UPPSALA
UNIVERSITET

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 1800

Rates and patterns of molecular evolution in avian genomes

PAULINA BOLÍVAR

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2019

ISSN 1651-6214
ISBN 978-91-513-0637-7
urn:nbn:se:uu:diva-381373

Dissertation presented at Uppsala University to be publicly examined in Evolutionary Biology Center, Norbyvägen 14, Uppsala, Tuesday, 11 June 2019 at 13:00 for the degree of Doctor of Philosophy. The examination will be conducted in English. Faculty examiner: Professor Mikkel Heide Schierup (Department of Bioscience, Aarhus University, Aarhus, Denmark).

Abstract

Bolívar, P. 2019. Rates and patterns of molecular evolution in avian genomes. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 1800. 51 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-513-0637-7.

Evolution is the change in inherited characteristics of a population through subsequent generations. The interplay of several evolutionary mechanisms determines the rate at which this change occurs. In short, genetic variation is generated though mutation, and the fate of these mutations in a population is determined mainly by the combined effect of genetic drift, natural selection and recombination. Elucidating the relative impact of these mechanisms is complex; making it a long-standing question in evolutionary biology. In this thesis, I focus on disentangling the relative roles of these evolutionary mechanisms and genetic factors in determining rates and patterns of evolution at the molecular level, by studying variation in the DNA sequence of multiple avian species, and in particular the collared flycatcher (*Ficedula albicollis*). Specifically, I aim to further our understanding regarding the impact of recombination rate on genome evolution, through its interaction with the efficacy of selection and through the process of GC-biased gene conversion (gBGC), which has been poorly characterized in birds. I demonstrate that gBGC has a pervasive effect on the genome of the collared flycatcher and other avian species, as it increases the substitution rate and affects interpretations of the impact of natural selection and adaptation. Interestingly, its effect is even stronger in neutrally evolving sites compared to sites evolving under selection. After accounting for gBGC, I disentangle the true impact of natural selection versus non-adaptive processes in determining rates of molecular evolution in the collared flycatcher genome, shedding light on the process of adaptation. Finally, I demonstrate the significant role of recombination through its impact on linked selection, along with mutation rate differences, in determining relative levels of genetic diversity and their relationship to the fast-Z effect across the avian phylogeny. This thesis urges future studies to account for the effect of recombination before interpreting patterns of selection in sequence evolution.

Keywords: Molecular evolution, recombination, GC-biased gene conversion, Hill-Robertson interference, effective population size, dN/dS, distribution of fitness effects, avian genomes, collared flycatcher, sex chromosomes

Paulina Bolívar, Department of Ecology and Genetics, Evolutionary Biology, Norbyvägen 18D, Uppsala University, SE-75236 Uppsala, Sweden.

ISSN 1651-6214
ISBN 978-91-513-0637-7
urn:nbn:se:uu:diva-381373 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-381373)

To my parents

List of papers

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

I Bolívar, P., Mugal, C. F., Nater, A. & Ellegren, H. (2016). Recombination rate variation modulates gene sequence evolution mainly via GC-biased gene conversion, not Hill-Robertson interference, in an avian system. *Molecular Biology and Evolution*, 33:216–227.

II Bolívar, P., Mugal, C. F., Rossi, M., Nater, A., Wang, M., Dutoit, L. & Ellegren, H. (2018). Biased inference of selection due to GC-biased gene conversion and the rate of protein evolution in flycatchers when accounting for it. *Molecular Biology and Evolution*, 35:2475–2486.

III Bolívar, P., Guéguen, L., Duret, L., Ellegren, H. & Mugal C. F. (2018). GC-biased gene conversion conceals the prediction of the nearly neutral theory in avian genomes. *Genome Biology*, 20:5.

IV Yazdi, H. P.*, Bolívar, P.*, Mugal, C. F., & Ellegren, H. (2018). Variation in the Z chromosome to autosomes ratio of genetic diversity across birds and its relationship to the fast-Z effect. *Manuscript.*

* These authors contributed equally to this work.

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 the thesis.

Dutoit, L., Mugal, C. F., Bolívar, P., Wang, M., Nadachowska-Brzyska, K., Smeds, L., Yazdi, H. P., Gustafsson, L. & Ellegren, H. (2018). Sex-biased gene expression, sexual antagonism and levels of genetic diversity in the collared flycatcher (*Ficedula albicollis*) genome. *Molecular Ecology*, 27:3572-3581.

Uebbing, S., Künstner, A., Mäkinen, H., Backström, N., Bolívar, P., Burri, R., Dutoit, L., Mugal, C. F., Nater, A., Aken, B., Flicek, P., Martin, F. J., Searle, S. M. J. & Ellegren, H. (2016). Divergence in gene expression within and between two closely related flycatcher species. *Molecular Ecology,* 25:2015-2028.

Smeds, L., Warmuth, V., Bolívar, P., Uebbing, S., Burri, R., Suh, A., Nater, A., Bureš, S., Garamszegi, L. Z., Hogner, S., Moreno, J., Qvarnström, A., Ružić, M., Sæther, S. A., Sætre, G. P., Török, J. & Ellegren H. (2015). Evolutionary analysis of the female-specific avian W chromosome. *Nature Communications,* 6:7330.

Smeds, L., Kawakami, T., Burri, R., Bolívar, P., Husby, A., Qvarnström, A., Uebbing, S. and Ellegren, H. (2014). Genomic identification and characterization of the pseudoautosomal region in highly differentiated avian sex chromosomes. *Nature Communications*, 5:5448.

Contents

Abbreviations

CO	Crossover
DNA	Deoxyribonucleic acid
DSB	Double-strand break
DSBR	Double-strand break repair
dHJ	Double Holliday-junction
d_N	Nonsynonymous substitution rate
d_N/d_S	Nonsynonymous to synonymous substitution rate ratio
d_S	Synonymous substitution rate
DFE	Distribution of fitness effects
Gbp	Giga base pairs
gBGC	GC-biased gene conversion
GC-conservative	S-to-S and W-to-W
GC*	Equilibrium GC content
HRI	Hill-Robertson interference
LHTs	Life-history traits
N_e	Effective population size
MK	McDonald-Kreitman
NCO	Non-crossover
PPI	Protein-protein interactions
RNA	Ribonucleic acid
s	Selection coefficient
S	Strong nucleotide (C and G)
SFS	Site frequency spectrum
SDSA	Synthesis-dependent strand annealing
SNP	Single nucleotide polymorphism
S-to-S	Strong-to-strong (C to G or G to C)
S-to-W	Strong-to-weak (C or G to A or T)
W	Weak nucleotide (A and T)
W-to-S	Weak-to-strong (A or T to C or G)
W-to-W	Weak-to-weak (A to T or T to A)
α	Proportion of adaptive nonsynonymous substitutions
ΔGC	The difference between current GC content and GC*
k	Neutral substitution rate
μ	Mutation rate
ω_{na}	Rate of nonadaptive substitutions
ω_a	Rate of adaptive substitutions

Introduction

A thorough understanding of the evolutionary process is crucial for the comprehension of life. Studying evolution can helps us decipher the history of living species, and allows us to describe and predict patterns that we observe in nature. It can also help us address more applied subjects, such as the characterization and conservation of biodiversity or the control and prevention of infectious diseases. Thus, evolution is a unifying principle of every discipline of biology, from biochemistry to phylogenetics.

Evolution is the change in inherited characteristics of a population through subsequent generations. The interplay of several evolutionary forces and genetic factors determine the rate at which this change occurs. In recent years, the advancement of genomic technologies has enabled researchers to study evolution in greater detail at the molecular level, thereby addressing evolutionary questions and theories that were previously put forward. The field of molecular evolution studies variation in the sequence composition of the genetic material (i.e. DNA and RNA) through time. To do so, molecular evolutionary studies compare and contrast the molecular data of several species, in a comparative genomics context, or alternatively, they study the molecular data within a single population, in a population genetics framework. The aims of evolutionary studies at the molecular level are twofold: to try to reconstruct the evolutionary histories and relationships of species or, like the work conducted in this thesis, to focus on understanding the mechanisms and factors that determine evolutionary change.

Mutation is the only mechanism that generates new genetic diversity. Other mechanisms do not create genetic variation, but determine the fate of mutations by affecting their probability of segregation in the population and eventual fixation, that is, their eventual presence in all individuals of the population. For example, finite populations are under the influence of stochastic and demographic processes, which will affect the probability of fixation of segregating mutations through the process of genetic drift. Natural selection will influence the fate of mutations in a deterministic way by increasing or decreasing the probability of fixation of variants that impact the fitness of individuals, that is, their ability to survive and reproduce. Further, by re-shuffling genetic variation, recombination will play a main role in evolutionary change as it can indirectly increase the efficacy of selection. Recombination can also directly influence the probability of fixation through the process of GC-biased gene conversion (gBGC). This process leads to the preferential transmission of guanine (G) and cytosine (C) nucleotides to the next generation, regardless of whether they are advantageous or not.

Elucidating the relative impact of the above mentioned evolutionary mechanisms that determine the dynamics of molecular evolution is a classic question in the field of evolutionary biology itself. All these evolutionary mechanisms interact with one another but their relative strength and impact on the evolutionary process is complex and varies drastically between species and along the genome. In this thesis, I focus on disentangling the relative roles of these evolutionary mechanisms in determining rates and patterns of evolution in bird genomes. I aim to further our understanding of molecular evolution, in particular, regarding the impact of recombination rate on genome evolution, through its interaction with the efficacy of selection and through the process of GC-biased gene conversion (gBGC), which has been poorly characterized in birds. I demonstrate that the impact of gBGC on rates of evolution is pervasive in the genome of the collared flycatcher (*Ficedula albicollis*) and other avian species. Interestingly, its effect is even stronger in neutrally evolving sites compared to sites evolving under selection (Papers I, II and III). After accounting for gBGC, I disentangle the relative importance of natural selection versus nonadaptive processes in determining rates of molecular evolution in the collared flycatcher genome, as this may shed light on the process of adaptation (Paper II). Finally, I confirm the pervasive role of recombination, along with mutation rate differences, in determining relative levels of genetic diversity and divergence among sex chromosomes and autosomes (Paper IV).

Determinants of the rate and patterns of molecular evolution

Variation in rates and patterns of molecular evolution can be observed not only between lineages, but also between different regions of the genome of one organism. Elucidating which factors are at play and how these factors determine regional genomic diversity, as well as intra- and inter-specific variation, is not an easy task. In the following chapter I aim to briefly explain the basic mechanisms that determine rates and patterns of molecular evolution.

Mutation

Mutations are changes in the DNA (or RNA) sequence that usually occur when a cell replicates its genetic information during the process of cell division. Point mutations are single base changes, where one of the four DNA nucleotides; adenine (A), cytosine (C), guanine (G) or thymine (T) changes to another. These changes in the DNA can be caused by several factors. The most common causes are errors in the replication machinery of the cell, which escape proofreading and enzyme repair mechanisms. However, mutations can also happen as a result of exposure to chemical or environmental mutagens, ultraviolet light, or oxidative radicals of the cell. In this thesis, I focus on single nucleotide mutations. Other types of mutations include insertions and deletions of one or more genetic bases. Furthermore, fissions, fusions or translocations of long stretches of a chromosome can occur at different scales. These mutations are most frequently a consequence of errors in the process of meiosis, when chromosomes recombine to form new gametes.

Mutations can occur in any cell of the organism. If mutations occur in somatic cells, they may affect the fitness of that particular individual. However, somatic mutations cannot be inherited and are therefore inconsequential for evolution. Only mutations that occur in the germ line can be inherited by the next generation and are a prerequisite for evolutionary change.

The rate at which new mutations appear in a population, the mutation rate, can vary substantially between species but also between different nucleotides and along genomic regions. For example, humans have an average mutation rate of 1.1×10^{-8} per site per generation (Roach *et al.*, 2010). In comparison, a direct estimates of the germ line mutation rate in the collared flycatcher lineage has been determined to be 4.6 × 10^{-9} per site and generation (Smeds *et al.*, 2016b). Variation in the mutation rate along the genome may be related to variation in several genomic characteristics. For example, the rate of transitions (mutations between two purines e.i. A or G or between two pyrimidines e.i. C or T) is higher than the rate of transvertions (mutations between one purine and one pyrimidine). Furthermore, a C followed by a G (CpG sites) will frequently be methylated in the DNA. These sites suffer from spontaneous deamination due to hydrolytic damage, which leads to C to T mutations. The C to T transition rate at methylated CpG sites is 10 fold higher than at unmethylated sites (Cooper & Gerber-Huber, 1985; Sved & Bird, 1990). Therefore, mutation rates can vary depending on the GC content and the proportion of CpG sites that are present in the genomic region or species of interest. Another explanation is related to variation in the rate of recombination; as it has been suggested that recombination may be mutagenic *per se* (Arbeithuber *et al.*, 2015; Halldorsson *et al.*, 2019; Hellmann *et al.*, 2003; Lercher & Hurst, 2002; Pratto *et al.*, 2014). Other factors that may contribute to such variation in the mutation rate are, for example, chromatin structure (Prendergast *et al.*, 2007) and replication timing (Lang & Murray, 2011; Sved & Bird, 1990) as these factors determine the accessibility and capability of the DNA repair machinery to identify replication errors.

Mutations are the main source of genetic diversity and can have different fitness effects. By altering the phenotype of individuals, selected mutations may have advantageous or disadvantageous effects and their consequences may vary from a mild to a drastic change in fitness. However, not all mutations have an effect on fitness. Mutations which do not result in a change in fitness are referred to as “neutrally evolving” or “neutral mutations”.

Genetic drift and effective population size

The fate of new mutations in a population is strongly determined by genetic drift, the process by which allele frequencies in a finite population change over time as a result of chance. The size of the population determines the impact of the stochastic process on the allele frequencies. Small populations are more strongly affected than large populations as a result of suffering from a larger effect of random sampling of mutations (Wright, 1931). Specifically, some individuals will not contribute their genetic material to the next generation, while others will contribute multiple times just by chance, leading to the loss of some genetic variants and the eventual random fixation of others. The change in allele frequencies from one generation to the next will be more drastic in small populations. In large populations, the effect of the stochastic process becomes less important and deterministic forces such as natural selection can act more efficiently.

The size of the population referred to here is not the census size but the effective population size (N_e). This is the size of an idealized population that would have the same allele frequency changes every generation as the real population of interest (Wright, 1931). This idealized population has specific characteristics such as random mating and constant size. Nonetheless, these features are rarely seen in natural populations. Non-random mating often occurs as a result of differential sex ratios or due to a large variance in reproductive success between individuals as a result of sexual selection. Also, fluctuations in population size are common. A population contraction, expansion or gene flow between populations will result in changes to N_e. This will in turn be reflected in an alteration of the allele frequencies.

Selection

When a mutation has a substantial impact on an organism's fitness, its probability of segregating and eventually getting fixed in the population will be determined by selection. The effect of selection is usually measured by the selection coefficient *s*, which is a measure of the relative fitness difference between an individual homozygous for the selected allele and an individual homozygous for the reference allele. When the new mutation provides an advantage to fitness, it may quickly increase in frequency and get fixed in the population through the action of positive selection, leading to adaptation. However, when the mutation provides a disadvantage to the individual's fitness, it's probability of segregating and becoming fixed will be reduced by the action of negative selection (also referred to as purifying selection). Another type of selection, balancing selection, can act to maintain both alleles in the population.

Nonsynonymous mutations and mutations in regulatory sequences are usually assumed to be targets of both negative and positive selection; since mutations in nonsynonymous sites change an amino acid of the protein that they code for, and may then change its biochemical properties. Similarly, mutations in regulatory sequences can modify patterns of gene expression and drastically affect the phenotype of the individual. These may potentially disrupt biochemical interactions and pathways, but may also, in some cases, be beneficial and lead to adaptive changes in the population. Synonymous mutations and mutations that occur in nonfunctional sites of the genome are instead usually assumed to be neutral, as they do not cause any change in protein sequence or patterns of gene expression. However, some evidence suggests that synonymous mutations may be under selection in the form of codon usage bias. Selection on codon usage has been suggested to be a result of transcriptional and translational selection. By affecting chromatin structures or translational efficacy, it provides an opportunity for natural selection to act upon (Kotlar & Lavner, 2006; Yannai *et al.*, 2018; Zhou *et al.*, 2016).

Recombination

Meiotic recombination is the process by which homologous chromosomes exchange DNA during meiosis (Figure 1). The process occurs after the formation of a double-strand break (DSB) when in order to repair the DNA damage, homologous sequences are identified and used as templates for repair. For this to happen, there is degradation (formally referred to as resection) of the broken strand to generate single-stranded DNA tails, one of which invades and pairs with the intact homologous sequence, forming a heteroduplex DNA structure. Then, the sequence information lost from the breakage is copied from the intact to the broken strand. Afterwards, the repair can follow different pathways, which are typically, double-strand break repair (DSBR) and synthesis-dependent strand annealing (SDSA). During DSBR, the second DSB end can be captured to form a structure called double Holliday-junction (dHJ), where the two molecules are joined together. The manner in which this junction is "dissolved" or "resolved" determines the outcome of the recombination event; crossover (CO) or non-crossover (NCO) (Sung & Klein, 2006; Szostak *et al.*, 1983; Youds & Boulton, 2011). The main difference is that the amount of DNA exchanged by the homologous chromosomes will be much larger in CO events. At least one CO is required for the proper segregation of chromosomes, making it fundamental for correct cell division. Alternatively, during SDSA, there is strand displacement and annealing, back to the other DSB end, followed by gap-filling DNA synthesis and ligation. This pathway leads only to NCO events, and the template DNA strand remains unchanged. Interestingly, the frequency of recombination events (both CO and NCO) is variable between individuals and populations, between sexes, across the genome, and between species (Stapley *et al.*, 2017a).

Recombination can impact rates of evolution through different processes. Crossover events break physical linkage between different genetic variants and create novel allele combinations in different genetic backgrounds. This way, interference between selected sites, a phenomenon referred to as Hill-Robertson interference (HRI) (Hill & Robertson, 1966), can be alleviated by recombination. Breaking linkage between two or more selected variants allows natural selection to act on each mutation independently, increasing its efficiency, which has a strong impact on genome evolution and adaptation (Felsenstein, 1974).

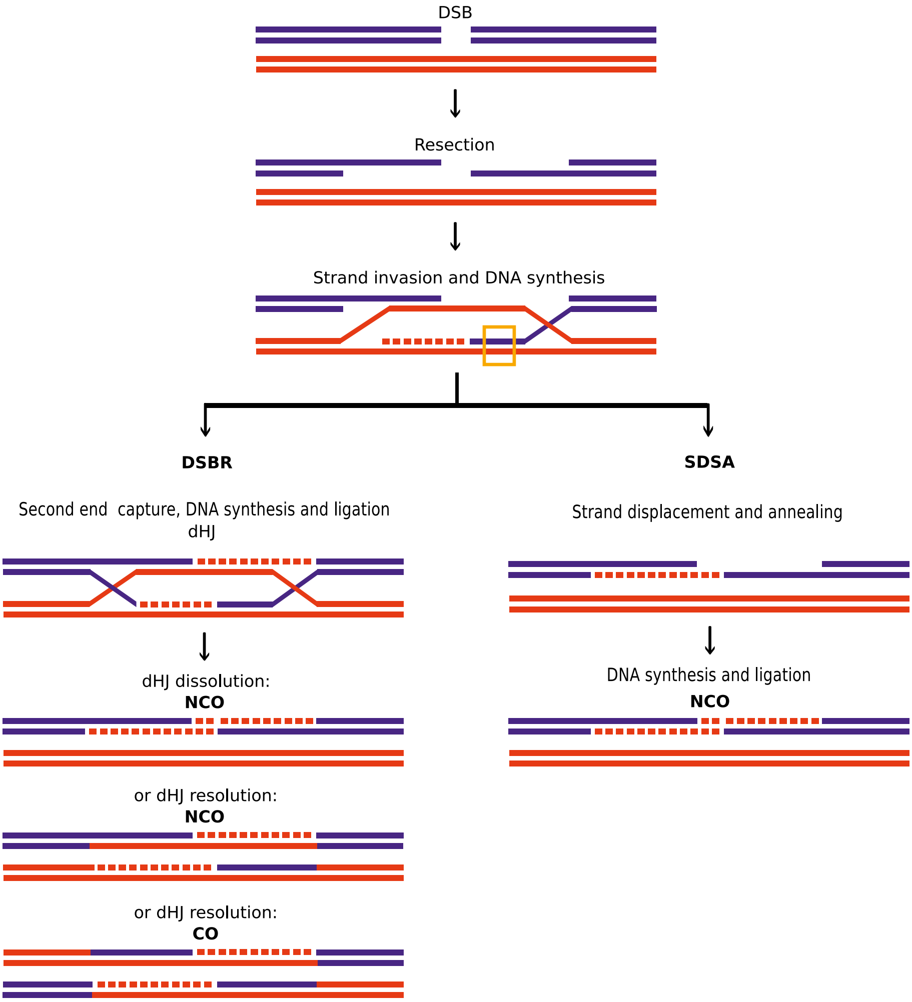
DSB
Resection
Strand invasion and DNA synthesis
DSBR
SDSA
Second end capture, DNA synthesis and ligation
dHJ
Strand displacement and annealing
dHJ dissolution:
NCO
DNA synthesis and ligation
NCO
or dHJ resolution:
NCO
or dHJ resolution:
CO

Figure 1. Meiotic recombination pathways to repair a DSB. Red and blue lines represent aligned homologous chromosomes. Meiotic recombination is initiated by the formation of a DSB. The repair initiates by resection of the broken strand. Then, one strand invades the complementary strand to use it as a template for DNA synthesis. Dashed lines indicate newly synthesized DNA. The orange square highlights one region (as an example) where heteroduplex DNA is formed. The repair can then follow different pathways; DSBR (*left*) and SDSA (*right*). DSBR leads to second end capture, followed by gap-repair DNA synthesis and ligation, forming of a dHJ. This junction can be “dissolved” or “resolved” into a NCO or a CO. SDSA leads to strand displacement and eventually a NCO, leaving the template strand unchanged. Adapted from (Sung & Klein, 2006).

Recombination can also alleviate the reduction in genetic diversity derived from the action of natural selection at linked neutral sites. If there is no recombination, all neutral mutations that are physically linked to a selected variant will share its fate. Specifically, neutral variants that are linked to a strongly deleterious mutation will be purged out of the population. Similarly, neutral variants that are linked to a positively selected mutation will "hitchhike" to fixation.

Besides breaking physical linkage, recombination can impact rates of evolution through gene conversion. After the DSB that initiates a recombination event, some stretches of the DNA sequence may be lost entirely on the broken strand. Also, the heteroduplex DNA formed to repair it may contain some mismatched base pairs if the two homologous chromosomes have different alleles. Hence, when the DNA sequence is copied from the intact chromosome to the broken one, there commonly is gene conversion; a unidirectional exchange of genetic information between them, which can occur during CO and NCO events. There is evidence that gene conversion at sites that are heterozygous for a "strong" (S; with strong referring to the number of hydrogen bonds between base pairs, i.e. three between G and C) and a "weak" nucleotide (W; two hydrogen bonds between A and T) transmits the S allele more frequently that the W one, in a process called GC-biased gene conversion (gBGC) (Duret & Galtier, 2009; Galtier *et al.*, 2001; Marais, 2003). The rate of gBGC events goes hand in hand with the rate of recombination and leads to the preferential fixation of S alleles.

Furthermore, recombination can also impact rates of evolution through other forms of meiotic drive events such as hotspot and indel drive. The first occurs when an individual is heterozygous for a recombinant and a non-recombinant allele and there is a higher transmission of the non-recombinant allele. The second is a biased transmission of indels in an indel/no indel polymorphism (for a review see Webster and Hurst 2012). Finally, as mentioned earlier, there is evidence that recombination may be mutagenic (Arbeithuber *et al.*, 2015; Halldorsson *et al.*, 2019; Hellmann *et al.*, 2003; Lercher & Hurst, 2002; Pratto *et al.*, 2014). Although indirect evidence has been found that recombination may be mutagenic in the collared flycatcher (Paper I), this thesis focuses mainly on the effect of recombination on rates of molecular evolution, mainly via HRI and gBGC. Therefore, I will be elaborating on these phenomena below.

Hill-Robertson Interference

Hill-Robertson interference (HRI) refers to a reduction in selection efficacy that occurs when selection acts in opposing directions on two or more linked variants (Hill & Robertson, 1966). When there is interference between linked sites, the fixation of an advantageous mutation may cause that one or several deleterious mutations to also be driven to fixation as a result of being linked together. In a similar manner, positively selected variants may be lost as a result of being linked to a strongly deleterious mutation. Recombination can alleviate this interference between sites. It re-shuffles genetic variation and breaks physical linkage among variants, creating new combinations of alleles in diverse genetic backgrounds. This will enhance the efficacy of natural selection, providing an evolutionary advantage. HRI may have a significant impact on the nonsynonymous and synonymous substitution rate. It can lead to the accumulation of slightly deleterious alleles, with the most pronounced consequences in regions of low recombination (Betancourt & Presgraves, 2002; McVean & Charlesworth, 2000).

GC-biased gene conversion

During the repair of a DSB, gene conversion can occur in many sites along a DNA sequence, also known as the conversion tract. Gene conversion is biased if some alleles are copied at a higher frequency than others. There is ample evidence to suggests that gene conversion is often GC biased, which means that it increases the transmission frequency of G and C over A and T alleles in sites that are heterozygote for a S and a W allele (Arbeithuber *et al.*, 2015; Duret & Galtier, 2009; Mugal *et al.*, 2015; Weber *et al.*, 2014a). This translates into an increased probability of fixation of W-to-S mutations, and a decreased probability of fixation of S-to-W mutations (Figure 2). Notably, GC-conservative mutations (S-to-S and W-to-W) are unaffected by gBGC.

By modifying the probability of fixation of W-to-S and S-to-W mutations, gBGC impacts the evolution of base composition. For instance, gBGC leads to a correlation between recombination rate and GC content (Duret & Galtier, 2009). This indirect evidence on the action of gBGC has now been reported in several and diverse taxa such as mammals (including primates), birds and reptiles but also plants, fungi and bacteria (Figuet *et al.*, 2015; Glémin *et al.*, 2014; Lassalle *et al.*, 2015; Lesecque *et al.*, 2013; Munch *et al.*, 2014; Pessia *et al.*, 2012; Weber *et al.*, 2014a). Direct evidence from yeast suggests that gBGC is only present in CO events (Lesecque *et al.*, 2013). However, more recent evidence suggests that, at least in humans and birds, gBGC can occur regardless of the outcome of the recombination event (i.e. CO or NCO) (Smeds *et al.*, 2016a; Williams *et al.*, 2015). Nonetheless, the extent of the impact of gBGC on rates of evolution along the genome is yet to be explored in most of these organisms.

gBGC is a nonadaptive process because S alleles are more frequently transmitted to the next generation regardless of their fitness effect. However, the impact of gBGC on allele frequency changes is similar to selection, because gBGC will increase the probability of fixation of some alleles over others (Nagylaki, 1983). Indeed, some studies have demonstrated that gBGC is responsible for the accelerated evolution of genes that were previously thought to have been under positive selection. Contrary to having positively selected substitutions, however, these genes contained several slightly deleterious S alleles that were driven to fixation by gBGC (Backström *et al.*, 2013; Berglund *et al.*, 2009; Galtier *et al.*, 2009; Ratnakumar *et al.*, 2010). Similarly, gBGC may increase the neutral substitution rate creating signatures similar to selection in codon usage (Galtier *et al.*, 2018).

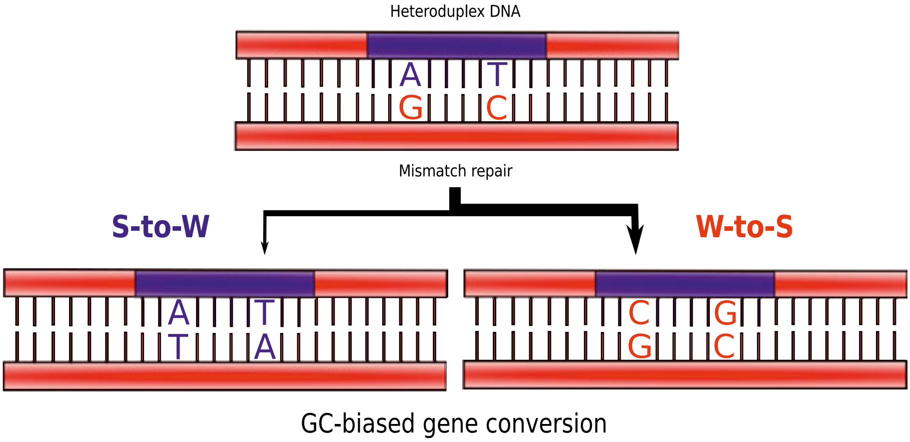

Figure 2. gBGC. The biased transmission of G and C alleles over A and T alleles during mismatch repair in heteroduplex DNA. In this example, there are two heterozygous sites/mismatches. These sites can be repaired towards the S or W allele. The thick black arrow represents a higher transmission of the S allele (compared to the expected 50:50 ratio), leading to a W-to-S fixation bias.

The neutral and nearly neutral theories of molecular evolution

The neutral theory of molecular evolution poses that the majority of mutations that can be observed within natural populations and between species are selectively neutral (Kimura, 1968, 1983). It predicts that negative selection is pervasive, but that negatively selected variants are rarely seen as polymorphisms (mutations segregating in a population) and rarely fixed as a result of their strong negative impact on an individual's fitness. The theory does not reject the importance of positive selection in the process of adaptation, but it suggests that positively selected mutations are relatively infrequent and, therefore, evolutionary change at the molecular level is mostly a result of neutral evolution though the action of random mutation and genetic drift.

One of the basic principles of the theory states that the fixation probability of a neutral mutation is equal to its frequency in the population (Kimura, 1968, 1983). For new mutations in a diploid organism, this frequency is $1/2N$, where N is the number of individuals in the population. The overall rate at which new mutations get fixed, k, is equal to the number of new mutations that enter the population, $2N\mu$, multiplied by the probability of a new mutation to become fixed, $1/2N$, meaning $k = 2N\mu \times 1/2N$ or $k = u$. Thus, the level of neutral genetic diversity within a species should be proportional to N, but the rate of fixation of neutral mutations is independent of N and is equal to the rate of mutation.

However, most mutations in a population are not either completely neutral or strongly deleterious. There exists a whole distribution of fitness effects where many mutations are weakly deleterious. The fate of these weakly deleterious mutations will not differ from the fate of neutral mutations, however, unless s is greater than $1/2N$ (Ohta, 1973, 1974, 1976). In other words, selection will only define the fate of new mutations if it can overcome the impact of genetic drift. If $s >> 1/2N$ selection will dominate over genetic drift, but if $s \approx 1/2N$, these mutations will be effectively neutral. Thus, selection is always more efficient in large populations where the impact of genetic drift is reduced. As the majority of polymorphisms in natural populations are nearly neutral, the rate of fixation of slightly deleterious mutations is predicted to be higher in small rather than large populations (Ohta, 1973, 1974, 1976). When N_e is small, slightly deleterious mutations may drift to fixation before selection is able to purge them from the population. This also implies that the fast evolving regions of the genome should show higher levels of polymorphism.

The importance of nearly neutral mutations was incorporated into the nearly neutral theory of molecular evolution, which provides a useful framework where a null hypothesis involving mutation and genetic drift can be tested against an alternative selective scenario (Ohta, 1992). If the null hypothesis is rejected, we can infer selection has taken place. Statistical methods that test if the neutral or nearly neutral theory properly describe evolution at the molecular level have been proposed and are widely used to detect regions or genes under positive selection (Eyre-Walker & Keightley, 2007; Keightley & Eyre-Walker, 2007; McDonald & Kreitman, 1991; Stoletzki & Eyre-Walker, 2011). While these methods are useful, our interpretations of selection may be biased if one or more assumptions are violated, or if they are oversimplified and fail to incorporate the effect of some mechanisms that describe the evolutionary process.

Molecular evolution of sex chromosomes

The molecular evolution of sex chromosomes differs from autosomes as a result of their unusual mode of inheritance. Consequently, comparing and contrasting rates and patterns of evolution between sex chromosomes and autosomes can be helpful for elucidating the role of different evolutionary forces on genome evolution. As birds have female heterogamety (males have ZZ and females have ZW sex chromosomes), the study of bird species offers a suitable system to compare to male heterogametic systems (where females have XX and males have XY sex chromosomes) and test theoretical predictions on sex chromosome evolution.

If the variance in offspring number for males and females is random, the N_e of Z (and X) chromosomes relative to the autosomes is expected to be 0.75 (Caballero, 1995) as there are three copies of the Z chromosome compared to four copies of any of the autosomes in the population. A deviation from the 0.75 expectation may occur as a result of variation in evolutionary forces between males and females. Firstly, the mutation rate differs between male and female germ lines. Even when the mutation rate per cell division is the same in males and females, male germ line mutation is usually higher as a result of a higher number of mitotic cell divisions in spermatogenesis compared to oogenesis (Bachtrog, 2008; Ellegren, 2007). In birds, the Z chromosome spends 2/3 of its time in the male germ line, while the W chromosome is inherited entirely through the female germ line. Consequently, the mutation rate per generation would likely be higher in the Z chromosome and lower in the W chromosomes relative to autosomes. This could lead to higher divergence rates on the Z chromosome and lower divergence rates on the W chromosome relative to autosomes, since the neutral substitution rate is determined solely by the mutation rate. Secondly, demographic changes such as population bottlenecks or expansions, will disproportionately reduce N_e and genetic diversity on sex chromosomes (Pool & Nielsen, 2007). Also, sex-specific migration and variance in reproductive success can affect (either increase or decrease) genetic diversity disproportionately on sex chromosomes compared to autosomes, as they effectively reduce N_e of one sex relative to the other (Caballero, 1995; Gillespie, 1974; Hammer *et al.*, 2008). For example, in systems where males disperse more often than females, or in a polygynous mating system (were few males mate with multiple females) the effective number of Z chromosomes is reduced. Finally, N_e is also influenced by the effect of selection at linked neutral sites. Because the Z chromosome recombines only in the homogametic sex, it has a reduced sex-averaged recombination rate compared to autosomes, which makes the effect of linked selection much greater (Hammer *et al.*, 2010).

Over a short evolutionary time scale a reduction in N_e reduces the levels of genetic diversity on the Z chromosome versus autosomes; however, over a long evolutionary time scale reduced N_e may translate into a fast-Z effect; a higher ratio of non-synonymous to synonymous substitution rates (d_N/d_S) on the Z chromosome compared to autosomes, as a consequence of a reduced efficacy of selection through HRI (Mank et al., 2007). Furthermore, positive selection may act more efficiently on the Z chromosome as a result of hemizygosity. In female heterogametic systems, recessive advantageous mutations will be exposed to selection in females, already at low frequencies, reducing their chance of being lost by genetic drift. Therefore, the Z chromosome may possess a disproportionately large role in harboring adaptive mutations, which could also contribute to the fast-Z effect.

Methods

The methods used in this thesis help us identify and quantify the relative strengths of the mechanisms and factors that determine evolutionary change at two different timescales. One is a long evolutionary timescale, where we analyze and compare DNA variation that has been fixed between different species. The other is a short timescale, where we analyze and compare DNA variation that is still segregating between individuals of a single population. These two types of data are referred to as divergence and diversity data, respectively. Each provides unique information and we use different statistical methods for their analysis. Importantly, the combined analysis of divergence and diversity information can provide further insights on the evolution of a particular lineage of interest. In the following chapter, I describe a few of the basic principles of the most important or recurrent methodologies used in this thesis. The particulars of the models used can be found within each chapter.

Models of molecular evolution

To infer rates and patterns of substitution, the first step is to identify and align homologous (assumed to have a single common ancestor) sequences from different species. After homology for each site in the DNA (or amino acid) sequence has been identified, differences between species can be detected. These differences are mutations that have been fixed in different lineages, becoming substitutions. Following this, probabilistic models can be built to describe the rate of change between nucleotides (or codons or amino acids) over time. These fixed nucleotide differences between species provide the essential information needed to estimate the evolutionary distance between them measured as the expected number of substitutions per site on a particular branch of a phylogenetic tree (or the expected number of substitutions per codon in codon models). This distance is a function of the mutation rate and the time of divergence.

To estimate these parameters, the most common probabilistic models are based on continuous-time Markov chains. They assume that the evolutionary process is stochastic and memory-less. This means that from each state, the sequence can change to another state with a certain probability, which depends only on the current state and not on past states. In analyses of nucleotide data these states are the A, C, G and T bases. The substitution matrix $Q = (q_{ij})$ describes the instantaneous rate of change between states i and j. We can translate these relative rates into transition probabilities over evolutionary time t through matrix exponentiation $P(t) = e^{Qt}$. Transition probabilities are estimated for each site in the sequence alignment and these probabilities are multiplied to obtain the likelihood of the data given the model. The aim of this analysis is to identify the model that best describes the data. Maximum likelihood is used to identify the best-fit model by exploring multiple combinations of model parameters (Nielsen, 2005; Yang, 1998).

In a similar way, codon substitution models are used to model nucleotide substitutions in protein coding genes. In this case, the substitution matrix describes the rate of change between the 61 codons in the genetic code (Goldman & Yang, 1994; Muse & Gaut, 1994) rather than the four DNA bases. These models commonly estimate the ratio of nonsynonymous over synonymous substitution rates (ω, also referred to as d_N/d_S) in order to model differences in the substitution rate between selected and neutrally evolving mutations. These models assume that synonymous sites evolve under neutrality, as substitutions in these sites do not change the amino acid sequence of the protein for which they code. In contrast, it is assumed that nonsynonymous sites evolve under selective pressure, as substitutions in these sites do change the amino acid sequence. Therefore, these models estimate a parameter that reflects the relative difference in the rate of evolution of sites that are only influenced by mutation rate and genetic drift (i.e. synonymous sites) and sites that are also affected by selection (i.e. nonsynonymous sites). Said differently, the rate of nonsynonymous changes can be normalized by the synonymous rate, which reflects local variation in mutation rate. The basic interpretation of d_N/d_S estimates is that if a gene is free of selective constraint, $d_N/d_S = 1$. If $d_N/d_S > 1$, the gene has evolved under positive selection, whereas if $d_N/d_S < 1$ then negative selection has acted upon the gene. However, it is important to acknowledge that $d_N/d_S < 1$ does not mean that no positive selection has taken place. This is due to the fact that the majority of mutations in functional sites of the genome are under negative selection (i.e. have a deleterious effect) (Li, 1997), thereby masking the signatures of positive selection. This problem can be solved by comparing alternative models, with and without positive selection, in order to identify genes or regions that were targets of positive selection. Additionally, as a result of the majority of mutations being deleterious, d_N/d_S can also be seen as a measure of selective constraint, where stronger constraint translates into lower d_N/d_S. Thus, high d_N/d_S may reflect the action of positive selection but also the effect of relaxed selective constraint, which results in higher accumulation of slightly deleterious nonsynonymous mutations. Mean values of d_N/d_S across the genome can therefore be used to compare evolutionary trends in different lineages as well as to compare genes with different characteristics or located in different genomic regions.

Models of molecular evolution can accommodate realistic and interesting assumptions regarding the rate of change between different nucleotides by optimizing distinct parameters. Examples include the transition-to-transversion rate ratio or the rate of S-to-W, W-to-S and GC-conservative mutations.

Estimation of genetic diversity

Estimating the genetic diversity of a population requires the comparison of DNA sequences between individuals of the same species. The first step in obtaining estimates of genetic diversity is the identification of polymorphisms within a population, most commonly by looking for single nucleotide polymorphisms (SNPs) across individuals. SNPs provide the raw information for estimating several measures of genetic diversity. Besides the number of polymorphisms in the sample, information on their frequency in the population is valuable. This information is summarized in the site frequency spectrum (SFS), which is a distribution of the derived (or minor) alleles in the sample. From the SFS we can estimate several measures of genetic diversity.

In this thesis, we mainly estimated nucleotide diversity as measured by θ_W and π. Nucleotide diversity as measured by θ_W is the number of polymorphisms scaled by the harmonic number of the sample size (Watterson, 1975). Nucleotide diversity as measured by π is the average number of nucleotide differences between any two alleles in the sample (Nei & Li, 1979). The sequence length usually normalizes these two summary statistics, so the estimate is a measure of genetic diversity per site. θ_W and π are expected to be the same in a neutrally evolving population. Any deviation from neutrality assumptions is reflected in differences between them.

Sometimes it is not possible to analyze sequences from several individuals. In such cases, a proxy for genetic diversity can be obtained by estimating heterozygosity in one diploid individual (Lewontin & Hubby, 1966). In this case, the information on allele frequencies in the population is lost and one can only make inferences based on the number of polymorphic sites.

The distribution of fitness effects and the rate of adaptation

A sophisticated way to disentangle and quantify adaptive and nonadaptive forces in determining rates of evolution is to combine divergence and diversity information in a McDonald-Kreitman (MK) framework (McDonald & Kreitman, 1991). The main assumptions remaining that only mutation rate and genetic drift determine neutrally evolving divergence and diversity levels, while mutation rate, genetic drift and selection determine the selected divergence and diversity levels. It is also assumed that there is independence among polymorphic sites (no-linkage). Finally, it is assumed that advantageous mutations reach fixation too quickly to be observed as polymorphisms and will therefore only contribute to divergence measures (although see Tataru *et al.*, 2017). As a result, the relationship between synonymous and nonsynonymous polymorphisms present in the population sample reflects the strength of negative selection and represents a "fair" reference point for a d_N/d_S expectation under neutrality (McDonald & Kreitman, 1991; Smith & Eyre-Walker, 2002). We can then compare the observed d_N/d_S with the inferred expectation and attribute the difference to positive selection.

The most recent methods derived form the MK framework account for the continuum of selective effects, the distribution of fitness effects (DFE), to estimate the expected d_N/d_S under neutrality (ω_{na}) (Eyre-Walker & Keightley, 2007, 2009; Keightley & Eyre-Walker, 2007). The DFE reflects the relative importance of selection and genetic drift in determining the probability of fixation of new nearly neutral mutations (Eyre-Walker & Keightley, 2007). To estimate the DFE, the population mutation rate is inferred from the neutral SFS to control for the effects of demography or any other factors that may influence the SFS under neutrality and under selection with the same magnitude. Then, a distribution of selection coefficients (usually a gamma distribution) is fitted to the selected SFS under selection to calculate two parameters by maximum likelihood; the shape and the scale of the distribution. The DFE is a continuous distribution but the available methods discretize this distribution, so that the interpretations regarding the deleterious effect of mutations are more robust (Eyre-Walker & Keightley, 2007, 2009; Keightley & Eyre-Walker, 2007). This way, we are able to determine the proportions of sites that are evolving within each selection category, i.e. strongly deleterious, deleterious or effectively neutral.

Assuming that evolutionary forces remain constant over time in the lineage of interest, the DFE helps us estimate an expected value of ω (d_N/d_S) for nonadaptive substitutions (ω_{na}). The difference between the observed (ω) and the expected substitution rate ratio (ω_{na}) can be attributed to the rate of adaptation (ω_a). Thus, ω_a measures the rate of adaptive evolution relative to the nearly neutral reference. The proportion of amino acid substitutions that are adaptive (α) can then be derived as the ratio of ω_a and ω_{na} (Eyre-Walker & Keightley, 2009; Smith & Eyre-Walker, 2002).

Gene expression analysis

To determine how selection has shaped patterns of sequence evolution one may also use complementary approaches to DNA-based analyses. One strategy is to investigate gene-specific characteristics such as patterns of gene expression and protein-protein interactions (PPI) (Pal *et al.*, 2006; Zhang & Yang, 2015). There is ample evidence to suggest that gene expression level, the level of pleiotropy, and sex-biased expression are determinants of the strength of purifying selection in various taxa (Ellegren & Parsch, 2007; Krylov *et al.*, 2003; Pal *et al.*, 2001). To obtain a gene expression estimate for each gene, RNA-seq reads are mapped to a reference genome. The gene length and the sequencing depth are then used to normalize the read counts. This allows us to estimate the relative abundance of transcripts and perform differential expression analyses, in order to ascertain which transcripts have different abundances between two groups. This approach allows us to understand the link between rates and patterns of molecular evolution and gene expression patterns. For example, we can compare adaptive and nonadaptive rates of evolution between highly and lowly expressed genes to assess how selection shapes patterns of gene expression levels. Also, we can analyse the differential expression of sex-biased genes and unbiased genes to make inferences on sexual selection.

Study system: the avian genome

Birds are fascinating organisms. They are an extremely diverse group that has adapted to every environment on the planet. They have fascinated scientists and amateurs alike; with Darwin famously deriving his evolutionary theory from observations on pigeons breeding, as well as Galapagos finches. Avian genomes share unique characteristics that distinguish them from other taxa and make them ideal genetic study systems. Bird genomes are the smallest among the amniotes. Their genome sizes range from 0.91 to 1.3 Gbp, while the genomes of mammals and non-avian reptiles typically range from 1.0 to 8.2 Gbp (Gregory, 2005). Gene length in birds also appears to have been reduced, compared to other amniotes (Jarvis *et al.*, 2014). The majority of bird genomes contain a very low proportion of repeat elements, ranging from only 4 – 10% compared to other taxa such as mammals, where repeat elements represent 35 – 50% of genomic sequence (Feschotte & Pritham, 2007; Jarvis *et al.*, 2014; Kidwell, 2002). The bird karyotype is relatively stable and consists of a few macrochromosomes and several microchromosomes with the average chromosome number being close to 40 homologous pairs (Ellegren, 2013). Macro- and microchromosomes differ on several genetic characteristics. Microchromosomes have been reported to show higher GC content, gene density and recombination rate but lower

d_N/d_S (Axelsson *et al.*, 2005). In general, birds have a relatively high recombination rate compared to other amniote species (Stapley *et al.*, 2017b). Interestingly, these rates vary greatly in different genomic regions but are also strikingly conserved thought the phylogeny (Kawakami *et al.*, 2014). Finally, contrary to most model species, birds have female heterogamety. Thus, the study of sex chromosomes in birds can provide unique insights into sex chromosome evolution as it allows us to compare female heterogametic systems with the more studied male heterogametic ones. All these characteristics make bird genomes interesting candidates within which to explore the impact of evolutionary mechanisms on genome evolution (Ellegren, 2013).

In the first part of this thesis (Papers I and II), I based my studies on the genome of the collared flycatcher (*Ficedula albicollis*); a small migratory passerine species from the Old World. This species has been intensely studied in the wild. Individuals return to the same forest and occupy artificial nest boxes every breeding season. Birds can be ringed and scientists are then able to track their development; as well, registering family relationships (Gustafsson *et al.*, 1994; A. Qvarnström *et al.*, 2016). As a result, their ecology and behavior is well understood (A. Qvarnström *et al.*, 2010; Sæther *et al.*, 2007). More recently, huge genomic resources were generated with the publication of a high quality genome, followed by the re-sequencing of hundreds of individuals (Burri *et al.*, 2015; Ellegren *et al.*, 2012). The construction of a large pedigree has also made it possible to estimate recombination rates along the genome, which was crucial for compiling this thesis (Kawakami *et al.*, 2014). Very few other bird species have such an array of resources to facilitate their study.

In 2014, dozens of avian genomes were simultaneously published along with new insights into avian evolution (Jarvis *et al.*, 2014). However, this was still a relatively poorly explored avian dataset, which provided me with an exciting opportunity to explore our hypotheses using a comparative genomics approach (Papers III and IV).

Research aims

The aim of my doctoral thesis is to investigate the mechanisms and factors that determine rates and patterns of molecular evolution in avian species. Specifically, I study the impact of recombination rate on genome evolution, through its interaction with the efficacy of selection and through the process of GC-biased gene conversion (gBGC). I evaluate how gBGC interacts with selection to determine evolutionary change, and how ignoring its effects may bias our interpretations of sequence data. I also aim to characterize the true strength of selection and identify factors that determine constraint in the collared flycatcher lineage. Finally, I focus on how the interaction between evolutionary mechanisms differentially impact sex chromosomes compared to autosomes in birds. Specifically, this thesis aims to:

Paper I — Examine the relative and combined effect of HRI and gBGC mediated through recombination rate variation on rates of molecular evolution and inferences of natural selection along the collared flycatcher genome.

Paper II — Assess the impact of gBGC, gene expression level, sex-biased expression and the number of PPI on rates of adaptive and nonadaptive evolution in the collared flycatcher genome.

Paper III — Explore whether the impact of gBGC conceals the correlation between life-history traits and d_N/d_S in the avian clade.

Paper IV — Investigate the determinants of the relative levels of Z chromosome to autosomes genetic diversity and their relationship to the fast-Z effect across the avian phylogeny.

Summary of papers

Paper I – Recombination rate variation modulates gene sequence evolution mainly via GC-biased gene conversion, not Hill-Robertson interference, in an avian system

The ratio of the nonsynonymous to synonymous substitution rates (*d_N/d_S)* is a widely used measure of the strength of selection acting on protein coding genes. However, d_N/d_S estimates of two genes subject to similar selection pressure can turn out to be very different if they are located in different recombination landscapes. Reduced recombination can impact d_N/d_S via a the reduction in the efficacy of selection as a result of linkage, a phenomenon referred to as HRI (Hill & Robertson, 1966). Alternatively, recombination may impact d_N/d_S by means of gBGC, a biased repair mechanism that leads to an increased transmission of GC over AT alleles (Duret & Galtier, 2009). As a consequence, gBGC affects substitution rates in the same way as directional selection, but unlike selection, it acts regardless of the fitness effect of mutations.

In this study we explored the impact of recombination rate variation via HRI and gBGC on inferences of natural selection along the collared flycatcher genome. To do so, we estimated substitution rates of >8000 genes for four different mutation categories independently; W-to-S, S-to-W (which are favored and disfavored by gBGC, respectively), S-to-S and W-to-W (which are unaffected by gBGC). We observed a negative relationship between d_N/d_S and recombination rate, which at first glance may be interpreted as a consequence of HRI. However, if HRI was determinative of genome-wide patterns of d_N/d_S, we would expect to also observe a negative correlation between recombination rate and d_N for all mutation categories, which we did not observe. On the contrary, the results showed several typical signatures of gBGC. Firstly, W-to-S substitution rates (d_N and d_S) were positively correlated, while S-to-W rates were negatively correlated, with recombination rate. Secondly, we observed a positive correlation between recombination rate and current GC content. Finally, analyses of diversity data confirmed the role of gBGC; the SFS showing a right skew and higher proportion of high-frequency derived variants for the W-to-S class, and a higher proportion of low-frequency derived alleles for the S-to-W class. This held true for both selected and neutrally evolving sites. We therefore concluded that gBGC was the underlying mechanism leading to a negative relationship between recombination and d_N/d_S. This was an unexpected result, as previous studies have suggested that, in mammals, gBGC leads to a higher d_N/d_S in genes located in high recombination regions (Backström *et al.*, 2013; Galtier & Duret, 2007; Galtier *et al.*, 2009; Ratnakumar *et al.*, 2010). We argued that the higher impact of gBGC on synonymous substitutions, compared to nonsynonymous substitutions, may lead to a negative relationship between recombination and d_N/d_S in the collared flycatcher. To better understand the consequences of recombination via gBGC on rates of molecular evolution, we provided a simple analytical description of its impact on substitution rates. We identified the equilibrium GC content (GC*) and the distance to the equilibrium GC (ΔGC) as important determinants of the impact of gBGC on substitution rates. In conclusion, the generally high and stable, yet simultaneously heterogeneous recombination landscape in birds may have allowed gBGC to show a particularly strong impact on substitution rates; even more so for neutrally evolving mutations. This study underlines the importance of investigating different groups of organisms to gain a better understanding of the general mechanism by which gBGC interacts with natural selection to determine rates of molecular evolution; and provides strong evidence against interpreting selection signatures based on d_N/d_S without properly accounting for gBGC.

Paper II – Biased inference of selection due to GC-biased gene conversion and the rate of protein evolution in flycatchers when accounting for it

Understanding the relative roles and interactions of different evolutionary forces and genetic factors that determine the rates and patterns of sequence evolution is a long-standing question in molecular evolution. Mainly, variation in μ, N_e, and s will influence rates of evolution (Charlesworth, 2009; Ohta, 1992). However, the local recombination rate may also influence rates and patterns or molecular change via HRI and gBGC. While there is clear evidence that signatures of gBGC are pronounced in the collared flycatcher genome and that they bias d_N/d_S estimates (Paper I), it is still unclear how gBGC may impact inferences on the distribution of fitness effects (DFE) and the rate of adaptive substitutions. Furthermore, little is known on the relative impact of several other factors such as biochemical protein properties in determining rates of adaptive and nonadaptive protein evolution.

In this study, we used a MK-derived approach (Eyre-Walker & Keightley, 2009; Keightley & Eyre-Walker, 2007) to assess the impact of gBGC, gene expression level, sex-biased expression and the number of PPI on estimates of the DFE, d_N/d_S, the rate of adaptive evolution (ω_a) and the proportion of amino acid substitutions fixed by positive selection (α) in the collared flycatcher (*Ficedula albicollis*) lineage, since its split from the zebra finch (*Taeniopygia guttata*) lineage. We showed that all measures of selection were strongly influenced by gBGC, with this being especially true at the time scale of fixed differences. Specifically, d_N/d_S was 27% higher when estimated using all changes compared to GC-conservatives only (0.144 vs. 0.113), α was 33% lower (0.180 vs. 0.270), and ω_a was 22% lower (0.025 vs. 0.032). This indicates that, in this lineage, gBGC lead to a significant underestimation of the amount of adaptive evolution.

We found a strong relationship between measures of purifying selection and gene expression level as well as with the number of PPI. These factors were also positively correlated with ω_a, and α, which suggests that highly expressed genes and genes that are part of several protein complexes have a higher rate of adaptation and a larger fraction of adaptive substitutions. We observed that both female- and male-biased genes have higher rates of adaptation, compared to unbiased genes, but only male-biased genes seemed to evolve under weaker selective constraint.

In conclusion, this study highlights the importance of taking gBGC into account when analyzing genome-wide patterns of selection; especially when making comparisons between taxa where the strength of gBGC may vary. We further show that individual protein properties like gene expression level, the number of PPI and sex-biased gene expression, are important determinants of both the strength of negative and positive selection in protein coding genes in the collared flycatcher.

Paper III – GC-biased gene conversion conceals the prediction of the nearly neutral theory in avian genomes

A core prediction of the nearly neutral theory is that the efficacy of natural selection increases with N_e (Ohta, 1992). Small populations should accumulate a larger proportion of slightly deleterious mutations compared to large populations. This is because the efficacy of selection depends on a balance between the strength of random genetic drift (determined by N_e) and the selection coefficient of new mutations. This prediction has been corroborated by independent observations in diverse taxa, where life-history traits (LHTs) (commonly used as proxies for N_e) are strongly correlated with measures of selection efficacy, such as the d_N/d_S ratio (Figuet *et al.*, 2016; Nabholz *et al.*, 2013; Popadin *et al.*, 2007). Surprisingly, several studies have failed to detect a correlation between LHTs and the d_N/d_S ratio in avian taxa (Figuet *et al.*, 2016; Nabholz *et al.*, 2013; Weber *et al.*, 2014b).

In this study, we explored the role of gBGC in concealing the prediction of the nearly neutral theory in birds. We analysed the relationship between d_N/d_S and LHTs in birds based on coding sequence alignments between 47 avian species (Jarvis *et al.*, 2014). To distinguish the impact of gBGC from the impact of selection, we applied a substitution model that accounts for non-stationary base composition and allows estimating d_N/d_S separately for substitution categories that are affected (W-to-S and S-to-W) and unaffected (GC-conservative) by gBGC (Guéguen & Duret, 2017). LHTs showed no correlation with d_N/d_S in birds when analysing all substitution categories together. However, we observed a strong positive correlation between LHTs and d_N/d_S when analysing GC-conservative substitutions independently. Hence, our results suggest that the impact of gBGC on estimates of substitution rates blurs the correlation between d_N/d_S and LHTs in birds. Furthermore, we observed that estimates of d_N/d_S are consistently lower for GC-rich genes compared to GC-poor genes, but the relationship between LHTs and the GC-conservative substitution rate is robust to variation in local GC content. Finally, the magnitude of the impact of gBGC on d_N/d_S varies between lineages. We hypothesized that this is potentially related to the distance to the equilibrium GC content, which is (in most avian taxa) larger for synonymous than nonsynonymous changes.

In conclusion, our study illustrates that accounting for gBGC is important to make correct inferences of selection. We confirmed that birds are not an exception to the prediction of the nearly neutral theory; the efficacy of selection increases with N_e.

Paper IV – Variation in the Z chromosome to autosomes ratio of genetic diversity across birds and its relationship to the fast-Z effect

Given their unique mode of inheritance, sex chromosomes differ from autosomes in several aspects including N_e, mutation and recombination rates. As a consequence, they differ in levels of genetic diversity and divergence. Understanding the mechanisms underlying differences between sex chromosome and autosomes can help us recognize sex-specific demographic and selective evolutionary events and eventually, further our comprehension of the evolutionary process. When the variance in offspring number for males and females is equal, N_e and, therefore, the levels of nucleotide diversity on the Z or X chromosomes (in female and male heterogametic systems, respectively) relative to the autosomes (Z:A diversity) is expected to be 0.75 (Caballero, 1995). However, a deviation of the expected 0.75 may occur as a result of sex-differences in mutation rate and N_e (Bachtrog, 2008; Ellegren, 2007). Over evolutionary timescale, these differences may also translate into differences in rates of sequence divergence. Frequently, this turns into a higher d_N/d_S on the Z chromosome in birds, a phenomenon formally known as the fast-Z effect (Counterman *et al.*, 2004; Mank *et al.*, 2007).

In this study, we analyzed genome-wide data in males from 32 avian species across an avian phylogeny to elucidate the evolutionary mechanisms that shape molecular evolutionary patterns on the Z chromosome compared to autosomes. While we observe large variation in levels of genetic diversity among members of the avian phylogeny, the mean of the distribution was not significantly different from the expected 0.75. Since the majority of the studied species are socially monogamous, we argue that unequal variance in reproductive success in males and females is not a strong determinant of the observed range of Z:A diversity. We observed an increased mutation rate on the Z chromosome compared to autosomes. The male to female mutation bias varied between species and was positively correlated to Z:A genetic diversity. This supports male mutation bias as an important determinant of the relative levels of diversity between the Z chromosome and autosomes. Furthermore, a negative correlation between Z:A diversity and N_e, coupled with a reduction in Z:A diversity in regions with a higher density of targets of selection, points toward a strong prevalence of linked selection on the Z chromosome compared to autosomes. In addition, we report a fast-Z effect in the majority of species. Interestingly, we observe no correlation between the extent of the fast-Z effect and levels of Z:A diversity, which suggests that genetic drift alone might not be enough to explain higher rates of evolution on the Z chromosome in birds.

Concluding remarks and future prospects

In this thesis, I perform a series of studies on the determinants of the rates and patterns of molecular evolution in the collared flycatcher and other avian species. These studies provide strong evidence of the pervasive impact of recombination rate variation on the evolutionary dynamics of avian genomes. More precisely, I confirm that gBGC is a key determinant of evolutionary rates and distribution of GC content along avian genomes, and may also lead to spurious signatures of selection. Thus, it is important that future studies account for the effect of gBGC before interpreting patterns of sequence divergence and polymorphism. A re-examination of previously identified genes and regions under selection in bird species as well as other taxa should be encouraged, these results demonstrate that gBGC may lead to false positives and in some cases, its effect may be wrongly interpreted as a signature of HRI. This work suggests that even though HRI can lead to the fixation of slightly deleterious variants and reduce the rate of adaptation, its impact is stronger in non-recombining or very low recombining regions and does not determine genome-wide patterns of d_N/d_S in the collared flycatcher. Further studies should confirm if this is also true for other bird species as well as other organisms.

Furthermore, whilst this thesis describes the patterns produced by gBGC in avian genomes, there are still several open questions regarding the reasons it is so pervasive in birds; especially in neutrally evolving sites. It is possible that the particular dynamics of the recombination landscape in birds allows for gBGC to have distinctly strong effect on avian genome evolution, by acting on the same genomic regions persistently over a long period of time (Mugal *et al.*, 2013). Additionally, I hypothesized that GC* and ΔGC play a major role; however, further studies are warranted to test this hypothesis or propose alternative explanations.

Although efforts have been made in this direction (Capra *et al.*, 2013; Duret & Arndt, 2008; Glémin *et al.*, 2015; Lartillot, 2012), additional investigations are needed in order to incorporate gBGC in a null model of molecular evolution (Galtier & Duret, 2007). This thesis strongly suggests that this model should accommodate differences in the impact of gBGC on neutrally evolving sites compared to sites under selection. This would facilitate the correct identification, and quantification of the strength, of gBGC and selection in molecular evolutionary data. However, the forces underlying the differential impact of gBGC on functional and nonfunctional sites are still challenging to understand theoretically (Duret & Galtier, 2009). Thus, future theoretical research that focuses on understanding the differential impact of gBGC on neutral and selected substitution rates will be of huge significance.

There is still a long way to go in characterizing the evolutionary causes and consequences of gBGC. It has been hypothesized that gBGC is an evolutionary response to a high rate of CpG mutations, as a fixation bias in favor of GC alleles could potentially attenuate the effect of a higher mutation rate toward AT alleles (Birdsell, 2002). However, gBGC may promote the fixation of slightly deleterious S alleles, which can increase the mutation load of a population and have drastic evolutionary consequences. Thus, understanding the interplay between these mutation and fixation biases and their overall impact on the mean fitness of the population could ultimately help us understand the evolution of gBGC. If, in the long run, gBGC decreases the mean fitness of the population, natural selection would likely act to reduce its strength. Conversely, if mutational bias (favoring AT alleles) was responsible for a lowering of the mean fitness of the population, then selection could potentially act to increase the strength of gBGC, as a counterbalance. This implies that selection would act to regulate the strength of gBGC, and as a consequence influence the evolution of the recombination rate itself.

Finally, these results provide new insights into the determinants of the relative levels of diversity and divergence of the Z chromosome compared to autosomes in birds. Recombination through the effect of linked selection and gBGC, in conjunction with higher mutation rates in males, determine relative levels of genetic diversity and divergence. These results suggest that the larger impact of genetic drift on the Z chromosome compared to autosomes is not sufficient to explain a higher d_N/d_S in sex-linked genes, and indicate that positive selection might also contribute to the fast-Z effect in birds. Thus, a thorough re-examination of the factors that contribute to the fast-Z effect in different female heterogametic systems is necessary.

In summary, a combination of the genomic analyses of several more avian and non-avian species in conjunction with theoretical models and analyses could lead to answers to these and several other open questions. Fortunately, now is an exciting moment to do research in the field of molecular evolution, as the number and quality of genomic resources is increasing and already providing valuable information that will help us test new hypotheses and theories.

Svensk sammanfattning

En grundlig förståelse för den evolutionära processen är avgörande för förståelsen av liv. Att studera evolution kan hjälpa oss att dechiffrera levande arters historia och beskriva och förutsäga mönster som vi observerar i naturen. Det kan också hjälpa oss i mer tillämpade ämnen, såsom karaktärisering och bevarande av biologisk mångfald eller kontroll och förebyggande av infektionssjukdomar. Således är evolutionen en förenande princip för varje disciplin av biologi, från biokemi till fylogenetik.

Under de senaste åren har framskridandet av genomteknik gjort det möjligt att studera evolution på molekylär nivå och besvara frågor och teorier som tidigare inte varit möjliga. Målen för evolutionära studier på molekylär nivå är tvåfaldiga: att försöka härleda släktskap mellan arter eller, som i denna avhandling, fokusera på att förstå mekanismer och faktorer som bestämmer evolutionär förändring. Detta kan göras genom att jämföra molekylära data från flera arter eller inom en enda population.

Enkelt uttryckt är evolution förändringen i den genetiska uppbyggnaden hos en befolkning genom tiden. Samspelet mellan flera mekanismer och genetiska faktorer bestämmer i vilken hastighet denna förändring sker. Mutation är den enda mekanismen som genererar ny genetisk mångfald. Mutationshastigheten är den takt som nya varianter kommer in i befolkningen. Andra mekanismer skapar inte mångfald, men bestämmer mutationers öde genom att påverka deras sannolikhet att segregera i befolkningen och så småningom fixeras. Exempelvis är ändliga populationer utsatta för stokastiska och demografiska processer, vilket kommer att påverka sannolikheten för fixering av segregerande mutationer genom genetisk drift. Å andra sidan påverkar selektion sannolikheten för fixering på ett deterministiskt sätt genom att öka eller minska fixeringssannolikheten av varianter som påverkar individens förmåga, det vill säga deras förmåga att överleva och reproducera. Vidare spelar även rekombination en viktig roll i utvecklingen. Genom omblandning av genetisk variation kan rekombination indirekt öka urvalseffektiviteten. Å andra sidan kan rekombination påverka sannolikheten för fixering genom en process som kallas *gene conversion*. Detta innebär att vissa specifika alleler oftare förs vidare till nästa generation på grund av en enkelriktad överföring av genetiskt material mellan homologa kromosomer. Omfattande bevis tyder på att denna process är *GC-biased*, och inte påverkas av fitnesseffekter.

Alla dessa mekanismer interagerar med varandra. Deras relativa styrka och inverkan på den evolutionära processen är komplex och varierar drastiskt mellan arter och även mellan olika genomiska regioner. Att förstå de relativa effekterna av dessa mekanismer och andra genetiska faktorer vid bestämning av evolutionära hastigheter och mönster är ett viktigt mål i många evolutionära studier, inklusive denna avhandling. Speciellt att identifiera regioner eller gener under selektion och att karaktärisera styrkan hos urvalsprocessen har varit några av de mest populära metoderna under de senaste åren, eftersom de kan kasta ljus över anpassningsprocessen.

Denna avhandling fokuserar på att reda ut de här mekanismernas relativa roll när det gäller att bestämma evolutionära hastigheter och mönster i fågelgenom. Jag inriktade mig först på att förstå hur variationer i rekombinationshastighet påverkar molekylär evolution och selektion längs med genomet hos halsbandsflugsnappare (*Ficedula albicollis*). Vi visade att effekten av *GC-biased gene conversion* (förkortat gBGC) på evolutionshastigheter är stark, speciellt i neutrala regioner jämfört med positioner under selektion, vilket kan påverka tolkningen av selektion (kapitel I och II). Efter att ha tagit hänsyn till gBGC reder vi ut den relativa betydelsen av naturligt urval jämfört med icke-adaptiva processer vid bestämning av den molekylära evolutionshastigheten i halsbandsflugsnappargenomet och identifierade andra faktorer, såsom genuttrycksmönster, som bestämmer selektiv begränsning (kapitel II). Senare visade vi att gBGC är utbrett i fågelgenom och att det döljer sambandet mellan N_e och selektionsstyrkan hos fåglar (kapitel III). Slutligen undersökte vi de faktorer som bestämmer relativa nivåer av genetisk mångfald och divergens mellan Z-kromosomen och autosomer hos fåglar. Vi fann att *male-biased* mutation och effekten av länkad selektion är de mest framträdande faktorerna som driver dessa skillnader (kapitel IV).

Resumen en español

Un entendimiento profundo del proceso evolutivo es crucial para la comprensión de la vida. Estudiar evolución nos permite reconstruir la historia de las especies y describir y predecir patrones que observamos en la naturaleza. También puede ayudarnos a abordar temas más aplicados, como la caracterización y conservación de la biodiversidad o el control y la prevención de enfermedades infecciosas. Por lo tanto, la evolución es un principio unificador de todas las disciplinas de la biología, desde la bioquímica hasta la filogenética.

En los últimos años, el avance de las tecnologías genómicas ha permitido estudiar la evolución a nivel molecular facilitando abordar preguntas y teorías presentadas anteriormente. La disciplina de la evolución molecular tiene dos objetivos principales. El primero es descifrar la historia evolutiva y relaciones filogenéticas de las especies. El segundo se centra en la comprensión de los mecanismos y factores que determinan el proceso evolutivo. Para ello, se basa en la comparación de datos moleculares de varias especies o, alternativamente, se enfoca en estudiar los datos moleculares pertenecientes a distintos individuos de una sola población, utilizando la teoría de genética de poblaciones.

La evolución es el cambio en la composición genética de una población a través del tiempo. La interacción de varios mecanismos y factores genéticos determinan la tasa a la que se produce este cambio y los patrones que genera. La mutación es el principal mecanismo que genera nueva diversidad genética. Otros mecanismos no crean diversidad genética, sino que determinan el destino de las mutaciones al afectar su probabilidad de segregación en la población y, finalmente, su probabilidad de fijación. Por ejemplo, las poblaciones naturales son víctimas de procesos estocásticos y demográficos, lo que afecta la probabilidad de fijación de mutaciones segregantes a través del proceso de deriva genética. Por otro lado, la selección natural influye en la probabilidad de fijación de las nuevas mutaciones de una manera determinista al aumentar o disminuir la probabilidad de fijación de variantes que afectan el "fitness" de los individuos, es decir, su capacidad para sobrevivir y reproducirse. Además, la recombinación también desempeña un papel principal en la evolución. Al redistribuir la variación genética, la recombinación puede aumentar indirectamente la eficacia de la selección natural. Por otro lado, la recombinación puede influir en la probabilidad de fijación de las mutaciones a través del proceso de conversión génica. Esta es la transmisión unidireccional de material genético entre cromosomas homólogos. Este proceso es sesgado al reparar preferentemente con los allelos GC, y su acción no reconoce los efectos en el “fitness”.

Todos estos mecanismos interactúan entre sí. Su fuerza relativa y su impacto en el proceso evolutivo es complejo y varía drásticamente entre las especies y también entre las diferentes regiones del genoma. Comprender el impacto relativo de estos mecanismos y otros factores genéticos para determinar las tasas y los patrones de evolución es un objetivo importante en varios estudios evolutivos. En particular, la identificación de regiones o genes bajo selección naturaly la caracterización de la fuerza del proceso de selección han sido una de las prácticas más populares en los últimos años, ya que proveen información sobre el proceso de adaptación de las especies.

Esta tesis se centra en comprender el papel relativo de estos mecanismos para determinar las tasas y los patrones de evolución en genomas de distintas especies de aves. La primera parte, trata de comprender el impacto de la variación de la tasa de recombinación en la evolución molecular y en las medidas de selección a lo largo del genoma del papamoscas acollarado (*Ficedula albicollis*). Los resultados del estudio, demuestran que el impacto de la conversión génica, que es preferentemente reparada con alelos GC (gBGC por sus siglas en inglés), es determinante en las tasas de evolución, y más aún en la evolución de mutaciones de efecto neutral en comparación con mutaciones seleccionadas, lo que puede sesgar las inferencias de selección (Paper I y II). Después de caracterizar la gBGC, nos centramos en la importancia relativa de la selección natural frente a los procesos no adaptativos para determinar las tasas de evolución molecular en el papamoscas acollarado e identificamos otros factores, como los patrones de expresión génica, que determinan la presión selectiva. Posteriormente, demostramos que la gBGC prevalece en los genomas de las aves y que oculta la relación entre el tamaño efectivo de una población y la fuerza de selección en estos animales (Paper III). Finalmente, investigamos los factores que determinan los niveles relativos de diversidad y divergencia genética entre los cromosomas sexuales y los autosomas en aves. Encontramos que una mayor tasa de mutación en machos y la reducción en la variación genética neutral ligada a sitios bajo selección direccional son los factores más prominentes que determinan estas diferencias (Paper IV).

Acknowledgements

I could not have finished this PhD without the guidance and unconditional support of many people. It is hard to mention everyone and even harder to describe in a couple of pages how each of you have impacted my academic and personal development during these years. So, I will try to personally thank each of you, but I will also do my best to briefly write a few thankful words in this little book.

To my supervisor Hans, thank you for letting me embark in this PhD journey. I first came to your lab as a master student and was soon positively impressed with your passion and attitude towards work. You were always having fun in every meeting and journal club discussion and you always made the time to answer questions and discuss with me. During my PhD, you were always encouraging me to come up with new questions and explanations. I always felt that you trusted me and knew that I would get to the right answers (even when sometimes it took a bit long). You always made me feel that my health and my personal wellbeing were the most important things. Thank you for being supportive.

To my co-supervisor Carina, thank you for always finding the time to discuss with me, answer questions and endlessly comment on my manuscripts. You always pushed me to give an extra effort and to go out of my comfort zone with new analyses. I think this reflects in my scientific output and I am really grateful for that. I am glad that I could count on you during these years.

I want to truly thank all my co-authors. You made this thesis possible through your hard work and amazing teamwork. Alex Nater, you were always available for questions and revisions and that was key for getting this work published. Ludo, thank you for always being so positive and happy to collaborate, working with you was always easy. Also, I thank you for being a thoughtful friend who supported me in the difficult days. Mi, thank you for always make time to teach me statistics while enjoying a nice cup of tea. Homa, thank you for being such a hard worker and for letting me crash at your place when I did not have a home. Matteo, it was very cool working with you, thank you for always putting an extra effort. I want to specially thank Laurent Guéguen and Laurent Duret. Thank you for collaborating with us and being always very positive. Also, thank you for your patience while I was absent. I admire you and your work very much.

I want to specially thank Linnea for always helping me trouble shoot and further my coding skills. You are not only an amazing colleague, but also a true and kind friend. You were sincerely supportive when I needed it the most and I will always be grateful for that. To Vera, Taki, Krysia, Alex S, Verena, Reto, Sev, Robert, Rory, Agnes, Marty, Toby and all other current and past members of the Ellegren lab. Your work always set a really high standard and some of it was the base for my PhD thesis. Your combined knowledge and personalities made our seminars and lab discussions inspiring and fun, always raising new questions. I wish you all great success.

For a truly amazing level of scientific discussion, I also have to thank the members of the Wolf and Suh labs with whom I shared lab meetings and seminars for a long time. I also learnt a lot from you. To all members of the department of evolutionary biology and people form EBC for providing an encouraging and very special scientific environment. I hope this never changes and you all keep this positive and inspiring atmosphere.

I specially want to thank Martin Lascoux, Sylvain Glémin, Douglas Scofield, Simon Whelan, Niclas Backström, Mattias Jakobsson, Jochen Wolf, Anna Rosling and other PIs in the department of ecology and genetics. I admire you all so much. Thank you for having the time to discuss with me about my project, giving me advice, answering questions and organizing courses for PhD students.

Thank you to Frida, Annette and Linn for your help and support with administrative issues.

I want to thank the organizers of the Erasmus Mundus Master in Evolutionary Biology (MEME) program as if it was not for the experience, I would have never been able to do this PhD. I want to thank also the professors who let me into their labs as part of that programme, John Parsh, Daniel Hartl and Hans Ellegren. Similarly, I would like to thank Ricardo Hernandez, Horacio Bach, Nestor Martinez and all other professors that supported and guided me during my bachelor degree, encouraging me to further my studies abroad.

To my office mates and dear friends Hector, TJ, Marcin, Torsten and Merce, thank you for making my day to day so fun. To all my closest friends at EBC; Caro, Sarai, Fede, Lore, Sergio, Foteini, Willian, Ana Cristina, Pili, Maria, Dyma, Venkat and all others. Thank you for everything that we shared. To my friends outside of EBC, Sarita, Angelica, Sonja, Emilia, Pau-Pau thank you for supporting me in different moments during this long endeavour. To my dear friends from Mexico, Elisa, Mosi, Eli, Carli, Adriana, Mariana, Marina, Ata, Enrique, Karla, Adina and everyone else (sorry I can't mention you all here!). I want to thank you for always being there for me. I know you will always be there no matter what.

I want to thank all my family for your support and encouragement over the years. I miss you all every day, but I know the long distance and time will never come between us. I want to specially thank Mayo, my favourite artist, for drawing this amazing cover. I also want to thank my second family, Leal Ramirez, your support and help during these years was invaluable.

I mostly want to thank my parents. I dedicate this thesis to you because you have always encouraged me to do what I enjoy, even if it means I am far away from home. Thank you for all the sacrifices you made that allowed me to be here today. Dad, you are the most passionate person I know, thank you for showing me how much you enjoy your profession. Mom, you are the bravest and most loving person I know. You have given me the courage and strength to search for new opportunities and never give up. Whenever I fail or succeed you are always there to cheer for me. No matter how far away I am, you are always most close to me.

Finally, I want to thank my two favourite people in the whole world. Alberto, thank you for your love, encouragement and patience during all these PhD years. Nothing would be the same without you. I am so grateful to have you in my life. To my daughter Ania, you are the best thing that happened during this PhD and in all my life. You are my biggest motivation and my greatest joy.

References

Arbeithuber, B., Betancourt, A. J., Ebner, T., & Tiemann-Boege, I. (2015). Crossovers are associated with mutation and biased gene conversion at recombination hotspots. *Proceedings of the National Academy of Sciences of the United States of America, 112*, 2109–2114.

Axelsson, E., Webster, M. T., Smith, N. G. C., Burt, D. W., & Ellegren, H. (2005). Comparison of the chicken and turkey genomes reveals a higher rate of nucleotide divergence on microchromosomes than macrochromosomes. *Genome Research, 15*, 120–125.

Bachtrog, D. (2008). Evidence for Male-Driven Evolution in Drosophila. *Molecular Biology and Evolution, 25*, 617–619.

Backström, N., Zhang, Q., & Edwards, S. V. (2013). Evidence from a house finch (*Haemorhous mexicanus*) spleen transcriptome for adaptive evolution and biased gene conversion in passerine birds. *Molecular Biology and Evolution, 30*, 1046–1050.

Berglund, J., Pollard, K. S., & Webster, M. T. (2009). Hotspots of biased nucleotide substitutions in human genes. *Plos Biology, 7*, 45-62.

Betancourt, A. J., & Presgraves, D. C. (2002). Linkage limits the power of natural selection in *Drosophila*. *Proceedings of the National Academy of Sciences of the United States of America, 99*, 13616–13620.

Birdsell, J. A. (2002). Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. *Molecular Biology and Evolution, 19*, 1181–1197.

Burri, R., Nater, A., Kawakami, T., Mugal, C. F., Olason, P. I., Smeds, L*., et al.* (2015). Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of *Ficedula* flycatchers. *Genome Research, 25*, 1656–1665.

Caballero, A. (1995). On the effective size of populations with separate sexes, with particular reference to sex-linked genes. *Genetics, 139*, 1007–1011.

Capra, J. A., Hubisz, M. J., Kostka, D., Pollard, K. S., & Siepel, A. (2013). A model-based analysis of GC-biased gene conversion in the human and chimpanzee genomes. *Plos Genetics, 9*, e1003684.

Charlesworth, B. (2009). Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics, 10*, 195–205.

Cooper, D. N., & Gerber-Huber, S. (1985). DNA methylation and CpG suppression. *Cell Differentiation, 17*, 199–205.

Counterman, B. A., Ortíz-Barrientos, D., & Noor, M. A. F. (2004). Using comparative genomic data to test for fast-X evolution. *Evolution, 58*, 656-660.

Duret, L., & Arndt, P. F. (2008). The impact of recombination on nucleotide substitutions in the human genome. *Plos Genetics, 4*, e1000071.

Duret, L., & Galtier, N. (2009). Biased gene conversion and the evolution of mammalian genomic landscapes. *Annual Review of Genomics and Human Genetics, 10*, 285–311.

Ellegren, H. (2007). Characteristics, causes and evolutionary consequences of male-biased mutation. *Proceedings of the Royal Society B-Biological Sciences, 274*, 1–10.

Ellegren, H. (2013). The evolutionary genomics of birds. *Annual Review of Ecology, Evolution, and Systematics, 44*, 239–259.

Ellegren, H., & Parsch, J. (2007). The evolution of sex-biased genes and sex-biased gene expression. *Nature Reviews Genetics, 8*, 689–698.

Ellegren, H., Smeds, L., Burri, R., Olason, P. I., Backström, N., Kawakami, T.*, et al.* (2012). The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature, 491*, 756–760.

Eyre-Walker, A., & Keightley, P. D. (2007). The distribution of fitness effects of new mutations. *Nature Reviews Genetics, 8*, 610–618.

Eyre-Walker, A., & Keightley, P. D. (2009). Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Molecular Biology and Evolution, 26*, 2097–2108.

Felsenstein, J. (1974). The evolutionary advantage of recombination. *Genetics, 78*, 737–756.

Feschotte, C., & Pritham, E. J. (2007). DNA transposons and the evolution of eukaryotic genomes. *Annual Review of Genetics, 41*, 331–368.

Figuet, E., Ballenghien, M., Romiguier, J., & Galtier, N. (2015). Biased gene conversion and GC-content evolution in the coding sequences of reptiles and vertebrates. *Genome Biology and Evolution, 7*, 240–250.

Figuet, E., Nabholz, B., Bonneau, M., Carrio, E. M., Nadachowska-Brzyska, K., Ellegren, H.*, et al.* (2016). Life history traits, protein evolution, and the nearly neutral theory in amniotes. *Molecular Biology and Evolution, 33*, 1517–1527.

Galtier, N., & Duret, L. (2007). Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends in Genetics, 23*, 274–277.

Galtier, N., Duret, L., Glemin, S., & Ranwez, V. (2009). GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. *Trends in Genetics, 25*, 287–287.

Galtier, N., Piganeau, G., Mouchiroud, D., & Duret, L. (2001). GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics, 159*, 907–911.

Galtier, N., Roux, C., Rousselle, M., Romiguier, J., Figuet, E., Glémin, S.*, et al.* (2018). Codon usage bias in animals: disentangling the effects of natural selection, effective population size, and GC-biased gene conversion. *Molecular Biology and Evolution, 35*, 1092-1103.

Gillespie, J. H. (1974). Nautural selection for within-generation variance in offspring number. *Genetics, 76*, 601–606.

Glémin, S., Arndt, P. F., Messer, P. W., Petrov, D., Galtier, N., & Duret, L. (2015). Quantification of GC-biased gene conversion in the human genome. *Genome Research, 25*, 1215–1228.

Glémin, S., Clément, Y., David, J., & Ressayre, A. (2014). GC content evolution in coding regions of angiosperm genomes: A unifying hypothesis. *Trends in Genetics, 30*, 263–270.

Goldman, N., & Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution, 11*, 725–736.

Gregory, T. R. T. (2005). The animal genome size database. Retrieved November 10, 2018, from http://www.genomesize.com

Guéguen, L., & Duret, L. (2017). Unbiased estimate of synonymous and non-synonymous substitution rates with non-stationary base composition. *Molecular Biology and Evolution, 35*, 734–742.

Gustafsson, L., Nordling, D., Andersson, M. S., Sheldon, B. C., & Qvarnstrom, A. (1994). Infectious diseases, reproductive effort and the cost of reproduction in birds. *Philosophical transactions of the Royal Society of London B-Biological Sciences, 346*, 323–331.

Halldorsson, B. V., Palsson, G., Stefansson, O. A., Jonsson, H., Hardarson, M. T., Eggertsson, H. P., *et al.* (2019). Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science, 363*, eaau1043.

Hammer, M. F., Mendez, F. L., Cox, M. P., Woerner, A. E., & Wall, J. D. (2008). Sex-biased evolutionary forces shape genomic patterns of human diversity. *Plos Genetics, 4*, e1000202.

Hammer, M. F., Woerner, A. E., Mendez, F. L., Watkins, J. C., Cox, M. P., & Wall, J. D. (2010). The ratio of human X chromosome to autosome diversity is positively correlated with genetic distance from genes. *Nature Genetics, 42*, 830–831.

Hellmann, I., Ebersberger, I., Ptak, S. E., Pääbo, S., & Przeworski, M. (2003). A neutral explanation for the correlation of diversity with recombination rates in humans. *American Journal of Human Genetics, 72*, 1527–1535.

Hill, W. G., & Robertson, A. (1966). The effect of linkage on limits to artificial selection. *Genetics Research, 8*, 269–294.

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

Kawakami, T., Smeds, L., Backström, N., Husby, A., Qvarnstrom, A., Mugal, C. F., *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, 23*, 4035–4058.

Keightley, P. D., & Eyre-Walker, A. (2007). Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics, 177*, 2251–2261.

Kidwell, M. G. (2002). Transposable elements and the evolution of genome size in eukaryotes. *Genetica, 115*, 49–63.

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature, 217*, 624–626.

Kimura, M. (1983). *The neutral theory of molecular evolution*. Cambridge: Cambridge University Press.

Kotlar, D., & Lavner, Y. (2006). The action of selection on codon bias in the human genome is related to frequency, complexity, and chronology of amino acids. *BMC Genomics, 7*, 67.

Krylov, D. M., Wolf, Y. I., Rogozin, I. B., & Koonin, E. V. (2003). Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Research, 13*, 2229–2235.

Lang, G. I., & Murray, A. W. (2011). Mutation rates across budding yeast chromosome VI are correlated with replication timing. *Genome Biology and Evolution, 3*, 799–811.

Lartillot, N. (2012). Phylogenetic patterns of GC-biased gene conversion in placental mammals and the evolutionary dynamics of recombination landscapes. *Molecular Biology and Evolution, 30*, 489–502.

Lassalle, F., Périan, S., Bataillon, T., Nesme, X., Duret, L., & Daubin, V. (2015). GC-Content evolution in bacterial genomes: the biased gene conversion hypothesis expands. *Plos Genetics, 11*, e1004941.

Lercher, M. J., & Hurst, L. D. (2002). Human SNP variability and mutation rate are higher in regions of high recombination. *Trends in Genetics, 18*, 337–340.

Lesecque, Y., Mouchiroud, D., & Duret, L. (2013). GC-biased gene conversion in yeast is specifically associated with crossovers: molecular mechanisms and evolutionary significance. *Molecular Biology and Evolution, 30*, 1409–1419.

Lewontin, R. C., & Hubby, J. L. (1966). A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila Pseudoobscura*. *Genetics, 54*, 595–609.

Li, W. H. (1997). *Molecular Evolution*: Sinauer Press.

Mank, J. E., Axelsson, E., & Ellegren, H. (2007). Fast-X on the Z: rapid evolution of sex-linked genes in birds. *Genome Research, 17*, 618–624.

Marais, G. (2003). Biased gene conversion: implications for genome and sex evolution. *Trends in Genetics, 19*, 330–338.

McDonald, J. H., & Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature, 351*, 652–654

McVean, G. A., & Charlesworth, B. (2000). The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. *Genetics, 155*, 929–944.

Mugal, C. F., Arndt, P. F., & Ellegren, H. (2013). Twisted signatures of GC-biased gene conversion embedded in an evolutionary stable karyotype. *Molecular Biology and Evolution, 30*, 1700–1712.

Mugal, C. F., Weber, C. C., & Ellegren, H. (2015). GC-biased gene conversion links the recombination landscape and demography to genomic base composition: GC-biased gene conversion drives genomic base composition across a wide range of species. *Bioessays, 37*, 1317–1326.

Munch, K., Mailund, T., Dutheil, J. Y., & Schierup, M. H. (2014). A fine-scale recombination map of the human-chimpanzee ancestor reveals faster change in humans than in chimpanzees and a strong impact of GC-biased gene conversion. *Genome Research, 24*, 467–474.

Muse, S. V., & Gaut, B. S. (1994). A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution, 11*, 715–724.

Nabholz, B., Uwimana, N., & Lartillot, N. (2013). Reconstructing the phylogenetic history of long-term effective population size and life-history traits using patterns of amino acid replacement in mitochondrial genomes of mammals and birds. *Genome Biology and Evolution, 5*, 1273–1290.

Nagylaki, T. (1983). Evolution of a finite population under gene conversion. *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences, 80*, 6278–6281.

Nei, M., & Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America, 76*, 5269–5273.

Nielsen, R. (2005). Molecular signatures of natural selection. *Annual Review of Genetics, 39*, 197–218.

Ohta, T. (1973). Slightly deleterious mutant substitutions in evolution. *Nature, 246*, 96–98.

Ohta, T. (1974). Mutational pressure as the main cause of molecular evolution and polymorphism. *Nature, 252*, 351–354.

Ohta, T. (1976). Role of very slightly deleterious mutations in molecular evolution and polymorphism. *Theoretical Population Biology, 10*, 254–275.

Ohta, T. (1992). The nearly neutral theory of molecular evolution. *Annual Review of Ecology and Systematics, 23*, 263–286.

Pal, C., Papp, B., & Hurst, L. D. (2001). Highly expressed genes in yeast evolve slowly. *Genetics, 158*, 927–931.

Pal, C., Papp, B., & Lercher, M. J. (2006). An integrated view of protein evolution. *Nature Reviews Genetics, 75*, 337–348.

Pessia, E., Popa, A., Mousset, S., Rezvoy, C., Duret, L., & Marais, G. A. (2012). Evidence for widespread GC-biased gene conversion in eukaryotes. *Genome Biology and Evolution, 4*, 675–682.

Pool, J. E., & Nielsen, R. (2007). Population size changes reshape genomic patterns of diversity. *Evolution, 61*, 3001–3006.

Popadin, K., Polishchuk, L. V., Mamirova, L., Knorre, D., & Gunbin, K. (2007). Accumulation of slightly deleterious mutations in mitochondrial protein-coding genes of large versus small mammals. *Proceedings of the National Academy of Sciences of the United States of America, 104*, 13390–13395.

Pratto, F., Brick, K., Khil, P., Smagulova, F., Petukhova, G. V., & Camerini-Otero, R. D. (2014). Recombination initiation maps of individual human genomes. *Science, 346*, 1256442.

Prendergast, J. G., Campbell, H., Gilbert, N., Dunlop, M. G., Bickmore, W. A., & Semple, C. A. (2007). Chromatin structure and evolution in the human genome. *BMC Evolutionary Biology, 7*, 72.

Qvarnström, A., Ålund, M., McFarlane, S. E., & Sirkiä, P. M. (2016). Climate adaptation and speciation: particular focus on reproductive barriers in *Ficedula* flycatchers. *Evolutionary Applications, 9*, 119–134.

Qvarnström, A., Rice, A. M., & Ellegren, H. (2010). Speciation in *Ficedula* flycatchers. *Philosophical transactions of the Royal Society of London B-Biological Sciences , 365*, 1841–1852.

Ratnakumar, A., Mousset, S., Glemin, S., Berglund, J., Galtier, N., Duret, L., *et al.* (2010). Detecting positive selection within genomes: the problem of biased gene conversion. *Philosophical Transactions of the Royal Society B-Biological Sciences, 365*, 2571–2580.

Roach, J. C., Glusman, G., Smit, A. F., Huff, C. D., Hubley, R., Shannon, P. T., *et al.* (2010). Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science, 328*, 636–639.

Sæther, A. S., Sætre, G. P., Borge, T., Wiley, C., Svedin, N., Andersson, G., *et al.* (2007). Sex chromosome-linked species recognition and evolution of reproductive isolation in flycatchers. *Science, 318*, 95–97.

Smeds, L., Mugal, C. F., Qvarnstrom, A., & Ellegren, H. (2016a). High-resolution mapping of crossover and non-crossover recombination events by whole-genome re-sequencing of an avian pedigree. *Plos Genetics, 12*, e1006044.

Smeds, L., Qvarnström, A., & Ellegren, H. (2016b). Direct estimate of the rate of germline mutation in a bird. *Genome Research, 26*, 1211–1218.

Smith, N. G., & Eyre-Walker, A. (2002). Adaptive protein evolution in *Drosophila*. *Nature, 415*, 1022–1024.

Stapley, J., Feulner, P. G. D., Johnston, S. E., Santure, A. W., & Smadja, C. M. (2017a). Recombination: the good, the bad and the variable. *Philosophical transactions of the Royal Society of London B-Biological Sciences, 372*, 20170279.

Stapley, J., Feulner, P. G. D., Johnston, S. E., Santure, A. W., & Smadja, C. M. (2017b). Variation in recombination frequency and distribution across eukaryotes: patterns and processes. *Philosophical transactions of the Royal Society of London B-Biological Sciences, 372*, 20160455.

Stoletzki, N., & Eyre-Walker, A. (2011). Estimation of the Neutrality Index. *Molecular Biology and Evolution, 28*, 63–70.

Sung, P., & Klein, H. (2006). Mechanism of homologous recombination: mediators and helicases take on regulatory functions. *Nature Reviews Molecular Cell Biology, 7*, 739–750.

Sved, J., & Bird, A. (1990). The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. *Proceedings of the National Academy of Sciences of the United States of America, 87*, 4692–4696.

Szostak, J. W., Orr-Weaver, T. L., Rothstein, R. J., & Stahl, F. W. (1983). The double-strand-break repair model for recombination. *Cell, 33*, 25–35.

Tataru, P., Mollion, M., Glémin, S., & Bataillon, T. (2017). Inference of distribution of fitness effects and proportion of adaptive substitutions from polymorphism data. *Genetics, 207*, 1103–1119.

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology, 7*, 256–276.

Weber, C. C., Boussau, B., Romiguier, J., Jarvis, E. D., & Ellegren, H. (2014a). Evidence for GC-biased gene conversion as a driver of between-lineage differences in avian base composition. *Genome Biology, 15*, 549.

Weber, C. C., Nabholz, B., Romiguier, J., & Ellegren, H. (2014b). Kr/Kc but not dN/dS correlates positively with body mass in birds, raising implications for inferring lineage-specific selection. *Genome Biology, 15*, 542.

Webster, M. T., & Hurst, L. D. (2012). Direct and indirect consequences of meiotic recombination: implications for genome evolution. *Trends in Genetics, 28*, 101–109.

Williams, A. L., Genovese, G., Dyer, T., Altemose, N., Truax, K., Jun, G.*, et al.* (2015). Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. *Elife, 4*, e04637.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics, 16*, 97–159.

Yang, Z. (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Molecular Biology and Evolution, 15*, 568–573.

Yannai, A., Katz, S., & Hershberg, R. (2018). The codon usage of lowly expressed genes is subject to natural selection. *Genome Biology and Evolution, 10*, 1237–1246.

Youds, J. L., & Boulton, S. J. (2011). The choice in meiosis – defining the factors that influence crossover or non-crossover formation. *Journal of Cell Science, 124*, 501–513.

Zhang, J. Z., & Yang, J. R. (2015). Determinants of the rate of protein sequence evolution. *Nature Reviews Genetics, 167*, 409–420.

Zhou, Z., Dang, Y., Zhou, M., Li, L., Yu, C.-h., Fu, J.*, et al.* (2016). Codon usage is an important determinant of gene expression levels largely through its effects on transcription. *Proceedings of the National Academy of Sciences of the United States of America, 113*, E6117–E6125.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 1800

Editor: The Dean of the Faculty of Science and Technology

Distribution: publications.uu.se
urn:nbn:se:uu:diva-381373

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2019