UPPSALA
UNIVERSITET

*Digital Comprehensive Summaries of Uppsala Dissertations
from the Faculty of Science and Technology* 2061

The exploration and evolution of the avian genomic dark matter

VALENTINA PEONA

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2021

ISSN 1651-6214
ISBN 978-91-513-1264-4
URN urn:nbn:se:uu:diva-450756

Dissertation presented at Uppsala University to be publicly examined in Ekmansalen, EBC, Norbyvägen 14, Uppsala, Thursday, 30 September 2021 at 14:00 for the degree of Doctor of Philosophy. The examination will be conducted in English. Faculty examiner: Professor Rachel O'Neill (Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT, USA).

Abstract

Peona, V. 2021. The exploration and evolution of the avian genomic dark matter. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 2061. 84 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-513-1264-4.

The development and improvement of genome sequencing technologies in the last decade revolutionised the entire field of biology with genome assemblies of virtually any organism. Despite this tremendous progress, complex genomic regions are systematically missing from genome assemblies and form the so-called "genomic dark matter". The presence of genomic dark matter entails that such regions cannot be fully studied and the effects and/or functions thereof (if any) on the organisms remain hidden. Therefore, it is key to be able to explore those dark genomic corners to fully understand the evolution and physiology of organisms without biasing the interpretations. In this thesis, I contribute to the understanding of the use of new sequencing technologies to assemble complex genomic regions and to investigate the evolution of such regions throughout the avian phylogeny. First, I assessed the best combination of technologies and assembly methods to maximise the resolution of genomic dark matter using genomic data from the paradise crow. This included testing for the presence of repetitive elements, GC-rich regions, G-quadruplex motifs, non-recombining sex chromosomes, and microchromosomes. Then, the high-quality assemblies for the paradise crow and other birds allowed the discovery that the avian W chromosome features more than half of potentially active transposable elements (TEs), especially endogenous retroviruses, of the genome. This characteristic makes the W chromosome potentially "toxic" for females. The female-biased accumulation of active TEs could also play a role in the origin of genetic incompatibilities and be an explanatory variable for Haldane's rule in birds. Next, I investigated the genetic variability of birds-of-paradise chromosomes originating from structural rearrangements with a special focus on the W chromosome. The analysis revealed more genetic variability than previously reported suggesting that all sources of genetic variability should be considered to understand the evolution of sex-limited chromosomes. Finally, I explored the evolution of another main component of avian genomic dark matter, satellite DNA, throughout the phylogeny of birds-of-paradise and closely related crow species. I found that the avian satellitome evolves in different modes in the two groups and a more comprehensive species sampling is necessary to establish which evolutionary mode is the most prevalent in birds. Altogether, the results of this thesis provide a case study for how to investigate the most complex genomic regions, highlight their possible evolutionary roles, and therefore showcase the necessity for the field to shed light into the dark corners of genomes. Mind the gap!

Keywords: evolutionary genomics, genome assembly, sex chromosomes, birds, satellite DNA, transposable elements, endogenous retroviruses

Valentina Peona, Department of Organismal Biology, Systematic Biology, Norbyv. 18 D, Uppsala University, SE-75236 Uppsala, Sweden.

ISSN 1651-6214
ISBN 978-91-513-1264-4
URN urn:nbn:se:uu:diva-450756 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-450756)

To imperfection

List of Papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I. Peona, V.*, Weissensteiner, M.H.*, Suh, A., 2018. How complete are "complete" genome assemblies?—an avian perspective. *Molecular Ecology Resources* **18**, 1188–1195.

II. Peona, V., Blom, M.P.K., Xu, L., Burri, R., Sullivan, S., Bunikis, I., Liachko, I., Haryoko, T., Jønsson, K.A., Zhou, Q., Irestedt, M., Suh, A., 2021. Identifying the causes and consequences of assembly gaps using a multiplatform genome assembly of a bird-of-paradise. *Molecular Ecology Resources* **21**, 263–286.

III. Peona, V., Palacios-Gimenez, O.M., Blommaert, J., Liu, J., Haryoko, T., Jønsson, K.A., Irestedt, M., Zhou, Q., Jern, P., Suh, A., 2021. The avian W chromosome is a refugium for endogenous retroviruses with likely effects on female-biased mutational load and genetic incompatibilities. *Philosophical Transactions of the Royal Society B* **376**, 20200186.

IV. Peona, V., Blom, M.P.K., Frankl-Vilches, C., Milá, B., Ashari, H., Thébaud, C., Benz, B.W., Christidis, L., Gahr, M., Irestedt, M., Suh, A., 2021. The hidden structural variability in avian genomes. *Manuscript.*

V. Peona, V., Kutschera, V.E., Blom, M.P.K., Irestedt, M., Suh, A., 2021. Satellite DNA evolution in Corvides inferred from short and long reads. *Manuscript.*

*equal contributions

Reprints were made with permission from the respective publishers.

Additional Papers

The following papers were published during the course of my doctoral studies but are not part of this thesis.

I. Ricci, M.*, **Peona, V.***, Guichard, E., Taccioli, C., Boattini, A., 2018. Transposable elements activity is positively related to rate of speciation in mammals. *Journal of Molecular Evolution* **86**, 303–310.

II. Guichard, E.*, **Peona, V.***, Malagoli Tagliazucchi, G., Abitante, L., Jagoda, E., Musella, M., Ricci, M., Rubio-Roldán, A., Sarno, S., Luiselli, D., Pettener, D., Taccioli, C., Pagani, L., Garcia-Perez, J.L., Boattini, A., 2018. Impact of non-LTR retrotransposons in the differentiation and evolution of anatomically modern humans. *Mobile DNA* **9**, 28.

III. Xu, L., Auer, G., **Peona, V.**, Suh, A., Deng, Y., Feng, S., Zhang, G., Blom, M.P.K., Christidis, L., Prost, S., Irestedt, M., Zhou, Q., 2019. Dynamic evolutionary history and gene content of sex chromosomes across diverse songbirds. *Nature Ecology and Evolution* **3**, 834–844.

IV. Weissensteiner, M.H., Bunikis, I., Catalán, A., Francoijs, K.-J., Knief, U., Heim, W., **Peona, V.**, Pophaly, S.D., Sedlazeck, F.J., Suh, A., Warmuth, V.M., Wolf, J.B.W., 2020. Discovery and population genomics of structural variation in a songbird genus. *Nature Communications* **11**, 3403.

V. Gemmell, N.J., Rutherford, K., Prost, S., Tollis, M., Winter, D., Macey, J.R., Adelson, D.L., Suh, A., Bertozzi, T., Grau, J.H., Organ, C., Gardner, P.P., Muffato, M., Patricio, M., Billis, K., Martin, F.J., Flicek, P., Petersen, B., Kang, L., Michalak, P., Buckley, T.R., Wilson, M., Cheng, Y., Miller, H., Schott, R.K., Jordan, M.D., Newcomb, R.D., Arroyo, J.I., Valenzuela, N., Hore, T.A., Renart, J., **Peona, V.**, Peart, C.R., Warmuth, V.M., Zeng, L., Kortschak, R.D., Raison, J.M., Zapata, V.V., Wu, Z., Santesmasses, D., Mariotti, M., Guigó, R., Rupp, S.M., Twort, V.G., Dussex, N., Taylor, H., Abe, H., Bond, D.M., Paterson, J.M., Mulcahy, D.G., Gonzalez, V.L., Barbieri, C.G., DeMeo, D.P., Pabinger, S., Van Stijn, T., Clarke, S., Ryder, O., Edwards, S. V, Salzberg, S.L., Anderson, L., Nelson, N., Stone, C., Stone, C., Smillie, J., Edmonds, H., 2020. The tuatara genome reveals ancient features of amniote evolution. *Nature* **584**, 403–409.

VI. Christmas, M.J., Jones, J.C., Olsson, A., Wallerman, O., Bunikis, I., Kierczak, M., **Peona, V.**, Whitley, K.M., Larva, T., Suh, A., Miller-Struttmann, N.E., Geib, J.C., Webster, M.T., 2021. Genetic barriers to historical gene flow between cryptic species of alpine bumblebees revealed by comparative population genomics. *Molecular Biology and Evolution* **38**, 3126–3143.

VII. Nguyen, D., **Peona, V.**, Unneberg, P., Suh, A., Jern, P., Johannesson, H., 2020. Transposon- and genome dynamics in the fungal genus Neurospora: insights from nearly gapless genome assemblies. *bioRxiv* 2020.09.27.311811.

VIII. Huang, Z., Furo, I., **Peona, V.**, Liu, J., Gomes, A.J.B., Cen, W., Huang, H., Zhang, Y., Chen, D., Ting, X., Chen, Y., Zhang, Q., Yue, Z., Suh, A., de Oliveira, E.H.C., Xu, L., 2021. Recurrent chromosome reshuffling and the evolution of neo-sex chromosomes in parrots. *bioRxiv* 2021.03.08.434498.

IX. Robert, A., **Peona, V.**, Ottenburghs, J., 2021. Digest: Population genomics reveals convergence toward melanism in different island populations. *Evolution* **75**, 1582–1584.

*equal contributions

Contents

Abbreviations

BOP	Birds-of-paradise
CR1	Chicken repeat 1
ERV	Endogenous retrovirus
G4	G-quadruplex
HMW	High-molecular weight
HOR	Higher order repeat
HTS	High-throughput sequencing
LINE	Long interspersed element
LTR	Long terminal repeat
NGS	Next Generation Sequencing
ORF	Open reading frame
satDNA	Satellite DNA
SNP	Single nucleotide point mutation
SV	Structural variant
T2T	Telomere-to-telomere consortium
TE	Transposable element
TSD	Target site duplication
VGP	Vertebrate Genome Project

1. Introduction

Repetitive elements are ubiquitous features of eukaryotic genomes that can make up a significant portion of genomes (Sotero-Caio *et al.*, 2017). Repetitive elements are usually divided into tandem repeats (microsatellites, minisatellites, and satellites) and interspersed elements such as transposable elements (TEs) that are able to actively move throughout the host genome. Transposable elements in particular can be considered as parasitic elements given their potential disruptiveness and tendency to increase in copy number by not following a Mendelian pattern of inheritance (Orgel and Crick, 1980). Transposable element insertions, especially on a short time scale, can be highly detrimental for the organism if they disrupt genes and gene networks (Klein and O'Neill, 2018). Though, on a longer evolutionary time scale, individual TE copies can lead to important evolutionary features (Broecker and Moelling, 2019; Schrader and Schmitz, 2019; Domínguez *et al.*, 2020; Senft and Macfarlan, 2021). Indeed, TEs played essential roles in evolution, for example their co-option in the development of placenta in mammals (Emera and Wagner, 2012), the V(D)J immune system in vertebrates (Kapitonov and Koonin, 2015), and telomeres in *Drosophila* (Pardue and DeBaryshe, 2003). They also act as regulatory elements of gene expression and chromatin state (Chuong, Elde and Feschotte, 2017) while representing a rich source of genomic variation (Schrader and Schmitz, 2019). Their disruptiveness triggered the evolution of a variety of defence mechanisms that work through histone modification, posttranscriptional silencing, and DNA hypermutation (Galagan and Selker, 2004; Deniz, Frost and Branco, 2019; Ozata *et al.*, 2019).

Nowadays, repetitive elements remain generally understudied with respect to other sources of genetic variation, mainly because of the intrinsic difficulty in investigating them. Their genomic characterisation (e.g., discovery, categorisation, distribution) heavily relies upon the quality of genome assemblies; quality that is hampered by repeats themselves (Sedlazeck *et al.*, 2018). This is a vicious cycle in which repeats are systematically underrepresented in genome assemblies, therefore are not fully taken into consideration during analysis, and the possible roles of repeats in biological phenomena are overlooked. This results in repeats mainly being considered as an assembly problem and masked away during analysis, increasing the impression that repeats do not have any effect on the biology of the genome. The effect of single insertions can range from being completely deleterious to strongly beneficial, but the key point to recognise is that repeats have the potential to influence many biological phenomena (e.g., selection, methylation patterns, gene expression (Lerat *et al.*, 2019). Therefore, it is important to take repetitive elements into consideration when performing genomic analysis (Slotkin, 2018).

Genome assemblies are often incomplete and miss many repetitive regions along with extremely AT-rich or GC-rich regions, such as telomeres, centromeres, the multicopy gene family of the Major Histocompatibility Complex (MHC), and the degenerate non-recombining sex chromosomes (Y/W). The systematic underrepresentation of these regions gave them the evocative name of genomic "dark matter" (Johnson *et al.*, 2005; Sedlazeck *et al.*, 2018): missing sequences whose nature is not well characterised. In this thesis, I study the evolution of repetitive elements in avian genomes, focusing on resolving previously elusive dark matter constituted by transposable elements and satellite DNA that are especially enriched on the W chromosome and at the centromeres. In my papers, I mostly use genomic data from birds-of-paradise (Corvides: Paradisaeidae family), other publicly available Corvides species, as well as chicken, zebra finch, emu, kakapo, Anna's hummingbird and blue-capped cordon-bleu. My choice to study repeats in bird genomes is based on the fact that bird genomes are overall repeat-poor. Although it may sound paradoxical, the depletion of repeats makes high-quality genome assemblies easier to achieve as well as a faithful representation of repeats. Not all the regions of bird genomes are "easy" to investigate, indeed centromeres, GC-rich microchromosomes, and the non-recombining W chromosome are all very challenging to assemble and study (Kapusta and Suh, 2017). Among birds, birds-of-paradise are a family of 40 species that evolved the most spectacular and diverse phenotypes in relatively short time (~15 million years) (Irestedt *et al.*, 2009) under strong sexual selection while subject to multiple events of hybridisation (Blom and Irestedt, 2021). Such framework provides a good chance to investigate the turnover of both tandem and interspersed repeats between species and sexes.

As mentioned above, genome assembly quality and completeness are key to study repeats (and many other genomic features), but there is still confusion about what quality and completeness mean. In **Paper I**, I focus on this problem by describing the state-of-the-art of genome completeness measures while examining the actual completeness of "complete genome" assemblies available. In **Paper II**, I continue to dig into the issue of assembly quality by testing the efficiency of the currently most used sequencing and scaffolding technologies in resolving genomic dark matter of the paradise crow (Paradisaeidae, birds-of-paradise). While testing these technologies, I also investigate the most common causes for assembly fragmentation. Starting in **Paper III**, I use the insights from the previous papers to investigate the evolution of what was previously dark matter in (avian) genome assemblies. In **Paper III**, I explore the potentially toxic transposable element activity specific to female individuals stemming from the W chromosome using a combination of genomic, transcriptomic, and proteomic data from six avian species. In **Paper IV**, I continue to investigate the evolution of the repetitive content of the W chromosome together with its structural changes in species of birds-of-paradise and estrildid finches to assess the presence of previously "hidden" genetic variability. Lastly, in **Paper V**, I explore the diversity, turnover, and structure of satellite DNA throughout the birds-of-paradise phylogeny and in the closely related *Corvus* genus (Corvides: Corvidae family) using a combination of short-read and long-read sequencing libraries and genome assemblies.

2. Genome sequencing and assembly

Sequencing and dark matter

Since the discovery of DNA as the heredity material and its double helix structure, sequencing DNA has become a central endeavour for biologists to shed light on the nature of genes and genomes. Sequencing DNA has never been an easy task though. The first attempts involved the transcription of DNA into small pieces of RNA, resulting in long processing times. For instance, it took three years to sequence 76 nucleotides of the yeast alanine tRNA in 1965 (Holley *et al.*, 1965). In the following decade, Sanger and colleagues revolutionised DNA sequencing (Sanger, Nicklen and Coulson, 1977). Sanger sequencing was still a slow (compared to modern technologies), time-consuming, and an expensive process that required a thorough manual curation. Indeed, the correction of the sequencing data was achieved through manual inspection of electropherograms, and therefore was a process prone to subjectivity and inefficient for large scale projects. Sanger sequencing reached its maximum development in the 1990's and early 2000's with the Human Genome Project (Lander *et al.*, 2001). During this colossal project, automatised and streamlined ways to sequence, read, and correct the data were developed. For example, the automated Phred scoring system (Ewing and Green, 1998; Ewing *et al.*, 1998) to quantify the quality of each base read was introduced in those years and would be used, although slightly modified, for the sequencing technologies to come.

Sanger sequencing largely remained the main sequencing technology until the advent of high-throughput technologies (HTS) (Shendure *et al.*, 2017), also known as Next Generation Sequencing (NGS), that vastly simplified, sped up, and reduced the prices of sequencing (**Figure 2.1**). Indeed, NGS has allowed even small laboratories to sequence and assemble virtually any genome without necessarily a consortium behind the projects. As a consequence, the number of species sequenced increased enormously in the last decade (**Figure 2.2**) as absolute values and rate. Numerous sequencing platforms can be ascribed to the category of NGS (e.g., Solexa/Illumina, Roche 454, IonTorrent), but for the scopes of this thesis I will take the Illumina platforms and protocols as representative of the group. NGS is based on the sequencing of small DNA fragments from whole genome shotgun libraries that results in the collection (library) of millions of short reads of 75-150 bp each (Shendure *et al.*, 2017).

Below, I briefly describe the library preparation and sequencing process of an Illumina platform to highlight the steps from which the genomic "dark matter" mainly originates.

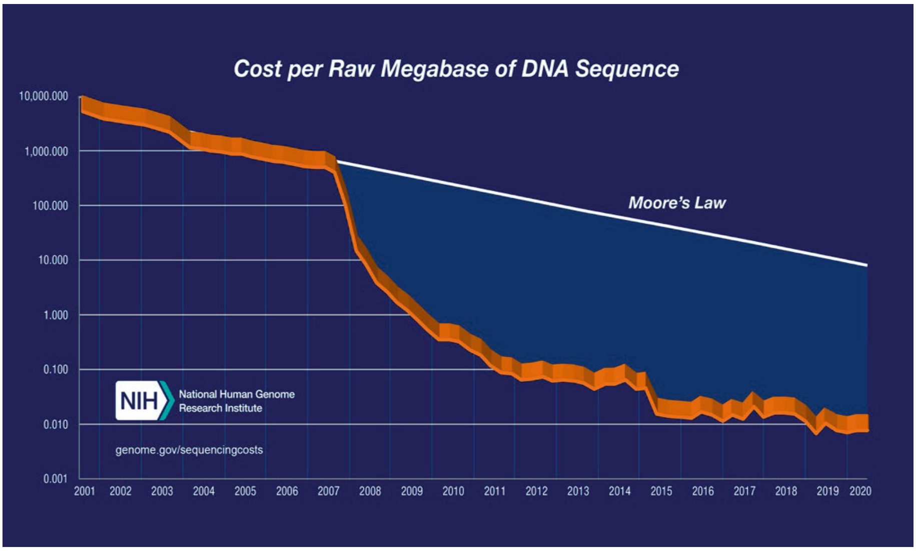


Figure 2.1 Sequencing cost in US dollars per raw megabase of DNA from September 2001 to August 2020. Graph made freely available by the National Human Genome Research Institute (NHGRI; https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data).

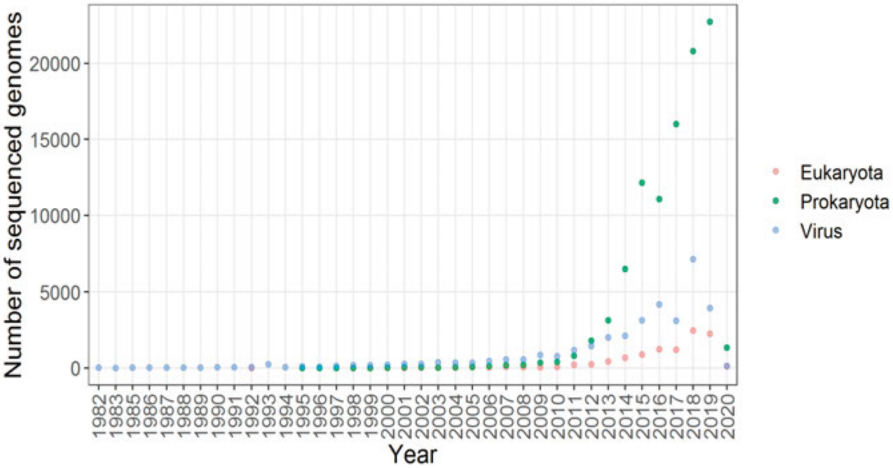


Figure 2.2 Number of sequenced genomes per year over the last forty years. Data collected from the National Center for Biotechnology Information (NCBI; https://www.ncbi.nlm.nih.gov/) on 04 February 2020.

Whole-genome shotgun library preparation and PCR amplification

The term “whole-genome shotgun” indicates that the extracted DNA of interest is randomly sheared into small pieces by sonication or enzymatic digestion. The library is then filtered to select fragments of a certain size suitable for the sequencing platform. The obtained DNA library consisting of millions of short fragments is then ligated to both ends with adapters (short DNA sequences) necessary as primers for the following amplification step. After ligation, the library is ready to be loaded onto the flow cells of the sequencing platform. A flow cell is a glass slide with multiple lanes, the surface of which is covered with oligonucleotides complementary to the adapters used for the library preparation. The DNA fragments flow through the cell and hybridise with the complementary adapters on the flow cell surface; the molecules that do not hybridise with their complementary oligonucleotides are washed away before the amplification. On the flow cell surface, the DNA fragments are replicated through a “bridge PCR” amplification to form small clusters (clonal groups) of the same sequence (Quail *et al.*, 2008). The amplification step is key to yield high sequencing accuracy but at the same time may introduce erroneous bases or be non-homogeneous throughout the library because of issues related to base composition (Kozarewa *et al.*, 2009; Oyola *et al.*, 2012).

Sequencing by synthesis (SBS)

Once the amplification is done, the DNA sequences attached to the flow cells are turned into single-stranded molecules and the actual sequencing can start. The sequencing relies on cycles of incorporation of fluorescently labelled deoxynucleotides (dNTP). At every cycle, a set of chain terminator dNTPs (A/C/T/G) is added to the flow cells and only one dNTP at the time will be incorporated to the nascent single-stranded molecules starting from the adapters/primers. At the end of each cycle, the fluorescent emission of each cluster is recorded by a digital camera. The wavelength and intensity of the signal are analysed to respectively identify the newly incorporated nucleotides (base calling) and to quantify the quality of every base called. The fluorescent label is then enzymatically cleaved and a new cycle can start. Most of the Illumina platforms generate reads of 75-150 base pairs long. As each molecule needs to be synthesised through nucleotide incorporation in order to be sequenced, this method is called sequencing by synthesis (SBS) (Ambardar *et al.*, 2016). The ultimate product of the sequencing process is a FASTQ file of millions of short reads where the bases are recorded together with their quality scores based on the Phred scale. The method described above generates single-read data, namely for each DNA fragment only one end is sequenced. This method can be modified in order to obtain paired-end and mate-pair read libraries that are useful to increase the genomic coverage and also for scaffolding purposes, because both ends of each DNA fragment yield a respective short read.

Short reads and genomic dark matter

During the process of Illumina sequencing, there are two steps that may introduce sequencing errors and prevent a faithful representation of the genome.

First, PCR does not easily amplify genomic regions that are extremely rich in AT or GC bases. This issue arises from the fact that GC-rich regions tend to have higher melting temperatures than the rest of the genome and are not as accessible with standard PCR protocols. Conversely, AT-rich regions require a lower melting and extension temperature than the rest of the genome (Su *et al.*, 1996). This biased amplification problem is not exclusive of the Illumina library protocol, indeed PCR problems always arose as exemplified by the *Plasmodium falciparum* sequencing project (Dame *et al.*, 1996). The *P. falciparum* genome is extremely AT-rich (~70%) and the researchers involved in the project needed to design specific amplification protocols to be able to effectively amplify and Sanger sequence the genome (Su *et al.*, 1996). Similarly, also extremely GC-rich regions are systematically missing from short-read libraries due to amplification issues. Several extraction and amplification protocols have been developed to overcome this problem, but they do not solve the two sides of the problem (AT and GC richness) at once (Kozarewa *et al.*, 2009; Oyola *et al.*, 2012).

Second, the sequencing by synthesis step may also bias the general genomic representation when dealing with GC-rich sequences. It has been proposed (Martin-Gallardo *et al.*, 1992; Stein, Takasuka and Collings, 2009) that single-stranded DNA tertiary structures may hamper the correct incorporation of nucleotides, therefore introducing sequencing errors or causing the sequencing process to stop.

Finally, the underrepresentation of AT-rich and GC-rich regions can lead to erroneous conclusions (e.g., regarding genome size estimation based on kmers) and incomplete genome assemblies (e.g., incorrect absence of genes (Lovell *et al.*, 2014; Hron *et al.*, 2015; Botero-Castro *et al.*, 2017). Although it is known that NGS technologies are biased, the exact nature of this genomic dark matter left un-sequenced is far less known and thus the main topic of **Paper I** and **Paper II**.

Long reads and genomic dark matter

The advent of long-read sequencing has been a major sequencing revolution that has significantly advanced genome assemblies towards their completeness. Currently, there are two long-read sequencing platforms that implement very different approaches: Pacific Biosystems (PacBio) and Oxford Nanopore Technologies. Both technologies require a library preparation substantially different from the procedure adopted for short-read sequencing. At the heart of the library preparation, there is the extraction of high-molecular weight (HMW) DNA that will allow to take full advantage of the sequencing step. With the proper HMW DNA library, PacBio and Nanopore can yield read lengths of 10-100 kb, the latter occasionally up to a megabase and limited in theory by DNA fragment size (Payne *et al.*, 2019). While the resulting long reads are usually less accurate than short reads, read accuracy is continuously improving both for PacBio and Nanopore (Wenger *et al.*, 2019).

PacBio platforms sequence the DNA molecule by detecting luminous signals in real time produced by the incorporation of fluorescently-labelled nucleotides (Korlach *et al.*, 2010) while Nanopore relies on the detection of voltage fluctuations caused by the passage of the different nucleotides through a voltage-sensitive artificial pore (Deamer, Akeson and Branton, 2016). Since the library preparation for PacBio and Nanopore does not require any amplification step, base composition biases due to PCR amplification are being minimised. Nonetheless, there are still base composition biases present in the reads mainly due to the secondary/tertiary DNA structures that may affect sequencing. For example, it has been suggested that stem-loop secondary structures at inverted repeats may biophysically interfere with the sequencing of Nanopore platforms during the translocation of single-stranded DNA through the pores (Spealman, Burrell and Gresham, 2019). Tertiary structures involved in sequencing issues are mainly non-B DNA structures. Non-B DNA structures are DNA conformations alternative to the canonical right-handed double-helix that form in the presence of particular repeated motifs and base composition patterns (e.g., G-quadruplexes, cruciforms, Z-DNA) (Choi and Majima, 2011). Indeed, a variety of non-B DNA structures may introduce sequencing errors in PacBio reads (Guiblet *et al.*, 2018). This sequencing inaccuracy is a cause of genomic dark matter since some regions may be greatly affected and thus not confidently assembled. For example, inaccuracy linked to tertiary structures is one of the causes of some GC-rich avian genes not being fully represented in long-read genome assemblies (Beauclair *et al.*, 2019).

Assemblies and dark matter

Basics of the assembly process (contigs, scaffolds) and genomic dark matter

The ultimate goal of genome sequencing is to faithfully reconstruct entire chromosomes as an uninterrupted string of nucleotide bases. The assembly of reads is an incredibly hard process similar to solving a gigantic jigsaw puzzle with no reference image and made up of millions of pieces (reads). The assembly process with either short or long reads is basically divided into three main steps: 1) assembly of contigs; 2) scaffolding of contigs; 3) gap-filling.

The first step of contig formation is based on finding partially overlapping reads that represent contiguous stretches of sequences (Wajid and Serpedin, 2012). The relationship between all the overlapping reads is called "assembly graph". Ideally, in the assembly graph all reads are present once and each one is related unambiguously to the next one to form the linear representation of the genome; ideally there are as many unambiguous paths (Eulerian paths) in the graph as there are chromosomes in the genome. In reality, the assembly graph is not a simple linear set of relationships between reads, but it is complex and convoluted in which reads have multiple connections with one another. The result of such complex assembly graph is the presence of multiple Eulerian paths that each represent fragments of our genome and ultimately are the so-called contigs. The fragmentation of the graph is translated into the introduction of gaps in the genome assembly. Intuitively, homogeneous repeats pose a serious problem for the assembly process, since the reads originated from them introduce multiple overlaps and ambiguities in the assembly graph (Wajid and Serpedin, 2012).

After the primary contigs are assembled, contigs can be linked to one another using long-range information that provides physical evidence for separate contigs to belong to the same molecule. The scaffolding process then results in a set of contigs linked to one another and separated by stretches of N nucleotides (explicit gaps), together called "scaffolds" (Wajid and Serpedin, 2012). The nature of the sequences between two contigs (the dark matter within the N gaps) belonging to the same scaffold is not known and the scaffolding process does not add any additional sequence information. In order to link contigs into scaffolds, long-range information is needed and there are several sources of such information including mate-pair read libraries (Wetzel, Kingsford and Pop, 2011), BAC (Bacterial Artificial Chromosome) clones (Liu *et al.*, 2009), linkage maps (Peñalba *et al.*, 2020), radiation hybrid maps (Bickhart *et al.*, 2017), linked reads (Zheng *et al.*, 2016), optical maps (Weissensteiner *et al.*, 2017), and chromosome conformation maps (Kadota *et al.*, 2020). Usually, the scaffolding process needs a first phase in which the long-range information is mapped onto the assembly and then the strength (order) and directionality (orientation) of the relationship between contigs is evaluated, while estimating the distance between contigs (gap size) when known from the scaffolding information. In the case of mate-pair reads, these are usually incorporated directly into the assembly process where the long-range information (insert size of several kb) they carry helps to solve the assembly graph and estimate gap sizes.

Finally, the gap-filling process tries to bridge gaps directly on the assembly (Salmela et al. 2016) and to further connect separated contigs/scaffolds. Gap-filling can either use short or long reads, but yields better results with long reads (English *et al.*, 2012). In **Paper II**, I apply a round of gap-filling with long reads during the curation of the paradise crow multiplatform assembly. Although gap-filling helped to close some gaps, it is not sufficient nor very helpful to scaffold the assembly. However, when gap-filling with long reads is applied on a short-read draft assembly (**Paper II**), it closes thousands of gaps but does not improve scaffolding.

The assembly process is currently not able to correctly assemble each chromosome of eukaryotic genomes from telomere-to-telomere, except for some small fungal genomes (Faino *et al.*, 2015; Thomma *et al.*, 2016), some nematodes (de la Rosa *et al.*, 2021), and most recently a human cell line (Nurk *et al.*, 2021). In addition to the fact that input libraries can be originally incomplete and not representative of the entire content of the genome (see **Sequencing and dark matter** section), the assembly process is not able to generate contigs reaching from one chromosome end to the other because of ambiguities intrinsic to the genome and the libraries generated from it. These ambiguities are caused by the repetitive elements present in the genome. When the reads are not long enough to completely span individual repeats or repeat arrays and anchor them to unique genomic regions, the assembly gets fragmented, gaps are introduced, and information on the nature of the DNA within those gaps is lost. A most famous example is given by centromeres that are made of megabase-scale stretches of highly homogeneous tandem repeats (Jain *et al.*, 2018; Miga, 2020; Miga *et al.*, 2020; Logsdon *et al.*, 2021). In this case, is not possible to differentiate reads belonging to same repeat or to different units of the repeat array, therefore what usually happens is that the repeats collapse into a single or a few repeat units. It is clear that the length and accuracy of the reads is key to distinguish and correctly assemble all these repeat loci. Indeed, long-read sequencing technologies are an invaluable tool at the moment to resolve this genomic dark matter. In **Paper II**, I expand on this topic by quantifying and characterising the unresolved dark matter inherent in different assembly approaches in comparison with a manually curated reference assembly that implements long reads.

The year 2021 has revealed immense efforts in the generation of high-quality genome assemblies from consortia like the Vertebrate Genome Project (Rhie *et al.*, 2021) and the T2T Consortium (Nurk *et al.*, 2021). The Vertebrate Genome Project combines multiple sequencing and scaffolding technologies to obtain high-quality genome assemblies similarly to how the technologies were combined in **Paper I**. The T2T Consortium aims to develop an assembly process achieving gapless genomes and first started with filling the last gaps of the human genome. Thanks to the implementation of highly accurate long reads, ultra-long reads, virtually homozygous cell lines, virtually unlimited source of DNA, and Hi-C data, earlier this year the T2T Consortium was able to, for the first time, provide completely assembled human chromosomes including all centromeres (Nurk *et al.*, 2021). Hopefully, these types of methodologies and genome assemblies will be soon available for all model organisms, and maybe even non-model organisms.

Long-range information: paired-end reads, mate-pair reads, linked reads and chromosome conformation maps

For the purpose of this thesis, I will describe only the three types of long-range information that I implement in my genome assembly comparisons of **Paper II**.

Paired-end and mate-pair reads

Previously, I described the basic sequencing on an Illumina platform in which single-end reads are produced. This type of library is not useful for scaffolding since it lacks any long-range information. It possible to add long-range information to such libraries by sequencing both ends of the molecules that are attached to the flow cells (paired-end reads). Furthermore, the distance between the two paired-end reads (insert size) is estimated during the shearing of the extracted DNA and fragment length selection. Insert sizes have an upper limit of ~1 kb for technical reasons, but it is possible to obtain paired-end reads from DNA fragments longer than 1 kb by preparing a mate-pair library (Van Nieuwerburgh *et al.*, 2012). A mate-pair library is a set of long DNA molecules that have been sheared to the length of 500 bp while preserving both extremities with a maximum insert size of 20 kb. For example, in **Paper II** we used data generated from a library of mate-pair reads with an insert size of 8 kb (Prost *et al.*, 2019).

The combination of multiple short-read libraries with different insert sizes allows for improved scaffolding. Since the insert size between the reads is known for each library, it is possible to estimate the gap size between contigs (van Heesch *et al.*, 2013).

Linked reads

Linked reads are sets of paired-end short reads that are known to belong to the same DNA molecule even if hundreds of kb apart (Bell *et al.*, 2017). There are various methodologies to get linked read libraries and I will give a brief overview of the concept based on the Chromium protocol developed by 10X Genomics (Weisenfeld *et al.*, 2017). The key difference between a regular paired-end read library and a linked-read one is that sets of paired-end reads originally belonging to the same long DNA molecule maintain this valuable information through special labelling prior to sequencing. First, the extracted DNA fragments (preferentially high-molecular weight DNA) are separated from one another into different oil droplets (theoretically one molecule per droplet, but in reality, several fragments may end up in the same droplet). Then, inside the droplet, the molecule is sheared and the resulting fragments are ligated to short oligonucleotides called "barcodes" that are unique for each droplet (Bell *et al.*, 2017). The barcoded fragments are then sequenced on an Illumina machine. Resultant paired-end reads with the same barcode are considered to belong to the same input DNA molecule. This information can be used to order and orient contigs in scaffolds and to efficiently produce low-cost *de-novo* assemblies (Armstrong *et al.*, 2018).

In the past few years, linked-read libraries have been demonstrated to be extremely valuable for more than just scaffolding. Indeed, linked reads can be used for resolving haplotypes to correct assemblies (Weisenfeld *et al.*, 2017), phase genomes (Zheng *et al.*, 2016), estimate recombination rate (Dréau *et al.*, 2019), detect structural variation (Marks *et al.*, 2019), and even find tissue-specific genome differences (Kinsella *et al.*, 2019).

Chromatin conformation capture maps

Chromatin–chromatin interactions are important genomic features that are studied to uncover the genomic locations of regulatory elements and the overall three-dimensional organisation of genomes (Dekker, Marti-Renom and Mirny, 2013). Several molecular techniques have been developed to study the three-dimensional chromosome structure, all based on chromosome conformation capture (3C) methods. The most recent and high-throughput 3C technique, Hi-C (Lieberman-Aiden *et al.*, 2009), has also been found to be extremely useful in scaffolding genomes (Kadota *et al.*, 2020).

Hi-C is a very powerful technique able to provide a chromosome-level assembly for virtually any kind of genome (Ghurye *et al.*, 2017; Peichel *et al.*, 2017; Dudchenko *et al.*, 2018). Briefly, the Hi-C library is a paired-end library where the reads belong to linearly distant loci (even megabases apart) that are in very close proximity in the three-dimensional nuclear space. To get such information, the Hi-C methodology relies on the fact that the 3D genomic architecture is mainly mediated by CTCF protein–DNA interactions that form chromatin loops. The protocol starts with the fixation of the DNA in its native chromatin conformation through cross-linking of protein–DNA interactions; then the DNA strands hanging from the DNA-protein complex are cut with restriction enzymes, ligated to one another, and sequenced on an Illumina machine (**Figure 2.3**). The so-obtained read pairs stem from the chimeric DNA molecule comprising each end of the CTCF protein–DNA interaction. These reads are then mapped to the assembly for hierarchical scaffolding, and with sufficient coverage it is possible to obtain scaffolds that encompass entire chromosomes. Empirical studies showed that the vast majority of CTCF protein–DNA interactions happen within the same chromosome and only a small fraction between chromosomes (Lieberman-Aiden *et al.*, 2009). Therefore, Hi-C is able to hierarchically order and orient contigs/scaffolds into comprehensive chromosome models. Since the linear distance between two paired-end reads (insert size) is too variable to be estimated beforehand, it is not possible to reliably estimate the distance between scaffolded contigs (gap size). Moreover, many tools have been developed to scaffold, correct, and phase diploid assemblies (Ghurye *et al.*, 2017; Dudchenko *et al.*, 2018).

Next to the “true” Hi-C, Dovetail Genomics developed the CHiCAGO protocol, a modified Hi-C protocol to be applied on already extracted DNA (Paajanen *et al.*, 2019). The extracted DNA does not maintain the native chromatin conformation since the proteins bound to the DNA are removed during DNA extraction, therefore the chromatin state is re-established *in-vitro*. The majority of the protein–DNA interactions restored in this technique will likely be between loci belonging to the same molecule. Since extracted DNA generally consists of fragments that are much shorter than entire chromosomes, the interactions are expected to occur on a smaller scale with respect to what is expected in the native chromatin state (Paajanen *et al.*, 2019). Therefore, the proximity ligation maps derived from the CHiCAGO technique will be useful to scaffold contigs and correct misassemblies, but unlike Hi-C, will not be able to scaffold entire chromosomes.

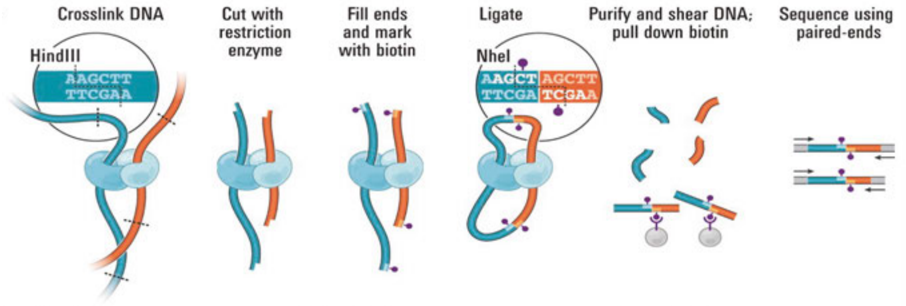

Crosslink DNA
Cut with restriction enzyme
Fill ends and mark with biotin
Ligate
Purify and shear DNA; pull down biotin
Sequence using paired-ends
HindIII
AAGCTT
TTCGAA
NheI
AAGCT AGCTT
TTCGA TCGAA

Figure 2.3 Schematic representation of the Hi-C library preparation. Image from (Lieberman-Aiden *et al.*, 2009). Reprinted with permission from AAAS.

3. Repeats

Repetitive elements are DNA sequences that are present in the genome in multiple copies. This broad definition encompasses tandem repeats like microsatellites and satellites, transposable elements, and also multicopy gene families like the MHC cluster and olfactory genes. Repetitive elements are ubiquitously present in eukaryotic genomes and usually in great abundance (Wicker *et al.*, 2007; Sotero-Caio *et al.*, 2017).

Repeats constitute a major portion of genomic dark matter (**Paper I** and **II**). Although repeats hamper the completion of genome assemblies, their correct assembly and annotation is of utmost importance because of the various effects they may exert on host genome evolution (Slotkin, 2018; Lerat *et al.*, 2019).

In the following sections, I will briefly cover the most salient aspects of tandem repeats and interspersed repeats.

Tandem repeats

Tandem repeats are highly homogenous DNA sequences arranged in arrays. The monomers that constitute the minimal repetitive units of tandem repeats can range from a minimum of 1 bp to several kilobases (Weissensteiner and Suh, 2019). The length of monomers dictates the categorisation of tandem repeats into microsatellites (1-6 bp) (Ellegren, 2004), minisatellites (6-100 bp) (Vergnaud and Denoeud, 2000), and satellites (>100 bp). The length of such tandem repeat arrays can range from hundreds of base pairs over kilobases to megabases.

Microsatellites and minisatellites have been extensively used in population genetics (Balloux and Lugon-Moulin, 2002), conservation biology (Moss, Piertney and Palmer, 2003), and forensic biology (e.g., DNA fingerprinting) (Ballantyne *et al.*, 2010). Their extreme variability (between species, populations and individuals) and the fact they are considered to evolve neutrally make them handy genetic markers. The functions, if any, of microsatellite and minisatellite DNA remain largely unknown but some lines of evidence indicate that some may play a role in the regulation of transcription factor binding and gene expression (Li *et al.*, 2002), and genetic disorders (Boulay *et al.*, 2018). Even though microsatellites in general do not seem to have a clear function, in vertebrates the particular $(TTAGGG)_n$ hexamer is the essential component of telomeres (Meyne *et al.*, 1990).

Microsatellite and minisatellite monomers are usually arranged in a head-to-tail fashion while satellite DNA can show a vast variety of arrangements. Stretches of satellite DNA can show complex hierarchical arrangements where arrays consist of multi-monomeric repeat units, called Higher Order Repeats (HOR) (Miga, 2019). HORs can often be found at centromeres, such as the most famous HOR example, the human alpha-satellite (Willard and Waye, 1987). Depending on the size of the HOR monomer, the characterisation of its structure can be very challenging using short reads (Lower *et al.*, 2018).

Different satellite DNAs can be present in the same genome (the set of satellites forms a library) and evolve independently from one another. Closely related species may share the same library but often a species-specific accumulation and origination of satellite is observed (Salser *et al.*, 1976; Palacios-Gimenez *et al.*, 2020). Indeed, satellites are one of the fastest evolving components of the genome, and significant differences in sequence, abundance and physical chromosomal location may act as reproductive barriers (Ferree and Barbash, 2009). It has been proposed that satellite DNA can affect chromosome behaviour in hybrids, leading to hybrid incompatibility by disrupting chromosome alignment during meiosis, altering chromosome heterochromatinisation, and by involvement in meiotic drive mechanisms (Ferree and Prasad, 2012). Recent studies on *Drosophila melanogaster* (Shatskikh *et al.*, 2020) showed that satellite DNA from the different chromosomes cluster to form chromocenters (densely packed chromatin structure) during cell divisions that likely prevent the dispersion of chromosomes during the nuclear assembly (that would lead to the formation of non-functional micronuclei and cell death (Jagannathan, Cummings and Yamashita, 2018). In a more recent paper (Jagannathan and Yamashita, 2021), the authors investigated the possible link between the correct formation of chromocenters and satellite DNA composition during hybridisation of *D. melanogaster*, *D. simulans*, and *D. mauritania*. They found that mismatches between satellite DNA binding proteins and satellite DNA sequences in hybrids lead to ineffective chromocenter clustering and consequent hybrid incompatibilities. These studies suggest that satellite DNA plays a role in the maintenance of genomic integrity as well as in establishing reproductive barriers.

Tandem repeats represent a major component of the genomic dark matter in short-read assemblies, but I show that they are mostly assembled in long-read assemblies (**Paper II**) with the exception of extremely long arrays.

Interspersed repeats

Genomes are dynamic entities with a fluid organisation that changes in space (e.g., between different kinds of cells) and in time (e.g., during the organism's development and during evolution). A major factor that contributes to the fluidity of genomes are interspersed elements such as transposable elements. Transposable elements are mobile elements able to move from one genomic locus to another (or even to another genome through horizontal transfer) (Schaack, Gilbert and Feschotte, 2010; Suh *et al.*, 2016; Zhang *et al.*, 2020).

Transposable elements can be divided in two main classes based on the mode of transposition they adopt: either "copy-and-paste" or "cut-and-paste". Class I elements move through an RNA intermediate that is reverse transcribed into a new genomic locus (copy-and-paste). Class II elements excise themselves from the original location and insert into a new locus (cut-and-paste). The latter mode of transposition is called conservative since it does not per se increase the number of copies of the element in the genome. Class II elements therefore seem to take advantage of the genome replication timing to generate multiple copies of themselves (Craig *et al.*, 2015).

Finally, transposable elements can be autonomous or non-autonomous. Autonomous elements are elements that encode all the proteins needed for transposition. Non-autonomous elements are elements that lack some or all of these, and instead rely on proteins from the autonomous elements for their transposition (Bowen and Jordan, 2002). Consequently, non-autonomous elements can replicate as long as there are intact protein machineries that recognise them.

Class I

Transposable elements belonging to this category are also called "retrotransposons" since their RNA intermediate needs to be reverse transcribed into DNA before inserting back into the host genome. Retrotransposons can be further divided into LTR and non-LTR elements depending on the presence or absence of identical Long Terminal Repeats at their 5' and 3' extremities. The presence of LTRs or the absence thereof is linked to their particular mode of replication.

LTR elements transpose through a replicative retrotransposition mechanism, and include endogenous retroviruses (ERVs) and exogenous retroviruses (XRVs). Endogenous retroviruses are virus-like sequences that lost the ability to leave the cell. ERVs are therefore very similar to retroviruses and they share many viral genes with their exogenous counterpart. Endogenous retroviruses have multiple ORFs that contain gag and pol genes encoding the proteins of the capsid, reverse transcriptase, ribonuclease H, and integrase. LTR elements are first transcribed in the nucleus, then their RNA is recognised by the viral proteins they encode for and captured inside a virus-like particle in the cytoplasm (Havecker, Gao and Voytas, 2004). Within the viral particle the mRNA is retrotranscribed into DNA (Kazazian, 2004). The new DNA sequence is then inserted into a new genomic location by the action of their integrase. Upon insertion, LTRs produce a short target site duplication (TSD).

LTR elements can be very long, even 10-12 kb, but they can often be found in a shorter version called solo LTR. Because of the presence of identical LTRs at both extremities of the element, these transposons are often subject to ectopic recombination occurring between the two repeats. The recombination event leads to the removal of the internal portion of the element and only one long terminal repeat remains as a solo LTR. The LTRs of ERVs contain regulatory motifs to initiate their own transcription and can influence the transcription of nearby genes as well. Indeed, ERV promoters may have been largely co-opted by the host for gene regulation in the human genome (Sundaram *et al.*, 2014). Since the regulatory elements are found on the LTRs, it means that both full-length and solo LTR elements can influence gene regulation (Thompson, Macfarlan and Lorincz, 2016). The presence and possible effects of full-length and solo LTRs in avian genomes is investigated in **Paper III** and **Paper IV**.

Non-LTR elements comprise a vast diversity of elements. The most important for the scope of this thesis are Long INterspersed Elements (LINEs) and their non-autonomous counterpart, the Short INterspersed Elements (SINEs). LINEs move through the Target Primed Reverse Transcription (TPRT) (Luan *et al.*, 1993) mechanism which is mediated by the endonuclease and reverse transcriptase proteins encoded in their two ORFs (Scott *et al.*, 1987; Singer *et al.*, 1993; Denli *et al.*, 2015). Once the element is transcribed, the RNA is transferred to the cytoplasm where is incorporated into a ribonucleoprotein particle (made of the proteins the element encodes for) and imported back into the nucleus. Upon recognition of the insertion site (target priming), the double-helix is nicked by the LINE endonuclease and the RNA is retrotranscribed (3’ → 5’). The insertion of a new LINE causes a target site duplication at the extremities of the element. The TPRT mechanism is not the most faithful transposition mechanism as the reverse transcription tends to stop prematurely, causing the 5' truncation of many insertions (Kazazian and Goodier, 2002).

Class II

DNA transposons move by using a single or double-stranded DNA intermediate (Chandler *et al.*, 2015). They are usually labelled as "cut-and-paste" elements (as I also did above) but in reality, the class encompasses elements with heterogeneous (sometimes cryptic) modes of transposition.

There are the classic "cut-and-paste" DNA transposons (e.g., Mariners, hAT, Harbingers) that fully excise themselves (double-stranded DNA) from one locus and insert to a new one. These elements, when autonomous, all encode for a transposase protein that is able to recognise the Terminal Inverted Repeats (TIRs) of their elements and initiate the transposition. A second category of Class II repeats are Helitrons which probably transpose through a rolling-circle mechanism. Then there are elements for which the mode of transposition is largely unknown: Polintons that encode many proteins likely related to double-stranded DNA viruses and Cryptons which encode for a tyrosine recombinase and may or may not leave TSDs behind (Feschotte and Pritham, 2007).

The importance of a curated repeat library

In order to accurately detect repetitive elements in any genome assembly, a well-curated repeat library is necessary. Repeat libraries can be retrieved from databases like Repbase (Bao, Kojima and Kohany, 2015) for already characterised genomes. In case the species of interest have not been previously analysed, the existing libraries from related species may under-annotate the assemblies especially if not closely related. This shortcoming in annotating repeats can be due to 1) the presence of novel repeats not shared with related species in the database; 2) the high divergence of repeats from available libraries that would lead to partial or no hits. It has been thoroughly demonstrated that a *de-novo* characterisation of repeats (for example through the use or RepeatModeler) drastically increases the accuracy of the repeat annotation (Platt, Blanco-Berdugo and Ray, 2016). Moreover, manual curation of the repeat library allows the characterisation of full-length elements and therefore improves and simplifies the classification of the repeats themselves (e.g., distinctive hallmarks like protein-coding domains or terminal motifs can be better represented). Recently, it has also been demonstrated that the thorough manual curation of repeat libraries of sister species synergically improves the repeat annotation of both species (Boman *et al.*, 2019). This suggests that a complete repeat annotation is never a trivial task, and no species should be left un-curated. For this reason, I manually curated the repeat libraries for several birds-of-paradise and crow species (**Paper II**, **IV**, and **V**) as well as emu, kakapo, and Anna's hummingbird (**Paper III**). In **Paper II**, I also demonstrate that the implementation of curated repeat libraries changes the overall annotation of a genome by uncovering portions of the genome previously not masked as repeats and by increasing the general amount of repeats, confirming the pattern reported previously (Platt, Blanco-Berdugo and Ray, 2016).

Although manual curation of repeats is key for a correct annotation, the choice of sequencing technologies is important to retrieve repeats in the first place and build a comprehensive library. In **Paper II**, I show that long-read assemblies allow the discovery and characterisation of a greater number of repetitive sequences with respect to short-read assemblies. Comparing the repeat libraries based either on Illumina or PacBio for the same bird-of-paradise individual, it is clear that many repeats were discovered only in the PacBio assembly because they were assembled (if at all) in too few copies in the Illumina assembly to be detected and curated.

4. Avian genomics

Genome size and karyotype

Bird genomes are small in size compared to the rest of land vertebrates (0.9-2.1 Gb) (Gregory *et al.*, 2007; Wright, Gregory and Witt, 2014), generally repeat poor (~10%) (Kapusta and Suh, 2017) and with compact genes due to reduced intron sizes (genes are 50% and 27% shorter than mammalian reptilian genes respectively; Zhang et al., 2014). Moreover, some studies (Hughes and Friedman, 2008; Lovell *et al.*, 2014) showed a dramatic reduction in number of genes and gene families present in birds with respect to mammals and reported hundreds to thousands of "missing genes". In more recent studies (Hron *et al.*, 2015; Botero-Castro *et al.*, 2017; Yin *et al.*, 2019), it has been shown that many of those missing genes are actually "hidden genes" that are hard to assemble and characterise due to their extreme GC content. Thanks to the new long-read technologies that are less biased towards extreme base composition, more and more "missing", or "hidden", genes are being found, assembled, and annotated in genome assemblies (Yin *et al.*, 2019).

Hidden genes are therefore another aspect of the genomic dark matter that is mainly due to technological limitations in sequencing extremely GC-rich regions of the genome. Recently, it was proposed that the problems in sequencing these regions may also be due to the tertiary structures the DNA forms during sequencing, as in the case of G-quadruplexes (G4s). G4 structures are non-B DNA structures that are formed in the presence of particular GC-rich motifs (Choi and Majima, 2011). As GC-rich sequences are a potential factor contributing to the hidden gene phenomenon (Beauclair *et al.*, 2019), I consider that some G4 motifs can be part of genomic dark matter. In **Paper II**, I show that long-read sequencing technologies and the assembly curation process lead to a better and more complete representation of such motifs with respect to short-read technologies.

Nearly all avian genomes are organised into macrochromosomes and microchromosomes (Kapusta and Suh, 2017). Macrochromosomes are large chromosomes with size that ranges from 40 Mb to ~200 Mb and comprise about 70% of the genome. On the other hand, microchromosomes, as the name indicates, are small chromosomes (<20 Mb) nearly indistinguishable under the microscope during karyotyping analysis (Burt, 2002). Finally, some studies use the term “intermediate chromosome” for size ranges from 20 to 40 Mb (Burt, 2002; Griffin and Burt, 2014), which I also use in this thesis when applicable. Macrochromosomes, intermediate chromosomes, and microchromosomes are certainly different in size but also differ in specific genomic features. For example, microchromosomes are more GC-rich than macrochromosomes and intermediate chromosomes, while exhibiting higher substitution rates and recombination intensity (Burt, 2002; Axelsson *et al.*, 2004).

About two thirds of bird species present a karyotype of 38-41 pairs of chromosomes (Degrandi *et al.*, 2020) and past studies comparing distant bird species (chicken and zebra finch) first highlighted a highly conserved synteny between avian chromosomes (Ellegren, 2010). Although some avian genomes do look highly syntenic, there are many exceptions to this observation. Indeed, avian chromosomes seem often subject to fast centromere repositioning (change in centromere location without altering the order of genetic markers) (Rocchi *et al.*, 2012) between species in macrochromosomes (Kiazim *et al.*, 2021) and microchromosomes alike (Westerberg, 2020; Vontzou, 2021). Various studies are finding many intrachromosomal (Skinner and Griffin, 2012; Zhang *et al.*, 2014; Farré *et al.*, 2016; Hooper and Price, 2017) and interchromosomal rearrangements (Coelho, Musher and Cracraft, 2019; Kretschmer *et al.*, 2020, 2021; Pinheiro *et al.*, 2021). In general, the finer the resolution, the more dynamic avian genomes look (Galbraith *et al.*, 2021). In **Paper IV**, I look at structural changes at short and long evolutionary timescales using genomic samples from individuals of the same species to species of different genera within the bird-of-paradise phylogeny. From the analysis of **Paper IV**, it seems that macrochromosomes and microchromosomes tend to accumulate structural changes in different ways.

Some microchromosomes can be part of genomic dark matter because their small size and base composition make them difficult to assemble and are thus tightly linked to the hidden gene problem. Avian autosomes show a recombination rate that negatively correlates with chromosome size (Burt, 2002; Backström *et al.*, 2010; Stapley *et al.*, 2010; Kawakami *et al.*, 2014, 2017), meaning that microchromosomes have a higher recombination rate than macrochromosomes. In birds, recombination is strongly linked to the GC-biased gene conversion phenomenon (Mugal, Arndt and Ellegren, 2013; Kawakami *et al.*, 2017; Bolívar *et al.*, 2019) for which GC-richer alleles are preferentially fixed in the population (Galtier *et al.*, 2009). Likely because of high rates of recombination and associated GC-biased gene conversion, microchromosomes became much more GC-rich than macrochromosomes. Base composition per se is an issue for NGS technologies, but extreme base composition may also carry additional molecular features such as non-B DNA structures. Indeed, microchromosomes are much denser in G4 motifs with respect to macrochromosomes as I show in **Paper II**, which adds another layer of difficulty in their sequencing and assembly. In **Paper II**, I show that only through the combination of cutting-edge sequencing and scaffolding technologies, it is possible to improve the assembly of microchromosomes.

Birds present a ZW sex determining system where the heterogametic sex is the female (ZW) and the homogametic one is the male (ZZ). The Z and W are heteromorphic in Neognathae while largely homomorphic in Palaeognathae except for tinamous (Zhou *et al.*, 2014; Xu, Wa Sin, *et al.*, 2019). In Neoaves, the W is non-recombining except for a short pseudoautosomal region (PAR) and is highly repetitive (~70% repetitive; **Paper III**). Conversely, the Z chromosome fully recombines in male individuals and it is much more similar to the autosomes in repeat content (~10% repetitive; **Paper III**). Therefore, the Z is mostly well-assembled for every bird species sequenced, while the highly repetitive W chromosome has always been challenging to assemble and is missing from the great majority of avian genome assemblies. Since the W chromosome is a major cause of assembly fragmentation, most of the avian genomes were sequenced from male individuals and thus very little was known on the evolution of this chromosome (Kapusta and Suh, 2017). Thanks to a combination of new long-read sequencing technologies and proximity ligation maps (a road map that we arrived at simultaneously and independently from the Vertebrate Genome Project (Rhie *et al.*, 2021)), we are able to generate an assembly of the W in a non-model bird (as shown in **Paper II**) comparable in quality and contiguity with the W chromosomes generated by the Vertebrate Genome Project itself (**Paper III**). More details about the repeat content and evolution of the sex chromosomes can be found in the following section and in **Section 5**.

The aforementioned genomic features (small size, low overall repeat content, and locally confined GC-rich and repeat-rich regions) make avian genomes the perfect combination of challenge for the new sequencing technologies to investigate the nature of genomic dark matter in this thesis and as a case study for complex eukaryote genomes in general.

Repeat content

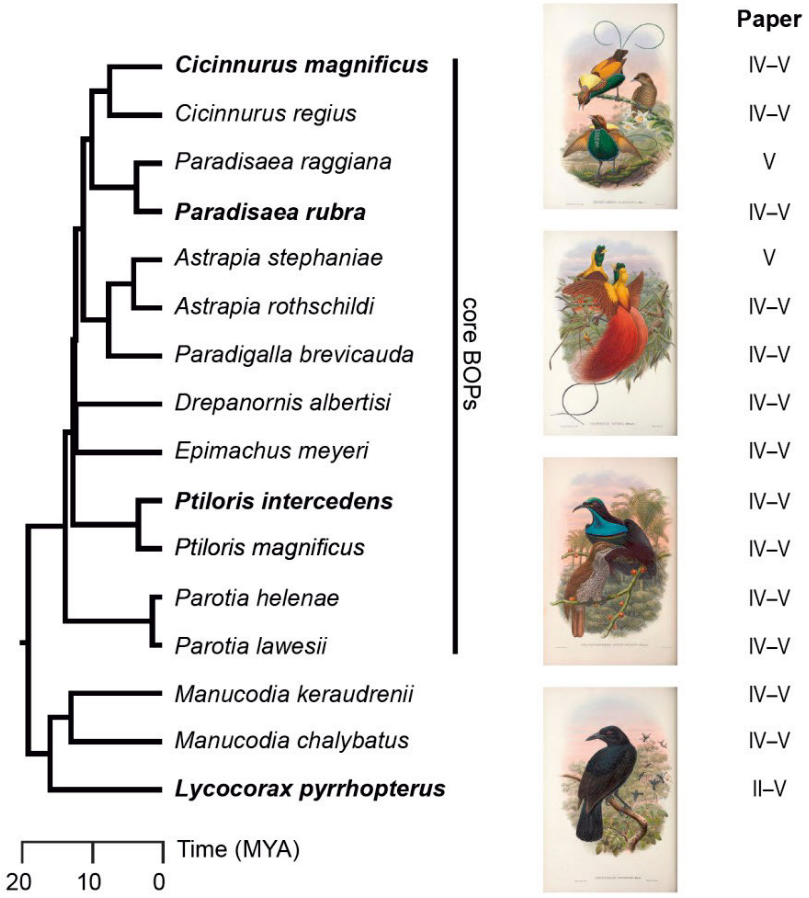
As mentioned before, avian genomes are generally repeat poor compared to other land vertebrates (Kapusta and Suh, 2017). In fact, most avian genomes are 10% repetitive with the exception of woodpeckers and relatives with a repeat content of ∼17–30% (Manthey, Moyle and Boissinot, 2018; Feng *et al.*, 2020). Even though this percentage of repeats has been first observed in the genomes sequenced in the Avian Phylogenomics Consortium (Zhang *et al.*, 2014) and the 10,000 Bird Genomes Consortium (Feng *et al.*, 2020) that used only short reads, it remains confirmed also in the new genome assemblies based on long reads (**Paper II**) (Rhie *et al.*, 2021). However, the repeat density on the W chromosome was previously estimated to be at least 50% (flycatcher and white-throated sparrow) (Davis, Thomas and Thomas, 2010; Smeds *et al.*, 2015) while the new assemblies show a W chromosome that is at least 70% repetitive (Bellott *et al.*, 2017; Rhie *et al.*, 2021). This extreme repetitiveness is initially investigated in **Paper II** and in more detail in **Paper III** and **IV**.

The fraction of interspersed repeats in avian genomes is mainly composed of retrotransposons: Chicken Repeat 1 (CR1; LINEs) and endogenous retroviruses (ERVs). CR1 are the most abundant repeats in birds and over 14 families of CR1 have been described so far (Kapusta and Suh 2017; but see Galbraith *et al.* 2021 for a new family nomenclature). Initial studies of the chicken and zebra finch genomes (Wicker, 2004; Warren *et al.*, 2010) did not find evidence for recent expansion of CR1 retrotransposons. Thanks to high-quality genome assemblies from a more comprehensive sample of birds, Galbraith *et al.* (2021) found evidence for such recent CR1 expansions whose insertions likely participate in the structural evolution of avian genomes. Full-length CR1 elements can be over 4 kb long but given their transposition via TPRT, the vast majority of copies are 5' truncated and often just a couple of hundreds bp long (Hillier *et al.*, 2004). Since the truncated copies lack the 5' end and are transpositionally "dead on arrival" (Kazazian and Goodier, 2002), it is not likely that they can exert any regulatory effect on the nearby genomic regions. However, truncated CR1 copies still contain their 3' end that consist of a hairpin and an octamer motif (Suh, 2015), it is possible that this may influence the transcription nearby.

The second most abundant type of interspersed repeats are ERVs, part of the larger group of LTR retrotransposons (Wicker *et al.*, 2007; Bolisetty *et al.*, 2012; Zhou *et al.*, 2014). Most ERVs in avian genomes are in the form of solo LTRs and very few can be found as full-length elements. ERVs have internal regulatory sequences in their LTRs, therefore full-length ERVs and solo LTRs can influence the expression of nearby genes. Because of the length (up to 10-12 kb) and their potential for regulatory effects even as solo LTRs, it is plausible that ERVs are more subject to selection and removal than CR1 elements. Moreover, the formation of solo LTRs is linked to the rate of recombination of the different chromosomes. Likely because of reduced efficacy of selection and rate of recombination, there is a marked accumulation of full-length elements on the W with respect to the rest of the genome as highlighted in **Paper III** and **IV**. In this regard, the W acts as a refugium for ERVs.

Birds-of-paradise

Birds-of-paradise (Corvides: Paradisaeidae) are a family of 40 species endemic of Papua New Guinea, some islands of the Indonesian Archipelago, and northern Australia (Irestedt *et al.,* 2009). These birds are most famous for their incredible and diverse phenotypes, colourful plumage, and mating dances that are likely the results of a long history of sexual selection (Irestedt *et al.*, 2009). The core clade of birds-of-paradise species (**Figure 4.1**) started diversifying ~15 million years ago (Irestedt *et al.*, 2009) while the deepest divergences of the Paradisaeidae family are ~20 million years old.



Paper
Cicinnurus magnificus IV–V
Cicinnurus regius IV–V
Paradisaea raggiana V
Paradisaea rubra IV–V
Astrapia stephaniae V
Astrapia rothschildi IV–V
Paradigalla brevicauda IV–V
Drepanornis albertisi IV–V
Epimachus meyeri IV–V
Ptiloris intercedens IV–V
Ptiloris magnificus IV–V
Parotia helenae IV–V
Parotia lawesii IV–V
Manucodia keraudrenii IV–V
Manucodia chalybatus IV–V
Lycocorax pyrrhopterus II–V
core BOPs
Time (MYA)
20
10
0

Figure 4.1 Phylogenetic tree of the species of the family Paradisaeidae (birds-of-paradise) sampled in this thesis. Key representative species (bold italics) are shown with paintings (Sharpe, 1891-1898). Roman numerals indicate thesis papers using genome data from each sampled species. Dated phylogeny obtained from Timetree.org (Kumar *et al.*, 2017) and paintings are in the public domain.

The papers of this thesis on the genome evolution of these fascinating birds-of-paradise are part of a large-scale project aimed at characterising their evolutionary history in terms of speciation, hybridisation, and sex chromosomes (Prost *et al.*, 2019; Xu, Auer, *et al.*, 2019; Blom and Irestedt, 2021) using short-read data. Prost et al. (2019) first focused on the speciation of these birds using three species (*Lycocorax pyrrhopterus*, *Astrapia rothschildi*, and *Ptiloris paradiseus*) and investigated gene gain and loss, and genes under positive selection. In particular, Prost et al. (2019) found that gene families that expanded the most are enriched in the Gene Ontology terms "startle response" and "olfactory receptor activity". Xu *et al.* (2019) focused on the evolution of genes on the sex chromosomes of 5 birds-of-paradise and 6 other songbird species, and, interestingly, some bird-of-paradise did not show a clear fast-Z pattern as expected under sexual selection.

In this thesis, I contribute to the study of birds-of-paradise evolution from the point of view of repetitive elements and structural variants after resolving as much genomic dark matter as possible with new sequencing technologies and a reference genome assembly of *L. pyrrhopterus*. In my papers, in total, I used genomic data from females and males of 16 species of birds-of-paradise (**Figure 4.1**) belonging to 10 genera of the major clades of birds-of-paradise: *Paradisaea raggiana, Paradisaea rubra*, *Cicinnurus magnificus*, *Cicinnurus regius*, *Astrapia rothschildi*, *Astrapia stephaniae*, *Epimachus meyeri*, *Ptiloris intercedens*, *Ptiloris magnificus*, *Drepanornis albertisi*, *Parotia helenae*, *Parotia lawesi*, *Manucodia chalybatus*, *Manucodia keraudrenii*, and *Lycocorax pyrrhopterus*. For all these species, linked-read libraries were produced, often in addition to short-read data. *Lycocorax pyrrhopterus* and *Ptiloris intercedens* were also sequenced with PacBio long reads. Finally, a Hi-C map and a CHiCAGO map were produced for *Lycocorax pyrrhopterus*.

Birds-of-paradise were traditionally considered closely related to bowerbirds (birds living in Papua New Guinea as well) (Gregory, 2020) because they share habitats, extraordinary sexual dimorphism, and breeding strategies. However, genetic markers placed birds-of-paradise far from bowerbirds and within the superfamily Corvides (Irestedt *et al.*, 2009; Jønsson *et al.*, 2016, Gregory, 2020). Being closely related to crows (Corvidae) among Corvides, in **Paper V** I use several species of the genus *Corvus* as outgroup to birds-of-paradise to investigate the evolution of satellite DNA sequences.

5. W chromosomes

ZW sex chromosomes

Birds have a ZW genetic sex determination system where females are the heterogametic sex (ZW) and males the homogametic sex (ZZ). It is well understood how the mammalian XY system works, in that the Y chromosome carries a male determinant (SRY gene) (Wallis, Waters and Graves, 2008) that activates the male developmental path during embryogenesis. In birds, sex determination seems to occur via a dosage-dependent process involving the Z chromosome. Smith and colleagues demonstrated that the key gene for sex development is the Z-linked gene DMRT1 (doublesex and mab-3-related transcription factor 1) (Smith *et al.*, 2009). The presence of DMRT1 on the Z and its absence on the W suggests that double dosage of DMRT1 suppresses the female developmental path (Flament *et al.*, 2011).

ZW evolution in birds

The Z and W chromosomes evolved from a pair of autosomes (Fridolfsson *et al.*, 1998) and followed different evolutionary paths in different bird clades (Zhou *et al.*, 2014). The two chromosomes remained largely homomorphic (similar in morphology) in ratites where they recombine across most of their lengths (Yazdi and Ellegren, 2014, 2018; Zhou *et al.*, 2014; Xu, Wa Sin, *et al.*, 2019; Yazdi, Silva and Suh, 2020). In the remaining Palaeognathae (i.e., tinamous) and Neognathae (including chicken and songbirds), the sex chromosomes are very heteromorphic (different in morphology) with a degenerated W that lost most of its genes and accumulated many repeats (Zhou *et al.*, 2014; Smeds *et al.*, 2015; Warren *et al.*, 2017; Xu, Wa Sin, *et al.*, 2019). In these birds, Z and W recombine only in a small region (PAR) which is the only homogametic region of the W. Furthermore, the remainder of the W is non-recombining and the Z fully recombines only in males, which results in a reduction in effective population size (N_e) of these chromosomes with respect to the autosomes and the PAR. A reduction in effective population size and recombination rate also implies a reduction in the efficacy of selection (Beukeboom and Perrin, 2014).

The reasons why the sex chromosomes drastically diverged through recombination suppression are still debated (Yazdi, Silva and Suh, 2020) and many models have been proposed (Charlesworth and Charlesworth, 1978; Rice, 1984, 1987). The underlying idea to all of these models is that one of the chromosomes acquires a Sex Determining Region (SDR) and a locus with sexually antagonistic effect, namely an allele beneficial for one sex but detrimental for the other (Wright *et al.*, 2016; Charlesworth, 2021). Reduction or cessation of recombination is expected to be favoured around the SDR and the sexually antagonistic locus that would thus become tightly linked. The suppression may expand across the chromosome as more sexually antagonistic loci arise, and once again linkage between these loci and the SDR would be selected for. Moreover, structural rearrangements like inversions can help to establish a strong linkage disequilibrium and to reduce recombination. For example, inversions spanning the sex determining region have been found on the human X chromosome (Lahn and Page, 1999). Although theoretically valid and robust, empirical studies failed to provide conclusive evidence for this model (Ironside, 2010; Ponnikas *et al.*, 2018). If recombination suppression proceeds in a stepwise manner, then specific involved regions should start to diverge between the sex chromosomes at discrete points in time. These discrete regions take the name of evolutionary strata, where more divergent regions constitute old events of recombination suppression initiation and less divergent regions represent more recent events. Evolutionary strata have been found in many species including birds (Handley, Ceplitis and Ellegren, 2004; Nam and Ellegren, 2008; Suh *et al.*, 2011; Yazdi and Ellegren, 2014; Smeds *et al.*, 2015). A study on the evolution of Z chromosome in ostrich (Yazdi and Ellegren, 2018) highlighted several inversions that may be linked to recombination suppression but also highlighted how some regions stopped recombining in Neognathae without the involvement of any inversion with respect to ostrich. This implies that if inversions were the cause of recombination cessation, they must have been happened on the W (Zhou *et al.*, 2014). Thanks to new sequencing and scaffolding technologies is now possible to get good assemblies of W chromosomes, and this may provide a unique opportunity to look for evolutionary strata and inversions on the W chromosome. It must be added that mechanisms other than inversions may have guided the recombination suppression (Ponnikas *et al.*, 2018). Although the model of recombination suppression through inversions is very intuitive, the absence of clear discrete boundaries between some evolutionary strata (Nam and Ellegren, 2008) on the avian sex chromosomes may suggest a gradual cessation of recombination.

Even though it is not clear how recombination suppression is initiated, there are many structural and molecular factors that can contribute to the establishment and expansion of the suppression (Wright *et al.*, 2016; Ponnikas *et al.*, 2018; Furman *et al.*, 2020). Structural variation, like the abovementioned inversions, and gradual expansion of heterochromatin may play an important role. For example, as repeats start to accumulate, they are silenced by confinement into a heterochromatic state (Slotkin and Martienssen 2007). It has been shown (Grandi et al. 2015; Lee and Karpen 2017; Quadrana et al. 2019) that repressive histone marks or DNA methylation can spill over to the neighbouring regions from the repeats themselves, thus likely expanding the heterochromatinisation to other parts of the chromosome. The presence of transposable elements and heterochromatin are often linked to reduced recombination rates (Bartolomé, Maside and Charlesworth, 2002; Zeng and Yi, 2014; Coulthard *et al.*, 2016) in a variety of organisms. On the other hand, repeats are often hotspots for chromosomal rearrangements (Levy-Sakin *et al.*, 2019; Weissensteiner and Suh, 2019). In **Paper IV**, I analyse the types and occurrences of structural rearrangements on autosomes, Z, and W.

W chromosomes as a transposable element refugium

Once recombination stops, the W chromosome starts to degenerate and accumulates repeats in its non-recombining region (Charlesworth, Charlesworth and Marais, 2005; Sigeman *et al.*, 2020). The tendency of non-recombining chromosomes to accumulate repeats is mainly due to the combination of low recombination rate and low N_e that decreases selection efficacy on these chromosomes. In Neognathae, the process of degeneration led to the sharp differentiation of Z and W, where the W has been reduced to a small chromosome that is >50-70% repetitive and with a few dosage-sensitive genes involved in housekeeping functions (Smeds *et al.*, 2015; Bellott *et al.*, 2017; Bellott and Page, 2021) (**Paper II**).

The difference between the genome-wide and the W-specific repeat content is striking, especially because there is a clearly differential accumulation of repeat families. As mentioned earlier, the most common repeats in avian genomes are CR1 and ERVs. CR1 elements are generally short due to 5'-truncation and have likely few disruptive effects because truncated copies lack regulatory elements, while ERVs can be very long and are expected to have regulatory effects both as full-length and solo LTR. CR1 are almost homogeneously distributed across the genome (**Paper II**), ERVs and in particular full-length ERVs are mostly present on the W (**Paper II** and **III**).

The repetitive and heterochromatic nature of the W makes this chromosome difficult to sequence and assemble, and thus one of the biggest sources of genomic dark matter (Tomaszkiewicz, Medvedev and Makova, 2017). This implies that the structure and repeat content of this chromosome is understudied and under-appreciated. The structure and repeat content of sex-limited chromosomes (Y/W; SLCs) have been shown in *Drosophila* to exert epistatic effects genome-wide as well as physiological effects on the individuals carrying the SLCs (Chippindale and Rice, 2001; Brown and O'Neill, 2010; Jiang, Hartl and Lemos, 2010; Kutch and Fedorka, 2018). Recent epigenomics studies on *Drosophila melanogaster* showed how the sole presence of the Y chromosome is toxic for male individuals because of its active content of transposable elements that shape the heterochromatinisation of the genome, causing premature ageing (Brown, Nguyen and Bachtrog, 2020a, 2020b).

Similarly, other studies in *Drosophila* identified structural variants on the Y chromosome that have genome-wide epistatic regulatory effects by modulating the genomic heterochromatin landscape (Lemos, Araripe and Hartl, 2008; Francisco and Lemos, 2014). Because of the regulatory effects these Y-linked variants have, they have been named Y-regulatory variants (YRV). These YRVs were detected to stem from differences in satellite DNA array lengths and other repetitive elements (Jiang, Hartl and Lemos, 2010). The repetitive content and structural variation on the Y chromosome of *Drosophila* seem to be interconnected and to have an effect on the heterochromatinisation of the genome and gene expression (Brown, Nguyen and Bachtrog, 2020a). Given these premises from *Drosophila* and the striking accumulation of transposable elements on the W chromosome of birds, I explored the potential toxic effect of the W (**Paper III**) as well as quantify the structural variability of this chromosome (**Paper IV**).

There is some empirical evidence that, in ZW systems, females have a shorter lifespan with respect to males (Clutton-Brock and Isvaran, 2007; Donald, 2007; Lambertucci *et al.*, 2012; Pipoly *et al.*, 2015; Xirocostas, Everingham and Moles, 2020). Since birds are becoming an important model system for biogerontology studies (Holmes and Harper 2018), the structure and repeat accumulation of the W chromosome are important aspects to further investigate given how Y-linked repeats negatively affect male lifespan in *Drosophila melanogaster* (Brown, Nguyen and Bachtrog, 2020b). Another aspect to take into consideration is the possible role played by the W repeats in Haldane's rule (Haldane, 1922). Haldane's rule predicts that in a hybridisation event, if there are sterile or inviable individuals, they belong to the heterogametic sex. This rule has been demonstrated for many taxa (Delph and Demuth, 2016). Interestingly, in some species of *Drosophila* hybrid sterility (hybrid dysgenesis) is caused by a mismatch between the transposable element repertoire of one species and the silencing mechanism of the other (Kidwell, Kidwell and Sved, 1977; Petrov *et al.*, 1995; Hill, Schlötterer and Betancourt, 2016). In *Drosophila* hybrids, the uncontrolled repeat activity leads to morphological and physiological aberrations (Hill, Schlötterer and Betancourt, 2016). In **Paper III**, I discuss the possibility that a female-specific load of potentially active ERVs may be partially responsible for heterogametic sterility seen in birds (Neubauer, Nowicki and Zagalska-Neubauer, 2014; Mořkovský *et al.*,

2018) and provide a possible additional molecular explanation for Haldane's rule in birds.

Since for birds there are not the same molecular tools to investigate transposable element activity as for *Drosophila*, in **Paper III** I take advantage of publicly available genomic, transcriptomic, and proteomic datasets to identify female-specific signatures of transposable element activity. I then formulated new quantitative measures and explanations for how the W chromosome represents a sex-specific load of potentially active transposable elements that can exert a genome-wide toxic effect and contribute to the sex differences in physiology and evolution.

6. Structural variants

The nature and effects of structural variants

Structural variants (SVs) are mutations that encompass any changes in position and orientation of DNA sequences that involve more than 50 bp and can be classified as balanced or unbalanced (Spielmann, Lupiáñez and Mundlos, 2018). Inversions, translocations, chromosome fissions, and fusions are considered balanced SVs because the quantity of DNA present in the genome remains unaltered by the rearrangement. On the other hand, insertions, deletions, duplications, expansion or contraction of repeat arrays, and polyploidisation are mutations that affect the copy number of genomic regions, thus the quantity of DNA between individuals carrying different alleles (unbalanced SVs).

Structural variants were first discovered as chromosomal inversion back in the rolling 20's of the last century (Sturtevant, 1921), but it took important technological advancements in sequencing to be able to detect them reliably and at large scales (e.g., genomic and taxonomic scales) (Wellenreuther *et al.*, 2019; Berdan *et al.*, 2021). Indeed, SVs are a part of the genomic variability that is harder to detect than SNPs because often there is an intrinsic difficulty to map them on a reference genome (Carvalho and Lupski, 2016; Tigano, 2020). For example, variability in repeat arrays with respect to a reference is almost impossible to reliably identify with short reads when the monomers and arrays are longer than single reads (Sedlazeck *et al.*, 2018). Similarly, the sequences of recent segmental duplications or of recent expansions of gene families often collapse into few or single contigs resulting in the underestimation of their copy number (Sedlazeck *et al.*, 2018). However, elevated haplotype variability poses instead the opposite problem, namely the false duplications of single copy regions. Recently, this kind of problem was highlighted by Vertebrate Genome Project reporting false gene duplications in bird genome assemblies due to haplotype divergence and in minor proportion to sequencing errors (Kim *et al.*, 2021; Ko *et al.*, 2021; Rhie *et al.*, 2021). Indeed, multi-platform assemblies that integrate long reads and long-range scaffolding data are helping in reconstructing and detecting SVs, but the methods are not infallible. A recent benchmark of SV detection comparing methods based on short and long reads highlighted how long reads generally outperform short reads except for large copy number variants (Zhao *et al.*, 2021). These SVs are better identified by depth-based approaches using short reads. To get the best results in identifying and validating SVs, it is necessary then to incorporate orthogonal types of evidence (e.g., long and short reads together with Hi-C).

A growing body of studies are linking structural variants to a plethora of genomic, physiological and macroevolutionary effects (Wellenreuther *et al.*, 2019; Shanta *et al.*, 2020; Berdan *et al.*, 2021; Zhang *et al.*, 2021). SVs are found to change the chromatin state of the genomic regions affected by the rearrangement and those close to it (Shanta *et al.*, 2020). Inversions and translocations, for example, can dislocate genes into heterochromatic regions and thus change their expression. Copy number variation of genes influence their expression and dosage with repercussions on the levels of dominance and penetrance of specific alleles. Transposable element insertions, inversions, and centromere shifts can shape the recombination rate locally or change the entire recombination landscape of chromosomes (Berdan *et al.*, 2021). Furthermore, SVs can change the rate of other types of mutations. For example, an inversion that reduces the recombination rate in a region can favour the accumulation of transposable elements (Kent, Uzunović and Wright, 2017) and triggers events of ectopic recombination (Kapusta, Suh and Feschotte, 2017; Kent, Uzunović and Wright, 2017; Jedlicka, Lexa and Kejnovsky, 2020), thus increasing the occurrence of insertions/deletions in the population.

While SVs were at first most studied in model organisms, they are now being studied in many non-model organisms leading to discoveries of fascinating underlying mechanisms for the evolution of phenotypic traits. One of the most famous examples of adaptation through natural selection certainly is the peppered moth *Biston betularia carbonaria*. The peppered phenotype is indeed given by an SV, in particular by the insertion of a DNA transposon within the *cortex* gene controlling wing pigmentation (Hof *et al.*, 2016). Hof *et al.* (2016) demonstrated that the TE insertion is responsible for upregulation of this gene. In *Philomachus pugnax* (ruff), a chromosomal inversion underlies the mating system of this species consisting of three different male phenotypes (Lamichhaney *et al.*, 2015). Similarly, inversions also underly the sex determination and mating systems of the fungus gnat by driving the development of two types of females (Urban *et al.*, 2020). A large copy number variation of several genes controlling colouration and thermal adaptation seem to be associated with the colour dimorphism and thermal adaptation in the seabird *Uria aalge* (Dorant *et al.*, 2020). Given all the effects found to be linked to structural variants and the difficulty in detecting them, it has been proposed that SVs may account for the missing heritability of complex phenotypic traits and diseases (Chakraborty *et al.*, 2019).

Although it would be naïve to assume that all SVs (as well as SNPs) have effects or can be adaptive, it is important to incorporate all types of mutations into evolutionary frameworks to understand how they can influence population genetic parameters and the evolution of traits among populations and species (Berdan *et al.*, 2021). For example, comprehensive models of neutral evolution can help finding more reliable signature of selection genome-wide, understanding how recombination rate changes in response to the different types of SVs, and understanding how the mutation rates of the different SVs influence one another. In the perspective of developing such comprehensive evolutionary framework, it is key to 1) collect measurements of mutation rates and population genetic effects of the different mutation types; 2) include such effects into theoretical models to get predictions of the evolutionary importance for the different types; 3) estimate the contribution to evolutionary outcomes of each mutation type. The first step toward such evolutionary framework is therefore to extend the detection of SVs in as many organisms as possible, study their rate of occurrence, diversity, and distribution across genomes and evolutionary timescales (Berdan *et al.*, 2021).

In **Paper IV**, I use a large genomic dataset of birds-of-paradise and estrildid finches to investigate the occurrence, diversity, and distribution of SVs at different taxonomic levels, namely within species, genera, and families. In the paper, I focus particularly on detecting the levels of structural variability in correspondence to those regions mostly consisting of genomic dark matter, namely the W chromosome. This chromosome is expected to harbour a low genetic variability with respect to the autosomes and Z because of its reduced effective population size and recombination rate (Charlesworth, Charlesworth and Marais, 2005; Irwin, 2018; Charlesworth, 2021). Studies on chicken and flycatcher (Berlin and Ellegren, 2004; Smeds *et al.*, 2015) found even less genetic variability than expected using short reads and SNPs as markers of diversity. Given the new and more complete W chromosome models of the paradise crow and zebra finch, in **Paper IV** I investigate the variability of such chromosome from the point of view of SVs to better understand if the occurrence of other mutations (SVs) is more prevalent than SNPs. **Paper II** and **III** show that the W accumulates more transposable elements and other repeats than the other chromosomes.

Assessing, or getting closer to, the real levels of genetic variability on the W (and genome-wide in general) is key to formulating better evolutionary models as well as quantitative models for understanding how evolutionary forces interact on the different genomic regions and understand the effects of cryptic variability. Of particular importance is to distinguish the lack of biological variability from the technical difficulty of detecting such variability because of sequencing and assembly methods. Methodological issues can introduce biases in downstream analyses and may, if undiscovered, become translated into biological interpretations. With the sampling and methods used in **Paper IV**, I find much more variability than previously discovered but note that this variability assessment is still limited to part of the W and to some SV types. I predict that my results are a conservative estimate of W variability and that much more will be revealed with even longer reads and even better assemblies.

7. Satellite DNA

Eukaryotic genomes are characterised by the presence of repetitive elements. The previous sections mostly focused on interspersed repeats but an important fraction of repetitive elements in genomes is formed by tandem repeats. As mentioned in **Section 3** and as I explore in **Paper II**, tandem repeats pose serious problems to the assembly of genomes. This entails that regions made of tandem repeat arrays, like centromeres and telomeres, are vastly underrepresented in genome assemblies. An important category of tandem repeats is satellite DNA (satDNA) that is often found associated with centromeres in animals and plants, and can have diverse effects genome-wide (Plohl, Meštrović and Mravinac, 2012; Larracuente, 2014; Hartley and O'Neill, 2019). For example, the presence/absence of satDNA arrays and their variance in length in a population can lead to epistatic effects (Jiang, Hartl and Lemos, 2010). Upon hybridisation, satDNA incompatibilities can arise between the two parental species and result into chromosome missegregation and cell death at meiosis (Dion-Côté and Barbash, 2017; Jagannathan and Yamashita, 2021). Given the potential genomic effects that satDNA can have on organisms, it is important to broaden the study of satDNA on as many organisms as possible to discover the evolutionary consequences satDNA can exert on genomes and vice versa.

satDNA is known to be fast evolving and often related species have very different satDNA content in terms of quantity of satDNA monomers and presence/absence of satDNA families. In general, the evolution of satDNA families between species follows the "library hypothesis" model (Salser *et al.*, 1976; Ruiz-Ruano *et al.*, 2016; Palacios-Gimenez *et al.*, 2020): Upon speciation, the daughter species inherit the entire collection of satDNA monomers (satellitome) from the ancestral species and these monomers will independently expand/contract in quantity and diverge in their sequence. In addition, new satDNA monomers can evolve from new sequences that invade the host genome like, for example, transposable elements. Transposable elements can directly provide seeds for satDNA arrays within their sequences (e.g., microsatellites) or can contribute to form arrays as a by-product of their preference to insert next to one another, like in the case of the *hobo* transposon in *Drosophila melanogaster* (McGurk and Barbash, 2018). One of the proposed explanations of why the satellitome (or at least part of it) evolves quickly between species is linked to the tendency of different centromere structures to compete and drive towards the egg pole during female meiosis (Malik, 2009; Iwata-Otsubo *et al.*, 2017; Drpic *et al.*, 2018). Such competition is predicted to happen on the size of the centromere arrays but also on the array composition that can both facilitate a strong association between the kinetochore and the spindle (Franke *et al.*, 2017; Hartley and O'Neill, 2019). A deep knowledge about satDNA monomers and evolution thereof can be extremely useful to detect the occurrence of the centromere drive phenomenon and to disentangle the true targets of selection in a genome.

Little is known about the content and evolution of the avian satellitome, collection of all the satellite DNA monomers present in one or more genomes. Indeed, so far, only few species or satDNA families have been studied (Shang *et al.*, 2010; Zlotina *et al.*, 2012; Weissensteiner *et al.*, 2017; Piégu *et al.*, 2018; Uno *et al.*, 2019; Westerberg, 2020; Vontzou, 2021) and a comprehensive overview of avian satellitome evolution is lacking. In **Paper V**, I widen the characterisation of satDNA in birds by using large genomic datasets covering a range of evolutionary timescales of birds-of-paradise and *Corvus* species.

Research aims

In this thesis, I explore the diversity and evolution of genomic regions that were previously inaccessible in genome assemblies and constituted genomic “dark matter”. These regions include highly repetitive genomic regions such as centromeres and W chromosomes. Specifically, I explore how current sequencing technologies can inform us about these structures, and then develop the appropriate combination of data and methods to investigate their evolution in birds-of-paradise and several other birds, including the model organisms chicken and zebra finch. Investigating the extent of hidden genetic variability is key for model and non-model organisms alike to better understand the genome evolution of species and populations. The specific aims for each of the thesis papers are described below:

Paper I – To give an overview of the completeness and incompleteness of available genome assemblies, and to discuss the main genomic and technological factors that influence genome assembly completeness.

Paper II – To test the strengths and limitations of current sequencing technologies in assembling genomes while providing new methods to evaluate assembly completeness.

Paper III – To explore the possible “toxicity” of the female-specific W chromosome of birds spanning avian diversity due to its stark accumulation of potentially active transposable elements.

Paper IV – To study the evolution of the W chromosome in birds-of-paradise from a structural and repetitive element point of view, and to contrast its hidden genetic variability to the Z chromosome and autosomes.

Paper V – To investigate the sequence and structural evolution of satellite DNA throughout the phylogeny of birds-of-paradise and other Corvides species considering short and long evolutionary timescales.

Summary of papers

Paper I

How complete are complete genome assemblies? — An avian perspective

Since the advent of Next Generation Sequencing at the beginning of 2000's, sequencing whole genomes has become affordable for any small lab, with the great result that the genomic data of thousands of species are now publicly available. Assembling genomes is far from a trivial task, indeed to produce reference genome assemblies for model organisms such as human, *Drosophila*, chicken, and *C. elegans*, a huge amount of money and time from numerous consortia have been put in place. Even with the efforts of consortia, these assemblies cannot be considered complete as they are still fragmented. For example, highly heterochromatic and repetitive genomic regions, such as centromeres and non-recombining sex chromosomes, are still missing from the assemblies of even model organisms. Even though the very high-quality assemblies of most model organisms cannot be considered complete, hundreds of genome assemblies of many species are often labelled as "complete genomes". This label is therefore misleading as final users may think to analyse unbiased genome sequences and to obtain unbiased results out of the analyses.

In this paper, we explored the completeness of publicly available avian genome assemblies by comparing assembly sizes and genome sizes (C-values) with sequencing technologies implemented to investigated the possible factors that affect assembly fragmentation. We found that most avian assemblies miss a significant portion of genome, which we predict to be mainly due to the presence of repeats and base composition issues coupled with sequencing technology choices. The common short-read sequencing protocols cause a non-uniform representation of the genomes. New long-read sequencing technologies are now able to span quite long repetitive regions and are less biased towards GC-rich and AT-rich regions, but taken alone are not able to give a complete overview of the structure of entire chromosomes. We discuss how thanks to scaffolding technologies such as optical mapping, linked reads, and chromosome conformation capture, it is finally possible to obtain chromosome-level assemblies for non-model organisms.

Paper II

Identifying the causes and consequences of assembly gaps using a multiplatform genome assembly of a bird-of-paradise

Complete genome assemblies are key to understand the depths of genome evolution. Nonetheless, nowadays no single sequencing technology alone is able to provide complete chromosome-level assemblies and we demonstrated in **Paper I** that most available assemblies are far from being complete. It is therefore important to establish the strengths and limitations of sequencing technologies in order to obtain reliable high-quality assemblies that are as unbiased as possible.

Repeats and base composition are the main factors that affect the quality of genomic data and assemblies. In this paper, we explored how currently available sequencing technologies behave with regard to these two factors. We did so by comparing the efficiency in assembling repeats, GC-rich regions and causes of assembly fragmentation using assemblies based on a single technology (Illumina/10X Genomics/PacBio) to a multiplatform reference assembly of the bird-of-paradise *Lycocorax pyrrhopterus* (paradise crow). In this multiplatform assembly we combined Illumina, 10X Genomics (linked reads), and PacBio sequencing data together with two chromosome conformation capture maps (Hi-C and Dovetail CHiCAGO).

This comparison allowed us to demonstrate that 1) it is possible to obtain a gold-quality chromosome-level assembly that includes the non-recombining W chromosome and most microchromosomes of a non-model organism for which only sub-optimal tissue samples were available; 2) many subfamilies of repeats (mainly endogenous retroviruses and satellite DNA) can be characterised only in long-read assemblies; 3) all current sequencing technologies tend to introduce gaps in assemblies with specific types of repeats; 4) a thoroughly curated multiplatform assembly helps to uncover a vast abundance of non-B DNA motifs hidden in assemblies based solely on short or linked reads.

As long as sequencing technologies are not able to sequence entire chromosomes from telomere-to-telomere (or nearly that), such a multiplatform approach to assemble genomes is required where the different strengths of technologies are combined to overcome their respective weaknesses. When combining technologies is not possible, it is still imperative to consider the completeness of assemblies during downstream analyses in order not to bias results as far as possible.

Paper III

The avian W chromosome is a refugium for endogenous retroviruses with likely effects on female-biased mutational load and genetic incompatibilities

It has been broadly observed that the non-recombining regions of Y and W chromosomes accumulate more repeats than the rest of the genome, irrespective how low the genome-wide repeat content is. We observed that in birds with highly heteromorphic sex chromosomes, the W chromosome has a repeat density of ~70% compared to the genome-wide density of ~10% and contains over half of all full-length (thus potentially active) endogenous retroviruses (ERVs) of the entire genome. Recently, studies on *Drosophila melanogaster* and *Drosophila miranda* showed that the sole presence of the Y chromosome and its active transposable elements exert a toxic effect on male individuals by shortening their lifespan with respect to females. In this paper, using high-quality chromosome-level assemblies for 6 species spanning the breadth of avian diversity, we investigated the possibility that the W chromosome can be similarly toxic by female-specific transposable element activity.

We collected genomic, transcriptomic, and proteomic data to specifically detect W-linked transposable element activity. We found that there are signatures of transposable element activity stemming from ERVs on the W chromosome, and that ERVs are more expressed and translated in females with respect to males. These results suggest that the W chromosome acts as a refugium of active elements with the possibility of an overall toxic effect. We also proposed the toxicity index as a quantitative estimate for the possible toxic level of any chromosome, so to be able to predict and test such toxic effects in further studies.

We propose that the degree of W-specific enrichment in ERVs over the rest of the genome may be an additional explanatory variable for the lifespan differences observed between sexes in birds as well as for Haldane's rule with consequences for reproductive isolation between species. We therefore suggest that the sequence content of the female-specific W chromosome can have effects far beyond its appreciated involvement in sex determination and gene dosage.

Paper IV

The hidden structural variability of avian sex chromosomes

Structural variants (SVs) are a vast source of genetic variability both within and between species. The term structural variants includes many types of DNA mutations that can be divided into balanced and unbalanced mutations. Inversions, translocations, chromosome fusions, and fissions are considered balanced SVs since the quantity of DNA within the genome does not change as a result of the rearrangement. On the other hand, the quantity of DNA present in the genome changes (increases or decreases) upon the occurrence of unbalanced SVs. These unbalanced SVs include duplications, segmental duplications, insertions, and deletions. Unlike point mutations (single-nucleotide polymorphisms, SNPs), SVs are far more difficult to precisely detect because of their variable size and often of their sequence content and context. For example, repetitive regions and elements are commonly subject to rearrangements: Tandem repeat arrays expand and contract because of unequal crossing over (among other mechanisms); active transposable elements jump into new genomic locations; highly similar repetitive elements trigger events of non-allelic homologous recombination resulting in the elimination of DNA tracts. Even when the SVs do not involve repetitive elements, they can be extremely challenging to detect. In general, SVs can remain undiscovered because of the incompleteness of genome assemblies and biases of the sequencing data.

In this paper, we used the multiplatform chromosome-level genome assembly of the paradise crow as a reference point to detect the occurrence, evolution, and distribution of insertions and deletions within the birds-of-paradise family. To do this, we produced a large set of linked-read draft genome assemblies of 13 additional birds-of-paradise species spanning the entire phylogeny of the family. This species sampling allowed us to investigate the occurrence of SVs between individuals of the same species, of the same genus and family. Then, we added long-read assemblies of zebra finch and another estrildid finch to the analysis to compare the SV distribution across chromosomes of species of a separate avian family.

The SV calling that used the paradise crow genome as reference revealed a non-uniform distribution of insertions and deletions across macrochromosomes (>40 Mb), intermediate chromosomes (>20 Mb and <40 Mb), microchromosomes (<20 Mb), and sex chromosomes. Microchromosomes were the densest in SVs and the female-specific W chromosome showed a density of SVs similar to the autosomes. The W chromosome is expected to harbour very little genetic variation with respect to the autosomes and Z, thus we also investigated the occurrence of these SVs at the population level using multiple individuals of the paradise crow. The levels of genetic diversity of SVs were very low for every chromosome taken into consideration and the W chromosome value distribution was not significantly different from the others. Next, we also assessed the levels of genetic diversity of SNPs on the different chromosomes and revealed more diversity on the W than expected.

Finally, we also investigated the evolution of the TE insertions on the W with respect to the autosomes and Z. We found that the new TE insertions accumulate more on the W and that old TE insertions tend to accumulate mutations faster on the W. The high content of young and homogeneous repeats on the W likely represents an important source of mutations for this chromosome, yet needs to be fully characterised. To conclude, these results suggest that the W chromosome is more variable than previously reported but more species and longer genomic reads are needed to quantify its true levels of variability.

Paper V

Satellite DNA evolution in Corvides inferred from long and short reads

Satellite DNA (satDNA) monomers can form highly homogeneous tandem arrays. Because of its homogeneous and repetitive nature, satDNA is one of the main causes for assembly fragmentation. Since satDNA is part of the genomic dark matter, genome assemblies are not a reliable source to study its diversity and evolution, and the current standard to characterise satDNA is to use raw short-read data. Given the paucity of studies about satDNA evolution in birds and the new availability of linked, short, and long reads for birds, we chose to investigate satDNA evolution using birds-of-paradise and crow species. The birds-of-paradise family diversified into a kaleidoscope of morphologies and mating behaviours while *Corvus* species maintained a dark monochrome plumage throughout their evolution as genus. These contrasting evolutionary paths and their relatively close phylogenetic relationship (both groups being part of Corvides superfamily) provide a unique opportunity to explore the evolutionary dynamics of the satellitome (collection of all the satDNA monomers) at different timescales of avian evolution.

satDNA is known to evolve quickly even between closely related species, with arrays and monomers expanding/contracting independently between species. In general, it is expected that the more distant two species are, the more different their satellitome is. In the context of birds-of-paradise and *Corvus* species, satDNA is expected to be more diverged in the former. However, our analyses revealed that in crows, satDNA families tend to show a fast turnover between species while birds-of-paradise satellitomes tend to be more similar between species. In order to understand which of these two modes of satellitome evolution is the most prevalent in birds, more species from other bird families must be investigated. In addition, we highlighted the presence of a surprisingly GC-rich avian satellitome with long (>1 kb) and short monomers alike. We also found key candidates for being centromeric satDNA families on the basis of the abundance, monomer size, and frequency of long arrays in long-read data.

Conclusions and Future perspectives

Life is transfer of information (Dawkins, 1976), biology is based on information especially in the form of DNA and RNA sequences. From the basic science of evolutionary biology to the most applied fields of biomedicine, researchers heavily rely on retrieving the correct and complete genome sequences of any organism, be it a magnificent bird-of-paradise or a deadly pathogen. Genomes are key to understand the evolution of life and, as Nick Lane brilliantly wrote: "*Genomes are the gateway to an enchanted land. The reams of code, 3 billion letters in our own case, read like an experimental novel, an occasionally coherent story in short chapters broken up by blocks of repetitive text, verses, blank pages, streams of consciousness: and peculiar punctuation*" (Lane, 2015).

Nowadays, genome assemblies are still full of blank pages, namely gaps in the sequences produced by the so-called "genomic dark matter". Genomic dark matter is formed by all those sequences that are systematically absent from genome assemblies like transposons, satellite DNA, and GC-rich regions because of the intrinsic difficulty in assembling them. While these sequences pose huge problems to the assembly process, they also provide an opportunity to benchmark the improvements of available technologies by measuring how much of this dark matter gets assembled. In general, it is important to understand both strength and limitations of technologies so to know where to be cautious with the biological interpretations and where to improve our technologies and methodologies further. My hope, with this thesis, is to have given some help to the readers to understand precisely that: what we can and cannot do with the sequencing data available for non-model organisms. The goal was not to undermine the value of any type of sequencing data but to raise awareness about how to use the data in order to avoid (or at least minimise) biases for downstream analyses and interpretations. Every new layer of genomic data and of new analyses is important to get closer and closer to true genetic variability within species and individuals.

More complete genomes allowed me to explore the repetitive content of the non-recombing W chromosome and highlighted its possible female-specific toxicity due to the stark accumulation of potentially active transposable elements. These results open up to questions like: Are sex-limited Y/W chromosomes bound to become toxic as they differentiate from the X/Z chromosome? Are there conditions for which this toxicity does not arise, such as the existence of particular molecular silencing mechanisms? Furthermore, assuming that the presence of structural variants in highly repetitive regions can have additional genome-wide epistatic effects, it would be interesting to directly experiment if toxic W chromosomes and different W haplotypes could act as an asymmetric reproductive barrier and/or have sex-specific fitness effects.

Usually, bird genomes are regarded to evolve slowly. Although there is evidence that the karyotype is rather stable through time as indicated by high synteny of chromosomes, I emphasise that this observation should not be taken as evidence for stability of all the regions of the genome alike. My results on the avian satellitome highlight that some genomic components can differentiate fast even between closely related species as it happens in *Corvus* species. Knowing the dynamics of satellite DNA can help in the future to understand the possible involvement in hybrid incompatibilities and the possible basis of centromere drive in birds and other vertebrates.

I am thankful to projects like the T2T for developing powerful methods to produce truly complete genome assemblies and I am hopeful that such methods will be soon available for non-human organisms as well. I eagerly look forward to the day in which each and every genome assembly will be complete but, until that day, don’t forget: Mind the gap!

Svensk Sammanfattning

Utvecklingen av genomsekvenseringsteknologier det senaste decenniet har revolutionerat hela det biologiska forskningsfältet genom att möjliggöra konstruktion och analys av i princip vilken organism som helst. Trots dessa betydande framsteg, har fullständiga genomkonstruktioner inte uppnåtts då komplexa regioner (så kallade ”mörk genomisk materia”) av genomet konsekvent saknats. Närvaron av mörk genomisk materia medför att sådana regioner och deras (eventuella) funktion inte kan uppdagas. För att kunna dra korrekta slutsatser av både evolutionära och fysiologiska studier utan att förvränga deras resultat är det viktigt att dessa mörka genomiska vrår kartläggs. I denna avhandling bidrar jag till att bredda förståelsen för hur nya sekvenseringsteknologier kan användas för att konstruera olika typer av komplexa genomiska regioner och undersöka evolutionen hos sådana regioner i fågelfylogenin. Först, använde jag olika typer av sekvenseringsdata från samma individ av paradiskråka (*Lycocorax pyrrhopterus*) för att avgöra den bästa kombinationen av teknologier och konstruktionsmetoder för att maximalt öka resolutionen hos de mörka genomiska materian. Detta inkluderade att undersöka andelen repetitiva element (transposabla element, multikopie- genfamiljer och satellit-DNA), GC-rika regioner, G-kvadraplex motiv, icke-rekombinerande könskromosomer och mikrokromosomer (kromosomer mindre än 20 Mb, typiska för fågelkaryotypen). När jag hade framställt en tillförlitlig genomreferens för paradiskråkan fokuserade jag på evolutionen av transposabla element och strukturella varianter på den icke-rekombinerande W-kromosomen (lik däggdjurens Y-kromosom) och evolutionen av satellit DNA på olika evolutionära tidsskalor. Referenskonstruktionen av paradiskråksgenomet genererat här och andra fåglar som en del av Vertebrate Genome Projekt, tillät mig att upptäcka att fåglars W-kromosom härbärgerar mer än hälften av alla fulllängds-, potentiellt aktiva, transposabla element, framförallt endogena retrovirus, som finns i genomet. Detta faktum gör W-kromosomen till ett refugium för aktiva transposabla element och kan utgöra en hon-specifik ”toxisk” kromosom och mutationslast. Det överskott av aktiva transposabla element i honor jämfört med hanar kan och spela en roll i uppkomsten av genetiska inkompatibiliteter vid hybridisering och vara en ytterligare förklaring för Haldanes regel hos fåglar. Därefter undersökte jag den genetiska variabiliteten hos paradisfåglars kromosomer som uppstår genom strukturella rearrangemang med ett speciellt fokus på W-kromosomen. Tidigare studier har påvisat en väldigt låg genetisk variation i de kodande delarna av W-kromosomen, men genom att inkludera icke-kodande regioner och strukturella varianter tillsammans med punktmutationer, var den genetiska variationen högre än vad som tidigare rapporterats. Dessa resultat antyder att mutationstakten och selektionstrycket av olika typer mutationer kan variera kraftigt längs med W-kromosomen och att alla källor till genetisk variation bör tas i beaktande för att förstå evolutionen av könsbegränsade kromosomer. Till sist använde jag olika typer av sekvenseringsdata för att undersöka evolutionen av en annan huvudkomponent av fåglarnas mörka genomiska materia: satellit-DNA. Jag undersökte detta i en fylogeni av paradisfåglar och närbesläktade kråkarter (Corvides). Jag upptäckte att fåglarnas satellitom evolverar på skiljda sätt i två olika grupper och ett mer fullständigt stickprov av arter krävs för att bestämma vilket sätt som är det vanligaste hos fåglar. Sammanfattningsvis, resultaten som presenteras i denna avhandling ger en fallstudie i hur man undersöka de mest komplexa genomiska regionerna, belyser deras möjliga evolutionära roller och uppvisar därför nödvändigheten för forskningsfältet att rikta sin strålkastare mot genomens allra mörkaste hörn och vrår. Se upp för klyftan!

Acknowledgements

I’ve dreamed to be a scientist since I was little. For some years I wanted to become a physicist but with time I got wiser and chose to become an evolutionary biologist. And now I cannot believe I’m writing the acknowledgements of my PhD thesis! This PhD has been the best I could hope for and more and I hope I can express all my gratitude for all the people who accompanied me in this marvellous journey.

I want to start with my dear lab. **Cormac** you were the first person to welcoming me in the lab and show me the department. You really made me feel at home. Thank you for all the laughs and for your lessons of English pronunciation, Marco will never forget the difference between “sheep” and “ship” (maybe). **Moos**, thank you for your kindness and guidance from the very first day. Your enthusiasm, support, happiness and smile helped me so much when I most felt like an imposter. **Anne-Marie,** thank you for starting the Slack channel of the lab! Thank you for your joy and enthusiasm, working with you has been a privilege. I really hope to work with you in the future and play Cards Against Humanity with you. **Jesper**, you’re a such a funny and bright person, thank you for all the complicated scientific discussions we had which were not always so easy to follow for me, but I always learned something new from you. Of course, thank you for giving the lab and TE Jamboree the memorable soundtrack “Hold the LINE”! And thank you for translating the Swedish summary. **James**, thank you for all the TE discussions, the music at the TE Symposium (you should’ve put Britney on though), the cheerful lunches and all the fikas around town trying all the possible kinds of pastries and kanelbulle. **Boel**, you’ve been a great and kind student to supervise, I’m very happy I could help you with your thesis. **Octavio**, thank you for all your satellite DNA knowledge and help you gave me for understanding these strange repetitive sequences that do not jump. I’m very happy we could work on some projects together, it’s been a lot of fun and I hope to keep working with you in the future. When Octavio arrived in the lab, I thought there could not exist another person as expert in satellite DNA as him but then you also arrived **Paco** and the mega-ultra satDNA expert was born: Pactavio! (Thanks Jesper for this nickname). Paco, your hard work and positive attitude have been a constant inspiration for me, thank you. **Pactavio** you’re the most amazing comic duo ever and “You’ve got crabs!” guys. **Julie**, first thank you for bringing “You’ve got crabs!” to the lab retreat, the laughs we had that evening make me still so happy. Second, thank you for being an amazing friend and neighbour. Your visits with Lilja in the garden, your cakes and the evenings spent ice-skating together at the river kept me sane throughout the pandemic. Thank you for being such a great person and scientist. **Niki**, thank you for all the great scientific discussions, you always bring up new and interesting points of view. I’m very happy you’ll start your PhD soon in Alex’s lab, it’ll be marvellous. **Ivar**, thank you for the happiness you bring to every meeting, for all the games played together and all the suggestions about fantasy and sci-fi books. You’re a very talented scientist and the true and only ENC detective. **Roberto**, thanks for sharing your incredible bioinformatics expertise with the lab. I may have lost some of the details (only a few!) but your suggestions were always very precious. You’re a brilliant researcher and I hope to meet you again soon. **Inês**, you just arrived in the lab, but your kindness and bioinformatics skills fast showed up. Thank you for the amazing Portuguese recipes and I hope we’ll be able to work together on the avian satellitome. **Augustin**, your physical presence in the lab was particularly short because of covid but your enthusiasm for science resonated in every Zoom meeting. Love your challenging questions! I’m happy you will join the lab once again soon. **Simone** you’re a brilliant and talented scientist, thanks for all the discussions at jamboree! You’re also a sweet, kind and supportive person, a special thank for the support when I got mansplained. And thank you for suggesting having lab cookouts, they are my favourite lab activity. I hope to finally hug you for the defence. **Matthias**, you weren’t technically part of the lab but you were. Thank you for introducing me to the structural variant and crow world. You’ve been like a big PhD brother to me, always ready to share wise pieces of advice about the PhD life. Special mention from the time in Naples so you won’t ever forget: “You’re a naked man with only a string of DNA!” Cremer vs. Bernardi.

Alex, I left you as the last person to thank in the lab because I find very difficult to write the right words to express how great it was to be your PhD student so, I will, instead, just tell you the most important thing I learned by being in your lab. Academia can be a lonely and psychologically violent place and most of us grow in such environment and think this is how it is, and we can do nothing about it. You showed me, and all the lab members, every day with your words and actions, that a kind, open, inclusive and supportive academia is possible. It takes an awful lot of strength to be kind as you are.

Hanna, thanks for being such an amazing person, PI, co-supervisor so full of happiness and enthusiasm, always ready to share a kind word with everyone. You’re a true model for me. Thank you for guiding Alex and me through this PhD and helping us setting realistic goals when we were a bit too ambitious. I look forward to seeing you singing at the karaoke again!

A big thank to the birds-of-paradise team: **Martin** and **Knud** your bird knowledge and enthusiasm always fascinated and inspired me. Thank you Knud for sampling BOPs Papua New Guinea in the years, without your hard work many of our projects wouldn't be possible. Martin, thank you for starting the BOP project with Alex and making my PhD project possible. I feel extremely lucky and grateful to have been able to work on these crazy birds.

PIs close to the lab. Thank you, **Dave**, for making me feel welcome in the great TE community and for being a mentor for me during your time in Uppsala. **Patric** and **Claudia**, thank you a lot for your scientific guidance and for giving me the opportunity to help you with the TE Symposium, it was a great experience. **Patric** you've also been a great collaborator who were always ready to help with analysis and manuscript revisions. The discussions about endogenous retroviruses with you have been essential these years. **Claudia**, your energy and enthusiasm are contagious and your research simply amazing, I hope to continue to discuss science with you in the future. **Reto**, thank you for all your help and feedback on the manuscripts and the great support all these years! I also want to thank **Melanie** and **Raphael**, although our time spent together in Uppsala was rather short, our joint lab retreat is one of my most peaceful and dear memory of my time here in Uppsala. Thank you, Marta **Farré Belmonte**, for being my halftime opponent, I had a lot of fun discussing with you.

I am also very grateful to the entire **TE Jamboree** group, discussing repeats with you is the best way to end the week. A special thank goes to **Anna Protasio** for the great discussions about TEs and transcription and for connecting people through your series of seminars.

In these years, my officemates have always been a bliss. **Taki**, thank you for being such a great friend and mentor. The time passed with you in the lab is one of the dearest memories I have from my PhD time. I (and Marco) miss you very much, you're the person who left Uppsala that I miss the most. I hope to see you and Maki soon. Dear **Marisol**, your arrival at the office was a true surprise, nobody told us anything but what a great surprise you've been! Thank you for all the ranting, breaks, laughs and support you gave me. You've been an awesome office mate, thank you. **Hannu**, our time shared in the office was very limited, nonetheless I have dear memories of your kindness, humour and company. Thank you for sharing with me your struggles and for letting me rant about my struggles as well. **Caesar**, **Faheema** and **Inga**, thank you for welcoming me in your office. Unfortunately, the pandemic hasn't allowed us to share much time, but I had a lot of fun with you. **Inga**, you are a strong, passionate and perseverant scientist and a model for all of us at Systematics. **Caesar**, thank you for all the rants and academic discussions we had. You've always been very friendly, positive, and supportive with me, thank you.

Faheema, you always bring a smile and happiness in the office, thank you for both the fun and serious scientific discussions.

EvoBio program: Thanks for making feel me at home. Thank you for all the lunches, fikas, program and IEG days. Also thank you all for having pancakes and pea-soup together every Thursday lunch and for sharing the after-lunch sleepiness caused by the pancakes.

Mercè, you're simply a tornado of laughs, energy and positivity. Being a teaching assistant with you has been great, you taught me a lot about teaching and softened me as a teacher. The days passed in those windowless computer labs with all kinds of students would have been a nightmare without you. You are a brilliant scientist, never stop until you found more and more unicorns. Thank you for the great atmosphere you contributed to create at EvoBio. **Lore**, I don't know where to start with you. Your unlimited knowledge of biology, your incredible intelligence and all the hard work you pull through always inspired me. More than once, I thought I wanted to be more like Lore. You're also a beautiful person and I want to thank you for all the support you gave me in these years. **Homa** you always inspired me as a scientist, for your passion and for your hard work, but it is your kind compassionate heart that inspired me the most. You were always ready to share a funny story to cheer me up. I won't ever forget when I was walking, sad, in the corridors because Taki was leaving Uppsala and you and Shadi made me laugh like crazy. **Shadi**, thank you for being such an amazing positive person and talented scientist. Thank you also for the help with placing the clues for the "Forkaplocalypse" treasure hunt! Talking about the treasure hunt, thank you **Agnese** for the great teamwork organising that event! Also thank you for the great discussions at journal clubs and seminars and all the fun ultimate frisbee matches we played in front of the department. **Venkat**, thank you also for the frisbee matches, you're a pro! Thank you for all the great parties (especially Halloween) and all the fun stories and discussions at lunch. Thank you, **Willian**, for your great scientific questions and discussions and also for your stubbornness. You and **Homa** together made the scariest pair of bio-mathematicians I have ever met (in a positive sense). Talking about math, I'd like to thank you **Agnes R**., you were the best at explaining mathematical models at journal clubs and you have the best dark humour ever (not about math though, you're very serious about that!). **Madee**, your popgen knowledge is incredible and your explanations at journal clubs have been very helpful for me. Your strength and feminist attitude helped me developing my own feminist conscience so let's smash the patriarchy! Thank you also for being a great friend, cat lover and baker. **Philipp**, you're a very kind person and talented scientist, I think I understood quantitative genetics only by discussing it with you at the book club. Thank you for all the fun chats at lunches, fikas and Fridays afterwork. **Karin**, I admire you a lot as a person and as a scientist. Thank you for our scientific and life discussions. **Jente**, I miss all your bird jokes, puns, and the Friday emails (**Aaron** had this Machiavellian plan to turn FAWNA into FUNGAL). Thanks for all your ornithological wisdom, all the Thursday pancakes, fikas, walks in the botanical gardens and laughs. By the way, since you left, I haven't seen the Uppsala dipper anymore! Coincidence?! **Paulina**, thank you for helping me with the development of a lab tutorial, I couldn't have done it without you! **Sergio**, you are just amazing and I'm really sorry we didn't have the chance to work together at EBC, your experiments on yeast super impressive! **Erik**, **Ghazal**, **Ludo**, **Mi**, **Krysha**, **Linnéa**, **Jonas**, **Zaenab**, **Toby**, **Roy**, **Paulina**, **Kerri**, thank you for all the fun and great scientific discussions we had together.

A big thanks to the senior researchers and PIs at EvoBio for the great community you create there. **Hans**, thank you for welcoming me to EvoBio and for the great journal clubs you're your lab. Thanks, **Robert**, for all the cool stories you shared with us at lunch and fika and thank you for the bandy games you organised on Friday afternoons at Studenternas. It was so much fun! **Niclas**, thank you for always cheer us up with your humour and stories at lunches. Thanks, **Simon**, for helping our book club with the math behind Markov chains. **Anna**, thanks for your passion for science, parties, karaoke and fungi-related songs; without you, organising the IEG days wouldn't be the same. **Elina** and **Arild**, thank you for your positive presence and for the good discussions. **Carina**, you always impressed me a lot for your strength and originality. Your rigour and determination in your research are inspiring. Thank you for the cool discussions at lunch. **Doug** just thank you for all the times you fixed some tools on Rackham for me or just came up with a super cool bioinformatics solution to my problems. You are always so kind and available, you always found time for me when I knocked on your door. You are a great person and on top of that you keep saving the day to many of us PhDs. Thank you, **Verena**, for teaching me Snakemake for being a great bioinformatics advisor. I had a lot of fun with you at the bioinformatics meetings in Uppsala and Stockholm.

The work at EvoBio wouldn't run so smoothly without a great admin team. Thank you, **Annette** and **Frida**, your professionality, kindness, and availability made my academic life much easier and thank you for all the nice chats at fika.

<u>Animal ecology program</u>. **Foteini**, you're bright as your name says. Thanks for all your support, kindness and hugs. **Zuzana**, I really enjoyed the time passed discussing the extended evolutionary synthesis at Cambridge and for the days visiting it. Thanks for inspiring me with all your hard work in the lab and super interesting experiments. **Carolina**, your passion for speciation and flycatchers is truly inspiring and your research is amazing. **Elizabeth**, thank

you for helping me with some delicate academic struggles and for sharing precious advice and in general for having listened to my worries and rants about academia.

Plant Ecology. **Alessandro**, **Maria**, **Giulia** grazie per essere stati la mia piccola gang italiana all'EBC. Ho sempre potuto contare su di voi quando avevo bisogno di lamentarmi o dell'Italia o della Svezia o di entrambe.

SystBio program: Thanks for welcoming me in your program and making me feel at home. **Aaron**, you're the true image of passion in science, the fungal and TE communities are lucky to have you. Thanks for all the discussions at TE Jamboree and all the Fridays at FAWNA/FLORA/FUNGAL. Live long and prosper! **Mahwash**, we didn't have many occasions to interact but every time we did, I thought you were a very kind, passionate and knowledgeable person. **Ioana**, you're a volcano of energy and ideas. Thank you for all the passion you put in everything you do. A special thank for the seminar series! **Diem**, you're a special friend, always ready to help everyone and you did help me many times. You're a concentrate of wisdom and yarn and I'm very lucky I could both discuss with you for hours about life while learning how to knit from you. **Raquel**, your passion for sponges, taxonomy and biology is inspiring, you're great! I really enjoyed all the time we had the occasion to spend together, thank you for all the fun (and the memes). **Ivain**, **Jenny**, **Sanea**, **Iker**, **Brendan**, **Anneli**, **Jesper**, **Stella**, **Markus** thank you for making or having made EBC and the SystBio program an incredible place. A special thank for you **Martin** that made us feel home from the very first moment.

Human evolution program. **Luciana** you literally saved my PhD, without you probably I wouldn't have my PhD defence properly registered. Thank you for saving me and for all your kindness and positive energies you always irradiate to the world. **Gwenna**, **Alex**, **Mario**, **James**, **TJ**, thank you for all the lunches in fika room.

Götgatan: **Moa**, **Jesper** and **Crille**, thank you for being the best landlords ever, I'm looking forward to having another party with you (the best parties in town!). Thank you, **Gunnel**, for all the cool anecdotes about Uppsala, Sweden, birds and snakes.

Enyulla beta. Ragazzi che vi posso dire? Grazie per i meravigliosi anni dell'università, grazie per l'incondizionato supporto che mi avete sempre dato e per le scorpacciate di crescentine. Vabbè insomma non fatemi diventare troppo sdolcinata che poi Pè dice che sono phalsa. Vi voglio bene a tutti **Francy**, **Bru**, **Matte**, **Mengo**, **Loura**, **Pè**, **Ale**, **Mary** ma un po' di più alla mia BFF **Leanne (**perché Leanne è Leanne regaz).

Diana Didi Diduz, semplicemente grazie per essere una grandiosa amica che c'è sempre stata per me, per le picolezze e per le cose importanti. Sono davvero impedita con le parole quando devo scrivere alle persone importanti quindi ti becchi un ringraziamento monco. Grazie per gli incoraggiamenti, soprattutto quelli di questa estate quando non riuscivo a staccare la testa dal lavoro per lo stress. Vorrei tanto tornare al mare ad Ancona da te, Etienne e Kaylee! **Etienne**, grazie per essere un grande amico, sono così contenta che Cupido Ricci abbia fatto bene il suo lavoro e le mie persone preferite siano perfettamente assieme! Ok questa era sdolcinata ma spero abbiate capito che vi voglio semplicemente molto bene. Grazie Etienne per tutta la passione per i trasposoni che mi hai trasmessa negli anni.

Bolo Lab. **Boa**, grazie per aver ospitato Marco e me nel tuo laboratorio, grazie per aver creduto in noi e nei trasposoni che adesso sono diventati i tuoi migliori amici! **Cristian**, grazie per mettere sempre in discussione quello che crediamo di sapere di biologia, mi ha aiutato tanto a scoprire nuovi fenomeni o vedere quelli già conosciuti in un'altra ottica. **Piermassimo**, grazie per la tua gentilezza, disponibilità e supporto informatico quando Marco ed io ne avevamo più bisogno. Ce ne vorrebbero altri mille come te.

Family. **Carl Gustav**, you little devil in the shape of a fluffy cutie cat! Thank you for keeping me company during the entire pandemic and supervising me even though meowing like crazy at 5 a.m. in the garden is not that great! **Marco**, I don't really know how to thank you other than saying that without you I wouldn't be here in Uppsala writing this thesis and you know how much I love my PhD. Thank you.

Nadia e **Franco**, grazie per essere sempre così gentili, ospitali e divertenti, mi fate sempre sentire come a casa. **Zio Italo**, anche se sei in Italia, ti sentiamo sempre vicino grazie alle tue telefonate. Pochi hanno la capacità di illuminare le giornate alle persone come te con il tuo umorismo.

Mamma, **papà**, grazie per avermi sempre sostenuta e per aver creduto in me anche più di quanto io creda in me stessa. **Zio** grazie per aver sempre guardato Sailor Moon insieme a me, per essere sempre venuto a vedere tutte le mie partite di pallavolo e per tutto l'aiuto che mi hai dato in questi anni. Grazie per avermi avvicinata alla scienza, se mi sto dottorando è anche merito tuo. **Nonna**, **nonno** so che voi non avete mai capito bene cosa faccia una biologa ma poco importa perché anche senza saperlo mi avete sempre sostenuta, grazie! **Zio Gigi**, **zia Orietta**, **Laura**, **Luca** grazie anche a voi per essermi stati sempre vicini. Un grazie anche a **Federico**, **Emanuela**, **zia Giuseppina**, **zia Angela** che mi avete sempre dimostrato affetto e supporto.

Last but not least, I would like to thank all the scientists involved in the development of the COVID vaccines that are protecting people even from themselves. A true beacon of light and hope in these dark times.

References

Ambardar, S. *et al.* (2016) 'High Throughput Sequencing: An Overview of Sequencing Chemistry', *Indian Journal of Microbiology*. Springer, 56(4), pp. 394–404. doi: 10.1007/s12088-016-0606-4.

Armstrong, E. E. *et al.* (2018) 'Cost-effective assembly of the African wild dog (Lycaon pictus) genome using linked reads', *GigaScience*, 8(2), p. giy124. doi: 10.1093/gigascience/giy124.

Axelsson, E. *et al.* (2004) 'Male-biased mutation rate and divergence in autosomal, Z-linked and W-linked introns of chicken and turkey', *Molecular Biology and Evolution*, 21(8), pp. 1538–1547. doi: 10.1093/molbev/msh157.

Backström, N. *et al.* (2010) 'The recombination landscape of the zebra finch Taeniopygia guttata genome', *Genome Research*, 20(4), pp. 485–495. doi: 10.1101/gr.101410.109.

Ballantyne, K. N. *et al.* (2010) 'Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications', *The American Journal of Human Genetics*. Elsevier, 87(3), pp. 341–353.

Balloux, F. and Lugon-Moulin, N. (2002) 'The estimation of population differentiation with microsatellite markers', *Molecular Ecology*. Wiley Online Library, 11(2), pp. 155–165. doi: 10.1046/j.0962-1083.2001.01436.x.

Bao, W., Kojima, K. K. and Kohany, O. (2015) 'Repbase Update, a database of repetitive elements in eukaryotic genomes', *Mobile DNA*, 6(1), p. 11. doi: 10.1186/s13100-015-0041-9.

Bartolomé, C., Maside, X. and Charlesworth, B. (2002) 'On the abundance and distribution of transposable elements in the genome of Drosophila melanogaster', *Molecular Biology and Evolution*. Oxford University Press, 19(6), pp. 926–937. doi: 10.1093/oxfordjournals.molbev.a004150.

Beauclair, L. *et al.* (2019) 'Sequence properties of certain GC rich avian genes, their origins and absence from genome assemblies: Case studies', *BMC Genomics*. BioMed Central, 20(1), pp. 1–16. doi: 10.1186/s12864-019-6131-1.

Bell, J. M. *et al.* (2017) 'Chromosome-scale mega-haplotypes enable digital karyotyping of cancer aneuploidy', *Nucleic Acids Research*, 45(19), p. e162. doi: 10.1093/nar/gkx712.

Bellott, D. W. *et al.* (2017) 'Avian W and mammalian y chromosomes convergently retained dosage-sensitive regulators', *Nature Genetics*, 49(3), pp. 387–394. doi: 10.1038/ng.3778.

Bellott, D. W. and Page, D. C. (2021) 'Dosage-sensitive functions in embryonic development drove the survival of genes on sex-specific chromosomes in snakes, birds, and mammals', *Genome Research*, 31(2), pp. 198–210. doi: 10.1101/GR.268516.120.

Berdan, E. L. *et al.* (2021) 'Unboxing mutations: Connecting mutation types with evolutionary consequences', *Molecular Ecology*. John Wiley & Sons, Ltd, 30(12), pp. 2710–2723. doi: 10.1111/mec.15936.

Berlin, S. and Ellegren, H. (2004) ‘Chicken W: A genetically uniform chromosome in a highly variable genome’, *Proceedings of the National Academy of Sciences of the United States of America*, 101(45), pp. 15967–15969. doi: 10.1073/pnas.0405126101.

Beukeboom, L. W. and Perrin, N. (2014) *The Evolution of Sex Determination*, *The Evolution of Sex Determination*. Oxford University Press, USA. doi: 10.1093/acprof:oso/9780199657148.001.0001.

Bickhart, D. M. *et al.* (2017) ‘Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome’, *Nature Genetics*. Nature Publishing Group, 49(4), pp. 643–650. doi: 10.1038/ng.3802.

Blom, M. P. K. and Irestedt, M. (2021) ‘Genome-level phylogenomics of all species of birds-of-paradise’, *in preparation*.

Bolisetty, M. *et al.* (2012) ‘Unexpected diversity and expression of avian endogenous retroviruses’, *mBio*. Am Soc Microbiol, 3(5), pp. e00344-12. doi: 10.1128/mBio.00344-12.

Bolívar, P. *et al.* (2019) ‘GC-biased gene conversion conceals the prediction of the nearly neutral theory in avian genomes 06 Biological Sciences 0604 Genetics’, *Genome Biology*, 20(1), p. 5. doi: 10.1186/s13059-018-1613-z.

Boman, J. *et al.* (2019) ‘The genome of blue-capped cordon-bleu uncovers hidden diversity of ltr retrotransposons in zebra finch’, *Genes*. doi: 10.3390/genes10040301.

Botero-Castro, F. *et al.* (2017) ‘Avian genomes revisited: Hidden genes uncovered and the rates versus traits paradox in birds’, *Molecular Biology and Evolution*. Oxford University Press, 34(12), pp. 3123–3131. doi: 10.1093/molbev/msx236.

Boulay, G. *et al.* (2018) ‘Epigenome editing of microsatellite repeats defines tumor-specific enhancer functions and dependencies’, *Genes and Development*. Cold Spring Harbor Lab, 32(15–16), pp. 1008–1019. doi: 10.1101/gad.315192.118.

Bowen, N. J. and Jordan, I. K. (2002) ‘Transposable elements and the evolution of eukaryotic complexity’, *Current Issues in Molecular Biology*, pp. 65–76. doi: 10.21775/cimb.004.065.

Broecker, F. and Moelling, K. (2019) ‘Evolution of immune systems from viruses and transposable elements’, *Frontiers in Microbiology*, p. 51. doi: 10.3389/fmicb.2019.00051.

Brown, E. J., Nguyen, A. H. and Bachtrog, D. (2020a) ‘The drosophila Y chromosome affects heterochromatin integrity genome-wide’, *Molecular Biology and Evolution*. Edited by J. Parsch, 37(10), pp. 2808–2824. doi: 10.1093/molbev/msaa082.

Brown, E. J., Nguyen, A. H. and Bachtrog, D. (2020b) ‘The Y chromosome may contribute to sex-specific ageing in Drosophila’, *Nature Ecology and Evolution*, 4(6), pp. 853–862. doi: 10.1038/s41559-020-1179-5.

Brown, J. D. and O’Neill, R. J. (2010) ‘Chromosomes, conflict, and epigenetics: Chromosomal speciation revisited’, *Annual Review of Genomics and Human Genetics*. Annual Reviews, 11(1), pp. 291–316. doi: 10.1146/annurev-genom-082509-141554.

Burt, D. W. (2002) ‘Origin and evolution of avian microchromosomes’, *Cytogenetic and Genome Research*, 96(1–4), pp. 97–112. doi: 10.1159/000063018.

Carvalho, C. M. B. and Lupski, J. R. (2016) ‘Mechanisms underlying structural variant formation in genomic disorders’, *Nature Reviews Genetics*, 17(4), pp. 224–238. doi: 10.1038/nrg.2015.25.

Chakraborty, M. *et al.* (2019) ‘Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits’, *Nature Communications*, 10(1), p. 4872. doi: 10.1038/s41467-019-12884-1.

Craig, N. L., *et al.* (2015) *Mobile DNA III*. ASM Books.

Charlesworth, B. and Charlesworth, D. (1978) ‘A Model for the Evolution of Dioecy and Gynodioecy’, *The American Naturalist*. The University of Chicago Press, 112(988), pp. 975–997. doi: 10.1086/283342.

Charlesworth, D. (2021) ‘When and how do sex-linked regions become sex chromosomes?’, *Evolution*, 75(3), pp. 569–581. doi: 10.1111/evo.14196.

Charlesworth, D., Charlesworth, B. and Marais, G. (2005) ‘Steps in the evolution of heteromorphic sex chromosomes’, *Heredity*, 95(2), pp. 118–128. doi: 10.1038/sj.hdy.6800697.

Chippindale, A. K. and Rice, W. R. (2001) ‘Y chromosome polymorphism is a strong determinant of male fitness in Drosophila melanogaster’, *Proceedings of the National Academy of Sciences of the United States of America*. 2001/04/24. The National Academy of Sciences, 98(10), pp. 5677–5682. doi: 10.1073/pnas.101456898.

Choi, J. and Majima, T. (2011) ‘Conformational changes of non-B DNA’, *Chemical Society Reviews*. Royal Society of Chemistry, 40(12), pp. 5893–5909.

Chuong, E. B., Elde, N. C. and Feschotte, C. (2017) ‘Regulatory activities of transposable elements: From conflicts to benefits’, *Nature Reviews Genetics*. Nature Publishing Group, 18(2), pp. 71–86. doi: 10.1038/nrg.2016.139.

Clutton-Brock, T. H. and Isvaran, K. (2007) ‘Sex differences in ageing in natural populations of vertebrates’, *Proceedings of the Royal Society B: Biological Sciences*. Royal Society, 274(1629), pp. 3097–3104. doi: 10.1098/rspb.2007.1138.

Coelho, L. A., Musher, L. J. and Cracraft, J. (2019) ‘A multireference-based whole genome assembly for the obligate ant-following antbird, Rhegmatorhina melanosticta (Thamnophilidae)’, *Diversity*. 11(9), pp. 144. doi: 10.3390/d11090144.

Coulthard, A. B. *et al.* (2016) ‘Meiotic recombination is suppressed near the histone-defined border of euchromatin and heterochromatin on chromosome 2L of Drosophila melanogaster’, *Genome*. NRC Research Press, 59(4), pp. 289–294. doi: 10.1139/gen-2015-0171.

Dame, J. B. *et al.* (1996) ‘Current status of the Plasmodium falciparum genome project’, *Molecular and Biochemical Parasitology*. Elsevier, 79(1), pp. 1–12. doi: 10.1016/0166-6851(96)02641-2.

Davis, J. K., Thomas, P. J. and Thomas, J. W. (2010) ‘A W-linked palindrome and gene conversion in New World sparrows and blackbirds’, *Chromosome Research*, 18(5), pp. 543–553. doi: 10.1007/s10577-010-9134-y.

Dawkins, R. (1976) *The Selfish Gene*. Oxford University Press.

Deamer, D., Akeson, M. and Branton, D. (2016) ‘Three decades of nanopore sequencing’, *Nature Biotechnology*. Nature Publishing Group, 34(5), pp. 518–524. doi: 10.1038/nbt.3423.

Degrandi, T. M. *et al.* (2020) ‘Introducing the Bird Chromosome Database: An Overview of Cytogenetic Studies in Birds’, *Cytogenetic and Genome Research*, 160(4), pp. 199–205. doi: 10.1159/000507768.

Dekker, J., Marti-Renom, M. A. and Mirny, L. A. (2013) ‘Exploring the three-dimensional organization of genomes: Interpreting chromatin interaction data’, *Nature Reviews Genetics*. Nature Publishing Group, 14(6), pp. 390–403. doi: 10.1038/nrg3454.

Delph, L. F. and Demuth, J. P. (2016) ‘Haldane’s rule: Genetic bases and their empirical support’, *Journal of Heredity*, 107(5), pp. 383–391. doi: 10.1093/jhered/esw026.

Deniz, Ö., Frost, J. M. and Branco, M. R. (2019) ‘Regulation of transposable elements by DNA modifications’, *Nature Reviews Genetics*. England, 20(7), pp. 417–431. doi: 10.1038/s41576-019-0106-6.

Denli, A. M. *et al.* (2015) ‘Primate-Specific ORF0 Contributes to Retrotransposon-Mediated Diversity’, *Cell*, 163(3), pp. 583–593. doi: 10.1016/j.cell.2015.09.025.

Dion-Côté, A. M. and Barbash, D. A. (2017) ‘Beyond speciation genes: an overview of genome stability in evolution and speciation’, *Current Opinion in Genetics and Development*, 47, pp. 17–23. doi: 10.1016/j.gde.2017.07.014.

Domínguez, M. *et al.* (2020) ‘The impact of transposable elements on tomato diversity’, *Nature Communications*. The Company of Biologists Ltd, 11(1), pp. 4101–4114. doi: 10.1038/s41467-020-17874-2.

Donald, P. F. (2007) ‘Adult sex ratios in wild bird populations’, *Ibis*. John Wiley & Sons, Ltd, 149(4), pp. 671–692. doi: 10.1111/j.1474-919X.2007.00724.x.

Dorant, Y. *et al.* (2020) ‘Copy number variants outperform SNPs to reveal genotype–temperature association in a marine species’, *Molecular Ecology*. John Wiley & Sons, Ltd, 29(24), pp. 4765–4782. doi: 10.1111/mec.15565.

Dréau, A. *et al.* (2019) ‘Genome-wide recombination map construction from single individuals using linked-read sequencing’, *Nature Communications*. Nature Publishing Group, 10(1), pp. 1–11. doi: 10.1038/s41467-019-12210-9.

Drpic, D. *et al.* (2018) ‘Chromosome Segregation Is Biased by Kinetochore Size’, *Current Biology*. Elsevier, 28(9), pp. 1344-1356.e5. doi: 10.1016/j.cub.2018.03.023.

Dudchenko, O. *et al.* (2018) ‘The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000’, *bioRxiv*, p. 254797. doi: 10.1101/254797.

Ellegren, H. (2004) ‘Microsatellites: Simple sequences with complex evolution’, *Nature Reviews Genetics*, 5(6), pp. 435–445. doi: 10.1038/nrg1348.

Ellegren, H. (2010) ‘Evolutionary stasis: the stable chromosomes of birds’, *Trends in Ecology and Evolution*, 25(5), pp. 283–291. doi: 10.1016/j.tree.2009.12.004.

Emera, D. and Wagner, G. P. (2012) ‘Transposable element recruitments in the mammalian placenta: Impacts and mechanisms’, *Briefings in Functional Genomics*. Oxford University Press, 11(4), pp. 267–276. doi: 10.1093/bfgp/els013.

English, A. C. *et al.* (2012) ‘Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology’, *PLoS ONE*. Public Library of Science San Francisco, USA, 7(11), p. e47768. doi: 10.1371/journal.pone.0047768.

Ewing, B. *et al.* (1998) ‘Base-calling of automated sequencer traces using phred. I. Accuracy assessment’, *Genome Research*. Cold Spring Harbor Lab, 8(3), pp. 175–185. doi: 10.1101/gr.8.3.175.

Ewing, B. and Green, P. (1998) ‘Base-calling of automated sequencer traces using phred. II. Error probabilities’, *Genome Research*. Cold Spring Harbor Lab, 8(3), pp. 186–194. doi: 10.1101/gr.8.3.186.

Faino, L. *et al.* (2015) ‘Single-molecule real-time sequencing combined with optical mapping yields completely finished fungal genome’, *mBio*. Am Soc Microbiol, 6(4), pp. e00936-15. doi: 10.1128/mBio.00936-15.

Farré, M. *et al.* (2016) ‘Novel insights into chromosome evolution in birds, archosaurs, and reptiles’, *Genome Biology and Evolution*, 8(8), pp. 2442–2451. doi: 10.1093/gbe/evw166.

Feng, S. *et al.* (2020) ‘Dense sampling of bird diversity increases power of comparative genomics’, *Nature*. Nature Publishing Group, 587, pp. 252–257. doi: 10.1038/s41586-020-2873-9.

Ferree, P. M. and Barbash, D. A. (2009) ‘Species-specific heterochromatin prevents mitotic chromosome segregation to cause hybrid lethality in Drosophila’, *PLoS Biology*. Public Library of Science, 7(10), p. e1000234. doi: 10.1371/journal.pbio.1000234.

Ferree, P. M. and Prasad, S. (2012) ‘How Can Satellite DNA Divergence Cause Reproductive Isolation? Let Us Count the Chromosomal Ways’, *Genetics Research International*. Edited by V. Sollars. Hindawi Publishing Corporation, 2012, pp. 1–11. doi: 10.1155/2012/430136.

Feschotte, C. and Pritham, E. J. (2007) ‘DNA transposons and the evolution of eukaryotic genomes’, *Annual Review of Genetics*. Annual Reviews, 41(1), pp. 331–368. doi: 10.1146/annurev.genet.40.110405.090448.

Flament, S. A. *et al.* (2011) ‘Sex Determination and Sexual Differentiation in Amphibians’, *Hormones and Reproduction of Vertebrates - Volume 2*. Wiley Online Library, 278(7), pp. 1–19. doi: 10.1016/B978-0-12-374931-4.10001-X.

Francisco, F. O. and Lemos, B. (2014) ‘How Do Y-Chromosomes Modulate Genome-Wide Epigenetic States: Genome Folding, Chromatin Sinks, and Gene Expression’, *Journal of Genomics*. Ivyspring International Publisher, 2, pp. 94–103. doi: 10.7150/jgen.8043.

Franke, V. *et al.* (2017) ‘Long terminal repeats power evolution of genes and gene expression programs in mammalian oocytes and zygotes’, *Genome Research*, 27(8), pp. 1384–1394. doi: 10.1101/gr.216150.116.

Fridolfsson, A.-K. *et al.* (1998) ‘Evolution of the avian sex chromosomes from an ancestral pair of autosomes’, *Proceedings of the National Academy of Sciences*, 95(14), pp. 8147 LP – 8152. doi: 10.1073/pnas.95.14.8147.

Furman, B. L. S. *et al.* (2020) ‘Sex Chromosome Evolution: So Many Exceptions to the Rules’, *Genome Biology and Evolution*, 12(6), pp. 750–763. doi: 10.1093/gbe/evaa081.

Galagan, J. E. and Selker, E. U. (2004) ‘RIP: The evolutionary cost of genome defense’, *Trends in Genetics*. Elsevier, 20(9), pp. 417–423. doi: 10.1016/j.tig.2004.07.007.

Galbraith, J. D. *et al.* (2021) ‘Genome stability is in the eye of the beholder: recent retrotransposon activity varies significantly across avian diversity’, *bioRxiv*, p. 2021.04.13.439746. doi: 10.1101/2021.04.13.439746.

Galtier, N. *et al.* (2009) ‘GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates’, *Trends in Genetics*. Elsevier, 25(1), pp. 1-5. doi: 10.1016/j.tig.2008.10.011.

Ghurye, J. *et al.* (2017) ‘Scaffolding of long read assemblies using long range contact information’, *BMC Genomics*. BioMed Central, 18(1), pp. 1–11. doi: 10.1186/s12864-017-3879-z.

Gregory, T. R. *et al.* (2007) ‘Eukaryotic genome size databases’, *Nucleic Acids Research*, 35(SUPPL. 1), pp. D332–D338. doi: 10.1093/nar/gkl828.

Griffin, D. and Burt, D. W. (2014) ‘All chromosomes great and small: 10 Years on’, *Chromosome Research*, 22(1), pp. 1–6. doi: 10.1007/s10577-014-9413-0.

Guiblet, W. M. *et al.* (2018) ‘Long-read sequencing technology indicates genome-wide effects of non-B DNA on polymerization speed and error rate’, *Genome Research*. Cold Spring Harbor Lab, 28(12), pp. 1767–1778. doi: 10.1101/gr.241257.118.

Haldane, J. B. S. (1922) ‘Sex ratio and unisexual sterility in hybrid animals’, *Journal of Genetics*, 12(2), pp. 101–109. doi: 10.1007/BF02983075.

Handley, L. J. L., Ceplitis, H. and Ellegren, H. (2004) ‘Evolutionary strata on the chicken Z chromosome: Implications for sex chromosome evolution’, *Genetics*. Oxford University Press, 167(1), pp. 367–376. doi: 10.1534/genetics.167.1.367.

Hartley, G. and O’Neill, R. J. (2019) ‘Centromere repeats: Hidden gems of the genome’, *Genes*. doi: 10.3390/genes10030223.

Havecker, E. R., Gao, X. and Voytas, D. F. (2004) ‘The diversity of LTR retrotransposons’, *Genome Biology*, 5(6), p. 225. doi: 10.1186/gb-2004-5-6-225.

van Heesch, S. *et al.* (2013) ‘Improving mammalian genome scaffolding using large insert mate-pair next-generation sequencing’, *BMC Genomics*. BioMed Central, 14(1), pp. 1–11. doi: 10.1186/1471-2164-14-257.

Hill, T., Schlötterer, C. and Betancourt, A. J. (2016) ‘Hybrid Dysgenesis in Drosophila simulans Associated with a Rapid Invasion of the P-Element’, *PLoS Genetics*. Public Library of Science, 12(3), p. e1005920. doi: 10.1371/journal.pgen.1005920.

Hillier, L. W. *et al.* (2004) ‘Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution’, *Nature*, 432(7018), pp. 695–716. doi: 10.1038/nature03154.

Hof, A. E. V. t. *et al.* (2016) ‘The industrial melanism mutation in British peppered moths is a transposable element’, *Nature*, 534(7605), pp. 102–105. doi: 10.1038/nature17951.

Holley, W. R. *et al.* (1965) ‘Nucleotide Sequences in the Yeast Alanine Transfer Ribonucleic Acid.’, *The Journal of Biological Chemistry*. ASBMB, 240(5), pp. 2122–2128. doi: 10.1016/s0021-9258(18)97435-1.

Hooper, D. M. and Price, T. D. (2017) ‘Chromosomal inversion differences correlate with range overlap in passerine birds’, *Nature Ecology and Evolution*, 1(10), pp. 1526–1534. doi: 10.1038/s41559-017-0284-6.

Hron, T. *et al.* (2015) ‘Hidden genes in birds’, *Genome Biology*. Springer, 16(1), pp. 1–4. doi: 10.1186/s13059-015-0724-z.

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. 2681–2688. doi: 10.1093/molbev/msn207.

Irestedt, M. *et al.* (2009) ‘An unexpectedly long history of sexual selection in birds-of-paradise’, *BMC Evolutionary Biology*, 9(1), p. 235. doi: 10.1186/1471-2148-9-235.

Ironside, J. E. (2010) ‘No amicable divorce? Challenging the notion that sexual antagonism drives sex chromosome evolution’, *BioEssays*. Wiley Online Library, 32(8), pp. 718–726. doi: 10.1002/bies.200900124.

Irwin, D. E. (2018) ‘Sex chromosomes and speciation in birds and other ZW systems’, *Molecular Ecology*. John Wiley & Sons, Ltd, 27(19), pp. 3831–3851. doi: 10.1111/mec.14537.

Iwata-Otsubo, A. *et al.* (2017) ‘Expanded Satellite Repeats Amplify a Discrete CENP-A Nucleosome Assembly Site on Chromosomes that Drive in Female Meiosis’, *Current Biology*, 27(15), pp. 2365-2373.e8. doi: 10.1016/j.cub.2017.06.069.

Jagannathan, M., Cummings, R. and Yamashita, Y. M. (2018) ‘A conserved function for pericentromeric satellite DNA’, *eLife*. 7, p. e34122. doi: 10.7554/eLife.34122.

Jagannathan, M. and Yamashita, Y. M. (2021) ‘Defective satellite DNA clustering into chromocenters underlies hybrid incompatibility in Drosophila’, *Molecular Biology and Evolution*, p. 2021.04.16.440167. doi: 10.1093/molbev/msab221.

Jain, M. *et al.* (2018) ‘Linear assembly of a human centromere on the Y chromosome’, *Nature Biotechnology*, 36(4), pp. 321–323. doi: 10.1038/nbt.4109.

Jedlicka, P., Lexa, M. and Kejnovsky, E. (2020) 'What Can Long Terminal Repeats Tell Us About the Age of LTR Retrotransposons, Gene Conversion and Ectopic Recombination?', *Frontiers in Plant Science*. Frontiers Media S.A., 11, p. 644. doi: 10.3389/fpls.2020.00644.

Jiang, P. P., Hartl, D. L. and Lemos, B. (2010) 'Y not a dead end: Epistatic interactions between Y-linked regulatory polymorphisms and genetic background affect global gene expression in Drosophila melanogaster', *Genetics*, 186(1), pp. 109–118. doi: 10.1534/genetics.110.118109.

Johnson, J. M. *et al.* (2005) 'Dark matter in the genome: Evidence of widespread transcription detected by microarray tiling experiments', *Trends in Genetics*. Elsevier, 21(2), pp. 93–102. doi: 10.1016/j.tig.2004.12.009.

Jønsson, K. A. *et al.* (2016) 'A supermatrix phylogeny of corvoid passerine birds (Aves: Corvides)', *Molecular Phylogenetics and Evolution*, 94, pp. 87–94. doi: 10.1016/j.ympev.2015.08.020.

Kadota, M. *et al.* (2020) 'Multifaceted Hi-C benchmarking: what makes a difference in chromosome-scale genome scaffolding?', *Gigascience*. Oxford University Press, 9(1), p. giz158.

Kapitonov, V. V. and Koonin, E. V. (2015) 'Evolution of the RAG1-RAG2 locus: Both proteins came from the same transposon', *Biology Direct*, 10(1), p. 20. doi: 10.1186/s13062-015-0055-8.

Kapusta, A. and Suh, A. (2017) 'Evolution of bird genomes—a transposon's-eye view', *Annals of the New York Academy of Sciences*. John Wiley & Sons, Ltd, 1389(1), pp. 164–185. doi: 10.1111/nyas.13295.

Kapusta, A., Suh, A. and Feschotte, C. (2017) 'Dynamics of genome size evolution in birds and mammals', *Proceedings of the National Academy of Sciences of the United States of America*, 114(8), pp. E1460–E1469. doi: 10.1073/pnas.1616702114.

Kawakami, T. *et al.* (2014) 'A high-density linkage map enables a second-generation collared flycatcher genome assembly and reveals the patterns of avian recombination rate variation and chromosomal evolution', *Molecular Ecology*. John Wiley & Sons, Ltd, 23(16), pp. 4035–4058. doi: 10.1111/mec.12810.

Kawakami, T. *et al.* (2017) 'Whole-genome patterns of linkage disequilibrium across flycatcher populations clarify the causes and consequences of fine-scale recombination rate variation in birds', *Molecular Ecology*. John Wiley & Sons, Ltd, 26(16), pp. 4158–4172. doi: 10.1111/mec.14197.

Kazazian, H. H. (2004) 'Mobile Elements: Drivers of Genome Evolution', *Science*, 303(5664), pp. 1626–1632. doi: 10.1126/science.1089670.

Kazazian, H. H. and Goodier, J. L. (2002) 'LINE drive: Retrotransposition and genome instability', *Cell*. Elsevier, 110(3), pp. 277–280. doi: 10.1016/S0092-8674(02)00868-1.

Kent, T. V., Uzunović, J. and Wright, S. I. (2017) 'Coevolution between transposable elements and recombination', *Philosophical Transactions of the Royal Society B: Biological Sciences*. Royal Society, 372(1736), p. 20160458. doi: 10.1098/rstb.2016.0458.

Kiazim, L. G. *et al.* (2021) 'Comparative Mapping of the Macrochromosomes of Eight Avian Species Provides Further Insight into Their Phylogenetic Relationships and Avian Karyotype Evolution', *Cells*. 10(2), p. 362 doi: 10.3390/cells10020362.

Kidwell, M. G., Kidwell, J. F. and Sved, J. A. (1977) 'Hybrid dysgenesis in Drosophila melanogaster: a syndrome of aberrant traits including mutation, sterility and male recombination', *Genetics*, 86(4), pp. 813–833. doi: 10.1093/genetics/86.4.813.

Kim, J. *et al.* (2021) 'False gene and chromosome losses affected by assembly and sequence errors', *bioRxiv*, p. 2021.04.09.438906. doi: 10.1101/2021.04.09.438906.

Kinsella, C. M. *et al.* (2019) 'Programmed DNA elimination of germline development genes in songbirds', *Nature Communications*, 10(1), p. 5468. doi: 10.1038/s41467-019-13427-4.

Klein, S. J. and O'Neill, R. J. (2018) 'Transposable elements: genome innovation, chromosome diversity, and centromere conflict', *Chromosome Research*. Chromosome Research, 26(1–2), pp. 5–23. doi: 10.1007/s10577-017-9569-5.

Ko, B. J. *et al.* (2021) 'Widespread false gene gains caused by duplication errors in genome assemblies', *bioRxiv*, p. 2021.04.09.438957. doi: 10.1101/2021.04.09.438957.

Korlach, J. *et al.* (2010) 'Real-Time DNA Sequencing from Single Polymerase Molecules', *Methods in Enzymology*. American Association for the Advancement of Science, 472(5910), pp. 431–455. doi: 10.1016/S0076-6879(10)72001-2.

Kozarewa, I. *et al.* (2009) 'Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes', *Nature Methods*. Nature Publishing Group, 6(4), pp. 291–295. doi: 10.1038/nmeth.1311.

Kretschmer, R. *et al.* (2020) 'Chromosomal Analysis in Crotophaga ani (Aves, Cuculiformes) Reveals Extensive Genomic Reorganization and an Unusual Z-Autosome Robertsonian Translocation', *Cells*. doi: 10.3390/cells10010004.

Kretschmer, R. *et al.* (2021) 'Interspecies Chromosome Mapping in Caprimulgiformes, Piciformes, Suliformes, and Trogoniformes (Aves): Cytogenomic Insight into Microchromosome Organization and Karyotype Evolution in Birds', *Cells*. 10(4), p. 826 doi: 10.3390/cells10040826.

Kumar, S. *et al.* (2017) 'TimeTree: A Resource for Timelines, Timetrees, and Divergence Times', *Molecular biology and evolution*, 34(7), pp. 1812–1819. doi: 10.1093/molbev/msx116.

Kutch, I. C. and Fedorka, K. M. (2018) 'Y-chromosomes can constrain adaptive evolution via epistatic interactions with other chromosomes', *BMC Evolutionary Biology*, 18(1), p. 204. doi: 10.1186/s12862-018-1327-6.

de la Rosa, P. M. G. *et al.* (2021) 'A telomere-to-telomere assembly of Oscheius tipulae and the evolution of rhabditid nematode chromosomes', *G3: Genes, Genomes, Genetics*, 11(1). doi: 10.1093/G3JOURNAL/JKAA020.

Lahn, B. T. and Page, D. C. (1999) 'Four evolutionary strata on the human X chromosome', *Science*. American Association for the Advancement of Science, 286(5441), pp. 964–967. doi: 10.1126/science.286.5441.964.

Lambertucci, S. A. *et al.* (2012) 'Large-Scale Age-Dependent Skewed Sex Ratio in a Sexually Dimorphic Avian Scavenger', *PLoS ONE*. 2012/09/27. Public Library of Science, 7(9), pp. e46347–e46347. doi: 10.1371/journal.pone.0046347.

Lamichhaney, S. *et al.* (2015) 'Structural genomic changes underlie alternative reproductive strategies in the ruff (Philomachus pugnax)', *Nature Genetics*, 48(1), pp. 84–88. doi: 10.1038/ng.3430.

Lander, E. S. *et al.* (2001) 'Initial sequencing and analysis of the human genome', *Nature*, 409(6822), pp. 860–921. doi: 10.1038/35057062.

Lane, N. (2015) *The vital question*. Profile Books LTD.

Larracuente, A. M. (2014) 'The organization and evolution of the Responder satellite in species of the Drosophila melanogaster group: Dynamic evolution of a target of meiotic drive', *BMC Evolutionary Biology*, 14(1), p. 233. doi: 10.1186/s12862-014-0233-9.

Lemos, B., Araripe, L. O. and Hartl, D. L. (2008) ‘Polymorphic Y chromosomes harbor cryptic variation with manifold functional consequences’, *Science*, 319(5859), pp. 91–93. doi: 10.1126/science.1148861.

Lerat, E. *et al.* (2019) ‘On the importance to acknowledge transposable elements in epigenomic analyses’, *Genes*. Multidisciplinary Digital Publishing Institute, 10(4), p. 258. doi: 10.3390/genes10040258.

Levy-Sakin, M. *et al.* (2019) ‘Genome maps across 26 human populations reveal population-specific patterns of structural variation’, *Nature Communications*. Nature Publishing Group, 10(1), pp. 1–14. doi: 10.1038/s41467-019-08992-7.

Li, Y. C. *et al.* (2002) ‘Microsatellites: Genomic distribution, putative functions and mutational mechanisms: A review’, *Molecular Ecology*. Wiley Online Library, 11(12), pp. 2453–2465. doi: 10.1046/j.1365-294X.2002.01643.x.

Lieberman-Aiden, E. *et al.* (2009) ‘Comprehensive mapping of long-range interactions reveals folding principles of the human genome’, *Science*. American Association for the Advancement of Science, 326(5950), pp. 289–293. doi: 10.1126/science.1181369.

Liu, Y. *et al.* (2009) ‘Bos taurus genome assembly’, *BMC Genomics*. BioMed Central, 10(1), pp. 1–11. doi: 10.1186/1471-2164-10-180.

Logsdon, G. A. *et al.* (2021) ‘The structure, function and evolution of a complete human chromosome 8’, *Nature*, 593(7857), pp. 101–107. doi: 10.1038/s41586-021-03420-7.

Lovell, P. V. *et al.* (2014) ‘Conserved syntenic clusters of protein coding genes are missing in birds’, *Genome Biology*, 15(12), p. 565. doi: 10.1186/s13059-014-0565-1.

Lower, S. S. *et al.* (2018) ‘Satellite DNA evolution: old ideas, new approaches’, *Current Opinion in Genetics and Development*, 49, pp. 70–78. doi: 10.1016/j.gde.2018.03.003.

Luan, D. D. *et al.* (1993) ‘Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: A mechanism for non-LTR retrotransposition’, *Cell*, 72(4), pp. 595–605. doi: 10.1016/0092-8674(93)90078-5.

Malik, H. S. (2009) ‘The centromere-drive hypothesis: a simple basis for centromere complexity.’, in Ugarkovic, D. (ed.) *Progress in molecular and subcellular biology*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 33–52. doi: 10.1007/978-3-642-00182-6_2.

Manthey, J. D., Moyle, R. G. and Boissinot, S. (2018) ‘Multiple and Independent Phases of Transposable Element Amplification in the Genomes of Piciformes (Woodpeckers and Allies)’, *Genome Biology and Evolution*, 10(6), pp. 1445–1456. doi: 10.1093/gbe/evy105.

Marks, P. *et al.* (2019) ‘Resolving the full spectrum of human genome variation using Linked-Reads’, *Genome Research*. Cold Spring Harbor Lab, 29(4), pp. 635–645. doi: 10.1101/gr.234443.118.

Martin-Gallardo, A. *et al.* (1992) ‘Automated DNA sequencing and analysis of 106 kilobases from human chromosome 19q13.3’, *Nature Genetics*. Nature Publishing Group, 1(1), pp. 34–39. doi: 10.1038/ng0492-34.

McGurk, M. P. and Barbash, D. A. (2018) ‘Double insertion of transposable elements provides a substrate for the evolution of satellite DNA’, *Genome Research*, 28(5), pp. 714–725. doi: 10.1101/gr.231472.117.

Meyne, J. *et al.* (1990) ‘Distribution of non-telomeric sites of the (TTAGGG)n telomeric sequence in vertebrate chromosomes’, *Chromosoma*. Springer, 99(1), pp. 3–10. doi: 10.1007/BF01737283.

Miga, K. H. (2019) ‘Centromeric satellite DNAs: Hidden sequence variation in the human population’, *Genes*, 10(5), p. 352. doi: 10.3390/genes10050352.

Miga, K. H. (2020) ‘Centromere studies in the era of “telomere-to-telomere” genomics’, *Experimental Cell Research*, 394(2), p. 112127. doi: 10.1016/j.yexcr.2020.112127.

Miga, K. H. *et al.* (2020) ‘Telomere-to-telomere assembly of a complete human X chromosome’, *Nature*, 585(7823), pp. 79–84. doi: 10.1038/s41586-020-2547-7.

Mořkovský, L. *et al.* (2018) ‘Genomic islands of differentiation in two songbird species reveal candidate genes for hybrid female sterility’, *Molecular Ecology*. John Wiley & Sons, Ltd, 27(4), pp. 949–958. doi: 10.1111/mec.14479.

Moss, R., Piertney, S. B. and Palmer, S. C. F. (2003) ‘The use and abuse of microsatellite DNA markers in conservation biology’, *Wildlife Biology*. BioOne, 9(4), pp. 243–250. doi: 10.2981/wlb.2003.011.

Mugal, C. F., Arndt, P. F. and Ellegren, H. (2013) ‘Twisted signatures of GC-biased gene conversion embedded in an evolutionary stable karyotype’, *Molecular Biology and Evolution*, 30(7), pp. 1700–1712. doi: 10.1093/molbev/mst067.

Nam, K. and Ellegren, H. (2008) ‘The chicken (Gallus gallus) Z chromosome contains at least three nonlinear evolutionary strata’, *Genetics*. Oxford University Press, 180(2), pp. 1131–1136. doi: 10.1534/genetics.108.090324.

Neubauer, G., Nowicki, P. and Zagalska-Neubauer, M. (2014) ‘Haldane’s rule revisited: Do hybrid females have a shorter lifespan? Survival of hybrids in a recent contact zone between two large gull species’, *Journal of Evolutionary Biology*. John Wiley & Sons, Ltd, 27(6), pp. 1248–1255. doi: 10.1111/jeb.12404.

Van Nieuwerburgh, F. *et al.* (2012) ‘Illumina mate-paired DNA sequencing-library preparation using Cre-Lox recombination’, *Nucleic Acids Research*. Oxford University Press, 40(3), pp. e24–e24.

Nurk, S. *et al.* (2021) ‘The complete sequence of a human genome’, *bioRxiv*, p. 2021.05.26.445798. doi: 10.1101/2021.05.26.445798.

Orgel, L. E. and Crick, F. H. C. (1980) ‘Selfish DNA: The ultimate parasite’, *Nature*. Nature Publishing Group, 284(5757), pp. 604–607. doi: 10.1038/284604a0.

Oyola, S. O. *et al.* (2012) ‘Optimizing illumina next-generation sequencing library preparation for extremely at-biased genomes’, *BMC Genomics*. Springer, 13(1), pp. 1–12. doi: 10.1186/1471-2164-13-1.

Ozata, D. M. *et al.* (2019) ‘PIWI-interacting RNAs: small RNAs with big functions’, *Nature Reviews Genetics*. Springer US, 20(2), pp. 89–108. doi: 10.1038/s41576-018-0073-3.

Paajanen, P. *et al.* (2019) ‘A critical comparison of technologies for a plant genome sequencing project’, *GigaScience*. Oxford University Press, 8(3), p. giy163. doi: 10.1093/gigascience/giy163.

Palacios-Gimenez, O. M. *et al.* (2020) ‘Eight Million Years of Satellite DNA Evolution in Grasshoppers of the Genus Schistocerca Illuminate the Ins and Outs of the Library Hypothesis’, *Genome Biology and Evolution*, 12(3), pp. 88–102. doi: 10.1093/gbe/evaa018.

Pardue, M. Lou and DeBaryshe, P. G. (2003) ‘Retrotransposons Provide an Evolutionarily Robust Non-Telomerase Mechanism to Maintain Telomeres’, *Annual Review of Genetics*. Annual Reviews, 37(1), pp. 485–511. doi: 10.1146/annurev.genet.38.072902.093115.

Payne, A. *et al.* (2019) ‘Bulkvis: A graphical viewer for Oxford nanopore bulk FAST5 files’, *Bioinformatics*. Oxford University Press, 35(13), pp. 2193–2198. doi: 10.1093/bioinformatics/bty841.

Peichel, C. L. *et al.* (2017) ‘Improvement of the Threespine Stickleback Genome Using a Hi-C-Based Proximity-Guided Assembly’, *Journal of Heredity*. Oxford University Press US, 108(6), pp. 693–700. doi: 10.1093/jhered/esx058.

Peñalba, J. V. *et al.* (2020) ‘Genome of an iconic Australian bird: High-quality assembly and linkage map of the superb fairy-wren (Malurus cyaneus)’, *Molecular Ecology Resources*. Wiley Online Library, 20(2), pp. 560–578. doi: 10.1111/1755-0998.13124.

Petrov, D. A. *et al.* (1995) ‘Diverse transposable elements are mobilized in hybrid dysgenesis in Drosophila virilis’, *Proceedings of the National Academy of Sciences of the United States of America*, 92(17), pp. 8050–8054. doi: 10.1073/pnas.92.17.8050.

Piégu, B. *et al.* (2018) ‘But where did the centromeres go in the chicken genome models?’, *Chromosome Research*, 26(4), pp. 297–306. doi: 10.1007/s10577-018-9585-0.

Pinheiro, M. L. S. *et al.* (2021) ‘Chromosomal painting of the sandpiper (Actitis macularius) detects several fissions for the Scolopacidae family (Charadriiformes)’, *BMC ecology and evolution*, 21(1), p. 8. doi: 10.1186/s12862-020-01737-x.

Pipoly, I. *et al.* (2015) ‘The genetic sex-determination system predicts adult sex ratios in tetrapods’, *Nature*, 527(7576), pp. 91–94. doi: 10.1038/nature15380.

Platt, R. N., Blanco-Berdugo, L. and Ray, D. A. (2016) ‘Accurate transposable element annotation is vital when analyzing new genome assemblies’, *Genome Biology and Evolution*, 8(2), pp. 403–410. doi: 10.1093/gbe/evw009.

Plohl, M., Meštrović, N. and Mravinac, B. (2012) ‘Satellite DNA evolution’, in *Genome Dynamics*, pp. 126–152. doi: 10.1159/000337122.

Ponnikas, S. *et al.* (2018) ‘Why Do Sex Chromosomes Stop Recombining?’, *Trends in Genetics*, 34(7), pp. 492–503. doi: 10.1016/j.tig.2018.04.001.

Prost, S. *et al.* (2019) ‘Comparative analyses identify genomic features potentially involved in the evolution of birds-of-paradise’, *GigaScience*, 8(5), p. giz003. doi: 10.1093/gigascience/giz003.

Quail, M. A. *et al.* (2008) ‘A large genome center’s improvements to the Illumina sequencing system’, *Nature Methods*. Nature Publishing Group, 5(12), pp. 1005–1010. doi: 10.1038/nmeth.1270.

Rhie, A. *et al.* (2021) ‘Towards complete and error-free genome assemblies of all vertebrate species’, *Nature*, 592(7856), pp. 737–746. doi: 10.1038/s41586-021-03451-0.

Rice, W. R. (1984) ‘Sex Chromosomes and the Evolution of Sexual Dimorphism’, *Evolution*. [Society for the Study of Evolution, Wiley], 38(4), p. 735. doi: 10.2307/2408385.

Rice, W. R. (1987) ‘The Accumulation of Sexually Antagonistic Genes as a Selective Agent Promoting the Evolution of Reduced Recombination between Primitive Sex Chromosomes’, *Evolution*. [Society for the Study of Evolution, Wiley], 41(4), p. 911. doi: 10.2307/2408899.

Rocchi, M. *et al.* (2012) ‘Centromere repositioning in mammals’, *Heredity*, 108(1), pp. 59–67. doi: 10.1038/hdy.2011.101.

Ruiz-Ruano, F. J. *et al.* (2016) ‘High-throughput analysis of the satellitome illuminates satellite DNA evolution’, *Scientific Reports*, 6(1), p. 28333. doi: 10.1038/srep28333.

Salser, W. *et al.* (1976) ‘Investigation of the organization of mammalian chromosomes at the DNA sequence level’, *Federation Proceedings*, 35(1), pp. 23–35. Available at: http://europepmc.org/abstract/MED/1107072.

Sanger, F., Nicklen, S. and Coulson, A. . (1977) ‘DNA sequencing with chain-terminating’, *Proc Natl Acad Sci USA*. National Acad Sciences, 74(12), pp. 5463–5467.

Schaack, S., Gilbert, C. and Feschotte, C. (2010) ‘Promiscuous DNA: Horizontal transfer of transposable elements and why it matters for eukaryotic evolution’, *Trends in Ecology and Evolution*, 25(9), pp. 537–546. doi: 10.1016/j.tree.2010.06.001.

Schrader, L. and Schmitz, J. (2019) ‘The impact of transposable elements in adaptive evolution’, *Molecular Ecology*, 28(6), pp. 1537–1549. doi: 10.1111/mec.14794.

Scott, A. F. *et al.* (1987) ‘Origin of the human L1 elements: Proposed progenitor genes deduced from a consensus DNA sequence’, *Genomics*, 1(2), pp. 113–125. doi: 10.1016/0888-7543(87)90003-6.

Sedlazeck, F. J. *et al.* (2018) ‘Piercing the dark matter: Bioinformatics of long-range sequencing and mapping’, *Nature Reviews Genetics*, 19(6), pp. 329–346. doi: 10.1038/s41576-018-0003-4.

Senft, A. D. and Macfarlan, T. S. (2021) ‘Transposable elements shape the evolution of mammalian development.’, *Nature reviews. Genetics*. doi: 10.1038/s41576-021-00385-1.

Shang, W. H. *et al.* (2010) ‘Chickens possess centromeres with both extended tandem repeats and short non-tandem-repetitive sequences’, *Genome Research*, 20(9), pp. 1219–1228. doi: 10.1101/gr.106245.110.

Shanta, O. *et al.* (2020) ‘The effects of common structural variants on 3D chromatin structure’, *BMC Genomics*. BMC Genomics, 21(1), pp. 1–10. doi: 10.1186/s12864-020-6516-1.

Sharpe, R. B. (1891–1898) *Monograph of the Paradiseidae, or birds of paradise and Ptilonorhynchidae, or bower-birds.*, *Monograph of the Paradiseidae, or birds of paradise and Ptilonorhynchidae, or bower-birds.* 2 volumes. London: H. Sotheran & Co. doi: 10.5962/bhl.title.109350.

Shatskikh, A. S. *et al.* (2020) ‘Functional Significance of Satellite DNAs: Insights From Drosophila’, *Frontiers in Cell and Developmental Biology*, p. 312. doi: 10.3389/fcell.2020.00312.

Shendure, J. *et al.* (2017) ‘DNA sequencing at 40: Past, present and future’, *Nature*. Nature Publishing Group, 550(7676), pp. 345–353. doi: 10.1038/nature24286.

Sigeman, H. *et al.* (2020) ‘Genomics of an avian neo-sex chromosome reveals the evolutionary dynamics of recombination suppression and sex-linked genes’, *bioRxiv*, p. 2020.09.25.314088. doi: 10.1101/2020.09.25.314088.

Singer, M. F. *et al.* (1993) ‘LINE-1: a human transposable element’, *Gene*, 135(1–2), pp. 183–188. doi: 10.1016/0378-1119(93)90064-A.

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. doi: 10.1038/hdy.2011.99.

Slotkin, R. K. (2018) ‘The case for not masking away repetitive DNA’, *Mobile DNA*. BioMed Central, 9(1), pp. 1–4. doi: 10.1186/s13100-018-0120-9.

Smeds, L. *et al.* (2015) ‘Evolutionary analysis of the female-specific avian W chromosome’, *Nature Communications*, 6(1), p. 7330. doi: 10.1038/ncomms8330.

Smith, C. A. *et al.* (2009) ‘The avian Z-linked gene DMRT1 is required for male sex determination in the chicken’, *Nature*, 461(7261), pp. 267–271. doi: 10.1038/nature08298.

Sotero-Caio, C. G. *et al.* (2017) ‘Evolution and diversity of transposable elements in vertebrate genomes’, *Genome Biology and Evolution*, 9(1), pp. 161–177. doi: 10.1093/gbe/evw264.

Spealman, P., Burrell, J. and Gresham, D. (2019) ‘Nanopore sequencing undergoes catastrophic sequence failure at inverted duplicated DNA sequences’, *bioRxiv*, p. 852665. doi: 10.1101/852665.

Spielmann, M., Lupiáñez, D. G. and Mundlos, S. (2018) 'Structural variation in the 3D genome', *Nature Reviews Genetics*, 19(7), pp. 453–467. doi: 10.1038/s41576-018-0007-0.

Stapley, J. *et al.* (2010) 'Pronounced inter- and intrachromosomal variation in linkage disequilibrium across the zebra finch genome', *Genome Research*, 20(4), pp. 496–502. doi: 10.1101/gr.102095.109.

Stein, A., Takasuka, T. E. and Collings, C. K. (2009) 'Are nucleosome positions in vivo primarily determined by histone-DNA sequence preferences?', *Nucleic Acids Research*. Oxford University Press, 38(3), pp. 709–719. doi: 10.1093/nar/gkp1043.

Sturtevant, A. H. (1921) 'A Case of Rearrangement of Genes in Drosophila', *Proceedings of the National Academy of Sciences*. National Academy of Sciences, 7(8), pp. 235–237. doi: 10.1073/pnas.7.8.235.

Su, X. Z. *et al.* (1996) 'Reduced extension temperatures required for PCR amplification of extremely A+T-rich DNA', *Nucleic Acids Research*. Oxford University Press, 24(8), pp. 1574–1575. doi: 10.1093/nar/24.8.1574.

Suh, A. *et al.* (2011) 'Retroposon insertions and the chronology of avian sex chromosome evolution', *Molecular Biology and Evolution*. Oxford University Press, 28(11), pp. 2993–2997. doi: 10.1093/molbev/msr147.

Suh, A. (2015) 'The Specific Requirements for CR1 Retrotransposition Explain the Scarcity of Retrogenes in Birds', *Journal of Molecular Evolution*. Springer, 81(1–2), pp. 18–20. doi: 10.1007/s00239-015-9692-x.

Suh, A. *et al.* (2016) 'Ancient horizontal transfers of retrotransposons between birds and ancestors of human pathogenic nematodes', *Nature Communications*, 7(1), p. 11396. doi: 10.1038/ncomms11396.

Sundaram, V. *et al.* (2014) 'Widespread contribution of transposable elements to the innovation of gene regulatory networks', *Genome Research*, 24(12), pp. 1963–1976. doi: 10.1101/gr.168872.113.

Thomma, B. P. H. J. *et al.* (2016) 'Mind the gap; seven reasons to close fragmented genome assemblies', *Fungal Genetics and Biology*, 90, pp. 24–30. doi: 10.1016/j.fgb.2015.08.010.

Thompson, P. J., Macfarlan, T. S. and Lorincz, M. C. (2016) 'Long Terminal Repeats: From Parasitic Elements to Building Blocks of the Transcriptional Regulatory Repertoire', *Molecular Cell*, 62(5), pp. 766–776. doi: 10.1016/j.molcel.2016.03.029.

Tigano, A. (2020) 'A population genomics approach to uncover the CNVs, and their evolutionary significance, hidden in reduced-representation sequencing data sets', *Molecular ecology*, 29(24), pp. 4749–4753. doi: 10.1111/mec.15665.

Tomaszkiewicz, M., Medvedev, P. and Makova, K. D. (2017) 'Y and W Chromosome Assemblies: Approaches and Discoveries', *Trends in Genetics*, 33(4), pp. 266–282. doi: 10.1016/j.tig.2017.01.008.

Uno, Y. *et al.* (2019) 'Molecular cytogenetic characterization of repetitive sequences comprising centromeric heterochromatin in three Anseriformes species', *PLoS ONE*. Edited by R. Stanyon, 14(3), p. e0214028. doi: 10.1371/journal.pone.0214028.

Urban, J. M. *et al.* (2020) 'Single-molecule sequencing of long DNA molecules allows high contiguity de novo genome assembly for the fungus fly, Sciara coprophila', *bioRxiv*, p. 2020.02.24.963009. doi: 10.1101/2020.02.24.963009.

Vergnaud, G. and Denoeud, F. (2000) 'Minisatellites: Mutability and genome architecture', *Genome Research*. Cold Spring Harbor Lab, 10(7), pp. 899–907. doi: 10.1101/gr.10.7.899.

Vontzou, N. (2021) *Comparative genomics of satellite DNA and putative centromere positions in birds*. Masters thesis. Stockholm University.

Wajid, B. and Serpedin, E. (2012) 'Review of General Algorithmic Features for Genome Assemblers for Next Generation Sequencers', *Genomics, Proteomics and Bioinformatics*. Elsevier, 10(2), pp. 58–73. doi: 10.1016/j.gpb.2012.05.006.

Wallis, M. C., Waters, P. D. and Graves, J. A. M. (2008) 'Sex determination in mammals - Before and after the evolution of SRY', *Cellular and Molecular Life Sciences*. Springer, 65(20), pp. 3182–3195. doi: 10.1007/s00018-008-8109-z.

Warren, W. C. *et al.* (2010) 'The genome of a songbird', *Nature*, 464(7289), pp. 757–762. doi: 10.1038/nature08819.

Warren, W. C. *et al.* (2017) 'A new chicken genome assembly provides insight into avian genome structure', *G3: Genes, Genomes, Genetics*, 7(1), pp. 109–117. doi: 10.1534/g3.116.035923.

Weisenfeld, N. I. *et al.* (2017) 'Direct determination of diploid genome sequences', *Genome Research*, 27(5), pp. 757–767. doi: 10.1101/gr.214874.116.

Weissensteiner, M. H. *et al.* (2017) 'Combination of short-read, long-read, and optical mapping assemblies reveals large-scale tandem repeat arrays with population genetic implications', *Genome Research*, 27(5), pp. 697–708. doi: 10.1101/gr.215095.116.

Weissensteiner, M. H. and Suh, A. (2019) 'Repetitive DNA: The Dark Matter of Avian Genomics', in Kraus, R. H. S. (ed.) *Avian Genomics in Ecology and Evolution*. Cham: Springer International Publishing, pp. 93–150. doi: 10.1007/978-3-030-16477-5_5.

Wellenreuther, M. *et al.* (2019) 'Going beyond SNPs: The role of structural genomic variants in adaptive evolution and species diversification', *Molecular Ecology*. John Wiley & Sons, Ltd, 28(6), pp. 1203–1209. doi: 10.1111/mec.15066.

Wenger, A. M. *et al.* (2019) 'Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome', *Nature Biotechnology*, 37(10), pp. 1155–1162. doi: 10.1038/s41587-019-0217-9.

Westerberg, I. (2020) *Deciphering the formation of evolutionary new centromeres in a microchromosome of birds*. Masters thesis. Uppsala University.

Wetzel, J., Kingsford, C. and Pop, M. (2011) 'Assessing the benefits of using mate-pairs to resolve repeats in de novo short-read prokaryotic assemblies', *BMC Bioinformatics*. Springer, 12(1), pp. 1–14. doi: 10.1186/1471-2105-12-95.

Wicker, T. (2004) 'The repetitive landscape of the chicken genome', *Genome Research*, 15(1), pp. 126–136. doi: 10.1101/gr.2438004.

Wicker, T. *et al.* (2007) 'A unified classification system for eukaryotic transposable elements', *Nature Reviews Genetics*, 8(12), pp. 973–982. doi: 10.1038/nrg2165.

Willard, H. F. and Waye, J. S. (1987) 'Hierarchical order in chromosome-specific human alpha satellite DNA', *Trends in Genetics*, 3(C), pp. 192–198. doi: 10.1016/0168-9525(87)90232-0.

Wright, A. E. *et al.* (2016) 'How to make a sex chromosome', *Nature Communications*, 7(1), p. 12087. doi: 10.1038/ncomms12087.

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*. Royal Society, 281(1779), p. 20132780. doi: 10.1098/rspb.2013.2780.

Xirocostas, Z. A., Everingham, S. E. and Moles, A. T. (2020) 'The sex with the reduced sex chromosome dies earlier: A comparison across the tree of life', *Biology Letters*. Royal Society, 16(3), p. 20190867. doi: 10.1098/rsbl.2019.0867.

Xu, L., Auer, G., *et al.* (2019) ‘Dynamic evolutionary history and gene content of sex chromosomes across diverse songbirds’, *Nature Ecology and Evolution*, 3(5), pp. 834–844. doi: 10.1038/s41559-019-0850-1.

Xu, L., Wa Sin, S. Y., *et al.* (2019) ‘Evolutionary Dynamics of Sex Chromosomes of Paleognathous Birds’, *Genome Biology and Evolution*, 11(8), pp. 2376–2390. doi: 10.1093/gbe/evz154.

Yazdi, H. P. and Ellegren, H. (2014) ‘Old but Not (So) degenerated-slow evolution of largely homomorphic sex chromosomes in ratites’, *Molecular Biology and Evolution*. Oxford University Press, 31(6), pp. 1444–1453. doi: 10.1093/molbev/msu101.

Yazdi, H. P. and Ellegren, H. (2018) ‘A genetic map of ostrich Z chromosome and the role of inversions in avian sex chromosome evolution’, *Genome Biology and Evolution*. Oxford University Press, 10(8), pp. 2049–2060. doi: 10.1093/gbe/evy163.

Yazdi, H. P., Silva, W. T. A. F. and Suh, A. (2020) ‘Why do some sex chromosomes degenerate more slowly than others? The odd case of ratite sex chromosomes’, *Genes*, pp. 1–13. doi: 10.3390/genes11101153.

Yin, Z. T. *et al.* (2019) ‘Revisiting avian “missing” genes from de novo assembled transcripts’, *BMC Genomics*, 20(1), p. 4. doi: 10.1186/s12864-018-5407-1.

Zeng, J. and Yi, S. V. (2014) ‘Specific modifications of histone tails, but not DNA methylation, mirror the temporal variation of mammalian recombination hotspots’, *Genome Biology and Evolution*. Oxford University Press, 6(10), pp. 2918–2929. doi: 10.1093/gbe/evu230.

Zhang, G. *et al.* (2014) ‘Comparative genomics reveals insights into avian genome evolution and adaptation’, *Science*, 346(6215), pp. 1311–1320. doi: 10.1126/science.1251385.

Zhang, H. H. *et al.* (2020) ‘Horizontal transfer and evolution of transposable elements in vertebrates’, *Nature Communications*, 11(1), p. 1362. doi: 10.1038/s41467-020-15149-4.

Zhang, L. *et al.* (2021) ‘How Important Are Structural Variants for Speciation?’, *Genes*, p. 1084. doi: 10.3390/genes12071084.

Zhao, X. *et al.* (2021) ‘Expectations and blind spots for structural variation detection from long-read assemblies and short-read genome sequencing technologies’, *American Journal of Human Genetics*, 108(5), pp. 919–928. doi: 10.1016/j.ajhg.2021.03.014.

Zheng, G. X. Y. *et al.* (2016) ‘Haplotyping germline and cancer genomes with high-throughput linked-read sequencing’, *Nature Biotechnology*. Nature Publishing Group, 34(3), pp. 303–311. doi: 10.1038/nbt.3432.

Zhou, Q. *et al.* (2014) ‘Complex evolutionary trajectories of sex chromosomes across bird taxa’, *Science*, 346(6215), p. 1246338. doi: 10.1126/science.1246338.

Zlotina, A. *et al.* (2012) ‘Centromere positions in chicken and Japanese quail chromosomes: De novo centromere formation versus pericentric inversions’, *Chromosome Research*, 20(8), pp. 1017–1032. doi: 10.1007/s10577-012-9319-7.

Genomes are the gateway to an enchanted land. The reams of code, 3 billion letters in our own case, read like an experimental novel, an occasionally coherent story in short chapters broken up by blocks of repetitive text, verses, blank pages, streams of consciousness: and peculiar punctuation.

Lane, 2015

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 2061

Editor: The Dean of the Faculty of Science and Technology

Distribution: publications.uu.se
urn:nbn:se:uu:diva-450756

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2021