Nucleotide Substitution Patterns in Vertebrate Genomes

CARINA FARAH MUGAL
Dissertation presented at Uppsala University to be publicly examined in Lindahlsalen, EBC, Norbyvägen 18C, Uppsala, Thursday, June 13, 2013 at 13:00 for the degree of Doctor of Philosophy. The examination will be conducted in English.

Abstract

The rates and patterns at which nucleotide substitutions occur vary significantly across the genome sequence of vertebrates. A prominent example is the difference in the rate of evolution of functional sequences versus nonfunctional (neutrally evolving) sequences, which is explained by the influence of natural selection on functional sequences. However, even within neutrally evolving sequences there is striking variation in the rates and patterns of nucleotide substitutions. Unraveling the underlying processes that induce this variation is necessary to understand the basic principles of variation in neutral substitution profiles, which in turn is crucial for the identification of regions in the genome where natural selection acts. This research question builds the main focus of the present thesis. I have studied the causes and consequences of variation in different patterns of nucleotide substitutions. In particular, I have investigated substitutional strand asymmetries in mammalian genes and could show that they result from the asymmetric nature of DNA replication and transcription. Comparative analysis of substitutional asymmetries then suggested that the organization of DNA replication and the level of transcription are conserved among mammals. Further, I have examined the variation in CpG mutation rate among human genes and could show that beside DNA methylation also GC content plays a decisive role in CpG mutability. In addition, I have studied the signatures of GC-biased gene conversion and its impact on the evolution of the GC isochore structure in chicken. By comparison of the results in chicken to previous results in human I found evidence that karyotype stability is critical for the evolution of GC isochores. Finally, beside the empirical studies, I have performed theoretical investigations of substitution rates in functional sequences. More precisely, I have explored the temporal dynamics of estimates of the ratio of non-synonymous to synonymous substitution rates $dN/dS$ in a phylogenetic-population genetic framework.

Keywords: nucleotide substitutions, mutation rate variation, strand asymmetries, CpG effect, GC-biased gene conversion, codon evolution

Carina Farah Mugal, Uppsala University, Department of Ecology and Genetics, Evolutionary Biology, Norbyväg 18 D, SE-752 36 Uppsala, Sweden.

© Carina Farah Mugal 2013

ISSN 1651-6214
urn:nbn:se:uu:diva-198787 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-198787)
THIS LITTLE BOOKLET I DEDICATE
TO MY FATHER

IQBAL ILLAHI MUGAL
List of papers

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


Reprints were made with permission from the publishers.
Beside the papers presented in this thesis, I am an author of the following papers.


Contents

1 Introduction .................................................................................................................. 9

2 Molecular evolution ...................................................................................................... 11
   2.1 Mutational changes of the genome ................................................................. 11
   2.2 Effects of mutations .......................................................................................... 13
   2.3 The fate of mutations ...................................................................................... 14

3 Nucleotide substitution patterns .................................................................................. 17
   3.1 Variation in neutral substitution patterns ......................................................... 17
      3.1.1 Mutation rate variation ........................................................................... 17
      3.1.2 Substitutional strand asymmetry ............................................................. 18
      3.1.3 DNA methylation and the CpG effect ...................................................... 19
      3.1.4 GC-biased gene conversion ...................................................................... 20
   3.2 Natural selection ................................................................................................. 23

4 Analysis methods .......................................................................................................... 24
   4.1 Comparative genomics ....................................................................................... 24
      4.1.1 Nucleotide substitution models ............................................................... 24
      4.1.2 Codon substitution models ...................................................................... 25
   4.2 Statistical data analysis ...................................................................................... 26
   4.3 Mathematical modeling ..................................................................................... 28

5 Research aims ............................................................................................................... 29
   5.1 General aims ...................................................................................................... 29
   5.2 Specific aims ...................................................................................................... 29

6 Paper summary ............................................................................................................. 30
   6.1 Paper I .............................................................................................................. 30
   6.2 Paper II ............................................................................................................. 31
   6.3 Paper III ............................................................................................................ 33
   6.4 Paper IV ............................................................................................................ 35

7 Prospects for the future ............................................................................................... 37

8 Summary in Swedish ..................................................................................................... 39

9 Acknowledgements ....................................................................................................... 41

References ......................................................................................................................... 45
1. Introduction

Nucleotide substitutions are base pair changes in the nucleotide sequence of a genome that have arisen in one individual of a lineage and have then spread and become fixed in the whole lineage. The rates and patterns at which nucleotide substitutions occur vary on multiple scales. There is significant variation found among lineages as well as across the genome sequence of a single lineage. A prominent example is the difference in the rate of evolution of functional sequences versus nonfunctional (neutrally evolving) sequences of the genome, which is explained by the influence of natural selection on functional sequences. However, even within neutrally evolving sequences there is striking variation in the patterns and rates of nucleotide substitutions [1–3]. Several population genetic factors and molecular processes can have an impact on the nucleotide substitution profile, either by affecting the rate at which changes in the nucleotide sequence occur or by influencing the probability that these changes become fixed in the whole lineage. Some examples are the effective population size, recombination, DNA replication and transcription [4–6]. Further, genomic features such as chromatin structure and DNA composition can play a role [7, 8]. The relevance and relative strength of each of these factors is still debated and has been shown to vary among lineages as well as among different regions of the genome. However, unraveling the importance of these various factors and their interactions is a critical issue in molecular evolution. It is necessary to understand the basic principles of variation in neutral substitution profiles, in order to identify regions in the genome where natural selection takes effect (e.g. genes and regulatory elements). More precisely, we have to adjust the baseline in order to distinguish signatures of natural selection from variation in neutral substitution profiles. Only then will we be able to correctly identify genomic regions evolving under selection. Thereupon, the identification of, for example, positively selected genes can help in understanding the processes involved in adaptive evolution and speciation. In addition, understanding how regulatory elements evolve is essential in understanding the complexities of gene expression regulation.

One way to study variation in nucleotide substitution patterns is by comparative genomic approaches, which rely on comparison of genomic sequences of divergent lineages. In order to perform comparative genomic studies we need 1) genomic sequence data and 2) methods to estimate substitution rates. The era of comparative genomics started to grow steadily with the release of more and more whole genome sequences. The first vertebrate genome to be sequenced and released was the human genome in 2001 [9]. Soon afterwards several other vertebrate genomes, like the mouse, rat, chicken and
chimp genomes were published [10–13]. Then, by the introduction of faster and cheaper sequencing technologies the availability of sequence data started to increase rapidly. At the same time as genome sequences were released, the development and improvement of sophisticated computational tools for sequence alignment and substitution rate estimation began to evolve.

The main focus of this thesis has been the study of variation in different patterns of nucleotide substitutions in vertebrates. During my studies, I have investigated the causes of variation in neutral substitution patterns in relation to molecular processes and genomic features, and have discussed their consequences on genome sequence composition. In addition, I have elaborated on the theory of estimation procedures of substitution rates in protein coding sequences within comparative genomic studies. More precisely, I have explored the temporal dynamics of a prominent measure of natural selection, the ratio of non-synonymous to synonymous substitution rates $dN/dS$. In order to perform these investigations I have applied comparative genomic tools, performed statistical data analysis and used mathematical modeling approaches.

For a better understanding of the actual study results, in the introductory part of this thesis I will first provide some underlying information on the molecular basis of nucleotide substitutions and their variation. I will further stress some population genetic aspects of nucleotide substitutions and discuss key relationships between substitution rates and population genetic parameters. Finally, I will describe how nucleotide substitutions can be estimated and which methods can be applied to study causes and consequences of variation in substitution patterns.
2. Molecular evolution

Mutational changes of the genome are the primary source of genomic variability. Genomic variability and its concomitant phenotypic variability constitute an ultimate requirement for evolution. On the other hand, with respect to a single individual, mutational changes are often harmful to its fitness. Therefore, precise machineries and repair mechanisms have evolved to prevent the occurrence of mutations and keep mutation rates at low levels. However, the cost of precision and repair hinder mutation rates from ultimately reaching zero. Consequently, the genetic material is not perpetuated with perfect fidelity and mutational changes provide the genomic variability needed to drive evolution.

In this chapter, I will first discuss the different types of mutations that occur and some of their main resources. Then, I will briefly review the effects of mutations on the individual and finally consider the fate of mutations in the population.

2.1 Mutational changes of the genome

Mutations are changes in the nucleotide sequence or structure of the genome. They occur on different scales, ranging from changes of a single nucleotide, to scales of kilobases as well as megabases. Some changes even involve whole chromosomes or genomes. Changes of a single nucleotide are classified as point mutations. Among these we can distinguish between nucleotide substitutions, where one of the four nucleotides adenine (A), cytosine (C), guanine (G) and thymine (T), is substituted by one of the other nucleotides, as illustrated in Figure 2.1, and insertions or deletions of a single nucleotide, together referred to as indels. However, indels are not restricted to a single nucleotide and can cause more drastic changes in the DNA, such as insertions or deletions of a few nucleotides or even longer stretches of nucleotide sequences, i.e. chromosome segments. Other drastic changes are chromosomal rearrangements such as translocations, where a chromosome segment is excised from one place and inserted into another place of the genome, and duplications, where a chromosome segment is copied into another place of the genome. If a chromosome undergoes breakage and is rearranged within itself such that a chromosome segment is reversed end to end, this rearrangement is referred to as an inversion. At the whole chromosome level, a chromosome can break and form two smaller chromosomes (chromosome fission), or two chromosomes can fuse and form one larger chromosome (chromosome fusion).
Figure 2.1. Single nucleotide substitutions are substitutions of one of the four nucleotides by any of the other three nucleotides. In total 12 substitutions are possible. Changes, where a purine (A or G) is replaced by another purine or a pyrimidine (C or T) by another pyrimidine, are referred to as transitions (blue arrows). Changes from a purine into a pyrimidine or vice versa are referred to as transversions (red arrows). Redrawn from http://en.wikipedia.org/wiki/Transition_(genetics).

the genome level, which represents the largest scale of a mutational change, the genome can be duplicated, referred to as whole genome duplication.

Large-scale mutational changes might be caused by nondisjunction of chromosome pairs during meiosis, by aberrant actions of the recombination process or by transposable element activity. Another important resource of mutations is inaccuracy in DNA replication. DNA replication inaccuracies mainly induce small-scale mutations, as for example incorporation of a wrong nucleotide in the newly synthesized DNA strand or slippage of the replication machinery resulting in short indels. Further, DNA undergoes spontaneous damage from the action of water in form of hydrolysis. The most frequent
hydrolytic damage is deamination. Deamination of C can induce a C → T substitution and deamination of A an A → G substitution. Hydrolysis can further lead to depurination, which produces an abasic site and often results in a G → T or A → T substitution. In addition, DNA can be damaged by oxidative radicals of metabolic origin or from some exogenous source and by radiation.

2.2 Effects of mutations

A mutation arises in a single cell of an individual. If the mutation arises in a somatic cell, it will affect the fitness of the respective cell and its descendant cells. If a mutation arises in an early developmental stage, a single mutation might affect a whole tissue or general developmental process and consequently the fitness of the respective individual. However, a somatic mutation cannot be transmitted to the offspring. In other words, it is not heritable and does not contribute to evolution. Only mutations that occur in germline cells that give rise to gametes are passed on to the offspring and contribute to evolution. I will therefore in the following only consider germline mutations and the effects of a mutation on the offspring carrying the mutant allele.

Mutational changes of the genome may alter the phenotype and thereby the fitness of the individual carrying the mutant allele. With respect to fitness, mutations can be classified into neutral, advantageous or deleterious changes. Neutral changes do not alter fitness. The fitness effect of advantageous and deleterious changes can vary from a mild change in fitness to as drastic fitness effects as being lethal. Large-scale mutational changes might induce difficulties in chromosome pairing during meiosis and thereby cause aneuploidy, and are as a consequence often strongly deleterious. The effects of these changes are not discussed any further. Instead we here focus on the effects of point mutations.

Point mutations that occur in nonfunctional regions of the genome are usually assumed to be neutral with respect to fitness. If point mutations occur in functional regions, such as regulatory elements or protein coding sequences, they might show a fitness effect and their fate in the population might be influenced by natural selection. Let us first consider changes in regulatory elements. Here, a point mutation might lead to reduction or improvement of function, or even loss or gain of function. Among mutations in protein coding sequences we can distinguish between frameshift mutations, missense mutations and synonymous versus non-synonymous changes. Frameshift mutations are induced by indels that are not divisible by 3 and lead to a change in the reading frame. These mutations generally cause non-functional proteins and are therefore often strongly deleterious. Missense mutations as well as synonymous and non-synonymous changes are caused by substitutions of single nucleotides. A change in the nucleotide sequence that leads to a translational stop codon is called missense mutation and is often considerably deleterious.
Figure 2.2. Generation 0 consists of $N$ diploid individuals with an initial frequency of the mutant allele represented by $x_0$. The $N$ individuals produce an infinite number of gametes, referred to as the gamete pool. Note that the frequency of the mutant allele in the gamete pool is equal to its frequency in the parental generation. From the initial gamete pool $2N$ gametes are sampled at random to form generation 1 of $N$ diploid individuals, who again produce an infinite number of gametes. This process is then repeated to form generation 2 and so forth.

The differentiation between synonymous and non-synonymous changes arises from the fact that the genetic code is degenerate. A change in the nucleotide sequence that does not lead to a change in the amino acid sequence is called synonymous. Synonymous mutations are often neutral with respect to fitness or show mild fitness effects due to, for example, preferential codon usage. Non-synonymous mutations change the amino acid sequence and might thereby reduce or improve the respective protein function and can show a wide range of fitness effects.

2.3 The fate of mutations

As discussed in the previous section, a mutational change of the genome might affect the fitness of the individual. But what is the fate of a mutation in the population? I will show in the following that the ultimate fate of the mutant allele is either loss or fixation in the population. Further, I will establish that its fate depends on the effects of the mutation on the individual’s fitness.

Let us assume that a population consists of $N$ sexually reproducing diploid individuals, as typical among vertebrates. Population size $N$ is assumed to be very large (in theory $N \rightarrow \infty$), and any mutant allele is not affected by further mutation, commonly referred to as infinite sites assumption. Let us further assume that mating occurs at random and that generations are non-overlapping. The $N$ diploid individuals produce an infinite number of gametes from which $2N$ gametes are sampled at random to form the next generation of $N$ diploid individuals, as illustrated in Figure 2.2. This model is standard within population genetics and is referred to as the Wright-Fisher model of reproduction [14]. Note that under random mating the sampling of one diploid individual is equivalent to sampling two haploids. The outcome of this sampling follows a binomial distribution, and its fluctuations from generation to generation are known as random genetic drift. If the mutant allele has a frequency of $x$, then
the probability that \( j \) individuals carry the mutant allele in the next generation is given by

\[
\text{Pr}(X = j) = \binom{2N}{j} x^j (1-x)^{2N-j}.
\]

Briefly, the fate of the mutant allele from generation to generation follows the Markov chain

\[
T_{ij} = \binom{2N}{j} \cdot \left( \frac{i}{2N} \right)^j \cdot \left( \frac{2N-i}{2N} \right)^{2N-j}.
\]

As \( j \) may take any value in \( \{0, 1, \ldots, 2N\} \), \( 2N + 1 \) different outcomes of this sampling are possible. Successive matrix multiplication will ultimately lead to \( j \in \{0, 2N\} \). These are the two absorbing states of this Markov chain and indicate loss or fixation of the mutant allele in the population due to random genetic drift.

In a continuous time approximation the Markov chain converges to a diffusion process as \( N \to \infty \) [15, 16]. Then, the allele frequency distribution is a continuous function \( \phi(y, x; t) \), where \( x \in (0, 1) \) denotes the allele frequency at time \( t \), and \( y \in (0, 1) \) denotes the initial allele frequency. The diffusion process can be described by a partial differential equation known as the Kolmogorov forward equation,

\[
\frac{\partial \phi(y, x; t)}{\partial t} = \frac{1}{2} \frac{\partial^2 [V(x)\phi(y, x; t)]}{\partial^2 x}.
\]

\( V(x) \) represents the probability that \( x \) changes because of random genetic drift and is described by the variance in allele frequency after one generation of sampling,

\[
V(x) = \frac{x \cdot (1-x)}{2N}.
\]

In order to allow changes in \( x \) due to some systematic force, such as natural selection, I introduce the probability \( M(x) \) that \( x \) changes because of some systematic force. Extension of Equation 2.3 by this systematic force leads to

\[
\frac{\partial \phi(y, x; t)}{\partial t} = -\frac{\partial [M(x)\phi(y, x; t)]}{\partial x} + \frac{1}{2} \frac{\partial^2 [V(x)\phi(y, x; t)]}{\partial^2 x}.
\]

Let us now specifically consider natural selection and let the mutant allele be sampled with a relative selective weight of \( 1 + s \), where \( s \) denotes the selection coefficient. Let the fitness effects be additive so that the equivalence of sampling of one diploid individual and two haploid individuals is still warranted. Then \( M(x) \) is given by

\[
M(x) = s \cdot x \cdot (1-x).
\]

Equation 2.5 has been formally solved by Motoo Kimura [15]. Based on this equation he then derived an expression for the probability of fixation of a
mutant allele with initial frequency $y$ [17].

$$p(y) = \frac{1 - e^{4Ny}}{1 - e^{4Ns}}. \quad (2.7)$$

The initial frequency of any newly arisen mutant allele $y = 1/2N$. Thus, Equation 2.7 can be simplified to

$$p_0 = \frac{1 - e^{2s}}{1 - e^{4Ns}}. \quad (2.8)$$

Now, let new mutations arise with a rate of $\mu$ per generation. Then the rate of mutations becoming fixed per generation is

$$u = 2N \cdot \mu \cdot p_0 = 2N \cdot \mu \cdot \frac{1 - e^{2s}}{1 - e^{4Ns}}. \quad (2.9)$$

In case of $s = 0$ we receive $p_0 = 1/2N$ and $u = \mu$. This means that in the absence of natural selection or any other systematic force the rate of mutations becoming fixed per generation is equal to the rate at which mutations arise.
3. Nucleotide substitution patterns

I will refer to the rate of mutations becoming fixed per generation as substitution rate \( u \), and I will only focus on the rates and patterns of single nucleotide substitutions. Equation 2.9 shows that substitution rate depends on mutation rate \( \mu \), population size \( N \) and the probability of fixation \( p_0 \). Thus \( u \) can vary due to variation in any of these parameters. In the absence of natural selection or any other systematic force that changes the allele frequency \( u = \mu \) and variation in substitution rate is equal to variation in mutation rate. For simplicity, I will start this chapter by discussing the absence of any systematic force. In subsection 3.1.4, I will then show a case of a systematic force other than natural selection that affects the probability of fixation of different kinds of nucleotide substitutions. In the next section of this chapter, section 3.2, I will briefly discuss natural selection.

3.1 Variation in neutral substitution patterns

3.1.1 Mutation rate variation

Figure 2.1 illustrates that 12 different nucleotide substitutions are possible. Any of the four nucleotides can be replaced by any of the other three nucleotides. Mutation rate per site is the sum of these 12 individual mutation rates weighted by the respective probabilities of the nucleotides at a certain site,

\[
\mu = \sum_{i=1}^{4} \sum_{j=1}^{4} p_i \cdot q_{ij},
\]

(3.1)

where \( i, j \in \{A, C, G, T\} \) represent indices over the four different nucleotides. \( p_i \) denotes the probability of nucleotide \( i \), and \( q_{ij} \) denotes the mutation rate from nucleotide \( i \) to nucleotide \( j \). Note the implicit assumption that nucleotide sites evolve independently of each other.

If we consider the average mutation rate of a particular region of the genome, equation 3.1 shows that mutation rate depends on the nucleotide frequencies \( p_i \), in other words base composition, of the genomic region and on the individual mutation rates \( q_{ij} \). Variation in mutation rate can hence be induced by variation in base composition and variation in individual mutation rates. Another source of variation in mutation rate across the genome is variation in the relative contribution of different mutation resources (as introduced in section 2.1) to the total number of mutations.
The majority of mutations are due to replication errors [18, 19], and it has been shown that replication timing during S-phase plays an important role in replication accuracy [6, 20]. Early replicating regions of the genome are less error prone than are late replicating regions. Another consequence of replication errors is a male-bias in mutation rate [21, 22]. The number of cell divisions and as a consequence the frequency of DNA replication is higher in the male germline than in the female germline. Thus, assuming that the mutation rate per cell generation is the same in males and females, mutation rates should be higher in males than in females.

Spontaneous DNA damage in form of hydrolysis makes up a second important resource of nucleotide substitutions. Hydrolytic DNA damage mainly occurs in a single-stranded state of the DNA [23], as found during DNA replication and transcription. As a consequence, mutation patterns will be enriched for hydrolytic DNA damage in transcribed regions as well as error prone regions during replication [24–26]. Mutation rates might thus not only depend on replication timing, but also on transcription levels and exon density [2, 3]. Other genomic features that might impact mutation rate are chromatin structure [8], distance to telomeres [2], GC content [2, 3, 7] and recombination [4, 5], as accessibility for mutations as well as DNA repair machineries varies dependent on these features. The relative contribution of each of the above listed molecular processes and genomic features on the rate of mutation and its variation across the genome remains to be revealed.

3.1.2 Substitutional strand asymmetry

Nucleotide substitutions can occur on both strands of the double-stranded DNA. If the rate of one of the 12 individual nucleotide substitutions is the same on both strands, this individual substitution rate is considered strand symmetric. Due to the self-complementary nature of DNA this definition is equivalent to equality of the individual nucleotide substitution rate and its complementary substitution rate on the same DNA strand. If for example C is substituted by T on one strand with rate $q_{CT}$, G is substituted by A with the same rate $q_{CT}$ on the complementary strand. As strand symmetry implies that nucleotide substitutions occur at the same rate on both strands, also on the first strand G is substituted by A with rate $q_{CT}$ and $q_{GA} = q_{CT}$. In case of a strand symmetric substitution pattern this equality of rates applies to all pairs of complementary rates. As a consequence, the 12 individual substitution rates are fully represented by 6 substitution rate parameters. Any disparity in complementary substitution rates is referred to as substitutional strand asymmetry. The exact pattern of asymmetry as well as the degree of disparity have been found to vary across the genome [25, 27].

Two processes, DNA replication and transcription, are thought to induce strand asymmetric substitution rates [25, 28]. In these two processes the
double-stranded DNA is unwound and the two complementary DNA strands are temporarily exposed in a single-stranded state. Asymmetry in the nucleotide substitution pattern might then arise due to asymmetric nature of DNA replication and transcription. In both processes the two complementary DNA strands are not treated equally. During DNA replication the sliding process of the replication machinery is restricted to 3' → 5'. Because of this constriction of the sliding direction and because of the anti-parallel nature of double-stranded DNA only one daughter strand, the leading strand, can be synthesized continuously. The other one, the lagging strand, has to be synthesized in fragments called Okazaki fragments. The synthesis of the lagging strand requires longer single-stranded parts of parental DNA (a previous leading strand) than the synthesis of the leading strand [29], which makes the leading strand more prone to spontaneous DNA damage. During transcription the coding strand is transiently exposed in single-stranded state, while the non-coding strand is partly protected by RNA polymerase and the nascent RNA. This makes the coding strand more vulnerable to DNA damage than the non-coding strand.

As discussed in section 2.1, spontaneous DNA damage predominantly induces C → T, A → G, G → T and A → T substitutions, which leads to an increase of these substitutions on the leading and coding strand compared to the lagging and non-coding strand, respectively. Note, that in case the coding strand of a transcribed region is the leading strand during DNA replication, the action of DNA replication and transcription enforce each other, while they oppose each other when the coding strand is the lagging strand.

3.1.3 DNA methylation and the CpG effect

Vertebrate genomes contain 5-methylcytosine (5mC) as a rare building block of the DNA, where 5mC constitutes a postsynthetic DNA modification of cytosine. Predominantly but not exclusively it occurs 5’ to a guanine, together referred to as CpG dinucleotide. One important role of 5mC is its function in gene expression regulation [30]. Aside from this function, the presence of 5mC in CpG dinucleotides also leaves considerable genomic signatures. Vertebrate genomes are depleted in CpG dinucleotides as a result of an about 10 times higher spontaneous mutation frequency of 5mC compared to C, known as the CpG effect [31, 32]. A combination of two factors is likely to explain the higher mutation frequency of 5mC, which I will briefly discuss in the following.

Both, C and 5mC are susceptible to spontaneous deamination in single-stranded DNA. However, 5mC is more liable to deamination than C, which promotes a higher mutation frequency of 5mC compared to C [33]. Moreover, deamination of C converts C into uracil (U), which does not constitute a building block of DNA. As a consequence, U can be easily detected and removed by uracil glycosylase with the subsequent replacement by the correct base. On the
other hand, deamination of 5mC converts 5mC into T, which is one of the four common building blocks of DNA. The mechanism to detect the deaminated 5mC as a wrong or mutated base is consequently less efficient, which reduces its rate of repair [34]. As such DNA methylation influences mutation rates of 5mC. Taking this one step further and considering that a large proportion of all mutations are CpG mutations, DNA methylation influences mutation rate variation across the genome as genomic regions high in 5mC content (CpG content) show higher total mutation rates than regions low in 5mC.

3.1.4 GC-biased gene conversion
GC-biased gene conversion (gBGC) is a process that is linked to meiotic recombination and implies short-scale, unidirectional exchanges between homologous chromosomes, where AT/GC heterozygotes produce more GC- than AT-gametes. I will first explain the molecular basis of meiotic recombination and the occurrence of gene conversion tracts. Then I will briefly discuss the impact of gBGC on the probability of fixation of the mutant allele and its consequences on base composition evolution. The exact molecular mechanism of gBGC is on the other hand still an open question and a matter of current research [35].

Meiotic recombination is a form of homologous recombination that occurs during meiosis and leads to genetic exchange between homologous chromosomes. Besides providing genetic variation, meiotic recombination constitutes an essential step during meiosis. It allows for the establishment of connections (chiasmata) between homologous chromosomes prior to the first nuclear division, which assists to properly align pairs of homologous chromosomes on the meiotic spindle. The pairing of homologous chromosomes requires at least one recombination event per chromosome pair and promotes proper segregation of chromosomes for the first meiotic division.

The various steps of meiotic recombination are schematically illustrated in Figure 3.1. Meiotic recombination is initiated by the programmed generation of double-stranded DNA breaks (DSBs) by the Spo11 protein [37]. Spo11 cuts the DNA at several locations approximately at the same time as homologous chromosomes start to align on the meiotic spindle. The DSBs are then further processed to generate single-stranded DNA tails with 3’-ends. These tails invade the parental chromosome to pair with its complementary strand in the homologous region of the intact duplex DNA molecule. This base pairing results in regions of new duplex DNA between DNA molecules from the two distinct parental chromosomes, called heteroduplex DNA and may contain some mismatched base pairs. After strand invasion the two DNA molecules become connected and form a cross structure called Holliday-junction (HJ). More precisely, two HJs are formed, which can move along the DNA by melting and formation of base pairs, a process referred to as branch migration.
Figure 3.1. Mechanism of meiotic recombination [36]. Green and blue lines indicate aligned DNA sequences of homologous chromosomes. Meiotic recombination is initiated by the formation of a double-stranded DNA break. Next, limited destruction of DNA strands leads to single-stranded DNA tails with 3’-ends. These tails invade the homologous chromosome and the DNA pairs with its complementary strand in the homologous chromosome. The complementary strand serves as a template for DNA synthesis. Sequence information lost from the green DNA molecule is replaced by sequence information present on the blue DNA molecule. Red dashed lines indicate newly synthesized DNA. Then DNA synthesis is followed by the formation of Holliday-junctions (HJs) and branch migration. Finally, recombination is finished by the resolution of the HJs, which can either result in a crossover or non-crossover product. In either case gene conversion tracts and regions of heteroduplex DNA are formed. Black dotted lines above the double-stranded DNA point to the regions of the DNA, where the gene conversion tracts are found.
Finally, cleavage of the HJs results in two separate duplex DNA molecules. The way the HJs are cut determines the outcome of the recombination event, which can either be a crossover or a non-crossover product. The amount of genetic information exchanged between chromosomes is larger in case of a crossover product. But regardless of the outcome, recombination leads to regions of heteroduplex DNA and unidirectional exchange of a stretch of DNA between chromosomes. In short, gene conversion is a byproduct of meiotic recombination [36].

Gene conversion is biased if some alleles are favored over others. Experimental as well as comparative genomic studies suggest that recombination-associated gene conversion is biased towards GC base pairs over AT base pairs [38–40]. Several hypothesis have been proposed regarding the molecular mechanism of gBGC. A widely accepted hypothesis suggests that the repair of mismatches in the heterodyplex DNA by Base Excision Repair (BER) is biased towards GC [41, 42]. However, the BER hypothesis has been questioned recently by the evidence of an alternative mechanism based on the activity of the mismatch repair machinery (MMR) [35]. Further work is required to assess the exact molecular mechanism of gBGC. But regardless of its mechanism, as a consequence of gBGC AT/GC heterozygotes produce more GC-than AT-gametes, which leads to a fixation bias for GC- alleles over AT-alleles in regions undergoing gBGC. gBGC systematically changes allele frequencies and acts similar to natural selection [43]. As gBGC is linked to meiotic recombination this results in a fixation bias for GC- alleles over AT-alleles in high-recombining regions. However, gBGC only impacts the probability of fixation of certain kinds of substitutions. Let W represent the ‘weak’ nucleotides A and T, and S represent the ‘strong’ nucleotides C and G. Then the probability of fixation of $S \rightarrow W$ substitutions,

$$p_{S\rightarrow W} = \frac{1 - e^{-2cr}}{1 - e^{-4Ncr}} > \frac{1}{2N},$$

(3.2)

and of $W \rightarrow S$ substitutions,

$$p_{W\rightarrow S} = \frac{1 - e^{-2cr}}{1 - e^{-4Ncr}} < \frac{1}{2N},$$

(3.3)

where $r$ denotes recombination rate and $c$ denotes a constant that specifies the strength of gBGC similar to the selection coefficient $s$. W $\rightarrow$ W and S $\rightarrow$ S nucleotide substitutions are not affected by gBGC and their probability of fixation equals $1/2N$. The genomic signature of such a fixation bias for certain nucleotide substitutions over others is an enrichment of GC content in high-recombining regions. Therefore, a positive correlation between GC content and meiotic recombination is expected [40, 44].
3.2 Natural selection

Natural selection may act in various ways. I will here only outline the action of positive and purifying selection and their consequences on substitution rates. The mutations under consideration show a fitness effect and are either advantageous or deleterious for the individual. Let us first consider a mutation that is advantageous for the individual. As a consequence, the individual that carries the mutant allele has an increased relative fitness \( w_1 = 1 + s > 1 \) compared to the wild-type \( w_0 = 1 \). Natural selection will then act as a systematic force in favor of individuals carrying the mutant allele, which will increase the probability of fixation for the mutant allele,

\[
p_0 = \frac{1 - e^{2s}}{1 - e^{4Ns}} > \frac{1}{2N},
\]

and \( u > \mu \). In other words, sites being affected by advantageous mutations will evolve at a faster rate than neutrally evolving sites. Such an increased rate of evolution is considered as a signature of positive selection. In contrast, if the individual that carries the mutant allele has a reduced relative fitness \( w_1 = 1 + s < 1 \) compared to the wild-type \( w_0 = 1 \), natural selection will act in disfavor for individuals carrying the mutant allele. Hence,

\[
p_0 = \frac{1 - e^{2s}}{1 - e^{4Ns}} < \frac{1}{2N},
\]

and \( u < \mu \). As such, sites being affected by deleterious mutations will evolve at a slower rate than neutrally evolving sites, a signature of purifying selection. In summary, positive selection acts to drive advantageous changes to fixation and thereby increases the rate of evolution. It is a mechanism that enables species to adapt to their environment. Purifying selection reduces the rate of evolution and affected sites are conserved over evolutionary time. Such selective constrained can be important to maintain the function of genes or regulatory sequences.
4. Analysis methods

4.1 Comparative genomics

Common to many comparative genomic studies is the aim to estimate distances between homologous nucleotide sequences. Usually, such studies rely on comparison between sequences of divergent lineages. In the simplest case, this can be a sequence comparison between two lineages and the estimation of a pairwise distance. More sophisticated methods also allow for the comparison between several sequences and the estimation of lineage specific distance measures. In any case, the distance $d$ is defined as the expected number of nucleotide substitutions per site that have accumulated between the sequences since their divergence,

$$d = u \cdot t,$$

where $t$ denotes the divergence time. In order to estimate the distance $d$ between homologous sequences, sequences have to be aligned and probabilistic models to describe the nucleotide substitution process are necessary. In this section, I will not discuss the various sequence alignment methods and will instead focus on underlying techniques and models for the estimation of $d$ given that sequences are properly aligned.

4.1.1 Nucleotide substitution models

As introduced in equation 3.1, mutation rate $\mu$ depends on the nucleotide frequencies $p_i$ and on the individual mutation rates $q_{ij}, i, j \in \{A, C, G, T\}$. Along the same lines, substitution rate $u$ depends on $p_i$ and $q_{ij}$, where $q_{ij}$ now represents the individual substitution rates. Given that nucleotide sites evolve independently of each other the nucleotide substitution process can be modeled as a Markov process with the four states $\{A, C, G, T\}$. The substitution rate matrix of the Markov process is given by,

$$Q = \begin{pmatrix}
q_{AC} & q_{AG} & q_{AT} \\
q_{CA} & q_{CG} & q_{CT} \\
q_{GA} & q_{GC} & q_{GT} \\
q_{TA} & q_{TC} & q_{TG}
\end{pmatrix},$$

where the diagonal elements are chosen such that each line of the rate matrix sums to zero. This is an essential characteristic for a Markov process due to
preservation of probability. Now, let the nucleotide frequencies of a sequence at time \( t \) be represented by a unity row vector,
\[
\vec{p}(t) = (p_A(t), p_C(t), p_G(t), p_T(t)).
\] (4.3)

Then, the evolution of the nucleotide frequencies over time is described by a differential equation,
\[
\frac{d}{dt} \vec{p}(t) = \vec{p}(t) \cdot Q.
\] (4.4)

Solving equation (4.4) leads to
\[
\vec{p}(t) = \vec{p}(0) \cdot P(t),
\] (4.5)
where
\[
P(t) = \exp(Q \cdot t)
\] (4.6)
is known as the transition-probability matrix. The matrix elements \( p_{ij}(t) \) describe the probability that nucleotide \( i \) present at time 0 is replaced by nucleotide \( j \) at time \( t \).

Comparison of sequence data allows to count the frequencies of replacements from \( i \) to \( j \), which can be used as an approximation for \( p_{ij}(t) \). We can then directly use equation 4.6, or extensions of equation 4.6, to estimate the individual substitution rates and finally \( d \) by a Maximum-likelihood (ML) approach. The complexity of the model, in other words the amount of parameters that are estimated by the ML approach, depends on the number of sequences that are compared and on the nucleotide substitution model that is implemented. In the simplest case two sequences are compared and all individual substitution rates are assumed to be equal. Then \( d \) is described by a single parameter. A less stringent substitution model leads to more parameters, where in case of an unrestricted substitution model all 12 individual substitution rates are allowed to differ. If several sequences are compared also the tree-topology and the various divergence times have to be taken into account. Furthermore, parameters can be allowed to differ between branches of the tree, so-called branch-specific models. The choice of the model depends on the one hand on the research question but also on the amount of information that can be extracted from the sequence data.

4.1.2 Codon substitution models

Nucleotide substitutions in protein coding regions are often modeled by codon substitution models, where substitutions from one codon to another are considered. Again the substitution process can be modeled as a Markov process, where the state space now is the sense codons in the genetic code [45, 46]. In vertebrates there are 61 sense codons and the substitution rate matrix \( Q \) describes the rate of change between these codons. One widely used model is
based on the model introduced by Goldman and Yang [45],

\[
q_{ij} = \begin{cases} 
0, & \text{if } i \text{ and } j \text{ differ at two or three codon positions}, \\
\pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transversion}, \\
\kappa \pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transition}, \\
\omega \pi_j, & \text{if } i \text{ and } j \text{ differ by a non-synonymous transversion}, \\
\omega \kappa \pi_j, & \text{if } i \text{ and } j \text{ differ by a non-synonymous transition}.
\end{cases}
\]

(4.7)

Here, \(\pi_j\) represents the equilibrium frequency of codon \(j\) and \(\kappa\) represents the transition/transversion rate ratio. The parameter \(\omega\) is introduced in order to distinguish between synonymous and non-synonymous changes. It represents the ratio of non-synonymous to synonymous changes. It represents the ratio of non-synonymous to synonymous changes. It represents the ratio of non-synonymous to synonymous changes.

Estimation of \(dN/dS\) can be performed by an ML approach and is one of the main applications of codon substitution models [47–49]. In short, the \(dN/dS\) ratio quantifies the mode and strength of selection acting at the protein level by comparing synonymous substitution rates (\(dS\)) - assumed to be neutral - to non-synonymous substitution rates (\(dN\)) which are exposed to selection as they change the amino acid composition of a protein. Unity of the ratio is generally taken to indicate neutrality, values exceeding unity are interpreted as selection promoting change (positive selection), and values less than one are usually taken as an indication for selection suppressing protein change (purifying selection).

4.2 Statistical data analysis

Statistical data analysis is implemented in order to make inferences from empirical data. In comparative genomic studies the data are usually sequence data and estimates of various genomic features or life history traits of the studied species. More specifically, in the study of variation in rates and patterns of nucleotide substitutions, estimates of substitution rates and patterns are investigated in relationship to genomic features or life history traits. The ultimate aim is to reveal relationships between rates and patterns of nucleotide substitutions and genomic features or life history traits, and then describe the nature and strength of the relationships in order to make biological inferences. Some of the common analysis methods to investigate relationships between variables are outlined in the following.

A straightforward measure to test for associations between variables is the covariance or correlation measure. As the covariance is dependent on the scale of the variables, computation of the correlation coefficient \(p_{XY}\) between two (random) variables \(X\) and \(Y\), which ranges between 0 and 1, is often preferred.
For a set of variables \( \mathbf{X} = (X_1, \ldots, X_p) \) one can compute the correlation matrix,

\[
\mathbf{C} = \begin{pmatrix}
1 & \rho_{X_1X_2} & \cdots & \rho_{X_1X_p} \\
\rho_{X_2X_1} & 1 & \cdots & \rho_{X_2X_p} \\
\vdots & \vdots & \ddots & \vdots \\
\rho_{X_pX_1} & \cdots & \rho_{X_pX_{p-1}} & 1
\end{pmatrix} \tag{4.8}
\]

As such, the correlation matrix \( \mathbf{C} \) is a simple way to summarize associations of data in high dimensions.

Another technique to examine data in high dimensions is by principal component analysis (PCA). Briefly, in PCA one looks for linear combinations of a \( p \)-variate random variable \( \mathbf{X} = (X_1, \ldots, X_p) \) such that the linear combinations capture the largest variances in \( \mathbf{X} \). These linear combinations, called principal components (PCs), can either directly be used to make inferences about the multivariate structure of \( \mathbf{X} \) or can serve as an intermediate step in a more complex analysis. The computation of the PCs is based on the covariance or correlation matrix. However, as mentioned above the covariance measure is not scale invariant. In a comparative genomic context, where variables often are represented by distinct genomic features or life history traits of different scales, it is therefore advisable to only consider PCA based on the correlation matrix. A particularly favorable characteristic of the PCs is that they are orthogonal to each other. This can be of advantage in downstream analysis as, for example, regression analysis with correlated explanatory variables, which will be discussed in the next paragraph.

Regression analysis is applied if there is biological evidence of causative relationships between variables. The ultimate goal is to assess the amount of variation in one or more response variables that is explained by one or more explanatory variables. Let us first consider the principles of multi-linear regression analysis and some of the underlying assumptions about the explanatory variables, the response variables and their relationships. First of all it is assumed that response variables are represented by a linear combination of the explanatory variables. Further, the errors of the regression model are assumed to be independent of the values of the explanatory variables, the response variables are assumed to be normally distributed and lack of multi-collinearity in the explanatory variables is necessary for standard least-squares estimation procedures. In a comparative genomic context many of these assumptions are often not full-filled and extended regression models or data transformation might be necessary. Data transformation is often applied to ensure approximately linear relationships between explanatory and response variables or to improve the error structure of the regression model. Generalized linear regression is an extension of linear regression, which is applied if response variables follow other than a normal distribution and is especially useful for count or binary data. However, neither data transformation nor generalized linear regression help to deal with multi-collinearity in the explanatory variables, a fre-
quent problem in comparative genomic studies. Many of the genomic features and life history traits used as candidate explanatory variables for substitution rate variation are strongly interrelated with each other. Here, principal component regression (PCR) or partial least square regression (PLSR) constitute alternative approaches. Both PCR and PLSR rely on PC decomposition of the explanatory variables. As the PCs are orthogonal to each other, subsequent linear regression analysis on the PCs allows to identify the number of independent effects on the response variables and helps to unravel the importance of the various explanatory variables.

4.3 Mathematical modeling

I want to provide a flavor of how mathematical modeling can assist to understand phenomenons that occur in nature. I will base the considerations on two quotations, "Essentially, all models are wrong, but some are useful." by George Box, and "Things should be made as simple as possible, but not any simpler." by Albert Einstein. The first quotation by George Box stresses the fact that it is impossible to describe reality in all its details by any mathematical model. Phenomenons in nature do not strictly follow mathematical laws, their structure and behavior is more complex and less certain. The task of a mathematician is to find models that capture the most important aspects of the phenomenons they want to study. Following the advice by Albert Einstein it is important to keep mathematical models as simple as possible and to not incorporate unnecessary aspects. A model that is saturated with parameters might capture many aspects of the phenomenon, but it is impractical for making inferences. On the other hand, it is important not to oversimplify and thereby miss to describe relevant aspects. Thus, a critical step is to identify main patterns and distinguish important aspects from nuisance. Here, one has to use one’s knowledge and experience to judge which aspects and patterns are important for the question of interest.
5. Research aims

5.1 General aims
The aim of my PhD studies was the investigation of variation in rates and patterns of nucleotide substitutions in vertebrates by a comparative genomic approach. The era of vertebrate comparative genomics started with the release of several mammalian genome sequences, and the quality and accessibility of mammalian genome sequences was by far superior to any other class of vertebrates. Therefore, I first focused on substitution patterns in mammalian genomes. However, the availability of genomic sequences has been increasing rapidly, which has opened up a door for refined comparative genomic analyses. Once several avian genomes were successfully sequenced and annotated, I complemented my studies with the investigation of substitution patterns in birds. Avian genomes show some specific characteristics, that distinguish them from mammals. Here, comparison of results in avian genomes with previous findings on mammalian genomes assisted to answer several questions in the field of comparative genomics. Finally, aside from comparative genomic studies applied to genome sequence data a sound understanding and improvement of the theoretical principles of substitution rate estimation is necessary. Therefore, another aim of my PhD studies was the investigation of a prominent measure of the strength of natural selection within comparative genomic studies by a comprehensive mathematical modeling approach.

5.2 Specific aims

I Comparative study of the conservation of nucleotide substitution rate and substitutional strand asymmetries in mammalian genes and elaboration on their causes and consequences. (Paper I)

II Investigation of the impact of DNA methylation and other genomic features on CpG mutation rate variation among human genes. (Paper II)

III Study of the signatures of GC-biased gene conversion and its impact on the evolution of the GC isochore structure in chicken. (Paper III)

IV Theoretical investigation of the temporal dynamics of the ratio of non-synonymous to synonymous substitution rates $dN/dS$ in a phylogentic-population genetic framework. (Paper IV)
6. Paper summary

6.1 Paper I

Conservation of neutral substitution rate and substitutional asymmetries in mammalian genes.

Mutation rate has been shown to vary substantially within mammalian genomes [3, 50]. Spatially this variation ranges from variation among sites, regions as well as chromosomes and has been related to various genomic and structural features. However, these genomic and structural features are often interrelated with each other, which makes it difficult to unravel the true causal variable(s) of mutation rate variation. As knowledge about the causes of mutation rate variation within genomes is of crucial importance to adjust models for molecular evolutionary analysis, several attempts have been made to unravel the true causative features of mutation rate variation. Here, substitution rate at presumably neutrally evolving sites has been used as a proxy of mutation rate [2, 50], and features such as the local GC content, recombination rate, sequence features and epigenetic modifications, have been suggested to determine mutation rate variation. Beside spatial variation in the rate of nucleotide substitutions, also variation in the strength of substitutional strand asymmetries has been observed [25, 28]. Substitutional strand asymmetries describe the tendency of nucleotide substitutions to preferentially occur on one specific strand of the double-stranded DNA and not on its complementary strand. There is evidence that both, DNA replication and transcription, can induce such strand asymmetries [25, 27, 51]. Now, conservation of substitution patterns among lineages, being it variation in nucleotide substitution rate or substitutional asymmetries, is likely to be affected by the degree of conservation of the underlying causative variables. Conservation of substitution patterns might therefore be interpreted as evidence for conservation of the underlying causative variables.

In this study we focused on the conservation of neutral substitution rates and in particular substitutional strand asymmetries between human and mouse. Further, we were interested in the relation between the degree of conservation in strand asymmetries and germline transcription levels. In order to perform these investigations we estimated neutral substitution rates and substitutional strand asymmetries of human and mouse orthologous genes based on transposable elements present in the intronic regions of the gene. Here, the restriction to lineage-specific transposable elements allowed for the estimation of lineage-specific substitution patterns. We performed principal component
analysis (PCA) to assess the degree of conservation of these patterns among human and mouse. Next, we evaluated the relative importance of DNA replication and transcription as well as mutation rate on substitutional asymmetries by model selection criteria applied to a set of linear regression models. Finally, we explored the degree of conservation in strand asymmetries in relation to conservation in germline transcription levels by interdependent correlation analysis.

Comparison of the strength in conservation of neutral substitution rates and substitutional strand asymmetries suggested that strand asymmetries show a stronger conservation over evolutionary time than neutral substitution rates. These results then indicated that substitution rates might be more influenced by short-lived processes, whereas strand asymmetries seem to be determined by more stable processes. Based on the regression analysis we then found that both, replication and transcription show a comparably strong impact on strand asymmetries in genes. Thus, it seems that both processes are significantly conserved over evolutionary time. However, we could only assess the conservation of transcription over time by comparison of human and mouse germline transcription levels. Here, the interdependent correlation analysis revealed that indeed a stronger conservation in transcription level led to a stronger conservation in strand asymmetries between human and mouse.

In conclusion, variation of substitutional asymmetries in mammalian genes is determined by superposing effects of transcription- and replication-induced mutation bias as well as the degree of conservation of these two processes. Hence, if genes are highly expressed over a significant evolutionary time-scale, mutation rates in these genes are likely to be strongly strand asymmetric and will leave asymmetric signatures of base composition on the coding versus the non-coding strand of the gene. This will not only affect the intronic regions of the gene but also other parts, such as 4-fold degenerate sites [52]. Therefore, if not taken into account properly strand asymmetries might affect the accuracy of several molecular evolutionary analysis as e.g. the estimation of codon usage bias.

6.2 Paper II

Substitution rate variation at human CpG sites correlates with non-CpG divergence, methylation level and GC content.

The rate of germline mutation is key to evolution and is also of significant importance for the origin of disease. The mutation rate varies considerably throughout the human genome and although a number of factors have been shown to contribute to this variation, much of it remains unexplained. The far highest rate of mutation is seen at CpG sites, where methylation of cytosine leads to a $C \rightarrow T$ transition upon spontaneous deamination [32, 53]. However,
using novel genome-wide approaches such as methylated DNA immunoprecipitation (MeDIP) and massive parallel bisulphite sequencing it has recently been shown that methylation levels of CpG sites vary across the genome [54–56]. Hence, if there is variation in methylation level among regions of the human genome, then methylation level should be an important factor to explain mutation rate variation at CpG sites.

In this study we used whole-genome DNA methylation data collected from human spermatozoa to study the impact of DNA methylation on CpG mutability by comparison of DNA methylation level to the local rate of CpG mutation in the human lineage. More precisely, we estimated human-specific CpG transition and transversion rate as well as CpH transition rate (where H is either A, C or T) in intronic regions of the genome for which DNA methylation level was available. We further estimated non-CpG divergence, GC content, transcription level and recombination rate, which we beside DNA methylation level used as possible explanatory variables for CpG and CpH mutability. We then performed logit-regression analysis to correctly account for the binomial error distribution of transition and transversion rates.

The regression analysis showed that variation in CpG transition rate was mainly explained by DNA methylation level, non-CpG divergence and GC content, where the relationship with DNA methylation and non-CpG divergence level was positive and negative with GC content. In particular in CpG islands (CGIs), where methylation level showed the strongest variation, it proved out that DNA methylation level strongly impacts CpG transition rate. In non-CGIs methylation level showed a much lower variation, and in these regions non-CpG divergence was the main explanatory variable of CpG transition rate. Variation in CpG transversion rate was mainly explained by non-CpG divergence and GC content. Here, the contrasting effects of methylation level on CpG transition versus transversion rate provided evidence that DNA methylation specifically increases the C → T transition rate. Analysis of variation in CpH transition rate served as a further negative control and showed that DNA methylation level only impacts C → T transition rate at CpG sites and not any other types of dinucleotides. Similar to CpG transversion rate, variation in CpH transition was mainly explained by non-CpG divergence and GC content. Finally, by a comparison of CpG transition rates on the X chromosome versus the autosomes we tested for a male-bias in mutation rate. The analysis showed that CpG transition rate was indeed male-biased. However, CpG-specific transition rates, i.e CpG transition rates after correction for general mutation rate variation, were no longer male-biased.

In summary, we concluded that genome-wide variation in methylation level has a strong impact on CpG mutability, providing a link between epigenetic modification and molecular evolution. However, not only DNA methylation but also other features such as non-CpG divergence and GC content impact CpG mutability. It has previously been suggested that GC content might influence the thermodynamic stability of double-stranded DNA, where genomic
regions with a low GC content are more likely to occur in a single-stranded state and thus be more prone to CpG mutations as regions rich in GC content [7]. This is well in agreement with our finding of a negative correlation between CpG mutation rate and GC content. The fact that non-CpG divergence is strongly correlated with CpG mutability suggests that those genomic features and processes that impact mutation rate in general also impact CpG mutation rate. As in general mutations mainly occur during DNA replication, this argues for that also CpG mutations at least partly occur during DNA replication and hence should show a male-bias in mutation rate [21, 22]. This is supported by our findings that CpG transition rate shows a male-bias. However, the absence of a male-bias in CpG-specific transition rates, proposes that methylation-induced mutations not only occur during DNA replication but also during other states where the DNA is found single-stranded, well in agreement with the negative correlation between mutation rate and GC content.

6.3 Paper III
Twisted signatures of GC-biased gene conversion embedded in an evolutionary stable karyotype.

Across a wide variety of vertebrates the composition of the genome can be characterized by a mosaic structure of DNA stretches rich in GC content adjacent to GC-poor stretches and vice versa, the so-called GC isochore structure [57–59]. The occurrence of this structure has been associated with meiotic recombination and the recombination-linked process called GC-biased gene conversion (gBGC) [40, 60]. gBGC favors the fixation of GC alleles over AT alleles and thereby has the potential to govern patterns of base composition within genomes in relation to recombination. Both mammalian and avian genomes show signatures of gBGC. However, in contrast to the situation in many mammalian genomes where heterogeneity in GC content is decreasing, recent comparative genomic studies suggest that the pronounced isochore structure of birds is being reinforced [61]. The identification of genomic features associated with this discrepancy will assist our general understanding of how gBGC impacts the evolution of base composition. In this study, we therefore in detail investigate the signatures of gBGC in the chicken genome (as one representative of a genome showing a reinforcement of the isochore structure) as a complement to detailed investigations of gBGC in the human genome (as one representative of a genomes showing an erosion of the isochore structure).

Using whole-genome alignments of chicken, turkey and zebra finch we estimated chicken-specific nucleotide substitution rates in intergenic regions of the genome partitioned into windows of 1Mb. For these windows we estimated total substitution rate and further classified substitution rates into the four sub-
categories $W \rightarrow S$, $S \rightarrow W$, $S \rightarrow S$ and $W \rightarrow W$, where $W$ stands for 'weak' nucleotide, i.e. A or T, and S stands for 'strong' nucleotide, i.e. C or G. For our analysis we restricted the data to a set of 880 windows for which estimates of the six genomic features, recombination rate, GC content, gene density, repeat density, CpG island (CGI) coverage and distance to telomere were available. In order to explore the relationships between substitution rates and these genomic features we performed standard multi-linear regression (MLR), partial correlation analysis and principal component regression (PCR). The statistical analysis was then complemented by a mathematical modeling approach, where we developed an expanded analytical model of gBGC adapted to describe the substitution patterns found in the chicken genome as a function of GC content and recombination rate. Finally, we investigated the conservation of signatures of gBGC between chicken and turkey in relation to karyotype stability, which was estimated based on evolutionary breakpoint data.

The MLR and partial correlation analysis identified a strong negative impact of GC content and the distance to telomere on total substitution rate as well as a strong positive impact of repeat density. The impact of recombination rate and gene density on total substitution rate was weak, where recombination rate showed an unexpected negative relationship with total substitution rate. CGI coverage showed only a negligible impact on substitution rate variation. However, as the genomic features used as candidate explanatory variables were strongly correlated with each other, conclusions about the strength of respective impacts on the response variable based on MLR and partial correlation analysis required some caution. Therefore, we in addition performed PCR. Even though this did not provide much information on the causative explanatory variable, one conclusion from the PCR was that the compound effect of all six genomic features led to a negative relationship between nucleotide substitution rate and recombination rate. This negative relationship appeared at first glance contradictory to a model of gBGC and encouraged more detailed investigations. Hence, we computed pair-wise correlation coefficients between recombination rate and total nucleotide substitution rate as well as the four sub-categories of nucleotide substitution rates. In agreement with a model of gBGC, the correlation was positive for $W \rightarrow S$ substitution rate and negative for $S \rightarrow W$ substitution rate. On the other hand, the negative correlations between recombination rate and $S \rightarrow S$, $W \rightarrow W$ and total substitution rates could not be explained by a model of gBGC. Supplemental PCR for the four sub-categories of substitution rate suggested that the local GC content itself, either directly or indirectly via interrelations to other genomic features, could cause these negative relationships by its strong negative impact on the rate of mutation and its strong correlation with recombination rate. Nevertheless, the impact of recombination rate on $W \rightarrow S$ substitution rate via the mechanism of gBGC was strong enough to counteract the otherwise negative relationship between substitution rate and recombination rate. Based on these results we then expanded the classical model of gBGC by incorporation of a negative im-
pact of the GC content on mutation rate. Model selection by a likelihood-ratio test clearly preferred the expanded model over the classical model of gBGC. We then hypothesized that the strong impact of GC content on nucleotide substitution pattern is specific to avian genomes due to their unusually slow rate of chromosomal evolution. Because of this, interrelations between the local GC content and recombination rate as well as other genomic features might build up over evolutionary time and therefore be more pronounced in the avian genome. This hypothesis was supported by the comparison of chicken- versus turkey-specific substitution patterns, which were found to be more conserved in evolutionary stable regions than in unstable regions.

In summary, our analysis provides support for that the reinforcement of the isochore structure in birds is directly related to karyotype stability, which is characteristic to avian genomes. We argue that the evolutionary stable karyotype constitutes an ideal environment for an evolutionary stable recombination landscape, which facilitates the build-up of signatures of gBGC leading to a reinforcement of the GC isochore structure. Moreover, this can also explain the unexpected negative correlation between nucleotide substitution rate and recombination rate as a secondary response of the strong negative correlation between GC content and mutation rate along with the strong interrelations between genomic features. The negative correlation between GC content and mutation rate might be caused by the relationship between GC isochores and replication timing [62, 63]. Another possibility is a causative relationship between GC content and mutation rate in relation to the thermodynamic stability of double-stranded DNA, which is increased in GC-rich regions and might be a rate-limiting factor of mutation [7].

6.4 Paper IV

Why time matters: Codon evolution and the temporal dynamics of $dN/dS$.

The role and the strength of natural selection versus neutral processes lies at the heart of many biological questions, and much effort has been devoted to measure the contribution of selection to DNA sequence evolution in a comparative genomic framework. One popular measure of selection used in a comparative genomic framework is the $dN/dS$ ratio. It is arguably the most common measure to compare the efficacy of selection across species [64] or to identify genes under positive selection for an evolutionary lineage of interest [48, 65]. Surprisingly, the $dN/dS$ measure has only vaguely been explored in a mathematical population genetic context despite of its common use and early introduction in the 90ies [45, 46]. Particularly for closely related lineages the measure has recently shown to be biased [66–68], and it remains
largely elusive how well the measure actually reflects selection across all relevant evolutionary time-scales.

We tried to fill this gap by firmly rooting the $dN/dS$ ratio in evolutionary theory. We first formulated an extended codon model that is anchored in population genetic theory and provided a missing link between population genetics and phylogenetics. In brief, we expanded on existing models of codon evolution and integrated the contribution of segregating ancestral and lineage-specific polymorphisms to codon sequence divergence. We then derived analytical expressions for the $dN/dS$ measure in a Poisson Random Field framework, which yielded fundamental insights into the dynamics of this measure across evolutionary time starting from a speciation event to deep phylogenetic divergence. The analytical results were then complemented by simulation studies.

In good agreement both analytical and simulation studies demonstrated that the $dN/dS$ ratio is indeed biased over a significant period of time through the contribution of segregating ancestral as well as lineage-specific polymorphisms. Further, our comprehensive mathematical description of $dN/dS$ as a ratio of two Poisson random variables allowed us to explore its statistical properties. These investigations showed that stochastic forces are not negligible. Estimates of $dN/dS$ will in particular be biased if the length of the gene sequence is short or if the mutation pressure on the gene is low. Moreover, not only the expected value of $dN/dS$ tends to be biased for such genes, also the measurement errors are particularly strong for the same set of genes. Thus, the stochastic nature of $dN/dS$ can by chance lead to extreme values, which then wrongly might be interpreted as a signal of natural selection.

Our results have significant implications for the interpretation of $dN/dS$ in particular for closely related lineages. In addition, our mathematical exploration of $dN/dS$ might form the basis for refined inference of selection not only from stereotypic genome-sequence data but incorporating also polymorphism information. Our theoretical investigations of $dN/dS$ constitute a refinement of the underlying theory of the widely used McDonald-Kreitman test and can improve the inference on the mode and strength of selection for closely related lineages. Given the rapidly increasing ease of acquiring whole-genome sequencing data not only for one stereotypic genome of a species, but for large population samples of a species as exemplified in the 1000 human genome project [69] and emerging population genomic studies in genetic non-model organisms [70], this is a timely contribution. Recent studies have contributed to this goal by introducing valuable extensions of traditional methodologies [71]. Our theoretical investigations of $dN/dS$ can provide an alternative approach. In addition, due to its flexibility, our codon model has the potential to be useful for other applications, e.g. to jointly infer the effect of GC-biased gene conversion which can mimic the signature of selection [72] and to quantify the selection pressure on silent mutations.
7. Prospects for the future

The increasing availability of genome sequence data for a wide range of species as well as advances in comparative genomic methods have brought about a substantial progress in the study of variation in rates and patterns of nucleotide substitutions. Several causative parameters of variation in neutral substitution profiles have been identified, and an obvious next step is to integrate these parameters into models of natural selection in order to tease apart the relative strength of neutral variation and variation due to natural selection. How do strand asymmetries, the CpG effect and gBGC affect estimates of natural selection? How do these substitution patterns affect estimates of codon usage bias? There is a long list of open questions, and more elaborated models and mechanistic development is required to answer these questions. For example, the results presented in Paper I indicate that substitutional strand asymmetries correlate with gene expression level, where substitutional asymmetries leave a more pronounced signature in highly expressed genes than moderately expressed genes. As a consequence, the estimation of codon usage bias could be flawed by substitutional strand asymmetries particularly in highly expressed genes. As these are exactly the genes expected to show a codon usage bias, it appears relevant to correctly account for strand asymmetries in models for codon usage.

Another example, which recently has attracted much attention, is the impact of gBGC on inferences of positive selection. gBGC has the potential to lead to an access of nucleotide substitutions within hotspots of meiotic recombination [72]. Thereby it can create genomic signatures of accelerated evolution as measured by elevated $dN/dS$ ratios, which initially have been mistaken as evidence for positive selection [73, 74]. Therefore, for unbiased inferences of positive selection the impact of gBGC needs to be properly accounted for [75, 76]. Recently, some first approaches in this direction have been made [77, 78], but further investigations on the interactions between gBGC and natural selection are necessary to develop refined methods to detect positive selection. Here, the theoretical framework developed in Paper IV can assist our understanding on how gBGC effects estimates of $dN/dS$ over evolutionary time. Further, the theoretical framework can help to address the impact of gBGC not only by theoretical investigations but also provides the basis for a refined null hypothesis for the inference of natural selection not only from stereotypic genome-sequence data but incorporating also polymorphism information. Besides, the same framework can be employed to correct for the impact of ancestral polymorphisms on estimates of natural selection.
for closely related lineages. Briefly, the theory presented in Paper IV can in several ways lead to a refinement to the widely used McDonald-Kreitman test [79] and thereby make use of the increasing availability of a combination of polymorphism and divergence data.

Aside from the development of more sophisticated methods to detect natural selection, another interesting research area is the study of the molecular mechanics of the evolution of GC isochores. This question is of interest because GC isochores have been shown to be involved in several key processes of genome maintenance and expression, such as genome structure, chromatin packing, replication timing and gene expression regulation [38, 62, 80]. The evolution of GC isochores has been explained via the mechanism of gBGC, and the results presented in Paper III suggest that signatures of gBGC will be blurred if the recombination landscape is frequently changing over time. Only an evolutionary stable recombination landscape allows for a reinforcement of the GC isochore structure. However, not only gBGC but also DNA methylation and the concomitant the CpG effect show an impact on GC content evolution [81, 82]. Expanded theoretical models are needed to disentangle the effects of gBGC and CpG mutations on the evolution of GC isochores. Here, a combination of genome sequence data and germline DNA methylation data can help to quantify the relative contribution of these two processes on GC content evolution. Moreover, such data can potentially allow to address the question if gBGC and the CpG effect act independently of each other and thus show additive effects, or if these two processes act interrelated with each other.

Finally, the results presented in Paper III suggest that the recombination landscape is affected by chromosomal rearrangements. Investigations of downstream consequences of an evolutionary changing recombination landscape were in particular focused on the GC isochore evolution. However, a changing recombination landscape might not only affect the GC isochore evolution but might also impact other genomic signatures, such as the genomic landscape of nucleotide diversity. Recombination has been found to correlate positively with nucleotide diversity, which has been attributed to the interaction between natural selection and genetic linkage [83]. Population genetic theory suggests that recombination promotes the efficiency of natural selection, whereas genetic linkage between loci can cause interference between responses to selection [84]. Thus, genomic regions affected by frequent chromosomal rearrangements and a concomitant changing recombination rate might experience different selection pressures than evolutionary stable regions, a speculation which motivates further analysis.

In general, a combination of empirical data and theoretical models are necessary to answer several open questions in molecular evolutionary studies. In light of the growing data resources, this is an inspiring period for theoretical investigations, as it allows for an immediate application and cross-validation of theoretical results.
8. Summary in Swedish

Substitutioner är bestående förändringar i arvsmassan som har uppstått genom att en mutation av en enskild bas i en individ har fixerats i hela populatio-
nen. Substitutionshastigheten, antalet substitutioner per tidsenhet, kan variera markant mellan olika arter och mellan olika regioner i arvsmassan. Ett fram-
stående exempel på denna variation är den stora skillnad i antalet substitu-
tioner som sker i funktionella jämfört med icke-funktionella delar av arvsmas-
san. Generellt fixeras mycket färre mutationer i funktionella regioner vilket leder till en lägre substitutionshastighet och det kan förklaras med naturligt urval (selektion) som verkar mot fixering av nyuppkomna mutationer. Förut-
tom variation mellan funktionella och icke-funktionella regioner av arvsmas-
san finns också betydande variation i substitutionshastighet mellan olika icke-
funktionella regioner. Det betyder att variation i substitutionshastighet inte bara beror på variation i selektion utan också på andra processer. Förutom
variation i substitutionshastighet finns också variation i substitutionsmönster; 
det finns 12 möjliga substitutioner som kan ske och variation i mönster bety-
der att vissa typer av substitutioner förekommer mer eller mindre frekvent i olika regioner av arvsmassan. För att kunna urskilja regioner av arvsmassan där selektion verkar eller har verkat är det därför viktigt att förstå de grundläg-
gande principerna för variation i icke-funktionella substitutionshastigheter och
substitutionsmönster.

Tidigare forskning har visat att flera populationsgenetiska och molekylära faktorer kan ha inverkan på substitutionshastigheten och substitutionsmön-
sterna. Några exempel på sådana faktorer är populationsstorlek, genetisk rekombi-
nation, transkription och replikation. Dessutom påverkar själva strukturen och
nukleotidkompositionen på arvsmassan substitutionshastigheten. Den rel-
ativa betydelsen av dessa faktorer och i vilken grad de påverkar varandra är
en kärnfråga inom evolutionära studier av arvsmassan som debatteras kon-
tinuerligt och den frågan är grunden för mina studier. Tyngdpunkten i min
avhandling ligger på studier av variation i substitutionshastighet och substi-
tutionsmönster i ryggradsdjur och mitt primära fokus ligger på att studera icke-
funktionella regioner, det vill säga regioner som troligtvis inte är påverkade av selektion.

Jag började med att undersöka de molekylära processer som påverkar skill-
nad i frekvens mellan olika typer av substitutioner på de två komplementära
strängarna som arvsmassan är uppbyggd av, så kallad strängasymmetri. Jag
fann att variationen i strängasymmetri till viss del kan förklaras av både tran-
skription och replikation. Genom att jämföra ett antal olika däggdjurs DNA-
sekvens kunde jag sedan visa att variationen i strängasymmetri bevaras mellan arter. Det betyder att även den regionala variationen i transkription och den globala organisationen av mekanismer som styr replikation bevaras mellan arter. Mitt nästa steg var att undersöka effekten av metylering på särskilda nukleotider i DNA-sekvensen på substitutionshastigheten i primater. Metylering är en process där en vätejon ersätts av en metylgrupp på en särskild plats i nukleotidmolekylen. Tidigare undersökningar har visat att mutationshastigheten är betydligt högre i de specifika positioner av arvsmassan där nukleotiden cytosin (C) är direkt efterföljd av nukleotiden guanin (G), så kallade CpG positioner, om dessa positioner har påverkats av metylering. I det här fallet visar mina studier att DNA-metylering inte är den enda faktorn som påverkar substitutionshastigheten på CpG-positioner, men att även till exempel den regionala kompositionen av nukleotider spelar en viktig roll.

Jag kompletterade mina studier av däggdjur med undersökningar av fåglar. Genom att studera olika artgrupper som däggdjur och fåglar kan man jämföra resultat vilket hjälper till i tolkningen av underliggande orsaker till variation i substitutionshastigheter och substitutionsmönster. Jag studerade olika typer av substitutioner som påverkar nukleotidkompositionen i en grupp hönsfåglar och jämförde mina resultat med studier gjorda på primater. Min slutsats från den jämförelsen var bland annat att stabiliteten i den globala organisationen av kromosomerna (karyotypen) har påverkat substitutionsmönstret och i förloppet också utvecklingen av nukleotidsammansättningen i arvsmassan.

Förutom empiriska undersökningar av substitutionshastigheter och substitutionsmönster i icke-funktionella regioner gjorde jag teoretiska studier av substitutionshastigheter i funktionella delar av arvsmassan. Jag utforskade hur stor påverkan divergenstiden mellan arter har på hur effektivt vi kan uppskatta selektion. Det har resulterat i nyvunnen kunskap om hur vi bör gå till väga för att på ett tillfredsställande sätt kunna uppskatta substitutionshastigheten och därmed selektion i funktionella delar av arvsmassan från empiriska data.
I like to observe and to analyze, and I often find myself trying to understand and describe patterns and processes that occur in nature. Trying to see the simple in the apparent complex. And I am very grateful to be given the opportunity to do what I enjoy as my work. Therefore, a special thanks to my supervisor Hans Ellegren. Thank you for recruiting me as a PhD student! And I thank you for making me part of a variegated and stimulating research team. It was and is great to work in such a lively environment, a nice mixture of people, even though heavily biased by birders ... Thank you as well for a lot of encouraging discussions, for giving me freedom in my work and for helping me to develop as a researcher. Thank you for being me a good supervisor, someone who I appreciate and who I like.

I also want to thank my two assistant supervisors, Ingemar and Jochen. Ingemar, I thank you for many good discussions and for your calm, precision and endurance in your work. I really enjoyed our collaboration, our regular meetings and our persistent try to find some balance between mathematics and biology. Jochen, I thank you for a lot of stimulating discussions, about science, statistics and whatsoever. And I thank you for being a strictly religious non-believer, for being someone who questions and tries to get to the bottom.

I thank current and former members of the Ellegren lab for sharing your thoughts, and in particular those who inspired me most in my research. Benoit, I thank you for many good discussions and for sharing your passion about work. Thank you for being me a really good colleague. Niclas, I thank you for helping me to understand how biologists think. And I thank you for constantly reminding me to eat more fat. I am not always doing my best ... Ammon, your thoughts and doubts during our group meetings and journal clubs inspired me a lot, and I thank you for that. Hannu, I thank you for being me a good person to talk to, not only about science but also about life up in the north ...

And Kristiina and Robert, my first office mates. Thank you for sharing office with me for quite some years and for making our office a place where I liked to be.

Peter, I want to thank you for integrating me into your lab during my research visit at the MPI in Berlin. I thank you for a good collaboration and for a really good time in Berlin.
But no PhD studies without undergraduate studies, and I want to thank my former supervisor Hennig von Grünberg for introducing me into science. I thank you for picking me out in a group of students, for your support and for encouraging me to continue with research. I am happy that I did so.

But there is life other than research. And for me life is often dancing, which is me as a source of peace and happiness. Dance has been me a great support during my years in Uppsala. Zala, I thank you for many nice ballet classes, and for being me an inspiring teacher and person. And then in Uppsala I came across tango, something I feel happy about, because of the dance and because of the people I met. Alain, I thank you for so often giving me warmth in the cold. It was a thought of you when I first felt a feeling of home here in Uppsala. Peter and Lotta, you always seem to be here for me, regardless if I am happy or if I am sad, if I need help or for some nice hours spent together at a milonga. I am very glad and grateful that I met you. Robert, sometimes you annoy me, because I know you are right and I don’t want to listen to you. But most of all you are me a good friend. Thank you! Janne, I thank you for a big bunch of little things. For my first nice tango, for your patience with my silence, for your hug when I was shaking in tears before leaving for Berlin, ... for a lot! Majid, I thank you for being me a friend, for being me someone to talk to and to listen to. Reza, somehow you are me an unfriendly friend, but I thank you for several reflections on myself. Anna-Maria, I thank you for being as you are. You naturally make me smile :) Josefin, I thank you for spoiling me like a child, but talking to me on equal terms. And Sebastian, I thank you for a lot of wonderful dances. _Inte mysigt, inte trevligt, bara härligt fint!

Well, there is even life other than research and dance?! And I am happy to have met two really good friends, Natassa and Qi. Yes, both of you told me, "You never have time, just for dancing you have time!". But sometimes I find time and I do not regret. Natassa, I thank you for sharing with me your passion and enthusiasm for whatever you do. I always enjoy our discussions and arguinngs, yes, even if I start to cry. And I thank you for shouting at me, when you are angry, and for laughing with me in good times. Qi, I thank you for so often giving me your ear and listening to me, and for sharing with me your thoughts about life as a neurotic. And I thank you for inspiring me to fly, when walking together through the streets of Uppsala ...

And no distance can break true friendship. I want to thank my friends outside of Sweden for being me friends. You are often with me in my thoughts and I am grateful to have you there.
And my family, even if far away, always close to me, most important to me...

Helga, I thank you for being me like a mother, for giving me a home and giving me your ear. Thank you for being here! Lisa and Britta, I thank you too for being here.

Gunilla, sometimes I wish you would think more of yourself than of others, but most of all I am glad and grateful to have a sister like you. You are someone special for me! And I thank you for making me an aunt of three wonderful children, who smile at me every day from the pictures on my wall.

Mamma, you often worry about the way I walk, but still you support me on many of my steps. I thank you for being here for me, for believing in me and for accepting me the way I am. This means a lot to me, and you mean a lot to me.

Pappa, I wish I would be able to thank you, but I cannot find the right words. You are with me in every step I take, in my tears and in my heart and always will be...
References


A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology.