Burhinus oedicnemus, Charadriiformes)--an atypical species with low diploid number', *Chromosome Research,* 17(1), pp. 99-113.

Nishida, C., Ishijima, J., Kosaka, A., Tanabe, H., et al. (2008) 'Characterization of chromosome structures of Falconinae (Falconidae, Falconiformes, Aves) by chromosome painting and delineation of chromosome rearrangements during their differentiation', *Chromosome Research,* 16(1), pp. 171-81.

Noor, M. A., Garfield, D. A., Schaeffer, S. W. and Machado, C. A. (2007) 'Divergence between the Drosophila pseudoobscura and D. persimilis genome sequences in relation to chromosomal inversions', *Genetics,* 177(3), pp. 1417-28.

O'Hare, T. H. and Delany, M. E. (2009) 'Genetic variation exists for telomeric array organization within and among the genomes of normal, immortalized, and transformed chicken systems', *Chromosome Research,* 17(8), pp. 947-64.

O'Connor, J. K. and Chiappe, L. M. (2011) 'A revision of enantiornithine (Aves: Ornithothoraces) skull morphology', *Journal of Systematic Palaeontology,* 9(1), pp. 135-157.

Ohno, S. (1970) *Evolution by Gene Duplication.* Springer Berlin Heidelberg.

Oliver, K. R. and Greene, W. K. (2009) 'Transposable elements: powerful facilitators of evolution', *BioEssays,* 31(7), pp. 703-714.

Olmo, E. (2008) 'Trends in the evolution of reptilian chromosomes', *Integrative and Comparative Biology,* 48(4), pp. 486-493.

Organ, C. L., Moreno, R. G. and Edwards, S. V. (2008) 'Three tiers of genome evolution in reptiles', *Integrative and Comparative Biology,* 48(4), pp. 494-504.

Organ, C. L., Shedlock, A. M., Meade, A., Pagel, M., et al. (2007) 'Origin of avian genome size and structure in non-avian dinosaurs', *Nature,* 446(7132), pp. 180-184.

Paten, B., Earl, D., Nguyen, N., Diekhans, M., et al. (2011) 'Cactus: Algorithms for genome multiple sequence alignment', *Genome Research,* 21(9), pp. 1512-1528.

Patterson, D. (2009) 'Molecular genetic analysis of Down syndrome', *Human Genetics,* 126(1), pp. 195-214.

Peichel, C. L., Sullivan, S. T., Liachko, I. and White, M. A. (2016) 'Improvement of the threespine stickleback (Gasterosteus aculeatus) genome using a Hi-C-based Proximity-Guided Assembly method', *bioRxiv*.

Peters, B. A., Kermani, B. G., Sparks, A. B., Alferov, O., et al. (2012) 'Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells', *Nature,* 487(7406), pp. 190-195.

Pevsner, J. (2015) *Bioinformatics and Functional Genomics.* Wiley.

Piacentini, L., Fanti, L., Specchia, V., Bozzetti, M. P., et al. (2014) 'Transposons, environmental changes, and heritable induced phenotypic variability', *Chromosoma,* 123(4), pp. 345-354.

Pigozzi, M. I. (2016) 'The Chromosomes of Birds during Meiosis', *Cytogenetic and Genome Research,* 150(2), pp. 128-138.

Poelstra, J. W., Vijay, N., Bossu, C. M., Lantz, H., et al. (2014) 'The genomic landscape underlying phenotypic integrity in the face of gene flow in crows', *Science,* 344(6190), pp. 1410-4.

Pokorna, M., Giovannotti, M., Kratochvil, L., Caputo, V., et al. (2012) 'Conservation of chromosomes syntenic with avian autosomes in squamate reptiles revealed by comparative chromosome painting', *Chromosoma,* 121(4), pp. 409-18.

Prauchner, C. A., Prestes Ade, S., Nogueira, C. W. and Rocha, J. B. (2013) 'Effects of diphenyl diselenide and diphenyl ditellurite on chicken embryo development', *Toxicology Mechanisms and Methods,* 23(9), pp. 660-4.

Price, T. D. (2002) 'Domesticated birds as a model for the genetics of speciation by sexual selection', *Genetica,* 116(2-3), pp. 311-27.

Prum, R. O., Berv, J. S., Dornburg, A., Field, D. J., et al. (2015) 'A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing', *Nature,* 526(7574), pp. 569-73.

Puerma, E., Orengo, D. J. and Aguadé, M. (2016) 'The origin of chromosomal inversions as a source of segmental duplications in the Sophophora subgenus of Drosophila', *Scientific Reports,* 6, pp. 30715.

Putnam, N. H., O'Connell, B. L., Stites, J. C., Rice, B. J., et al. (2016) 'Chromosome-scale shotgun assembly using an in vitro method for long-range linkage', *Genome Research,* 26(3), pp. 342-350.

Qiu, Q., et al. (2012) 'The yak genome and adaptation to life at high altitude', Nat Genet, 44(8), pp. 946-9.

Quick, J. and Loman, N. J. and Duraffour, S. and Simpson, J. T., et al. (2016) 'Real-time, portable genome sequencing for Ebola surveillance', *Nature,* 530(7589), pp. 228-232.

Quinlan, A. R. and Hall, I. M. (2010) 'BEDTools: a flexible suite of utilities for comparing genomic features', *Bioinformatics,* 26(6), pp. 841-842.

R CoreTeam (2015) 'R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, Vienna, 2012). http://www.R-project.org'.

Rands, C. M., Darling, A., Fujita, M., Kong, L., et al. (2013) 'Insights into the evolution of Darwin’s finches from comparative analysis of the Geospiza magnirostris genome sequence', *BMC Genomics,* 14(1), pp. 95.

Raynard, S., Niu, H. and Sung, P. (2008) 'DNA double-strand break processing: the beginning of the end', *Genes & development,* 22(21), pp. 2903-7.

Rhoads, A. and Au, K. F. (2015) 'PacBio Sequencing and Its Applications', *Genomics, Proteomics & Bioinformatics,* 13(5), pp. 278-289.

Richard, F., Lombard, M. and Dutrillaux, B. (2003) 'Reconstruction of the ancestral karyotype of eutherian mammals', *Chromosome Research,* 11(6), pp. 605-618.

Rieseberg, L. H. (2001) 'Chromosomal rearrangements and speciation', *Trends in Ecology & Evolution,* 16.

Rodionov, A. V., Chel'sheva, L. A., Solovei, I. V. and Miakoshina Iu, A. (1992) '[Chiasma distribution in the lampbrush chromosomes of the chicken Gallus gallus domesticus: hot spots of recombination and their possible role in proper dysjunction of homologous chromosomes at the first meiotic division]', *Genetika,* 28(7), pp. 151-60.

Romanov, M. N., Dodgson, J. B., Gonser, R. A. and Tuttle, E. M. (2011) 'Comparative BAC-based mapping in the white-throated sparrow, a novel behavioral genomics model, using interspecies overgo hybridization', *BMC Research Notes,* 4, pp. 211.

Romanov, M. N., Farré, M., Lithgow, P. E., Fowler, K. E., 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), pp. 1-18.

Rosenbloom, K. R., Armstrong, J., Barber, G. P., Casper, J., et al. (2015) 'The UCSC Genome Browser database: 2015 update', *Nucleic Acids Research,* 43(Database issue), pp. D670-81.

Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., et al. (2011) 'An integrated semiconductor device enabling non-optical genome sequencing', *Nature,* 475(7356), pp. 348-352.

Ruiz-Herrera, A., Castresana, J. and Robinson, T. J. (2006) 'Is mammalian chromosomal evolution driven by regions of genome fragility?', *Genome Biology,* 7(12), pp. R115.

Ruiz-Herrera, A., Farre, M. and Robinson, T. J. (2012) 'Molecular cytogenetic and genomic insights into chromosomal evolution', *Heredity*
108(1), pp. 28-36.

Sanger, F., Nicklen, S. and Coulson, A. R. (1977) 'DNA sequencing with chain-terminating inhibitors', *Proceedings of the National Academy of Sciences,* 74(12), pp. 5463-5467.

Scanes, C. G. (2014) *Sturkie's Avian Physiology.* Elsevier Science.

Schatz, M. C., Delcher, A. L. and Salzberg, S. L. (2010) 'Assembly of large genomes using second-generation sequencing', *Genome Research,* 20(9), pp. 1165-73.

Schier, A. F. and Shen, M. M. (2000) 'Nodal signalling in vertebrate development', *Nature,* 403(6768), pp. 385-389.

Schmid, M., Fernández-Badillo, A., Feichtinger, W., Steinlein, C., et al. (1988) 'On the highest chromosome number in mammals', *Cytogenetic and Genome Research,* 49(4), pp. 305-308.

Schmid, M. and Smith, J. and Burt, D. W. and Aken, B. L., et al. (2015) 'Third Report on Chicken Genes and Chromosomes 2015', *Cytogenetic and Genome Research,* 145(2), pp. 78-179.

Schneider, V. A., Chen, H. C., Clausen, C., Meric, P. A., et al. (2013) 'Clone DB: an integrated NCBI resource for clone-associated data', *Nucleic Acids Research,* 41(Database issue), pp. D1070-8.

Schreck, R. R. and Disteche, C. M. (2001) 'Chromosome banding techniques', *Current Protocols in Human Genetics,* Chapter 4, pp. Unit4.2.

Schubert, I. and Lysak, M. A. (2011) 'Interpretation of karyotype evolution should consider chromosome structural constraints', *Trends in Genetics,* 27(6), pp. 207-216.

Segura, J., Ferretti, L., Ramos-Onsins, S., Capilla, L., et al. (2013) 'Evolution of recombination in eutherian mammals: insights into mechanisms that affect recombination rates and crossover interference', *Proc Biol Sci,* 280.

Shaffer, H. B., Minx, P., Warren, D. E., Shedlock, A. M., et al. (2013) 'The western painted turtle genome, a model for the evolution of extreme physiological adaptations in a slowly evolving lineage', *Genome Biology,* 14(3), pp. R28.

Shang, W. H., Hori, T., Toyoda, A., Kato, J., et al. (2010) 'Chickens possess centromeres with both extended tandem repeats and short non-tandem-repetitive sequences', *Genome Research,* 20(9), pp. 1219-28.

Shapiro, M. D., Kronenberg, Z., Li, C., Domyan, E. T., et al. (2013) 'Genomic diversity and evolution of the head crest in the rock pigeon', *Science,* 339(6123), pp. 1063-7.

Sharma, A. K. and Sharma, A. (2001) 'Chromosome painting – principles, strategies and scope', *Methods in Cell Science,* 23(1), pp. 1-5.

Sharp, Andrew J., Locke, Devin P., McGrath, Sean D., Cheng, Z., et al. (2005) 'Segmental Duplications and Copy-Number Variation in the Human Genome', *American Journal of Human Genetics,* 77(1), pp. 78-88.

Shaw, C. J., Bi, W. and Lupski, J. R. (2002) 'Genetic proof of unequal meiotic crossovers in reciprocal deletion and duplication of 17p11.2', *American Journal of Human Genetics,* 71(5), pp. 1072-81.

Shetty, S., Griffin, D. K. and Graves, J. A. (1999) 'Comparative painting reveals strong chromosome homology over 80 million years of bird evolution', *Chromosome Research,* 7.

Shi, L., Guo, Y., Dong, C., Huddleston, J., et al. (2016) 'Long-read sequencing and de novo assembly of a Chinese genome', *Nature Communications,* 7, pp. 12065.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., et al. (2005) 'Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes', *Genome Research,* 15(8), pp. 1034-1050.

Skinner, B. M. and Griffin, D. K. (2012) 'Intrachromosomal rearrangements in avian genome evolution: evidence for regions prone to breakpoints', *Heredity,* 108(1), pp. 37-41.

Sneddon, T. P., Li, P. and Edmunds, S. C. (2012) 'GigaDB: announcing the GigaScience database', *GigaScience,* 1(1), pp. 11.

Speicher, M. R. and Carter, N. P. (2005) 'The new cytogenetics: blurring the boundaries with molecular biology', *Nat Rev Genet,* 6(10), pp. 782-792.

Stankiewicz, P. and Lupski, J. R. (2002) 'Genome architecture, rearrangements and genomic disorders', *Trends in Genetics,* 18(2), pp. 74-82.

Stringham, S. A., Mulroy, E. E., Xing, J., Record, D., et al. (2012) 'Divergence, convergence, and the ancestry of feral populations in the domestic rock pigeon', *Current Biology,* 22(4), pp. 302-8.

Sturtevant, A. H. (1913) 'The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association', *Journal of Experimental Zoology,* 14(1), pp. 43-59.

Sundaram, V., Cheng, Y., Ma, Z., Li, D., et al. (2014) 'Widespread contribution of transposable elements to the innovation of gene regulatory networks', *Genome Research,* 24(12), pp. 1963-1976.

Sybert, V. P. and McCauley, E. (2004) 'Turner's Syndrome', *New England Journal of Medicine,* 351(12), pp. 1227-1238.

Tamazian, G., Dobrynin, P., Krasheninnikova, K., Komissarov, A., et al. (2016) 'Chromosomer: a reference-based genome arrangement tool for producing draft chromosome sequences', *GigaScience,* 5(1), pp. 38.

Tanaka, M., Kasahara, H., Bartunkova, S., Schinke, M., et al. (1998) 'Vertebrate homologs of tinman and bagpipe: roles of the homeobox genes in cardiovascular development', *Developmental Genetics,* 22(3), pp. 239-49.

Tesler, G. (2002) 'GRIMM: genome rearrangements web server', *Bioinformatics,* 18(3), pp. 492-493.

The Gene Ontology Consortium (2015) 'Gene Ontology Consortium: going forward', *Nucleic Acids Research,* 43(D1), pp. D1049-D1056.

The Genomes Project Consortium (2010) 'A map of human genome variation from population scale sequencing', Nature, 467(7319), pp. 1061-1073.

Therneau, T., Atkinson, B. and Ripley, B. 2015. rpart: Recursive Partitioning and Regression Trees. R package version 4.1–10.

Thomas, G. H. (2015) 'Evolution: An avian explosion', *Nature,* 526(7574), pp. 516-517.

Thomas, J. W., Cáceres, M., Lowman, J. J., Morehouse, C. B., et al. (2008) 'The Chromosomal Polymorphism Linked to Variation in Social Behavior in the White-Throated Sparrow (Zonotrichia albicollis) Is a Complex Rearrangement and Suppressor of Recombination', *Genetics,* 179(3), pp. 1455-1468.

Tiersch, T. R. and Wachtel, S. S. (1991) 'On the Evolution of Genome Size of Birds', *Journal of Heredity,* 82, pp. 363-368.

Tomkinson, A. E., Vijayakumar, S., Pascal, J. M. and Ellenberger, T. (2006) 'DNA ligases: structure, reaction mechanism, and function', *Chemical Reviews,* 106(2), pp. 687-99.

Tucker, V., Cade, T. and Tucker, A. (1998) 'Diving speeds and angles of a gyrfalcon (Falco rusticolus)', *Journal of Experimental Biology,* 201(13), pp. 2061-2070.

Ullastres, A., Farré, M., Capilla, L. and Ruiz-Herrera, A. (2014) 'Unraveling the effect of genomic structural changes in the rhesus macaque - implications for the adaptive role of inversions', *BMC Genomics,* 15(1), pp. 530.

Uno, Y., Nishida, C., Tarui, H., Ishishita, S., et al. (2012) 'Inference of the Protokaryotypes of Amniotes and Tetrapods and the Evolutionary Processes of Microchromosomes from Comparative Gene Mapping', *PLoS ONE,* 7(12), pp. e53027.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., et al. (2012) 'Primer3--new capabilities and interfaces', *Nucleic Acids Res,* 40(15), pp. e115.

Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., et al. (2008) 'A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning', *Genome Research,* 18(7), pp. 1051-63.

Van Berkum, N. L., Lieberman-Aiden, E., Williams, L., Imakaev, M., et al. (2010) 'Hi-C: A Method to Study the Three-dimensional Architecture of Genomes', *Journal of Visualized Experiments,* (39), pp. 1869.

Van den Bussche, R. A., Longmire, J. L. and Baker, R. J. (1995) 'How bats achieve a small C-value: frequency of repetitive DNA in Macrotus', *Mammalian Genome,* 6(8), pp. 521-5.

Van Dijk, E. L., Auger, H., Jaszczyszyn, Y. and Thermes, C. (2014) 'Ten years of next-generation sequencing technology', *Trends in Genetics,* 30(9), pp. 418-426.

Viguera, E., Canceill, D. and Ehrlich, S. (2001) 'Replication slippage involves DNA polymerase pausing and dissociation', *The EMBO Journal,* 20(10), pp. 2587-2595.

Voelkerding, K. V., Dames, S. A. and Durtschi, J. D. (2009) 'Next-generation sequencing: from basic research to diagnostics', *Clinical Chemistry,* 55(4), pp. 641-658.

Völker, M., Backström, N., Skinner, B. M., Langley, E. J., et al. (2010) 'Copy number variation, chromosome rearrangement, and their association with recombination during avian evolution', *Genome Research,* 20.

Voskoboynik, A., Neff, N. F., Sahoo, D., Newman, A. M., et al. (2013) 'The genome sequence of the colonial chordate, Botryllus schlosseri', *eLife,* 2, pp. e00569.

Voss, S. R., Kump, D. K., Putta, S., Pauly, N., et al. (2011) 'Origin of amphibian and avian chromosomes by fission, fusion, and retention of ancestral chromosomes', *Genome Research,* 21(8), pp. 1306-12.

Wan, Q. H., Pan, S. K., Hu, L., Zhu, Y., et al. (2013) 'Genome analysis and signature discovery for diving and sensory properties of the endangered Chinese alligator', *Cell Research,* 23(9), pp. 1091-105.

Warren, W. C., Clayton, D. F., Ellegren, H., Arnold, A. P., et al. (2010) 'The genome of a songbird', *Nature,* 464(7289), pp. 757-62.

Warren, W. C., Hillier, L. W., Tomlinson, C., Minx, P., et al. (2016) 'A New Chicken Genome Assembly Provides Insight into Avian Genome Structure', *G3: Genes|Genomes|Genetics*.

Weiss, M. M., Hermsen, M. A., Meijer, G. A., van Grieken, N. C., et al. (1999) 'Comparative genomic hybridisation', *Molecular Pathology,* 52(5), pp. 243-251.

Werner, T. (2010) 'Next generation sequencing in functional genomics', Briefings in Bioinformatics, 11(5), pp. 499-511.

Wessler, S. R. (2006) 'Transposable elements and the evolution of eukaryotic genomes', *Proceedings of the National Academy of Sciences,* 103(47), pp. 17600-17601.

White, M. J. D. (1978) *Modes of speciation.* San Francisco: WH Freeman 455p.- Illus., maps, chrom. nos.. General (KR, 197800185).

Whitney, O., Pfenning, A. R., Howard, J. T., Blatti, C. A., et al. (2014) 'Core and region-enriched networks of behaviorally regulated genes and the singing genome', *Science,* 346(6215), pp. 1256780.

Williams, G. (2011) *Data mining with Rattle and R: The art of excavating data for knowledge discovery.* Springer Science & Business Media.

Williams, G. M., Duan, J. D., Brunnemann, K. D., Iatropoulos, M. J., et al. (2014) 'Chicken fetal liver DNA damage and adduct formation by activation-dependent DNA-reactive carcinogens and related compounds of several structural classes', *Toxicological Sciences,* 141(1), pp. 18-28.

Wilson, G. M., Flibotte, S., Missirlis, P. I., Marra, M. A., et al. (2006) 'Identification by full-coverage array CGH of human DNA copy number increases relative to chimpanzee and gorilla', *Genome Research,* 16(2), pp. 173-181.

Wong, A. K., Ruhe, A. L., Dumont, B. L., Robertson, K. R., et al. (2010) 'A Comprehensive Linkage Map of the Dog Genome', *Genetics,* 184(2), pp. 595-605.

Wood, T. E., Takebayashi, N., Barker, M. S., Mayrose, I., et al. (2009) 'The frequency of polyploid speciation in vascular plants', *Proceedings of the National Academy of Sciences,* 106(33), pp. 13875-13879.

Woolfe, A., Goodson, M., Goode, D. K., Snell, P., et al. (2004) 'Highly Conserved Non-Coding Sequences Are Associated with Vertebrate Development', *PLoS Biology,* 3(1), pp. e7.

Wright, N. A., Gregory, T. R. and Witt, C. C. (2014) 'Metabolic 'engines' of flight drive genome size reduction in birds', *Proceedings of the Royal Society B: Biological Sciences,* 281(1779).

Wurster, D. H. and Benirschke, K. (1970) 'Indian Muntjac, Muntiacus muntiak: A Deer with a Low Diploid Chromosome Number', *Science,* 168(3937), pp. 1364-1366.

Wuster, A. (2011) 'Is it cheaper to re-sequence a genome than to save it in computer memory?', Seqnomics. Available at: http://seqonomics.blogspot.co.uk.

Yunis, J. J. and Prakash, O. (1982) 'The origin of man: a chromosomal pictorial legacy', *Science,* 215(4539), pp. 1525-30.

Zhan, X., Pan, S., Wang, J., Dixon, A., et al. (2013) 'Peregrine and saker falcon genome sequences provide insights into evolution of a predatory lifestyle', *Nature Genetics,* 45(5), pp. 563-566.

Zhang, G., Cowled, C., Shi, Z., Huang, Z., et al. (2013) 'Comparative Analysis of Bat Genomes Provides Insight into the Evolution of Flight and Immunity', *Science,* 339(6118), pp. 456-460.

Zhang, G., Li, B., Li, C., Gilbert, M. T. P., et al. (2014a) 'Comparative genomic data of the Avian Phylogenomics Project', *GigaScience,* 3(1), pp. 26.

Zhang, G. and Li, C. and Li, Q. and Li, B., et al. (2014b) 'Comparative genomics reveals insights into avian genome evolution and adaptation', *Science,* 346(6215), pp. 1311-1320.

Zhang, J., Li, C., Zhou, Q. and Zhang, G. (2015) 'Improving the ostrich genome assembly using optical mapping data', *Gigascience,* 4, pp. 24.

Zhang, Q. and Edwards, S. V. (2012) 'The Evolution of Intron Size in Amniotes: A Role for Powered Flight?', *Genome Biology and Evolution,* 4(10), pp. 1033-1043.

Zhao, J., Bacolla, A., Wang, G. and Vasquez, K. M. (2010) 'Non-B DNA structure-induced genetic instability and evolution.', *Cellular and molecular life sciences,* 67, pp. 43-62.

Zheng, G. X. Y., Lau, B. T., Schnall-Levin, M., Jarosz, M., et al. (2016) 'Haplotyping germline and cancer genomes with high-throughput linked-read sequencing', *Nature Biotechnology,* 34(3), pp. 303-311.

Zinzow-Kramer, W. M., Horton, B. M., McKee, C. D., Michaud, J. M., et al. (2015) 'Genes located in a chromosomal inversion are correlated with territorial song in white-throated sparrows', *Genes, Brain, and Behavior,* 14(8), pp. 641-54.

Zlotina, A., Galkina, S., Krasikova, A., Crooijmans, R. P., et al. (2012) 'Centromere positions in chicken and Japanese quail chromosomes: de novo centromere formation versus pericentric inversions', *Chromosome Research,* 20(8), pp. 1017-32.

Zody, M. C., Garber, M., Adams, D. J., Sharpe, T., et al. (2006) 'DNA sequence of human chromosome 17 and analysis of rearrangement in the human lineage', *Nature,* 440(7087), pp. 1045-1049.

7 Appendices

Supplemental Table 1: Statistics of the Pekin duck original and RACA (2 rounds) genome assemblies.

	Scaffold assembly	RACA default	RACA 20X
Assembly statistics			
No. scaffolds/PCFs	2,368	173	175
Total length (Gbp)	1.08	1.00	1.00
N50 (Mbp)	1.26	28.31	21.32
Max. scaffold/PCF length (Mbp)	6.00	113.50	104.22
No. PCFs with only one scaffold	-	73	73
PCFs statistics			
No. used scaffolds by RACA	-	1,144	1,144
Max. no. scaffolds per PCF	-	142	124
Min. no. scaffolds per PCF	-	1	1
No. PCFs homologous to ref. chrs.	-	2	2[1]
No. putative chimeric scaffolds/joints	-	52/53	33/33
No. putative EBRs	-	7	26[2]

[1] Chicken homologous chromosome: 17 and 25.
[2] No chromosomal fusions detected

Supplemental Table 2: Statistics of emperor penguin original and RACA (2 rounds) genome assemblies.

	Scaffold assembly	RACA default	RACA 50X
Assembly statistics			
No. scaffolds/PCFs	682	94	94
Total length (Gbp)	1.25	1.22	1.22
N50 (Mbp)	5.08	40.25	40.25
Max. scaffold/PCF length (Mbp)	28.26	138.35	138.35
No. PCFs with only one scaffold	-	37	37
PCFs statistics			
No. used scaffolds by RACA	-	408	408
Max. no. scaffolds per PCF	-	37	37
Min. no. scaffolds per PCF	-	1	1
No. PCFs homologous to ref. chrs.	-	10	10[1]
No. putative chimeric scaffolds/joints	-	25/28	25/28
No. putative EBRs	-	9	10[2]

[1] Zebra finch homologous chromosome: 4A, 10, 17-19, and 21-25.
[2] Fusion of zebra finch chromosomes 1-1B

Supplemental Table 3: Statistics of Anna's hummingbird original and RACA (2 rounds) genome assemblies.

	Scaffold assembly	RACA default	RACA 583X
Assembly statistics			
No. scaffolds/PCFs	887	102	114
Total length (Gbp)	1.05	1.02	1.02
N50 (Mbp)	4.30	35.86	22.39
Max. scaffold/PCF length (Mbp)	23.31	96.88	86.39
No. PCFs with only one scaffold	-	33	42
PCFs statistics			
No. used scaffolds by RACA	-	431	431
Max. no. scaffolds per PCF	-	45	27
Min. no. scaffolds per PCF	-	1	1
No. PCFs homologous to ref. chrs.	-	4	4[1]
No. putative chimeric scaffolds/joints	-	65/102	34/39
No. putative EBRs	-	16	75[2]

[1] Chicken homologous chromosome: 11, 17, 21 and 25.
[2] No chromosomal fusions detected.

Supplemental Table 4: Statistics of chimney swift original and RACA (2 rounds) genome assemblies.

	Scaffold assembly	RACA default	RACA 239X
Assembly statistics			
No. scaffolds/PCFs	1,172	130	132
Total length (Gbp)	1.10	1.06	1.06
N50 (Mbp)	3.88	28.48	27.05
Max. scaffold/PCF length (Mbp)	19.54	82.12	73.27
No. PCFs with only one scaffold	-	50	52
PCFs statistics			
No. used scaffolds by RACA	-	524	524
Max. no. scaffolds per PCF	-	33	26
Min. no. scaffolds per PCF	-	1	1
No. PCFs homologous to ref. chrs.	-	7	7[1]
No. putative chimeric scaffolds/joints	-	41/48	23/24
No. putative EBRs	-	7	26[2]

[1] Chicken homologous chromosome: 9, 11, 17, 19, 22, 25 and Z.
[2] No chromosomal fusions detected.

Supplemental Table 5: Statistics of killdeer original and RACA (2 rounds) genome assemblies.

	Scaffold assembly	RACA default	RACA 239X
Assembly statistics			
No. scaffolds/PCFs	1,598	97	101
Total length (Gbp)	1.21	1.15	1.15
N50 (Mbp)	3.68	36.52	34.89
Max. scaffold/PCF length (Mbp)	21.92	112.59	112.59
No. PCFs with only one scaffold	-	44	26
PCFs statistics			
No. used scaffolds by RACA	-	527	527
Max. no. scaffolds per PCF	-	47	47
Min. no. scaffolds per PCF	-	1	1
No. PCFs homologous to ref. chrs.	-	5	5[1]
No. putative chimeric scaffolds/joints	-	18/18	9/9
No. putative EBRs	-	13	22[2]

[1] Zebra finch homologous chromosome: 8, 17, 19, 22 and 25.
[2] Fusion of zebra finch chromosomes 1-1B.

Supplemental Table 6: Statistics of rock pigeon original and RACA (2 rounds) genome assemblies.

	Scaffold assembly	RACA default	RACA 85X
Assembly statistics			
No. scaffolds/PCFs	1,081	150	134
Total length (Gbp)	1.10	1.07	1.07
N50 (Mbp)	3.15	34.54	22.17
Max. scaffold/PCF length (Mbp)	25.67	100.53	100.53
No. PCFs with only one scaffold	-	65	52
PCFs statistics			
No. used scaffolds by RACA	-	572	572
Max. no. scaffolds per PCF	-	59	56
Min. no. scaffolds per PCF	-	1	1
No. PCFs homologous to ref. chrs.	-	5[1]	5[1]
No. putative chimeric scaffolds/joints	-	78/109	17/18
No. putative EBRs	-	1	72[2]

[1] Chicken homologous chromosome: 11, 13, 17, 22 and 25.
[2] No chromosomal fusions detected.

Supplemental Table 7: Statistics of American crow original and RACA (2 rounds) genome assemblies.

	Scaffold assembly	RACA default	RACA 85X
Assembly statistics			
No. scaffolds/PCFs	1,156	89	88
Total length (Gbp)	1.08	1.03	1.03
N50 (Mbp)	7.08	53.50	53.50
Max. scaffold/PCF length (Mbp)	25.92	102.11	102.11
No. PCFs with only one scaffold	-	35	37
PCFs statistics			
No. used scaffolds by RACA	-	275	275
Max. no. scaffolds per PCF	-	23	23
Min. no. scaffolds per PCF	-	1	1
No. PCFs homologous to ref. chrs.	-	6	6[1]
No. putative chimeric scaffolds/joints	-	23/26	17/17
No. putative EBRs	-	7	16[2]

[1] Zebra finch homologous chromosome: 1B, 12, 14, 17, 21 and 24.
[2] No chromosomal fusions detected.

Supplemental Table 8: Statistics of common cuckoo original and RACA (2 rounds) genome assemblies.

	Scaffold assembly	RACA default	RACA 583X
Assembly statistics			
No. scaffolds/PCFs	900	119	137
Total length (Gbp)	1.15	1.11	1.11
N50 (Mbp)	2.99	33.40	20.85
Max. scaffold/PCF length (Mbp)	14.00	82.42	79.08
No. PCFs with only one scaffold	-	43	63
PCFs statistics			
No. used scaffolds by RACA	-	569	569
Max. no. scaffolds per PCF	-	51	40
Min. no. scaffolds per PCF	-	1	1
No. PCFs homologous to ref. chrs.	-	7	5[1]
No. putative chimeric scaffolds/joints	-	70/93	31/35
No. putative EBRs	-	9	67

[1] Zebra finch homologous chromosome: 14, 17, 19, 22 and 25.
[2] Fusion of zebra finch chromosomes 1-1B.

Supplemental Table 9: Statistics of little egret original and RACA (2 rounds) genome assemblies.

	Scaffold assembly	RACA default	RACA 85X
Assembly statistics			
No. scaffolds/PCFs	1,195	100	103
Total length (Gbp)	1.20	1.15	1.15
N50 (Mbp)	3.11	37.58	30.49
Max. scaffold/PCF length (Mbp)	14.64	103.52	103.52
No. PCFs with only one scaffold	-	30	35
PCFs statistics			
No. used scaffolds by RACA	-	601	601
Max. no. scaffolds per PCF	-	40	40
Min. no. scaffolds per PCF	-	1	1
No. PCFs homologous to ref. chrs.	-	6	7[1]
No. putative chimeric scaffolds/joints	-	34/34	17/17
No. putative EBRs	-	9	26[2]

[1] Zebra finch homologous chromosome: 1A, 10, 17, 19, 21, 22 and 25.
[2] Fusion of zebra finch chromosomes 1-1B and 9-14.

Supplemental Table 10: Statistics of peregrine falcon original and RACA (2 rounds) genome assemblies.

	Scaffold assembly	RACA default	RACA 583X
Assembly statistics			
No. scaffolds/PCFs	723	113	97
Total length (Gbp)	1.17	1.14	1.14
N50 (Mbp)	3.94	27.44	26.78
Max. scaffold/PCF length (Mbp)	18.33	110.36	73.36
No. PCFs with only one scaffold	-	43	29
PCFs statistics			
No. used scaffolds by RACA	-	478	478
Max. no. scaffolds per PCF	-	41	37
Min. no. scaffolds per PCF	-	1	1
No. PCFs homologous to ref. chrs.	-	8	6[1]
No. putative chimeric scaffolds/joints	-	72/85	23/23
No. putative EBRs	-	25	58[2]

[1] Zebra finch homologous chromosome: 4A, 9, 11, 14, 17 and 19.
[2] Fusion of zebra finch chromosomes 1-1B, 2-23 and 5-20.

Supplemental Table 11: Statistics of medium ground finch original and RACA (2 rounds) genome assemblies.

	Scaffold assembly	RACA default	RACA 85X
Assembly statistics			
No. scaffolds/PCFs	1,168	46	91
Total length (Gbp)	1.04	0.99	0.99
N50 (Mbp)	5.28	55.14	36.73
Max. scaffold/PCF length (Mbp)	30.50	113.25	94.17
No. PCFs with only one scaffold	-	6	6
PCFs statistics			
No. used scaffolds by RACA	-	373	373
Max. no. scaffolds per PCF	-	46	29
Min. no. scaffolds per PCF	-	1	1
No. PCFs homologous to ref. chrs.	-	21	6[1]
No. putative chimeric scaffolds/joints	-	21/23	13/14
No. putative EBRs	-	8	17[2]

[1] Zebra finch homologous chromosome: 4A, 8, 13-15 and 17.
[2] Fusion of zebra finch chromosomes 1-1B.

Supplemental Table 12: Statistics of the golden-collared manakin original and RACA (2 rounds) genome assemblies.

	Scaffold assembly	RACA default	RACA 583X
Assembly statistics			
No. scaffolds/PCFs	954	95	97
Total length (Gbp)	1.05	1.02	1.02
N50 (Mbp)	2.86	45.22	36.36
Max. scaffold/PCF length (Mbp)	12.47	96.35	95.91
No. PCFs with only one scaffold	-	26	29
PCFs statistics			
No. used scaffolds by RACA	-	566	566
Max. no. scaffolds per PCF	-	58	57
Min. no. scaffolds per PCF	-	1	1
No. PCFs homologous to ref. chrs.	-	3	3[1]
No. putative chimeric scaffolds/joints	-	30/34	11/11
No. putative EBRs	-	17	40[2]

[1] Zebra finch homologous chromosome: 17, 20 and 25.
[2] Fusion of zebra finch chromosomes 1-1B.

Supplemental Table 13: Statistics of the budgerigar original and RACA (2 rounds) genome assemblies.

	Scaffold assembly	RACA default	RACA 50X
Assembly statistics			
No. scaffolds/PCFs	1,138	84	86
Total length (Gbp)	1.08	1.04	1.04
N50 (Mbp)	11.41	46.54	38.57
Max. scaffold/PCF length (Mbp)	39.89	147.08	137.80
No. PCFs with only one scaffold	-	29	30
PCFs statistics			
No. used scaffolds by RACA	-	254	254
Max. no. scaffolds per PCF	-	22	17
Min. no. scaffolds per PCF	-	1	1
No. PCFs homologous to ref. chrs.	-	6	8[1]
No. putative chimeric scaffolds/joints	-	80/154	54/85
No. putative EBRs	-	13	81[2]

[1] Zebra finch homologous chromosome: 10, 11, 15, 17, 18, 20, 21 and 25.
[2] Fusions of zebra finch chromosomes 1-1B, 1-4, 3-Z, 4-1A, 5-14, 6-7, 8-9-4A and 8-9.

Supplemental Table 14: Statistics of the crested ibis original and RACA (2 rounds) genome assemblies.

	Scaffold assembly	RACA default	RACA 583X
Assembly statistics			
No. scaffolds/PCFs	1,479	101	105
Total length (Gbp)	1.20	1.15	1.15
N50 (Mbp)	5.35	41.93	38.73
Max. scaffold/PCF length (Mbp)	27.02	74.83	74.83
No. PCFs with only one scaffold	-	31	37
PCFs statistics			
No. used scaffolds by RACA	-	422	422
Max. no. scaffolds per PCF	-	26	26
Min. no. scaffolds per PCF	-	1	1
No. PCFs homologous to ref. chrs.	-	8	7[1]
No. putative chimeric scaffolds/joints	-	35/40	32/36
No. putative EBRs	-	10	14[2]

[1] Zebra finch homologous chromosome: 4A, 10, 14, 17, 19, 24 and 25.
[2] Fusion of zebra finch chromosomes 1-1B.

Supplemental Table 15: Statistics of the hoatzin original and RACA (2 rounds) genome assemblies.

	Scaffold assembly	RACA default	RACA 239X
Assembly statistics			
No. scaffolds/PCFs	1,620	158	163
Total length (Gbp)	1.20	1.14	1.14
N50 (Mbp)	2.94	35.40	25.07
Max. scaffold/PCF length (Mbp)	13.34	73.74	68.95
No. PCFs with only one scaffold	-	80	81
PCFs statistics			
No. used scaffolds by RACA	-	680	680
Max. no. scaffolds per PCF	-	46	34
Min. no. scaffolds per PCF	-	1	1
No. PCFs homologous to ref. chrs.	-	2	2[1]
No. putative chimeric scaffolds/joints	-	26/28	12/12
No. putative EBRs	-	8	24[2]

[1] Chicken homologous chromosome: 17 and 19.
[2] Fusion of zebra finch chromosomes 1-2 and 1-3.

Supplemental Table 16: Statistics of the downy woodpecker original and RACA (2 rounds) genome assemblies.

	Scaffold assembly	RACA default	RACA 85X
Assembly statistics			
No. scaffolds/PCFs	1,944	116	144
Total length (Gbp)	1.15	1.05	1.05
N50 (Mbp)	2.12	25.71	14.74
Max. scaffold/PCF length (Mbp)	8.70	86.58	45.67
No. PCFs with only one scaffold	10.00	193.56	193.56
PCFs statistics			
No. used scaffolds by RACA	-	755	755
Max. no. scaffolds per PCF	-	32	40
Min. no. scaffolds per PCF	-	1	1
No. PCFs homologous to ref. chrs.	-	7	2[1]
No. putative chimeric scaffolds/joints	-	102/131	16/16
No. putative EBRs	-	12	127[2]

[1] Zebra finch homologous chromosome: 11 and 17.
[2] Fusion of zebra finch chromosomes 1-1A, 1-1B, 1-5, 4-12 and 18-23.

Supplemental Table 17: Statistics of the Adélie penguin original and RACA (2 rounds) genome assemblies.

	Scaffold assembly	RACA default	RACA 239X
Assembly statistics			
No. scaffolds/PCFs	819	105	105
Total length (Gbp)	1.21	1.17	1.17
N50 (Mbp)	5.23	39.67	39.67
Max. scaffold/PCF length (Mbp)	22.86	104.22	94.42
No. PCFs with only one scaffold	-	36	38
PCFs statistics			
No. used scaffolds by RACA	-	433	433
Max. no. scaffolds per PCF	-	36	34
Min. no. scaffolds per PCF	-	1	1
No. PCFs homologous to ref. chrs.	-	5	5[1]
No. putative chimeric scaffolds/joints	-	32/39	25/30
No. putative EBRs	-	9	18[2]

[1] Zebra finch homologous chromosome: 7, 17, 19, 22 and 25.
[2] Fusion of zebra finch chromosomes 1-1B and 1-2.

Supplemental Table 18: Statistics of the ostrich original and RACA (2 rounds) genome assemblies.

	Scaffold assembly	RACA default	RACA 239X
Assembly statistics			
No. scaffolds/PCFs	1,179	100	136
Total length (Gbp)	1.22	1.17	1.17
N50 (Mbp)	3.64	37.95	28.09
Max. scaffold/PCF length (Mbp)	19.38	109.84	82.53
No. PCFs with only one scaffold	-	34	71
PCFs statistics			
No. used scaffolds by RACA	-	588	588
Max. no. scaffolds per PCF	-	65	50
Min. no. scaffolds per PCF	-	1	1
No. PCFs homologous to ref. chrs.	-	7	3[1]
No. putative chimeric scaffolds/joints	-	59/69	31/35
No. putative EBRs	-	13	41[2]

[1] Chicken homologous chromosome: 10, 17 and 19.
[2] No chromosomal fusions detected.

Supplemental Table 19: Statistics for the super-scaffold and PCFs adjacencies comparisons.

	Pekin duck		Ostrich		Rock pigeon	
Species	**Default PCFs**	**Adjusted PCFs**	**Default PCFs**	**Adjusted PCFs**	**Default PCFs**	**Adjusted PCFs**
Total adjacencies	1,133 (100%)	1,133 (100%)	433 (100%)	433 (100%)	415 (100%)	415 (100%)
Maintained	891 (79%)	891 (79%)	366 (84%)	350 (81%)	361 (87%)	357 (86%)
Not maintained	242 (21%)	240 (21%)	67 (15%)	83 (19%)	54 (13%)	58 (14%)
Missing[1]	146 (13%)	147 (13%)	37 (9%)	64 (15%)	36 (9%)	48 (12%)
Inconsistent[2]	96 (8%)	93 (8%)	30 (7%)	19 (4%)	18 (4%)	10 (2%)
Split in PCF	0	0	0	0	0	0
Split in map	6	13	0	0	0	2
Inverted	25	30	3	1	1	0
Diff. connection	65	50	27	18	17	8

[1] Both blocks present in the assemblies and represent extra connections.

[2] Both blocks present in the assemblies but are connected in different way. This can be due to the scaffold being putatively chimeric in one of the assemblies (Split in PCF or Split in map), the scaffold being connected in the inverse way (Inverted) or the existence of a completely different adjacency (Diff. connection).

Supplemental Table 20: Statistics for the PCFs and super-scaffold adjacencies comparisons.

	Pekin duck		Ostrich		Rock pigeon	
Species	**Default PCFs**	**Adjusted PCFs**	**Default PCFs**	**Adjusted PCFs**	**Default PCFs**	**Adjusted PCFs**
Total adjacencies	1,014 (100%)	993 (100%)	548 (100%)	478 (100%)	529 (100%)	454 (100%)
Maintained	893 (88%)	891 (90%)	370 (67%)	351 (74%)	362 (69%)	355 (78%)
Not maintained	121 (12%)	102 (10%)	178 (33%)	127 (27%)	167 (32%)	98 (22%)
Missing[1]	8 (1%)	3 (0%)	120 (22%)	87 (18%)	123 (23%)	86 (19%)
Inconsistent[2]	113 (11%)	99 (10%)	58 (11%)	40 (8%)	44 (8%)	12 (3%)
Split in PCF	20	7	30	19	23	1
Split in map	0	1	0	2	0	0
Inverted	39	37	6	1	5	1
Diff. connection	54	54	22	18	16	10

[1] Both blocks present in the assemblies and represent extra connections.

[2] Both blocks present in the assemblies but are connected in different way. This can be due to the scaffold being putatively chimeric in one of the assemblies (Split in PCF or Split in map), the scaffold being connected in the inverse way (Inverted) or the existence of a completely different adjacency (Diff. connection).

Supplemental Table 21: Chicken genome intervals corresponding to lineage-specific intrachromosomal EBRs identified in PCFs.

Chicken chromosome	Start (bp)	End (bp)	Size (bp)	Lineage
1*	**1,324,619**	**1,324,781**	**163**	Peregrine falcon
1	**7,914,166**	**7,914,167**	**2**	Peregrine falcon
1	**9,872,482**	**9,872,484**	**3**	Peregrine falcon
1	**21,932,380**	**21,936,303**	**3,924**	Downy woodpecker
1	**33,348,237**	**33,350,745**	**2,509**	Common cuckoo
1	**33,824,637**	**33,828,021**	**3,385**	Downy woodpecker
1	**35,392,489**	**35,394,306**	**1,818**	Common cuckoo
1	**38,055,896**	**38,057,449**	**1,554**	Downy woodpecker
1	38,853,269	39,044,295	191,027	Downy woodpecker
1	**39,389,690**	**39,389,690**	**1**	Downy woodpecker
1	**39,724,692**	**39,724,692**	**1**	Peregrine falcon
1	**40,123,986**	**40,124,810**	**825**	Downy woodpecker
1	**41,516,148**	**41,516,265**	**118**	Downy woodpecker
1	42,081,101	42,205,802	124,702	Downy woodpecker
1	**42,685,446**	**42,685,446**	**1**	Downy woodpecker
1	**44,020,767**	**44,022,569**	**1,803**	Downy woodpecker
1	**47,071,306**	**47,073,066**	**1,761**	Common cuckoo
1	**47,777,668**	**47,778,702**	**1,035**	Common cuckoo
1	**50,556,002**	**50,556,002**	**1**	Killdeer
1	**52,837,638**	**52,851,425**	**13,788**	Common cuckoo
1	**52,967,996**	**52,969,436**	**1,441**	Common cuckoo
1	**53,229,893**	**53,250,063**	**20,171**	Budgerigar
1	**54,853,256**	**54,853,256**	**1**	Rock pigeon
1	**54,885,460**	**54,885,803**	**344**	Medium ground finch
1	**56,529,702**	**56,529,912**	**211**	Medium ground finch
1	**56,805,853**	**56,805,893**	**41**	Common cuckoo
1	**58,454,505**	**58,455,171**	**667**	Rock pigeon
1	**58,565,360**	**58,573,289**	**7,930**	Common cuckoo
1	**59,807,065**	**59,834,895**	**27,831**	Downy woodpecker
1	**61,926,418**	**61,926,798**	**381**	Hoatzin
1	**62,234,293**	**62,234,311**	**19**	Anna's hummingbird
1	**62,318,387**	**62,323,552**	**5,166**	Budgerigar
1	**65,315,377**	**65,315,929**	**553**	Pekin duck
1	**65,445,425**	**65,445,859**	**435**	Common cuckoo
1	**66,859,181**	**66,859,181**	**1**	Golden-collared manakin
1	**67,000,350**	**67,014,237**	**13,888**	Golden-collared manakin
1	**67,312,349**	**67,312,349**	**1**	Medium ground finch
1	**67,331,644**	**67,331,856**	**213**	Downy woodpecker
1	**68,548,636**	**68,548,636**	**1**	Golden-collared manakin
1	69,682,077	69,986,593	304,517	Medium ground finch
1	**71,529,095**	**71,529,095**	**1**	Downy woodpecker
1	**72,413,659**	**72,434,324**	**20,666**	Killdeer
1	**73,158,774**	**73,161,389**	**2,616**	Downy woodpecker
1	**75,209,222**	**75,218,460**	**9,239**	Rock pigeon

Chicken chromosome	Start (bp)	End (bp)	Size (bp)	Lineage
1	**75,241,162**	**75,241,169**	**8**	Anna's hummingbird
1	**76,358,872**	**76,358,872**	**1**	Chimney swift
1	**77,425,720**	**77,452,619**	**26,900**	Hoatzin
1	**77,738,412**	**77,747,032**	**8,621**	Chicken
1	**78,767,337**	**78,767,337**	**1**	American crow
1	**82,312,639**	**82,312,639**	**1**	Little egret
1	**82,456,422**	**82,457,288**	**867**	Budgerigar
1	**82,645,407**	**82,645,408**	**2**	Little egret
1	**83,897,504**	**83,897,504**	**1**	Zebra finch
1	**83,978,044**	**83,985,031**	**6,988**	American crow
1	**86,730,431**	**86,812,265**	**81,835**	Budgerigar
1	**86,968,682**	**86,979,826**	**11,145**	Budgerigar
1	**86,980,066**	**86,980,066**	**1**	Rock pigeon
1	**87,456,636**	**87,456,702**	**67**	Rock pigeon
1	**87,943,749**	**87,943,749**	**1**	Chimney swift
1	**88,828,570**	**88,828,570**	**1**	Budgerigar
1	92,050,387	92,194,398	144,012	Downy woodpecker
1	**92,422,861**	**92,422,861**	**1**	Golden-collared manakin
1	93,639,934	93,788,586	148,653	Ostrich
1	**94,858,217**	**94,907,354**	**49,138**	Golden-collared manakin
1	**95,026,680**	**95,026,683**	**4**	Ostrich
1	**97,363,066**	**97,365,022**	**1,957**	Budgerigar
1	**98,960,107**	**98,962,865**	**2,759**	Rock pigeon
1	**104,196,498**	**104,198,856**	**2,359**	Peregrine falcon
1	**105,957,571**	**105,961,222**	**3,652**	Medium ground finch
1	121,247,582	121,508,290	260,709	Downy woodpecker
1	122,092,352	122,240,347	147,996	Downy woodpecker
1	**127,819,517**	**127,840,705**	**21,189**	Peregrine falcon
1	**131,341,766**	**131,343,747**	**1,982**	Downy woodpecker
1	**148,452,311**	**148,518,437**	**66,127**	Hoatzin
1	**150,963,319**	**150,965,218**	**1,900**	Hoatzin
1	**153,017,132**	**153,018,044**	**913**	Hoatzin
1	**154,606,088**	**154,606,091**	**4**	Hoatzin
1	**157,084,126**	**157,171,230**	**87,105**	Golden-collared manakin
1	**171,597,547**	**171,597,547**	**1**	Golden-collared manakin
1	**182,043,126**	**182,045,369**	**2,244**	Common cuckoo
1	**187,962,399**	**187,963,769**	**1,371**	Common cuckoo
1	**189,456,635**	**189,456,635**	**1**	Common cuckoo
1	**189,474,286**	**189,476,678**	**2,393**	Downy woodpecker
1	**189,686,791**	**189,695,360**	**8,570**	Common cuckoo
1	**192,817,426**	**192,817,426**	**1**	Golden-collared manakin
1	**194,821,153**	**194,821,381**	**229**	Zebra finch
1	**194,821,464**	**194,821,472**	**9**	Medium ground finch
1	**194,821,985**	**194,821,985**	**1**	American crow
1	**195,080,098**	**195,080,972**	**875**	Golden-collared manakin
2	**575,171**	**575,171**	**1**	Downy woodpecker

Chicken chromosome	Start (bp)	End (bp)	Size (bp)	Lineage
2	**740,124**	**771,016**	**30,893**	Downy woodpecker
2	**933,416**	**970,871**	**37,456**	Chicken
2	**3,302,429**	**3,339,177**	**36,749**	American crow
2	**4,047,517**	**4,051,130**	**3,614**	Chicken
2	**16,342,908**	**16,356,265**	**13,358**	Downy woodpecker
2	**16,814,606**	**16,816,230**	**1,625**	Downy woodpecker
2	**20,160,346**	**20,218,646**	**58,301**	Downy woodpecker
2	**22,519,923**	**22,522,434**	**2,512**	Downy woodpecker
2	**22,629,761**	**22,639,145**	**9,385**	Downy woodpecker
2	38,523,087	39,018,675	495,589	Downy woodpecker
2	**42,026,243**	**42,026,362**	**120**	Anna's hummingbird
2	**42,751,087**	**42,759,634**	**8,548**	Killdeer
2	**46,620,064**	**46,689,875**	**69,812**	Ostrich
2	**48,701,814**	**48,702,717**	**904**	Downy woodpecker
2	**49,112,877**	**49,126,739**	**13,863**	Medium ground finch
2	**49,527,118**	**49,527,424**	**307**	Common cuckoo
2	**51,759,577**	**51,759,797**	**221**	Rock pigeon
2	**51,765,593**	**51,766,698**	**1,106**	Downy woodpecker
2	**53,056,764**	**53,135,014**	**78,251**	Downy woodpecker
2	53,256,537	53,367,804	111,268	Downy woodpecker
2	**54,272,754**	**54,305,600**	**32,847**	Downy woodpecker
2	**55,410,564**	**55,416,892**	**6,329**	Ostrich
2	**55,642,735**	**55,643,492**	**758**	Killdeer
2	**56,145,819**	**56,146,033**	**215**	Ostrich
2	**60,298,302**	**60,299,930**	**1,629**	Hoatzin
2	**66,011,278**	**66,011,285**	**8**	Pekin duck
2	66,151,441	66,271,555	120,115	Pekin duck
2	**67,767,268**	**67,863,753**	**96,486**	Ostrich
2	**69,061,190**	**69,061,471**	**282**	Rock pigeon
2	**70,533,769**	**70,563,825**	**30,057**	Peregrine falcon
2	**71,510,804**	**71,511,287**	**484**	Hoatzin
2	**71,793,435**	**71,797,780**	**4,346**	Hoatzin
2	**72,379,117**	**72,383,528**	**4,412**	Rock pigeon
2	72,734,095	73,514,078	779,984	Downy woodpecker
2	**76,709,142**	**76,789,521**	**80,380**	Peregrine falcon
2	78,384,636	78,553,908	169,273	Hoatzin
2	**79,882,598**	**79,887,476**	**4,879**	Ostrich
2	**83,507,539**	**83,507,539**	**1**	Chimney swift
2	**89,783,693**	**89,784,127**	**435**	Ostrich
2	**90,260,464**	**90,288,914**	**28,451**	Ostrich
2	**94,331,459**	**94,331,459**	**1**	Pekin duck
2	**94,817,167**	**94,817,204**	**38**	Crested ibis
2	**117,911,790**	**117,952,433**	**40,644**	Downy woodpecker
2	**118,143,745**	**118,154,613**	**10,869**	Downy woodpecker
2	**119,950,140**	**119,972,203**	**22,064**	Downy woodpecker
2	**127,246,840**	**127,254,216**	**7,377**	Downy woodpecker

Chicken chromosome	Start (bp)	End (bp)	Size (bp)	Lineage
2	128,704,087	128,857,209	153,123	Downy woodpecker
2	**131,638,024**	**131,650,999**	**12,976**	Downy woodpecker
2	**135,975,490**	**136,020,100**	**44,611**	Downy woodpecker
2	**137,836,594**	**137,836,594**	**1**	Downy woodpecker
2	**145,288,679**	**145,289,895**	**1,217**	Peregrine falcon
2	**147,841,658**	**147,846,069**	**4,412**	Peregrine falcon
2	148,488,047	148,609,211	121,165	Downy woodpecker
3	**358,844**	**358,861**	**18**	Ostrich
3	**368,550**	**369,135**	**586**	Little egret
3	**3,034,702**	**3,034,702**	**1**	Anna's hummingbird
3	**4,376,061**	**4,442,399**	**66,339**	Medium ground finch
3	**5,595,577**	**5,596,227**	**651**	Downy woodpecker
3	**7,787,812**	**7,790,691**	**2,880**	Medium ground finch
3	**8,037,333**	**8,037,339**	**7**	Crested ibis
3	**8,468,065**	**8,487,821**	**19,757**	Zebra finch
3	**8,868,324**	**8,869,649**	**1,326**	Downy woodpecker
3	**9,833,952**	**9,843,064**	**9,113**	Downy woodpecker
3	**11,265,986**	**11,331,561**	**65,576**	Downy woodpecker
3	11,788,564	11,931,057	142,494	Pekin duck
3	**14,698,909**	**14,698,991**	**83**	Pekin duck
3	**14,907,931**	**14,908,026**	**96**	Anna's hummingbird
3	**16,099,443**	**16,099,443**	**1**	Anna's hummingbird
3	**16,277,069**	**16,282,622**	**5,554**	Pekin duck
3	**16,560,813**	**16,561,175**	**363**	Pekin duck
3	**17,044,186**	**17,044,287**	**102**	Medium ground finch
3	**17,204,230**	**17,204,377**	**148**	Anna's hummingbird
3	**18,085,092**	**18,085,129**	**38**	Little egret
3	**20,360,470**	**20,390,407**	**29,938**	Downy woodpecker
3	**21,255,139**	**21,255,624**	**486**	Pekin duck
3	**28,486,591**	**28,492,732**	**6,142**	Little egret
3	**34,684,038**	**34,707,599**	**23,562**	Common cuckoo
3	**37,224,845**	**37,224,884**	**40**	Ostrich
3	**40,468,377**	**40,468,377**	**1**	Peregrine falcon
3	**41,741,126**	**41,741,126**	**1**	Ostrich
3	**42,163,223**	**42,177,332**	**14,110**	Anna's hummingbird
3	**44,454,971**	**44,454,971**	**1**	Hoatzin
3	**44,938,858**	**45,008,487**	**69,630**	Little egret
3	**45,137,673**	**45,138,888**	**1,216**	Little egret
3	**51,455,834**	**51,493,028**	**37,195**	Anna's hummingbird
3	54,138,507	54,296,678	158,172	Chimney swift
3	**54,463,251**	**54,463,251**	**1**	Chimney swift
3	**70,642,853**	**70,676,555**	**33,703**	Peregrine falcon
3	**72,550,375**	**72,645,551**	**95,177**	Anna's hummingbird
3	**75,252,841**	**75,252,847**	**7**	Little egret
3	**80,928,188**	**80,928,188**	**1**	Anna's hummingbird
3	**81,830,467**	**81,830,467**	**1**	Downy woodpecker

Chicken chromosome	Start (bp)	End (bp)	Size (bp)	Lineage
3	**86,132,103**	**86,226,836**	**94,734**	Little egret
3	89,598,019	89,724,626	126,608	Hoatzin
3	**93,358,971**	**93,359,489**	**519**	Downy woodpecker
3	**96,056,488**	**96,058,210**	**1,723**	Common cuckoo
3	**96,521,843**	**96,540,436**	**18,594**	Chimney swift
3	**104,838,689**	**104,841,842**	**3,154**	Chimney swift
3	**104,884,266**	**104,889,327**	**5,062**	Common cuckoo
3	**106,856,698**	**106,856,887**	**190**	Peregrine falcon
3	**107,044,848**	**107,046,677**	**1,830**	Common cuckoo
3	109,477,426	109,606,728	129,303	Pekin duck
4	**2,019,931**	**2,019,931**	**1**	Rock pigeon
4	**2,168,787**	**2,169,536**	**750**	Downy woodpecker
4	**2,809,725**	**2,836,575**	**26,851**	Downy woodpecker
4	**2,878,046**	**2,891,784**	**13,739**	Zebra finch
4	**2,896,163**	**2,897,345**	**1,183**	American crow
4	**3,389,896**	**3,392,150**	**2,255**	Downy woodpecker
4	**3,568,049**	**3,569,038**	**990**	Downy woodpecker
4	**8,692,434**	**8,706,528**	**14,095**	Rock pigeon
4	**8,950,764**	**8,953,036**	**2,273**	Ostrich
4	**9,481,867**	**9,482,034**	**168**	Chicken
4	**10,385,714**	**10,387,281**	**1,568**	Golden-collared manakin
4	**10,431,665**	**10,431,665**	**1**	American crow
4	**11,241,217**	**11,245,088**	**3,872**	Common cuckoo
4	**12,172,096**	**12,172,697**	**602**	Common cuckoo
4	**12,319,271**	**12,319,271**	**1**	Hoatzin
4	**12,526,470**	**12,526,470**	**1**	Hoatzin
4	**13,341,463**	**13,345,322**	**3,860**	American crow
4	**13,376,069**	**13,376,069**	**1**	Zebra finch
4	**14,271,322**	**14,276,754**	**5,433**	Common cuckoo
4	**16,440,544**	**16,454,169**	**13,626**	Rock pigeon
4	**17,287,000**	**17,287,000**	**1**	Golden-collared manakin
4	**17,478,797**	**17,528,495**	**49,699**	Anna's hummingbird
4	17,984,483	18,146,735	162,253	Rock pigeon
4	**24,682,558**	**24,691,545**	**8,988**	Pekin duck
4	**25,024,582**	**25,030,089**	**5,508**	Zebra finch
4	**31,427,645**	**31,427,649**	**5**	Downy woodpecker
4	**33,783,259**	**33,784,601**	**1,343**	Ostrich
4	**34,771,968**	**34,811,614**	**39,647**	Rock pigeon
4	**41,846,597**	**41,861,041**	**14,445**	Anna's hummingbird
4	**44,089,766**	**44,090,417**	**652**	Anna's hummingbird
4	**44,679,114**	**44,748,236**	**69,123**	Chimney swift
4	**45,054,535**	**45,057,881**	**3,347**	Rock pigeon
4	**46,430,075**	**46,434,084**	**4,010**	Chimney swift
4	**48,576,319**	**48,577,424**	**1,106**	Rock pigeon
4	**48,580,539**	**48,582,365**	**1,827**	Hoatzin
4	**51,022,548**	**51,023,909**	**1,362**	Rock pigeon

Chicken chromosome	Start (bp)	End (bp)	Size (bp)	Lineage
4	**56,808,569**	**56,809,104**	**536**	Chimney swift
4	**56,811,785**	**56,814,560**	**2,776**	Anna's hummingbird
4	**59,605,527**	**59,624,845**	**19,319**	Anna's hummingbird
4	**60,644,106**	**60,644,798**	**693**	Peregrine falcon
4	**60,901,745**	**60,901,783**	**39**	Peregrine falcon
4	**72,988,972**	**73,026,637**	**37,666**	Downy woodpecker
4	**76,311,431**	**76,317,989**	**6,559**	Downy woodpecker
5	**761,364**	**761,366**	**3**	Common cuckoo
5	**1,384,513**	**1,400,436**	**15,924**	Downy woodpecker
5	**2,128,829**	**2,129,441**	**613**	Downy woodpecker
5	**4,588,471**	**4,589,857**	**1,387**	Common cuckoo
5	**7,671,466**	**7,673,370**	**1,905**	Budgerigar
5	**7,693,156**	**7,694,301**	**1,146**	Hoatzin
5	**7,886,229**	**7,886,242**	**14**	Budgerigar
5	**8,922,237**	**8,922,254**	**18**	Hoatzin
5	**11,920,700**	**11,924,676**	**3,977**	Peregrine falcon
5	**13,575,913**	**13,575,996**	**84**	Downy woodpecker
5	**14,133,276**	**14,134,077**	**802**	Anna's hummingbird
5	**15,291,760**	**15,296,906**	**5,147**	Anna's hummingbird
5	**15,641,676**	**15,673,140**	**31,465**	Budgerigar
5	**16,015,073**	**16,015,122**	**50**	Crested ibis
5	**17,109,172**	**17,109,172**	**1**	Budgerigar
5	**18,002,362**	**18,002,362**	**1**	Hoatzin
5	**18,791,571**	**18,796,928**	**5,358**	Budgerigar
5	18,938,336	19,627,327	688,992	Budgerigar
5	**20,683,043**	**20,683,043**	**1**	Budgerigar
5	**21,305,603**	**21,306,716**	**1,114**	Downy woodpecker
5	**21,673,016**	**21,673,807**	**792**	Common cuckoo
5	**22,198,049**	**22,198,068**	**20**	Peregrine falcon
5	**24,127,609**	**24,127,609**	**1**	Common cuckoo
5	**24,751,847**	**24,843,616**	**91,770**	Peregrine falcon
5	**25,463,669**	**25,483,787**	**20,119**	Budgerigar
5	**26,807,683**	**26,807,683**	**1**	Budgerigar
5	**27,253,920**	**27,255,545**	**1,626**	Anna's hummingbird
5	**28,319,875**	**28,326,198**	**6,324**	Downy woodpecker
5	**38,162,262**	**38,162,265**	**4**	Common cuckoo
5	**40,665,735**	**40,666,168**	**434**	Rock pigeon
5	**40,911,824**	**40,911,924**	**101**	Rock pigeon
5	**42,301,909**	**42,301,914**	**6**	Peregrine falcon
5	**43,542,798**	**43,542,798**	**1**	Killdeer
5	**51,438,383**	**51,487,667**	**49,285**	Common cuckoo
5	51,615,197	51,762,500	147,304	Common cuckoo
5	**55,173,177**	**55,173,807**	**631**	Killdeer
5	**55,426,754**	**55,437,970**	**11,217**	Common cuckoo
5	**56,370,596**	**56,371,718**	**1,123**	Pekin duck
5	57,177,811	57,326,175	148,365	Pekin duck

Chicken chromosome	Start (bp)	End (bp)	Size (bp)	Lineage
5	**57,433,788**	**57,461,454**	**27,667**	Pekin duck
5	**57,851,771**	**57,909,985**	**58,215**	Budgerigar
5	**58,107,431**	**58,109,471**	**2,041**	Common cuckoo
6	417,995	542,177	124,183	Common cuckoo
6	**2,058,827**	**2,126,276**	**67,450**	Rock pigeon
6	**6,072,764**	**6,073,950**	**1,187**	Rock pigeon
6	**12,716,102**	**12,716,491**	**390**	Rock pigeon
6	**25,354,032**	**25,354,032**	**1**	Common cuckoo
6	**29,927,592**	**29,936,107**	**8,516**	Peregrine falcon
6	**31,551,126**	**31,552,283**	**1,158**	Downy woodpecker
7	**290,521**	**295,591**	**5,071**	Common cuckoo
7	825,542	1,045,773	220,232	Common cuckoo
7	**1,369,022**	**1,406,101**	**37,080**	Rock pigeon
7	**2,045,976**	**2,052,574**	**6,599**	Downy woodpecker
7	**2,206,856**	**2,207,754**	**899**	Golden-collared manakin
7	**4,510,237**	**4,527,867**	**17,631**	Anna's hummingbird
7	**7,052,486**	**7,052,548**	**63**	Rock pigeon
7	**9,277,413**	**9,280,798**	**3,386**	Rock pigeon
7	**9,523,349**	**9,524,330**	**982**	Downy woodpecker
7	**10,733,844**	**10,784,677**	**50,834**	Downy woodpecker
7	**10,956,314**	**10,958,951**	**2,638**	Anna's hummingbird
7	**11,001,647**	**11,001,672**	**26**	Pekin duck
7	**12,024,636**	**12,033,468**	**8,833**	Anna's hummingbird
7	**12,806,084**	**12,851,916**	**45,833**	Anna's hummingbird
7	**13,397,507**	**13,403,118**	**5,612**	Rock pigeon
7	**21,350,970**	**21,350,970**	**1**	Budgerigar
7	**27,742,887**	**27,742,887**	**1**	Pekin duck
7	**27,805,172**	**27,805,204**	**33**	Zebra finch
7	**27,887,466**	**27,888,223**	**758**	Pekin duck
7	**30,730,753**	**30,730,753**	**1**	Downy woodpecker
8	**1,251,761**	**1,252,175**	**415**	Hoatzin
8	**1,337,923**	**1,338,893**	**971**	Little egret
8	**3,855,688**	**3,866,664**	**10,977**	Common cuckoo
8	**4,294,526**	**4,295,791**	**1,266**	Common cuckoo
8	**4,856,037**	**4,856,051**	**15**	Budgerigar
8	**6,056,932**	**6,057,001**	**70**	Budgerigar
8	**6,566,747**	**6,566,748**	**2**	Budgerigar
8	**7,053,877**	**7,082,088**	**28,212**	Chimney swift
8	**7,754,950**	**7,758,870**	**3,921**	Chimney swift
8	**8,648,899**	**8,738,923**	**90,025**	Ostrich
8	**10,595,445**	**10,597,333**	**1,889**	Hoatzin
8	**12,488,814**	**12,489,366**	**553**	Budgerigar
8	**12,745,940**	**12,746,680**	**741**	Common cuckoo
8	**14,951,943**	**14,956,243**	**4,301**	Budgerigar
8	**15,379,696**	**15,383,955**	**4,260**	Golden-collared manakin
8	**22,876,667**	**22,877,454**	**788**	Downy woodpecker

Chicken chromosome	Start (bp)	End (bp)	Size (bp)	Lineage
8	**27,565,916**	**27,568,285**	**2,370**	Downy woodpecker
8	**27,906,361**	**27,913,531**	**7,171**	Downy woodpecker
8	**27,971,511**	**27,973,433**	**1,923**	American crow
9	**798,704**	**882,024**	**83,321**	Anna's hummingbird
9	**907,413**	**909,349**	**1,937**	Medium ground finch
9	**1,054,635**	**1,057,658**	**3,024**	Chimney swift
9	**15,661,968**	**15,665,081**	**3,114**	Little egret
9	**20,151,338**	**20,238,734**	**87,397**	Downy woodpecker
9	**23,308,061**	**23,308,061**	**1**	Crested ibis
10	**899,404**	**899,873**	**470**	Common cuckoo
10	**2,429,267**	**2,429,267**	**1**	Peregrine falcon
10	**3,607,483**	**3,609,720**	**2,238**	Zebra finch
10	**6,364,433**	**6,373,124**	**8,692**	Common cuckoo
10	**7,419,216**	**7,419,647**	**432**	American crow
10	**7,420,556**	**7,420,556**	**1**	Rock pigeon
10	**11,159,134**	**11,162,024**	**2,891**	Downy woodpecker
10	**17,152,487**	**17,156,616**	**4,130**	Downy woodpecker
10	**19,561,265**	**19,568,428**	**7,164**	Downy woodpecker
11	**450,253**	**451,600**	**1,348**	Common cuckoo
11	**579,593**	**580,199**	**607**	Pekin duck
11	**1,642,022**	**1,642,907**	**886**	Common cuckoo
11	**2,655,635**	**2,701,405**	**45,771**	Rock pigeon
11	**2,777,908**	**2,781,479**	**3,572**	Hoatzin
11	**3,162,057**	**3,162,943**	**887**	Rock pigeon
11	**7,229,719**	**7,229,719**	**1**	Golden-collared manakin
11	**9,337,792**	**9,338,814**	**1,023**	Killdeer
11	**10,794,981**	**10,794,981**	**1**	Common cuckoo
11	**13,237,579**	**13,237,629**	**51**	Zebra finch
11	**14,594,769**	**14,594,785**	**17**	Rock pigeon
11	**14,855,343**	**14,856,626**	**1,284**	Downy woodpecker
11	**16,240,932**	**16,241,424**	**493**	Hoatzin
11	**17,094,551**	**17,103,456**	**8,906**	Zebra finch
11	**19,160,714**	**19,162,117**	**1,404**	Common cuckoo
12	**1,198,422**	**1,198,422**	**1**	Zebra finch
12	**1,352,683**	**1,353,042**	**360**	Peregrine falcon
12	**2,718,109**	**2,746,781**	**28,673**	Budgerigar
12	**2,912,454**	**2,922,124**	**9,671**	Budgerigar
12	**3,455,409**	**3,465,525**	**10,117**	Budgerigar
12	**3,466,651**	**3,466,651**	**1**	Chimney swift
12	**8,940,900**	**8,942,182**	**1,283**	Budgerigar
12	**14,333,129**	**14,335,867**	**2,739**	Chimney swift
12	**19,435,017**	**19,438,519**	**3,503**	Budgerigar
13	**377,359**	**378,569**	**1,211**	Golden-collared manakin
13	**382,240**	**382,273**	**34**	Common cuckoo
13	1,101,452	1,214,493	113,042	Anna's hummingbird
13	**2,187,716**	**2,189,266**	**1,551**	Downy woodpecker

Chicken chromosome	Start (bp)	End (bp)	Size (bp)	Lineage
13	**6,044,597**	**6,059,379**	**14,783**	Downy woodpecker
13	**7,522,358**	**7,523,236**	**879**	Rock pigeon
13	**7,768,752**	**7,768,893**	**142**	Ostrich
13	**10,816,579**	**10,817,280**	**702**	Budgerigar
13	**12,464,088**	**12,465,365**	**1,278**	Ostrich
13	**15,374,295**	**15,385,423**	**11,129**	Downy woodpecker
13	**15,967,876**	**15,972,117**	**4,242**	Common cuckoo
13	**16,164,641**	**16,166,145**	**1,505**	Rock pigeon
13	16,493,451	16,624,456	131,006	Common cuckoo
13	**17,498,859**	**17,503,648**	**4,790**	Rock pigeon
13	**17,576,536**	**17,576,554**	**19**	Ostrich
14	**147,260**	**148,688**	**1,429**	Budgerigar
14	**650,373**	**666,763**	**16,391**	Peregrine falcon
14	**937,549**	**937,874**	**326**	Chicken
14	**9,302,911**	**9,302,923**	**13**	Common cuckoo
14	**10,118,823**	**10,133,384**	**14,562**	Anna's hummingbird
14	**11,784,234**	**11,784,392**	**159**	Ostrich
14	**12,237,309**	**12,238,841**	**1,533**	Rock pigeon
14	**12,687,067**	**12,732,406**	**45,340**	Budgerigar
14	**13,216,528**	**13,228,098**	**11,571**	Anna's hummingbird
14	**13,393,288**	**13,399,010**	**5,723**	Ostrich
14	**13,789,331**	**13,789,360**	**30**	Pekin duck
14	**13,849,959**	**13,886,885**	**36,927**	Common cuckoo
15	**329,422**	**331,846**	**2,425**	Golden-collared manakin
15	**1,413,641**	**1,414,233**	**593**	Rock pigeon
15	**5,195,562**	**5,196,215**	**654**	Common cuckoo
15	**7,715,147**	**7,716,336**	**1,190**	Killdeer
15	**7,795,924**	**7,795,924**	**1**	Emperor penguin
17	**547,870**	**549,871**	**2,002**	Anna's hummingbird
17	**2,872,949**	**2,876,352**	**3,404**	Anna's hummingbird
17	**5,876,933**	**5,880,625**	**3,693**	Peregrine falcon
17	**6,771,735**	**6,772,976**	**1,242**	Budgerigar
17	**8,311,265**	**8,315,794**	**4,530**	Anna's hummingbird
17	**8,315,911**	**8,316,843**	**933**	Peregrine falcon
17	**10,244,185**	**10,247,259**	**3,075**	Anna's hummingbird
18	**2,939,704**	**2,972,051**	**32,348**	Common cuckoo
18	**2,981,417**	**2,983,622**	**2,206**	Downy woodpecker
18	3,331,733	3,449,560	117,828	Chimney swift
18	**3,644,093**	**3,653,222**	**9,130**	Ostrich
18	**4,437,561**	**4,437,561**	**1**	Peregrine falcon
18	**5,037,402**	**5,038,354**	**953**	Hoatzin
18	**6,545,499**	**6,545,499**	**1**	Zebra finch
18	**6,964,001**	**6,964,660**	**660**	Pekin duck
18	8,250,967	8,460,497	209,531	Zebra finch
18	**9,771,220**	**9,771,928**	**709**	Budgerigar
18	**10,054,990**	**10,056,308**	**1,319**	Common cuckoo

Chicken chromosome	Start (bp)	End (bp)	Size (bp)	Lineage
18	**10,251,803**	**10,251,842**	**40**	Chicken
18	**10,696,637**	**10,696,637**	**1**	Pekin duck
19	**380,851**	**400,944**	**20,094**	Downy woodpecker
19	2,753,538	2,860,369	106,832	Rock pigeon
19	**3,504,418**	**3,504,418**	**1**	Anna's hummingbird
19	**5,337,519**	**5,337,790**	**272**	Anna's hummingbird
19	**8,082,668**	**8,086,091**	**3,424**	Downy woodpecker
19	**9,797,203**	**9,798,132**	**930**	Crested ibis
20	**486,582**	**486,582**	**1**	Common cuckoo
20	1,670,748	1,795,324	124,577	Common cuckoo
20	**1,834,883**	**1,834,950**	**68**	Anna's hummingbird
20	**1,970,278**	**1,995,633**	**25,356**	Rock pigeon
20	2,710,064	3,061,871	351,808	Downy woodpecker
20	**4,166,600**	**4,170,591**	**3,992**	Downy woodpecker
20	**4,199,659**	**4,200,558**	**900**	Anna's hummingbird
20	**5,489,616**	**5,490,537**	**922**	Crested ibis
20	5,603,676	5,728,438	124,763	Anna's hummingbird
20	**9,076,643**	**9,076,689**	**47**	Downy woodpecker
20	**11,104,544**	**11,104,593**	**50**	Ostrich
20	**13,946,826**	**13,951,854**	**5,029**	Anna's hummingbird
20	**14,061,427**	**14,063,727**	**2,301**	Rock pigeon
21	**1,377,888**	**1,377,888**	**1**	Budgerigar
21	**1,410,931**	**1,411,405**	**475**	Anna's hummingbird
21	**1,798,837**	**1,798,837**	**1**	Budgerigar
21	2,447,419	2,708,370	260,952	Zebra finch
21	**4,129,442**	**4,129,447**	**6**	Anna's hummingbird
21	**4,201,583**	**4,201,995**	**413**	Chicken
21	**4,856,742**	**4,857,758**	**1,017**	Golden-collared manakin
21	**5,833,578**	**5,834,428**	**851**	Pekin duck
21	**6,159,931**	**6,160,089**	**159**	Pekin duck
21	**6,170,466**	**6,170,467**	**2**	Golden-collared manakin
22	910,586	1,020,267	109,682	Downy woodpecker
22	**1,154,360**	**1,159,142**	**4,783**	Downy woodpecker
22	1,342,571	1,639,314	296,744	Downy woodpecker
22	**2,565,550**	**2,638,093**	**72,544**	Budgerigar
23	**561,775**	**565,402**	**3,628**	Anna's hummingbird
23	**1,168,426**	**1,226,911**	**58,486**	Anna's hummingbird
23	**1,267,277**	**1,281,891**	**14,615**	Budgerigar
23	**1,400,414**	**1,410,639**	**10,226**	Anna's hummingbird
23	**2,526,093**	**2,547,933**	**21,841**	Budgerigar
23	**3,026,066**	**3,026,219**	**154**	Peregrine falcon
23	**3,358,793**	**3,360,279**	**1,487**	Anna's hummingbird
23	**5,288,932**	**5,289,432**	**501**	Anna's hummingbird
23	**5,408,474**	**5,408,507**	**34**	Anna's hummingbird
24	**392,608**	**394,381**	**1,774**	Rock pigeon
24	**404,140**	**417,948**	**13,809**	Common cuckoo

Chicken chromosome	Start (bp)	End (bp)	Size (bp)	Lineage
24	**656,964**	**657,377**	**414**	Zebra finch
24	**1,139,943**	**1,207,434**	**67,492**	Little egret
24	**1,791,741**	**1,794,597**	**2,857**	Golden-collared manakin
24	**3,792,334**	**3,792,334**	**1**	Rock pigeon
25	**1,235,301**	**1,287,268**	**51,968**	Peregrine falcon
25	**1,904,999**	**1,952,901**	**47,903**	Ostrich
26	**364,795**	**365,078**	**284**	Pekin duck
26	**467,124**	**500,614**	**33,491**	Anna's hummingbird
26	**852,362**	**852,789**	**428**	Golden-collared manakin
26	**1,145,839**	**1,238,280**	**92,442**	American crow
26	**1,459,471**	**1,460,419**	**949**	Golden-collared manakin
26	**1,838,545**	**1,840,577**	**2,033**	Anna's hummingbird
26	**2,024,352**	**2,025,596**	**1,245**	Budgerigar
26	**2,155,563**	**2,226,442**	**70,880**	Pekin duck
26	**3,198,705**	**3,199,592**	**888**	Budgerigar
26	**3,416,149**	**3,417,052**	**904**	Rock pigeon
26	**3,663,562**	**3,663,562**	**1**	American crow
26	**3,687,326**	**3,688,485**	**1,160**	Pekin duck
26	**4,017,849**	**4,019,971**	**2,123**	Chicken
26	**4,217,291**	**4,217,291**	**1**	Budgerigar
27	**2,299,452**	**2,299,637**	**186**	Common cuckoo
27	**2,836,702**	**2,836,702**	**1**	Common cuckoo
28	**845,610**	**848,352**	**2,743**	Peregrine falcon
28	**873,755**	**904,975**	**31,221**	Anna's hummingbird
28	**1,139,611**	**1,140,061**	**451**	Adélie penguin
28	**1,845,812**	**1,862,425**	**16,614**	Anna's hummingbird
28	**2,532,651**	**2,569,138**	**36,488**	Common cuckoo
28	**3,700,088**	**3,717,924**	**17,837**	Peregrine falcon
28	**4,206,421**	**4,206,421**	**1**	Anna's hummingbird
Z	**1,382,729**	**1,382,729**	**1**	American crow
Z	**1,814,337**	**1,815,275**	**939**	Zebra finch
Z	**10,901,988**	**10,961,111**	**59,124**	Common cuckoo
Z	**12,772,364**	**12,776,700**	**4,337**	Rock pigeon
Z	**13,184,755**	**13,210,257**	**25,503**	Rock pigeon
Z	21,571,901	21,814,536	242,636	American crow
Z	**23,604,768**	**23,611,231**	**6,464**	Hoatzin
Z	**25,352,918**	**25,356,012**	**3,095**	Ostrich
Z	26,983,713	27,340,770	357,058	Rock pigeon
Z	34,165,065	34,283,251	118,187	Chicken
Z	**34,509,829**	**34,510,098**	**270**	Downy woodpecker
Z	**37,099,078**	**37,099,078**	**1**	Adélie penguin
Z	**37,099,170**	**37,099,170**	**1**	Emperor penguin
Z	**37,099,204**	**37,099,204**	**1**	Peregrine falcon
Z	**37,099,767**	**37,099,767**	**1**	Hoatzin
Z	**37,161,839**	**37,161,839**	**1**	Killdeer
Z	**37,161,847**	**37,161,847**	**1**	American crow

Chicken chromosome	Start (bp)	End (bp)	Size (bp)	Lineage
Z	**37,163,617**	**37,164,423**	**807**	Downy woodpecker
Z	**37,183,225**	**37,185,622**	**2,398**	Common cuckoo
Z	**37,186,440**	**37,186,440**	**1**	Rock pigeon
Z	**37,200,122**	**37,200,122**	**1**	Little egret
Z	**39,647,535**	**39,667,465**	**19,931**	Common cuckoo
Z	**41,328,756**	**41,329,326**	**571**	Golden-collared manakin
Z	**43,744,576**	**43,745,641**	**1,066**	Killdeer
Z	**44,351,365**	**44,351,550**	**186**	Killdeer
Z	**45,136,959**	**45,136,959**	**1**	Killdeer
Z	**45,523,224**	**45,523,233**	**10**	Chicken
Z	**46,138,445**	**46,139,449**	**1,005**	Common cuckoo
Z	50,107,459	50,453,265	345,807	Medium ground finch
Z	51,147,812	51,347,420	199,609	Common cuckoo
Z	51,401,610	51,595,048	193,439	Killdeer
Z	**52,376,189**	**52,376,189**	**1**	Budgerigar
Z	52,650,853	52,895,681	244,829	Killdeer
Z	**53,291,606**	**53,293,189**	**1,584**	Chimney swift
Z	**54,816,611**	**54,816,611**	**1**	Medium ground finch
Z	**54,824,837**	**54,825,277**	**441**	American crow
Z	**54,862,334**	**54,862,351**	**18**	Golden-collared manakin
Z	**55,982,826**	**56,006,677**	**23,852**	American crow
Z	**63,566,609**	**63,567,188**	**580**	Common cuckoo
Z	**64,275,893**	**64,275,893**	**1**	Budgerigar
Z	**66,287,944**	**66,318,403**	**30,460**	Downy woodpecker
Z	**66,945,899**	**66,946,496**	**598**	Downy woodpecker
Z	**67,032,907**	**67,100,398**	**67,492**	Medium ground finch
Z	**70,873,959**	**70,953,909**	**79,951**	Little egret
Z	**72,102,303**	**72,102,304**	**2**	Common cuckoo

* In bold are the intervals used in the CNE analysis.

Supplemental Table 22: Chicken genome intervals corresponding to lineage-specific interchromosomal EBRs identified in PCFs.

Chicken chromosome	Start (bp)	End (bp)	Size (bp)	Chicken chromosome	Start (bp)	End (bp)	Size (bp)	EBR type	Lineage
1*	**36,332,912**	**36,333,121**	**210**	**4**	**52,395,090**	**52,395,090**	**1**	Fusion GGA1-4	Budgerigar
1	**74,596,605**	**74,645,439**	**48,835**	**4**	**34,451,491**	**34,451,491**	**1**	Fusion GGA1-4	Budgerigar
1	**147,010,625**	**147,014,027**	**3,403**	**5**	**13,365,336**	**13,365,336**	**1**	Fusion GGA1-5	Downy woodpecker
2	109,558,621	109,559,524	904	1	151,082,279	151,082,285	7	Fusion GGA2-1	Adélie penguin
2	120,100,254	120,100,254	1	23	4,406,850	4,406,850	1	Fusion GGA2-23	Peregrine falcon
3	**74,151,407**	**74,151,581**	**175**	**Z**	**79,904,594**	**79,907,393**	**2,800**	Fusion GGA3-Z	Budgerigar
4	9,445,510	9,582,560	137,051	8	28,728,002	28,767,244	39,243	Fusion GGA4-8	Budgerigar
4	**17,973,596**	**17,973,596**	**1**	**9**	**20,311,785**	**20,311,802**	**18**	Fusion GGA4-9	Budgerigar
4	**19,198,525**	**19,220,202**	**21,678**	NA	NA	NA	NA	Fusion GGA4	Chicken
4	**39,245,686**	**39,245,686**	**1**	**12**	**5,138,109**	**5,138,296**	**188**	Fusion GGA4-12	Downy woodpecker
5	**16,207,082**	**16,208,213**	**1,132**	**20**	**6,982,982**	**6,984,442**	**1,461**	Fusion GGA5-20	Peregrine falcon
6	**15,572,572**	**15,572,572**	**1**	**7**	**21,410,473**	**21,410,473**	**1**	Fusion GGA6-7	Budgerigar
7	**36,208,715**	**36,245,040**	**36,326**	6	8,644,379	8,814,195	169,817	Fusion GGA7-6	Budgerigar
8	**12,732,816**	**12,732,816**	**1**	**9**	**6,326,038**	**6,326,038**	**1**	Fusion GGA8-9	Budgerigar
8	**15,279,906**	**15,279,906**	**1**	**9**	**4,117,527**	**4,118,114**	**588**	Fusion GGA8-9	Budgerigar
9	4,323,360	4,324,043	684	14	7,421,465	7,465,922	44,458	Fusion GGA9-14	Little egret
9	**4,323,360**	**4,324,043**	**684**	**8**	**8,079,812**	**8,111,465**	**31,654**	Fusion GGA9-8	Budgerigar
9	**6,326,038**	**6,326,038**	**1**	**8**	**3,569,408**	**3,622,814**	**53,407**	Fusion GGA9-8	Budgerigar
9	**12,716,025**	**12,716,025**	**1**	**8**	**9,912,294**	**9,912,534**	**241**	Fusion GGA9-8	Budgerigar
14	14,458,809	14,459,897	1,089	5	6,929,776	8,136,232	1,206,456	Fusion GGA14-5	Budgerigar
18	**10,689,190**	**10,689,190**	**1**	23	1,590,960	1,778,610	187,651	Fusion GGA23-18	Downy woodpecker

* In bold are the intervals used in the CNE analysis; NA: not applicable.

Supplemental Table 23: Chicken genome intervals corresponding to lineage-specific intrachromosomal EBRs identified in chromosome assemblies.

Chicken chromosome	Start (bp)	End (bp)	Size (bp)	Lineage
1*	**1,324,368**	**1,324,800**	**432**	Peregrine falcon
1	**6,029,884**	**6,031,169**	**1,285**	Peregrine falcon
1	**7,914,166**	**7,914,732**	**566**	Peregrine falcon
1	**8,882,618**	**8,887,826**	**5,208**	Chicken
1	**9,219,766**	**9,224,154**	**4,388**	Chicken
1	**9,884,967**	**9,885,328**	**361**	Peregrine falcon
1	**26,812,569**	**26,815,922**	**3,353**	Peregrine falcon
1	**28,404,118**	**28,405,020**	**902**	Flycatcher
1	**28,973,698**	**28,974,040**	**342**	Flycatcher
1	**39,784,124**	**39,784,258**	**134**	Peregrine falcon
1	**47,452,522**	**47,453,443**	**921**	Peregrine falcon
1	**47,570,545**	**47,628,260**	**57,715**	Peregrine falcon
1	**54,853,254**	**54,857,707**	**4,453**	Pigeon
1	**56,803,036**	**56,811,099**	**8,063**	Zebra finch
1	**57,155,309**	**57,156,419**	**1,110**	Flycatcher
1	**57,489,115**	**57,491,098**	**1,983**	Flycatcher
1	**58,476,175**	**58,476,241**	**66**	Pigeon
1	**61,930,217**	**61,930,280**	**63**	Flycatcher
1	**62,681,919**	**62,696,997**	**15,078**	Flycatcher
1	**63,084,381**	**63,084,436**	**55**	Flycatcher
1	**64,154,506**	**64,154,603**	**97**	Flycatcher
1	**64,309,525**	**64,311,297**	**1,772**	Flycatcher
1	**65,274,220**	**65,280,569**	**6,349**	Flycatcher
1	**65,338,671**	**65,343,204**	**4,533**	Zebra finch
1	**65,434,145**	**65,441,412**	**7,267**	Flycatcher
1	**66,649,227**	**66,649,812**	**585**	Zebra finch
1	**69,390,834**	**69,391,367**	**533**	Flycatcher
1	69,938,074	70,043,043	104,969	Zebra finch
1	**71,592,439**	**71,593,557**	**1,118**	Flycatcher
1	**71,784,616**	**71,828,201**	**43,585**	Zebra finch
1	**72,413,660**	**72,414,700**	**1,040**	Chicken
1	**72,988,017**	**72,994,267**	**6,250**	Zebra finch
1	**73,158,211**	**73,168,020**	**9,809**	Chicken
1	**74,158,301**	**74,236,372**	**78,071**	Flycatcher
1	74,608,437	75,953,766	1,345,329	Chicken
1	79,266,404	80,323,581	1,057,177	Pigeon
1	**83,482,558**	**83,485,084**	**2,526**	Flycatcher
1	**83,918,384**	**83,934,204**	**15,820**	Zebra finch
1	**86,964,089**	**86,964,101**	**12**	Pigeon
1	**91,171,371**	**91,174,910**	**3,539**	Pigeon
1	**98,960,025**	**98,962,906**	**2,881**	Pigeon
1	**104,819,863**	**104,825,312**	**5,449**	Flycatcher

Chicken chromosome	Start (bp)	End (bp)	Size (bp)	Lineage
1	105,212,045	105,344,966	132,921	Zebra finch
1	**105,370,750**	**105,382,220**	**11,470**	Flycatcher
1	**105,563,181**	**105,564,585**	**1,404**	Flycatcher
1	**105,742,052**	**105,743,256**	**1,204**	Flycatcher
1	**115,150,914**	**115,151,711**	**797**	Pigeon
1	**115,309,400**	**115,312,620**	**3,220**	Pigeon
1	**127,819,431**	**127,841,128**	**21,697**	Peregrine falcon
1	**131,824,670**	**131,827,102**	**2,432**	Pigeon
1	**149,440,203**	**149,444,844**	**4,641**	Zebra finch
1	**179,001,619**	**179,001,763**	**144**	Zebra finch
2	**931,838**	**970,908**	**39,070**	Chicken
2	**4,047,548**	**4,051,400**	**3,852**	Chicken
2	**26,118,162**	**26,118,232**	**70**	Zebra finch
2	55,627,369	56,187,871	560,502	Flycatcher
2	59,440,356	64,180,770	4,740,414	Zebra finch
2	68,914,935	69,067,473	152,538	Pigeon
2	**70,347,550**	**70,349,506**	**1,956**	Peregrine falcon
2	72,395,280	72,502,961	107,681	Pigeon
2	72,631,037	73,494,981	863,944	Zebra finch
2	94,127,281	94,234,930	107,649	Zebra finch
2	**94,829,596**	**94,837,624**	**8,028**	Zebra finch
2	**120,100,222**	**120,100,272**	**50**	Peregrine falcon
2	**121,362,131**	**121,362,153**	**22**	Peregrine falcon
2	**121,563,369**	**121,563,397**	**28**	Peregrine falcon
2	**145,288,500**	**145,289,895**	**1,395**	Peregrine falcon
3	**2,395,028**	**2,406,457**	**11,429**	Chicken
3	**11,266,541**	**11,271,715**	**5,174**	Peregrine falcon
3	**11,571,165**	**11,619,462**	**48,297**	Chicken
3	**11,929,811**	**11,953,549**	**23,738**	Chicken
3	**12,514,375**	**12,514,413**	**38**	Zebra finch
3	**13,572,601**	**13,572,601**	**-**	Flycatcher
3	**14,195,505**	**14,195,923**	**418**	Flycatcher
3	**16,277,015**	**16,282,517**	**5,502**	Chicken
3	**17,015,103**	**17,057,419**	**42,316**	Zebra finch
3	**24,696,708**	**24,696,885**	**177**	Flycatcher
3	**26,123,868**	**26,126,691**	**2,823**	Zebra finch
3	**26,747,128**	**26,749,339**	**2,211**	Flycatcher
3	**28,440,655**	**28,440,817**	**162**	Zebra finch
3	**40,468,377**	**40,468,377**	**-**	Peregrine falcon
3	**70,642,811**	**70,676,596**	**33,785**	Peregrine falcon
3	**106,856,633**	**106,856,944**	**311**	Peregrine falcon
4	**2,020,905**	**2,022,164**	**1,259**	Pigeon
4	5,007,908	5,141,468	133,560	Pigeon
4	13,760,221	13,931,036	170,815	Pigeon

Chicken chromosome	Start (bp)	End (bp)	Size (bp)	Lineage
4	**16,440,524**	**16,454,171**	**13,647**	Pigeon
4	**20,643,486**	**20,643,751**	**265**	Peregrine falcon
4	**20,773,991**	**20,774,533**	**542**	Peregrine falcon
4	28,278,509	28,823,040	544,531	Zebra finch
4	**34,474,454**	**34,474,907**	**453**	Chicken
4	**36,711,371**	**36,716,488**	**5,117**	Chicken
4	**36,968,524**	**37,003,727**	**35,203**	Flycatcher
4	**38,056,016**	**38,057,837**	**1,821**	Flycatcher
4	38,688,788	38,817,468	128,680	Chicken
4	**39,418,367**	**39,438,868**	**20,501**	Pigeon
4	41,738,397	42,076,213	337,816	Chicken
4	**44,024,013**	**44,091,739**	**67,726**	Chicken
4	**46,254,077**	**46,254,681**	**604**	Zebra finch
4	**57,795,684**	**57,801,782**	**6,098**	Chicken
4	**60,643,171**	**60,660,962**	**17,791**	Chicken
4	**61,297,325**	**61,300,911**	**3,586**	Peregrine falcon
4	**68,580,126**	**68,581,362**	**1,236**	Peregrine falcon
4	75,558,167	76,214,491	656,324	Peregrine falcon
5	3,064,699	3,226,802	162,103	Chicken
5	3,505,172	3,607,787	102,615	Zebra finch
5	**3,920,467**	**3,921,806**	**1,339**	Zebra finch
5	**4,207,837**	**4,225,827**	**17,990**	Zebra finch
5	5,681,347	5,833,445	152,098	Chicken
5	**6,507,620**	**6,535,656**	**28,036**	Chicken
5	**11,920,644**	**11,924,718**	**4,074**	Peregrine falcon
5	**16,013,932**	**16,016,270**	**2,338**	Chicken
5	16,162,195	16,419,626	257,431	Chicken
5	**17,825,030**	**17,827,020**	**1,990**	Flycatcher
5	**18,002,051**	**18,002,505**	**454**	Chicken
5	**22,198,087**	**22,198,280**	**193**	Peregrine falcon
5	**40,665,557**	**40,666,511**	**954**	Pigeon
5	**42,299,980**	**42,301,918**	**1,938**	Peregrine falcon
5	**50,528,897**	**50,531,950**	**3,053**	Peregrine falcon
5	**57,848,407**	**57,863,325**	**14,918**	Peregrine falcon
6	**2,058,825**	**2,126,442**	**67,617**	Pigeon
6	**14,957,776**	**14,958,505**	**729**	Peregrine falcon
6	**17,123,426**	**17,170,023**	**46,597**	Peregrine falcon
6	**29,924,390**	**29,936,355**	**11,965**	Peregrine falcon
7	**999,837**	**1,005,793**	**5,956**	Peregrine falcon
7	**1,368,850**	**1,406,526**	**37,676**	Pigeon
7	**3,526,408**	**3,569,092**	**42,684**	Zebra finch
7	**4,497,339**	**4,501,801**	**4,462**	Zebra finch
7	**5,536,576**	**5,569,739**	**33,163**	Zebra finch
7	**9,278,452**	**9,280,798**	**2,346**	Pigeon

Chicken chromosome	Start (bp)	End (bp)	Size (bp)	Lineage
7	**9,813,362**	**9,813,362**	**-**	Chicken
7	**10,014,120**	**10,014,120**	**-**	Chicken
7	**17,863,268**	**17,864,755**	**1,487**	Flycatcher
7	**18,283,306**	**18,283,720**	**414**	Flycatcher
7	**27,805,170**	**27,805,281**	**111**	Zebra finch
8	**5,073,955**	**5,073,999**	**44**	Chicken
8	**7,401,061**	**7,405,478**	**4,417**	Pigeon
8	8,030,791	10,005,834	1,975,043	Chicken
8	18,511,434	19,999,999	1,488,565	Pigeon
8	22,246,010	23,004,261	758,251	Pigeon
8	**23,775,554**	**23,780,105**	**4,551**	Flycatcher
9	**4,116,694**	**4,118,012**	**1,318**	Zebra finch
9	**4,321,288**	**4,322,902**	**1,614**	Chicken
9	4,712,939	4,863,290	150,351	Chicken
9	**12,394,835**	**12,394,985**	**150**	Pigeon
9	**15,162,865**	**15,171,622**	**8,757**	Pigeon
9	**19,341,833**	**19,344,688**	**2,855**	Pigeon
10	**2,134,590**	**2,135,412**	**822**	Chicken
10	2,456,474	3,089,337	632,863	Peregrine falcon
10	**3,531,438**	**3,531,937**	**499**	Pigeon
10	**3,607,425**	**3,609,795**	**2,370**	Zebra finch
10	**3,609,846**	**3,648,687**	**38,841**	Flycatcher
10	**7,421,814**	**7,422,624**	**810**	Pigeon
10	**17,999,142**	**18,000,598**	**1,456**	Flycatcher
11	**338,764**	**348,457**	**9,693**	Chicken
11	**2,175,213**	**2,176,532**	**1,319**	Chicken
11	**2,314,761**	**2,315,690**	**929**	Chicken
11	**2,638,020**	**2,703,827**	**65,807**	Chicken
11	**3,161,385**	**3,162,962**	**1,577**	Pigeon
11	**5,983,232**	**5,992,779**	**9,547**	Zebra finch
11	**7,344,225**	**7,381,817**	**37,592**	Pigeon
11	**13,237,592**	**13,237,950**	**358**	Zebra finch
11	**14,586,964**	**14,586,982**	**18**	Pigeon
11	**17,092,230**	**17,103,474**	**11,244**	Zebra finch
12	**8,244,390**	**8,244,424**	**34**	Pigeon
12	**8,362,979**	**8,367,099**	**4,120**	Pigeon
12	**8,508,581**	**8,509,418**	**837**	Pigeon
12	**18,624,938**	**18,625,329**	**391**	Pigeon
13	**2,656,250**	**2,658,094**	**1,844**	Pigeon
13	**4,072,035**	**4,093,791**	**21,756**	Pigeon
13	**7,522,370**	**7,522,376**	**6**	Pigeon
13	**16,501,929**	**16,564,665**	**62,736**	Peregrine falcon
13	**17,498,657**	**17,504,153**	**5,496**	Pigeon
14	**6,916,159**	**6,926,191**	**10,032**	Flycatcher

Chicken chromosome	Start (bp)	End (bp)	Size (bp)	Lineage
14	**7,258,840**	**7,264,491**	**5,651**	Zebra finch
14	**8,708,224**	**8,722,870**	**14,646**	Pigeon
14	**8,900,773**	**8,904,651**	**3,878**	Pigeon
14	**13,597,365**	**13,671,981**	**74,616**	Chicken
14	**14,286,593**	**14,306,071**	**19,478**	Chicken
14	**14,431,346**	**14,459,945**	**28,599**	Chicken
14	**14,814,615**	**14,904,537**	**89,922**	Chicken
15	1,309,743	1,416,134	106,391	Pigeon
15	**7,846,096**	**7,847,734**	**1,638**	Chicken
15	9,909,742	10,030,648	120,906	Chicken
17	**8,315,856**	**8,316,842**	**986**	Peregrine falcon
18	3,489,677	4,326,327	836,650	Peregrine falcon
18	**4,442,787**	**4,445,807**	**3,020**	Pigeon
18	**4,592,395**	**4,600,415**	**8,020**	Pigeon
18	**4,812,654**	**4,815,509**	**2,855**	Chicken
18	**5,037,378**	**5,037,399**	**21**	Chicken
18	**6,003,931**	**6,011,826**	**7,895**	Pigeon
18	**6,541,981**	**6,543,165**	**1,184**	Zebra finch
18	**6,624,558**	**6,626,204**	**1,646**	Peregrine falcon
18	8,243,821	8,460,485	216,664	Zebra finch
18	**10,249,436**	**10,253,576**	**4,140**	Chicken
19	2,491,720	2,931,925	440,205	Pigeon
20	**937,650**	**943,963**	**6,313**	Flycatcher
20	1,642,252	4,945,945	3,303,693	Flycatcher
20	**11,104,290**	**11,105,020**	**730**	Chicken
21	**171,731**	**172,267**	**536**	Chicken
21	**2,447,743**	**2,520,267**	**72,524**	Zebra finch
21	**2,534,001**	**2,534,006**	**5**	Peregrine falcon
21	**2,780,061**	**2,780,375**	**314**	Peregrine falcon
21	**3,031,447**	**3,031,929**	**482**	Peregrine falcon
21	**4,201,584**	**4,202,188**	**604**	Chicken
21	**4,423,965**	**4,425,044**	**1,079**	Zebra finch
21	**4,947,252**	**4,948,468**	**1,216**	Flycatcher
22	**454,415**	**456,267**	**1,852**	Chicken
22	**2,750,596**	**2,783,326**	**32,730**	Chicken
23	**3,151,880**	**3,208,232**	**56,352**	Chicken
23	**4,554,121**	**4,554,183**	**62**	Peregrine falcon
23	4,802,961	5,045,626	242,665	Zebra finch
24	**657,347**	**660,000**	**2,653**	Zebra finch
24	**2,899,660**	**2,905,523**	**5,863**	Zebra finch
24	**3,194,369**	**3,194,441**	**72**	Peregrine falcon
24	**3,786,014**	**3,788,248**	**2,234**	Pigeon
24	**4,336,751**	**4,336,769**	**18**	Chicken
24	**5,498,139**	**5,570,339**	**72,200**	Chicken

Chicken chromosome	Start (bp)	End (bp)	Size (bp)	Lineage
24	**5,707,069**	**5,708,160**	**1,091**	Chicken
24	**6,111,047**	**6,114,494**	**3,447**	Peregrine falcon
26	1,618,988	2,124,876	505,888	Chicken
26	**2,427,305**	**2,427,873**	**568**	Pigeon
26	**2,581,525**	**2,589,502**	**7,977**	Zebra finch
26	**3,150,633**	**3,152,441**	**1,808**	Flycatcher
26	**3,416,031**	**3,417,505**	**1,474**	Pigeon
26	**4,017,799**	**4,019,987**	**2,188**	Chicken
27	**1,551,217**	**1,587,888**	**36,671**	Zebra finch
27	**1,660,471**	**1,662,243**	**1,772**	Pigeon
27	**2,323,210**	**2,330,928**	**7,718**	Pigeon
27	**2,828,063**	**2,837,768**	**9,705**	Pigeon
28	**4,082,833**	**4,084,257**	**1,424**	Flycatcher
Z	**1,814,281**	**1,815,225**	**944**	Zebra finch
Z	7,193,324	7,329,824	136,500	Flycatcher
Z	**12,772,165**	**12,776,763**	**4,598**	Pigeon
Z	19,484,554	22,168,262	2,683,708	Zebra finch
Z	23,034,483	23,646,027	611,544	Zebra finch
Z	**30,601,213**	**30,617,877**	**16,664**	Chicken
Z	**30,720,869**	**30,721,276**	**407**	Chicken
Z	36,523,733	36,904,832	381,099	Flycatcher
Z	**37,191,363**	**37,208,200**	**16,837**	Pigeon
Z	40,672,853	41,092,932	420,079	Zebra finch

* In bold are the interval used in the CNE analysis.

Supplemental Table 24: Chicken genome intervals corresponding to lineage-specific interchromosomal EBRs identified in chromosome assemblies.

Chicken chromosome	Start (bp)	End (bp)	Size (bp)	Chicken chromosome	Start (bp)	End (bp)	Size (bp)	EBR type	Lineage
1*	**74,604,570**	**74,642,939**	**38,369**	NA	NA	NA	NA	Fission GGA1	Peregrine falcon, flycatcher, zebra finch
2	**1**	**20,713**	**20,712**	ND	ND	ND	ND	Fusion GGA2-28	Peregrine falcon
2	48,696,437	64,084,268	15,387,831	NA	NA	NA	NA	Fission GGA2	Peregrine falcon
2	48,696,437	64,084,268	15,387,831	ND	ND	ND	ND	Fusion GGA2-21	Peregrine falcon
3	**33,171,474**	**33,191,237**	**19,763**	NA	NA	NA	NA	Fission GGA3	Peregrine falcon
4	**19,198,385**	**19,203,158**	**4,773**	NA	NA	NA	NA	Fusion GGA4	Chicken
5	**16,207,082**	**16,208,213**	**1,131**	**20**	**6,982,982**	**6,984,442**	**1,460**	Fusion GGA5-20	Peregrine falcon
5	**18,002,366**	**18,002,434**	**68**	NA	NA	NA	NA	Fission GGA5	Peregrine falcon
5	**59,525,950**	**59,526,594**	**644**	**10**	**2,429,267**	**2,429,267**	**-**	Fusion GGA5-10	Peregrine falcon
6	**7,240,813**	**7,240,915**	**102**	**17**	**5,876,933**	**5,880,929**	**3,996**	Fusion GGA6-17	Peregrine falcon
13	**1**	**2,563**	**2,562**	ND	ND	ND	ND	Fusion GGA7-13	Peregrine falcon
14	**7,458,981**	**7,459,764**	**783**	**12**	**19,868,308**	**19,897,011**	**28,703**	Fusion GGA14-12	Peregrine falcon
15	**3,695,551**	**3,702,477**	**6,926**	**19**	**1**	**66,909**	**66,908**	Fusion GGA15-19	Peregrine falcon
17	**10,394,199**	**10,454,150**	**59,951**	**5**	**24,832,742**	**24,843,616**	**10,874**	Fusion GGA17-5	Peregrine falcon
19	**9,942,178**	**9,962,510**	**20,332**	**18**	**4,437,561**	**4,437,561**	**-**	Fusion GGA19-18	Peregrine falcon
28	**4,042,843**	**4,042,855**	**12**	**14**	**650,373**	**666,763**	**16,390**	Fusion GGA28-14	Peregrine falcon
23	**4,406,830**	**4,406,830**	**-**	ND	ND	ND	ND	Fusion GGA23-21	Peregrine falcon
15	**3,695,551**	**3,702,477**	**6,926**	ND	ND	ND	ND	Fusion GGA15-4	Peregrine falcon

* In bold are the interval used in the CNE analysis; NA: not applicable; ND: not determined.