UPPSALA
UNIVERSITET

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 261

Comparative Genomics in Birds

ERIK AXELSSON

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2007

ISSN 1651-6214
ISBN 978-91-554-6770-8
urn:nbn:se:uu:diva-7432

Dissertation presented at Uppsala University to be publicly examined in Zootissalen, EBC, Villavägen 9, Uppsala, Friday, February 2, 2007 at 13:00 for the degree of Doctor of Philosophy. The examination will be conducted in English.

Abstract
Axelsson, E. 2007. Comparative Genomics in Birds. Acta Universitatis Upsaliensis. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 261. 62 pp. Uppsala. ISBN 978-91-554-6770-8.

To shed light on forces that shape the molecular evolution of bird genomes, and in turn avian adaptations, comparative analyses of avian DNA sequences are important. Moreover, contrasting findings in birds to those of other organisms will lend a clearer view on general aspects of molecular evolution. However, few such analyses have been conducted in birds. Progress is presented in this thesis.

Theories predict a reduction in the mutation rate of the Z chromosome as the harmful effects of recessive mutations are exposed in female birds. We find no evidence for this. Instead, the substitution rates of sex chromosomes and autosomes are largely compatible with expectations from male-biased mutation. This suggests that a majority of mutations arise during DNA replication.

Substitution rates also vary across chicken autosomes. For instance, microchromosomes accumulate ~20% more substitutions than macrochromosomes. We show that a majority of the autosomal variation in substitution rate can be accounted for by GC content, mainly due to the incidence of mutable CpG-dinucleotides.

Sequence comparisons also show that the pattern of nucleotide substitution varies in the chicken genome and this reinforces regional differences in base composition.

The level of selective constraint in at least some avian lineages is higher than in mammalian lineages as indicated by low d_N/d_S – ratios. Larger historical population sizes of birds relative to mammals could explain this observation. Within the avian genome, the d_N/d_S is lower for genes on micro- than macrochromosomes, potentially owing to a higher incidence of house-keeping genes in the former category.

Contrasting data on non-synonymous and synonymous substitution for divergence and polymorphism shows that positive selection has contributed more to the evolution of Z-linked than autosomal genes. This is likely explained by the full exposure of beneficial recessive mutations on Z when in female birds.

Keywords: Birds, molecular evolution, macrochromosomes, microchromosomes, mutation rate, substitution rate, base composition, positive selection, genetic drift, population size

Erik Axelsson, Department of Evolution, Genomics and Systematics, Norbyv. 18C, Uppsala University, SE-75236 Uppsala, Sweden

ISSN 1651-6214
ISBN 978-91-554-6770-8
urn:nbn:se:uu:diva-7432 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-7432)

List of papers

The thesis is based on the following papers, referred to by their roman capitals throughout the text.

I Axelsson E., Smith N.G.C., Sundström H., Berlin S. and Ellegren H. (2004) Male-biased mutation rate and divergence in autosomal, Z-linked and W-linked introns of chicken and turkey. *Mol. Biol. Evol.* 21:1538−1547

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

III Webster M.T., Axelsson E. and Ellegren H. (2006) Strong regional biases in nucleotide substitution in the chicken genome. *Mol. Biol. Evol.* 23(6):1203−1216

IV Mank J.E., Axelsson E. and Ellegren H. (2006) Fast-X on the Z: rapid evolution of sex-linked genes in birds. *Submitted manuscript*

V Axelsson E., Hultin-Rosenberg L., Brandström M., Zwalen M. and Ellegren H. (2006) Natural selection in protein-coding genes expressed in avian brain. *Manuscript*

Art work on front page “Tuppiga tuppar”, by Adriana Jerlström-Maj

AJM i Vasselhyttan
http://www.krumelur.tk/

Contents

Abbreviation

BGC	biased gene conversion
MK-test	McDonald Kreitman test
SNP	single nucleotide polymorphism
LRT	likelihood ratio test
d_N/d_S	the ratio of nonsynonymous substitutions divided by the ratio of synonymous substitutions
DSB	double strand break
GC3	GC-content at the third codon position
GC*	equilibrium GC-content
N	population size
s	selection coefficient
μ	mutation rate
CR1	chicken repeat 1 element
EST	expressed sequence tag
GO-term	gene ontology term
cDNA	complementary DNA

Introduction

With huge amounts of DNA-sequence data now available, the field of molecular evolution has moved into an exploratory phase, where theories and predictions, mostly from the 20th century, are tested. There are almost complete genome sequences of numerous mammalian, fish, insect, fungi and plant species accessible for analyses. In analogy to the exploration of the fauna and flora of new continents hundreds of years ago, biologists now work their way through un-explored genome sequences in order to determine how evolution shapes life. It is obvious that much knowledge in this respect will come from comparing observations from the genomes of different species, i.e. comparative genomics.

So far the genomes of birds have been poorly investigated. The near complete analysis of the chicken genome sequence (ICGSC 2004) advanced the field of avian genomics enormously, but access to one bird sequence, in contrast to the 22 analysed mammalian genome sequences, shows how this field lags behind. As the distinct autosomal chromosome classes and the reversed sex chromosome system of the avian genome (males ZZ, females ZW) provides a useful contrast to mammals, studies of molecular evolution in birds has the potential to increase our understanding of molecular evolution in general. In addition to doing that, this thesis also aims at illuminating avian adaptations at the molecular level.

In more detail I will discuss how and at what rate mutations are created, and to what extent mutation rates vary across different regions of genomes, or among different evolutionary lineages. Another focus of this thesis will be to better understand the causes and consequences of a heterogeneous base composition. I will also discuss how mutations become fixed in populations; to what extent is evolution neutral or driven by selection? Observations in birds will be compared to those of other organisms in order to better comprehend these and earlier observations, or to define features specific to the molecular evolution of birds.

The avian genome

The avian genome is distinct to that of other organisms in several respects. First, the genome size is unusually small. The total number of base pairs in the chicken genome amounts to ~1.1 Gb, which means there is roughly a 3-fold difference in the genome size of birds and mammals (ICGSC 2004). This size difference is not due to a smaller number of genes in chicken as it has been estimated that the chicken genome contains between 20 000 and 23 000 protein coding genes (ICGSC 2004). There is however a marked difference in the occurrence of repeats in avian and mammalian genomes. Interspersed repeats make up ~9% of the chicken genome while somewhere between 40% and 50% of a mammalian genome consists of such sequences (ICGSC 2004).

Furthermore, birds usually have many chromosomes, with the average avian genome consisting of 2n≈80 chromosomes (Griffin et al. unpublished). This is clearly different from most other organisms and represents an almost 2-fold difference when compared to our own species (2n=46). Moreover, the avian karyotype appears to be more stable than that of mammals as 63 % of all birds in which the karyotype has been determined have 2n=74–86 (Griffin et al. unpublished).

Another distinct feature of the bird genome, in addition to some reptilian genomes, is the microchromosomes. These are tiny chromosomes that range from a few 100 Kb to at most 20 Mb in size (ICGSC 2004). The chicken genome consists of 28 such chromosomes in addition to 5 large macrochromosomes and 5 intermediately sized chromosomes (2n=78). Microchromosomes are more GC-rich, more gene dense and recombine more frequently than the rest of the genome. The median recombination rate for microchromosomes is 6.4 cM per Mb, as compared to 2.8 cM per Mb for macrochromosomes (ICGSC 2004).

Finally, birds differ from many other organisms in having a reversed sex-chromosome system. Males carry two Z-chromosomes, while females carry one Z- and one W-chromosome (figure 1).

Mutation

Mutation is the ultimate source of genetic variation and as such the fundament for the variety of life forms on earth. Without mutations there would be no evolution. At the same time it is paradoxically clear that most mutations are harmful to the individual.

Mutations are defined as permanent changes in the genetic material and include various types of genetic alterations; whole genome duplications, chromosome translocations and inversions, transpositions, insertions and deletions, and on the smallest scale, replacements of one single nucleotide by another. Here I will focus on this last class of genetic alterations – the base substitutional mutations – here after simply referred to as mutations. Understanding how and at what rate mutations arise is for example crucial for finding regions of selective importance in genomes.

How do mutations arise?

DNA contains the heritable information required for all higher life forms to develop and reproduce. The ability to correctly transfer information, to new cells or across generations, is dependent on the complementary structure of DNA. Four nucleotides form base pairs in DNA such that adenine (A) pairs with guanine (G), and cytosine (C) pairs with thymine (T). Each time a cell divides, the DNA needs to be copied, or replicated, in a process where the complementary, double stranded molecule is split into two single stranded parts, or templates, onto which new nucleotides are base paired aided by a DNA polymerase. As vertebrate genomes usually consist of $\sim 10^9$ nucleotides, replication has to be done at a considerable speed. In mammals DNA polymerases has been estimated to incorporate 50 new nucleotides per second during replication (Alberts et al. 1994). This speed should however come at a cost of accuracy as it is possible for nucleotides to base pair in other ways apart from the correct G-C and A-T manner. If a mispairing between the template and the incoming base is accepted this could lead to the formation of a mutation.

In order to reduce their error frequency, many DNA polymerases are equipped with a proof reading function, which has the ability to replace a nucleotide mistakenly bound to the growing DNA sequence. In *Escherichia*

coli the rate at which errors are introduced by a variety of DNA polymerases, whose proofreading function has been destroyed, was estimated to vary between 10^{-4} and 10^{-6}. When proofreading on the other hand is functioning, error rates for the same polymerases drop to between 10^{-7} and 10^{-8} (Kunkel 2004). Eukaryotic DNA polymerases are likely to be at least as accurate as those of *E. coli*.

Would proof reading fail, leading to the incorporation of an erroneous base into the DNA, there are other mechanisms like the mismatch repair that can locate and correct the misinsertion.

In addition to errors being introduced during replication there are several other ways in which mutations can be created. DNA is thermodynamically unstable with spontaneous structural changes as a consequence. For example, 5000 purine bases, adenines and guanines, are lost per day in each human cell due to thermal disruption (Alberts et al. 1994). It is also a well established fact that methylated cytosine spontaneously and frequently deaminate to uracil. Furthermore, active metabolites and UV-radiation can cause DNA to alter base pairing. The relative contribution of each of the mentioned mutation inducers to the overall rate of germ line mutation is still a debated issue and I will return to discuss this issue soon.

Measuring mutation rates

The quantity of new mutations added to a genome is usually measured in the number of mutations per nucleotide site, either per year or per generation. In order to actually measure this quantity there are in principle two different procedures. One more direct, based on phenotypic changes across generations, and another indirect, based on DNA sequence comparisons mostly between species. The different approaches are here exemplified in human.

Using the first approach, the number of affected children born to unaffected parents for a range of autosomal or X-linked dominant disorders has been counted. As the parents are healthy and it is known that the trait is dominant, a diseased child must have received a new mutation. Thus by knowing the proportion of affected children it is possible to estimate the per locus per generation rate of mutation. In order to obtain a rate per nucleotide it is necessary to divide the obtained rate with the number of sites of the particular locus which can be expected to result in a phenotypic change when mutated.

The second indirect approach uses the fact that the probability of fixation for a neutral allele is equal to its frequency in the population (Kimura 1968). If new mutations are introduced at a rate u mutations per site and generation, then each site receives $2Nu$ new mutations in a diploid population of size N in each generation. As mentioned, the chance that a new allele completely

replaces the original nucleotide in the entire population is equal to its frequency, which is 1/2*N*. The overall rate of fixation, *k*, is thus equal to the number of new mutations per generation times the probability that these mutations will become fixated, or $2Nu$ x 1/2*N*, meaning $k = u$. In other words, for neutral DNA sequence the rate of fixation is equal to the rate of mutation. By comparing an orthologous stretch of DNA from e.g. human and chimpanzee, it is possible to estimate the proportion of fixed differences, or the sequence divergence. If the divergence time for the two species is known, dividing the sequence divergence with the divergence time gives the substitution rate, *k*, which is equal to *u*, the mutation rate.

Using the more direct approach, (Kondrashov 2003) estimated mutation rates from 20 different disease causing loci, arriving at a rate of 1.8×10^{-8} mutations per site and generation in human. Nachman and Crowell (Nachman and Crowell 2000) instead compared processed pseudogenes in human and chimpanzee, with which they indirectly estimated the mutation rate to 2.5×10^{-8}. The two different methods thus seem to lead to similar conclusions. In practice this means that every human is born with in the order of 100 and 200 new mutations.

Generation time and mutations

In table 1, indirect estimates (using the sequence comparison method) of mutation rates for different organisms are listed. When mutation rate estimates per site and year are compared across organisms an interesting pattern emerges: mutation rate is negatively correlated with generation time. This correlation depicts what has been termed the generation time hypothesis (Li 1997) and possibly sheds further light on how mutations are created. As every generation needs a new copy of the genetic material, shorter generation times should mean an increased number of replications measured per year. Given that replication cause mutations, a high mutation rate in organisms with short generation times could thus be interpreted as evidence for the importance of replication in causing genetic change.

There are however alternative explanations to the connection between generation time and mutation rate. It has been proposed that the high metabolic rates of short lived animals result in an increased exposure of DNA to mutation-inducing reactive metabolites. Furthermore, Kumar and Subramanian (Kumar and Subramanian 2002) suggested that the generation time effect is an artefact of compositional non equilibrium. According to this theory the high rate of mutations observed in, for example, the rodent lineage could be due to a high rate of chromosomal rearrangements in this lineage. Interestingly, following this last hypothesis and considering our observation of a more or less stable base composition in chicken (discussed in more de-

tail later), it might be expected that life history characteristics, such as generation time, should have no effect on mutation rates in birds. This does however not seem to be the case (Bartosch-Harlid et al. 2003).

Table1. Mutation rates per base pair and year or generation, μ, for organisms with different generation times, measured in years.

species	gen. time	μ^{-year}	$\mu^{-generation}$
D. psuedoobscura – D. Melanogaster	0.1	2.24×10^{-8}	2.24×10^{-9}
M. domesticus – R. Norvegicus	0.5	4.83×10^{-9}	2.41×10^{-9}
G. gallus – T. guttata	1 – 1.8	2.32×10^{-9}	3.24×10^{-9}
M. gallopavus – G . gallus	1.8	1.46×10^{-9}	2.63×10^{-9}
H. sapiens – P. troglodytes	20	1.02×10^{-9}	2.04×10^{-8}

Divergence estimates, k, from (Richards et al. 2005), (IRGSC 2004), (ICSAC 2005), chicken - turkey : **paper I** , chicken - zebra finch: **paper V**.

Males and mutations

There is evidence to suggest that males contribute more mutations to the next generation than females do. This, in turn, could be seen as evidence for a replication origin of mutations, as the nature of the male and female germ lines differ markedly. Whereas the female egg production is mostly finished before birth, the male sperm production continues through out the whole life. At the time of reproduction, human sperm will thus have a history of many more replications than egg cells. Consequently, if mutations are intruduced when DNA is being copied, a higher rate of mutations in males fits well with the many more male than female replications.

Haldane (Haldane 1947) was the first to estimate sex specific mutation rates. In a human population, he used the observed proportion of males affected by the X-linked disorder haemophilia, their fitness relative to other males, and the state of the disease causing locus in their mothers, to calculate the female mutation rate, assuming selection-mutation equilibrium. He concluded that the female mutation rate could well be an order of magnitude lower than the male mutation rate.

More recently, evidence for a male biased mutation rate in human has been provided through comparisons of divergence estimates from sex chromosomes or sex chromosomes and autosomes (Miyata et al. 1987), (Bohossian et al. 2000), (Makova and Li 2002). In contrast to autosomes, sex chromosomes, by definition, spend different amount of time in male and female germ lines. The mammalian Y-chromosome is always found in the male germ line, while X spends 1/3 of the time there. As DNA is replicated more often in male then female germ line, Y is expected to accumulate mutations at a higher rate than autosomes, which in turn should receive more

mutations than X. Thus, comparing the sequence divergence of neutrally evolving Y and X-linked loci (or X- or Y-linked loci to autosomal) and accounting for the difference in time spent in the two germ lines provides a means to calculate α, the male to female mutation rate ratio. For estimates of α based on comparisons including human and other primate sequence data, values obtained usually range from 3−6 (Makova and Li 2002), (ICSAC 2005). How well does these values correspond to the actual difference in number of cell divisions, and thus replications, between male and female germ line? As sperm production is a continuous process, this difference will depend on the average male age at reproduction. Assuming this has been 20 years in the human lineage leads to an estimated male to female germ line cell division ratio, also called C, of 6.25 (Hurst and Ellegren 1998). The fine correspondence between α and C supports the view that replications cause most human mutations.

Using DNA sequence comparisons α has been estimated in a number of different organisms, including mammals, birds, fish**,** flies and plants (table 2). From these studies, a male biased mutation rate seems to be a general phenomenon and measures of C, when available, are in fairly good agreement with α. As expected from the almost identical numbers of germ line replications in males and females Drosophila seems to lack a male bias (Bauer and Aquadro 1997).

An interesting pattern emerges when comparing α and generation time in table 2: the longer the generation time the higher α is. This fits into the realm of a replication origin of mutations since a longer generation time, in combination with a continuously dividing male germ line, adds to the difference in number of replications in male and female germ line.

In conclusion, there is quite some evidence for the existence of a male biased mutation rate in many organisms and this is in turn provides good support for replication causing the majority of mutations. There is however not a lack of challenges to this view as will be evident further on in this text.

Table 2. Comparison of the ratio of the number of cell divisions in male and female germ line at the average age of reproduction, C; the male to female mutation rate ratio, α; and the generation time, for different organisms.

Species	*C*	*α*	Gen. time
D. melanogaster – D. simulans	0.96	~1	0.1
M. musculus – R. norvegicus	2.0	1.9	0.5
G. gallus – M. gallopavo	4.45	2.47	1 – 1.8
Perissodactyla ssp.	–	3.88	4
Caprini ssp.	–	3.94	6 – 8 *[1]
Felidae ssp.	–	4.38	3 – 7
C. familiaris	–	4.8	5 – 10 *[2]
H. sapiens – P. troglodytes	6.25	5.25	20
S. latifolia – S. diocia	–	∞	–

*1: Female average age at reproduction. *2: Author's own estimation. Flies: (Bauer and Aquadro 1997), mouse – rat: (IRGSC 2004), chicken – turkey: **Paper I** , sheep - goat: (Lawson and Hewitt 2002), dog: (IDGSC 2005), human – chimp: (Makova and Li 2002), Silene: (Filatov and Charlesworth 2002), cats: (Pecon Slattery and O'Brien 1998), horses and rhinos: (Goetting-Minesky and Makova 2006).

Evolution of mutation rates

As mentioned earlier between 100 and 200 new mutations are added to a human genome every generation. What forces shape the magnitude of mutation rates? In order to function efficiently a protein is built from a specific succession of amino acids that usually tolerate little or no change. As mutations hit this sequence in a random like fashion, most of them will lead to an impaired protein function. Considering the harmful effects of most mutations, are individuals with a low mutation rate more fit than others? In other words, do mutation rates evolve?

In a constant environment there is no need for organisms to change and a mutation rate of zero, if possible, is expected. Environments are however rarely constant and in order for individuals to adapt to changed surroundings mutations are sometimes needed. It might thus be that the rate at which mutations are incorporated into the genome is determined by a balance between the deleterious effects of mutations and the organism's ability to adapt to new environments (Leigh 1970). A prediction from this theory is that organisms in rapidly changing environments should have high mutation rates. The environment of disease causing bacteria might fit to this description, as they are influenced by the dynamic nature of the host's immune system. Observations of relatively high frequencies of so called mutator alleles, which increase mutation rates, in pathogenic bacteria (LeClerc et al. 1996) is thus in agreement with this hypothesis. In asexual organisms like bacteria, a mutator allele would increase the chance of a beneficial mutation at a locus else-

where in the genome, with which the mutator allele then can hitch hike to high population frequencies (Leigh 1970), (Sniegowski et al. 1997). For sexual organisms, the situation is different as recombination will brake up the linkage between the mutator locus and the favoured allele, and in fact, theoretical work predicts that mutators invariably will be selected against in sexual organisms (Leigh 1970). Sexual organisms will hence due to the deleterious nature of most mutations and the effect of recombination, always favour a zero mutation rate.

Then why are mutation rates not equal to zero in sexual organisms? First it should be pointed out that replicating a large genome without introducing even a single error is likely a physiological impossibility (Nilsson and Snoad 2002). Asking whether mutation rates are at their physiological lower limit in sexual organisms, and if not why, might therefore be better questions. Instead of through a balance between adaptability and adapted ness, a trade-off between harmful effects of mutations and the cost of a higher fidelity during DNA-replication could explain mutation rates. Energy spent on avoiding extra mutations, for example via an improvement in the mismatch repair system, can potentially be of better use for other processes in the organism. At this time it is unclear whether mutation rates physiologically can be reduced further or the cost for this is too high, however evidence do point towards the latter case (Sniegowski et al. 2000).

If mutation rates are determined in a trade-off as described above, it has been argued that there might be regions in genomes of sexually reproducing organisms that should evolve at separate rates. McVean and Hurst (McVean and Hurst 1997) proposed that since recessive deleterious mutations on the X-chromosome are exposed in the hemizygous male, while masked on autosomes, a reduced mutation rate on X relative to that of autosomes should have evolved. Testing this hypothesis is somewhat problematic as a reduced divergence on X is also expected on the basis of the male biased mutation rate hypothesis. A difference in mutation rate between X and autosomes which is solely due to more mutations of male than female origin, can however at most lower the ratio of X to autosomal divergence to 0.67 (Miyata et al. 1987). In a study of rodents McVean and Hurst estimated this ratio to 0.62 and therefore suggested that the mutation rate of X might be reduced for adaptive reasons. In doing so, they did not exclude the existence of a male biased mutation rate but rather questioned the importance of such a bias.

In birds, males are homozygous with two Z-chromosomes, while females carry one Z- and one W-chromosome (*Fig 1*). In contrast to mammals, this different set of sex chromosomes offers an improved possibility to discriminate between the forces of a male biased mutation rate and a selectively reduced mutation rate. While more mutations in the male than female germ line should lead to a greater divergence for Z- than W-linked loci, the opposite is expected from the trade-off theory, where the hemizygous exposure of

Z in females should lead to a lower mutation rate on Z than W. Ellegren and Fridolfsson found evidence for a high Z- relative to W-linked divergence (Ellegren and Fridolfsson 1997) and therefore reemphasized the importance of the male-driven evolution theory.

Mammals		**Birds**	
♀	♂	♀	♂
AA	AA	AA	AA
XX	XY	ZW	ZZ

Figure 1. Schematic figure showing the reversed sex chromosome system of birds relative to mammals.

Support for a male origin of many mutations thus stands strong, but it would be premature to rule out the possibility that a combination of effects leads to the observed pattern of sex chromosomal divergences in birds and mammals. Aiming at shedding further light on this pattern, we compared divergences of Z-linked, W-linked and autosomal intron sequence in chicken and turkey (**paper I**). The existence of a reduced mutation rate on Z could potentially have led to a Z-linked divergence less than that of autosomes, while a high male mutation rate still could have kept the Z divergence well above that of the W-chromosome. However, from our divergence estimates, Z=10.99%, A=10.08% and W=5.74%, we were unable to find significant evidence for a reduced mutation rate on Z (**paper I**). Furthermore, should it anyhow exist it can only be a weak reduction. Assuming that a male bias in the mutation rate is the sole cause of divergence differences among the different chromosomal classes, we used autosomal and W-linked divergence to calculate the expected divergence of Z. From this we concluded that a mutation rate reduction on Z of 4.89% is compatible with the data, although this reduction is not statistically significant.

In conclusion, support for a specifically reduced mutation rate on the Z- and X-chromosomes is faint. Furthermore, instead of invoking the trade-off theory, high constraint on synonymous sites (Lu and Wu 2005) may explain why X-linked synonymous divergence is lower than expected.

Mutation rate variation

If time spent in male germ-line is decisive for the mutation rate, it might be expected that neutral sequences residing on autosomes should evolve at similar rates. The first indication that this is perhaps not the case was provided by Wolfe, Sharp and Li (Wolfe et al. 1989). Based on data on silent site divergence they argued that mammalian mutation rates vary among different regions of the genome. Support for this view has later been lent from numerous studies in several mammalian lineages (Matassi et al. 1999), (Lercher et al. 2001), (Smith et al. 2002), (Hardison et al. 2003), (IDGSC 2005).

When zooming in on genomes, from a regional to a smaller scale, it appears as if rate variations are smoothened out and finally replaced by similarities on a local scale. How far are local similarities in mutation rates extending? Lercher and colleagues (Lercher et al. 2001), in a comparison of the mouse and human genomes, observed that rates of substitution are more similar than expected on a scale that stretched from 1Mb up to whole chromosomes. In murids mutational similarities were found on scales from 100 kb up to 15 Mb, but the unit of mutation rate variation is probably not longer than 1 Mb (Gaffney and Keightley 2005). It was therefore argued that similarities over larger distances are due to the propagation of local similarities. In accordance with this, rate similarities on a 1-2 Mb scale has been noted in several other analyses (Hardison et al. 2003), (Webster et al. 2004), (IDGSC 2005), (2005). Hence in mammalian genomes regional rather than chromosomal location seems to influence the mutation rate of a particular sequence.

Yet, mutation rate variation can also be observed at even smaller scales, ranging from only one to a few base pairs in length (Silva and Kondrashov 2002). It is for example clear that neighbouring bases can affect the mutation rate of single nucleotides (Zhao and Boerwinkle 2002).

In the following I will discuss the effects of different mechanisms or conditions that have the potential to affect mutation rates in both birds and mammals. In doing that I will describe in more detail how mutation rates vary in relation to several genomic features and discuss how these observations match theories that attempt to explain mutation rate variation.

CpG dinucleotides

The most striking example of a neighbour effect on mutation rates is that of CpG dinucleotides. When the cytosine of a CpG-site is methylated it is deaminated to thymine at a rate that is almost 9 times faster than that of other mutations (ICSAC 2005).

Methylation of cytosine is a common feature of mammalian and most likely also of avian genomes. It has been estimated that there are 3×10^7

such modified nucleotides in the human genome and that these predominately are found within CpG dinucleotides on the coding strand (Li 1997). The reasons for methylation could be many, but it seems to play an important role in transcriptional repression. When promotor regions are methylated gene expression is hampered. Furthermore, methylation plays a role in both X-chromosome inactivation and genomic imprinting, and it could possibly act to block transcription of repeated sequences and retrotransposons (Trasler 1998).

In man, the fast rate of CpG-mutations means that roughly 25% of all germ line mutations are of this kind (ICSAC 2005). Interestingly, when comparing estimates of the male to female mutation rate ratio based on either sequence including all possible sites or CpG sites only, there is a marked difference: $\alpha_{\text{all sites}} = 6.37$ and $\alpha_{\text{CpG}} = 2.00$. This indicates that transitions in CpG dinucleotides are time-dependent, rather than dependent on the number of germ-line cell divisions (Hwang and Green 2004; Taylor et al. 2006).

Spatial variations in the distribution of CpG sites thus have the potential to explain autosomal mutation rate variation. For instance in our study of substitution rates in chicken chromosomes (**paper II**), the significantly elevated divergence of microchromosomes relative to macrochromosomes vanished after removing CpG-substitutions. Furthermore, in primates, higher rates of sequence divergence at synonymous sites than non coding sequence can be explained by a high incidence of CpG sites in exonic sequence (Subramanian and Kumar 2003).

However in the comparison of the human and chimpanzee genomes only 4% of the divergence rate variance could be explained by mutations in CpG-sites. In addition, CpG rates were found to correlate with non-CpG rates, suggesting that there exist additional mechanisms controlling these two mutation processes simultaneously (ICSAC 2005). In relation to this it might also be worth noting that mutation rates at CpG-sites in fact also vary between different genomic regions. As an example of this, note our detailed analysis of individual substitution rates in repetitive chicken sequences (**paper III**). Here we found a slight increase in CpG-mutability with GC-content and also detected an increase in the relative contribution of CpG mutations to the overall rate of substitution as the GC-content of the sequence increased. These variations in CpG-mutability could be due to regional differences in methylation patterns.

In conclusion, variation in the density of CpG-sites can explain some of the observed regional differences in substitution rates of both avian and mammalian genomes. Regional differences do however still remain after controlling for this (Smith et al. 2002), (Filatov 2004) (although excluding substitutions affected by CpG sites is difficult), meaning that additional mechanisms must be invoked to explain all mutation rate variation.

Recombination

In yeast there is experimental evidence that the repair of double strand breaks (DBS) caused in association with recombination is error prone (Strathern et al. 1995). Moreover, it has been shown that the rate of recombination correlates with substitution rates between human and chimpanzee (Hellmann et al. 2003) and between human and mouse (Lercher and Hurst 2002). For example, genes residing in the frequently recombining pseudoautosomal region show a higher divergence than genes located elsewhere on the X-chromosome (Filatov 2004). Recombination is however also correlated with GC-content (Kong et al. 2002), which in turn influence substitution rates, why conclusions on a mutagenic effect of recombination are difficult to draw. Hellman and co-workers (Hellmann et al. 2003) noted that after controlling for sequence context effects there is still a correlation between recombination and divergence at a 100 kb scale, but that this is lacking at a 3-Mb scale. They furthermore estimated that ~ 6% of the divergence between human and chimpanzee is explained by mutations with a recombination origin. This rather limited role for recombination in generating mutations was recently supported in a detailed analysis of the relationship between recombination and diversity (Spencer et al. 2006). Here it was shown that correlation between recombination and diversity only extend across a few thousand base pairs, in practice meaning that the mutagenic effect of recombination is strong enough to influence human diversity only in recombination hotspots. Similar results were presented by Arndt and colleagues (Arndt et al. 2005), who found little effect of recombination on individual substitution rates.

In chicken, neither intron divergence, nor synonymous site divergence was observed to correlate with recombination rate (**paper II**). However it should be emphasized that these analyses were carried out using crude recombination rate estimates.

Hence, it seems that recombination plays only a minor role in generating mutation rate variation.

Base composition and mutation rates

It is clear that mutation rate varies with base composition, and that GC-content is the best predictor of regional mutation rates in both mammals (Hardison et al. 2003), (Arndt et al. 2005) and birds (**paper III**). As a consequence, we might gain more insight into the causes of mutation rate variation from understanding how base composition varies and evolves.

When compared to other eukaryotes, mammalian and avian genomes exhibit large regional differences in base composition (Nekrutenko and Li 2000). This variation is above all characterised by the existence of many GC-rich regions. Regions like these, consisting of a relatively homogeneous

base composition, can extend for several hundreds of Kb of sequence before being interrupted by what is sometimes a sharp, sometimes a more gradual decline in GC-content. Transitions from high to low GC-regions are furthermore reflected in changes in several other genomic features so that gene density, expression pattern, replication timing, recombination and repeat density all correlate with base composition (Eyre-Walker and Hurst 2001). Interestingly, understanding the reasons for variation in base composition may therefore not only shed light on mutation rate variation, but also on how genomes are organised. However, considering the many interrelated parameters it is very difficult to separate causative relationships from factors related to each other through covariance with another feature.

There are many different theories on the evolution of the heterogeneity of base composition; some invoke selective arguments while others rely on neutral explanations. Bernardi (Bernardi et al. 1985) suggested that the elevated GC-levels of mammals and birds evolved as a response to an unstable environment for DNA in warm blooded animals. However, the finding that at least some reptile genomes also display a clearly heterogeneous base composition likely puts the origin of GC-rich regions before the emergence of endothermy, thereby questioning this argument (Hughes et al. 1999). It has also been proposed that GC-richness evolved in order to facilitate transcription in gene rich regions (Vinogradov 2005). In support of this it has been observed that chromatin structure of GC-rich regions is in an open state which is believed to affect transcriptional activity in a positive way (Gilbert et al. 2004).

Turning to neutral explanations, it has been suggested that patterns of mutation vary due to either differences in the efficiency of repair (Filipski 1987), varying timings for the onset of DNA replication (Wolfe et al. 1989) or to variation in the frequency of recombination. Apart from via a mutagenic effect, it has been proposed that recombination can influence base composition through a process called biased gene conversion (BGC). I shall soon return to discus these theories but let us first look in more detail at how mutation rates vary with GC-content.

Substitution rates vary with base composition

Mammals

In mammals positive correlations between rates of substitutions at four fold degenerate sites and GC content have been observed (Williams and Hurst 2000), (Kumar and Subramanian 2002), (IMGSC 2002). Moreover, even if divergence estimated from ancestral repeat sequences is correlated with that measured at four-fold degenerate sites it is obvious that the relationship between base composition and substitution rate differ depending on what sequence type is used (Hardison et al. 2003). Using repeat sequence a

more U-shaped relationship with GC is obtained meaning that both AT-rich and GC-rich sequences evolve relatively quickly. This finding is mirrored when analysing alignments of complete human and chimpanzee genomes, where AT-rich regions show a 10% higher divergence compared to the genome average, and GC-rich areas close to chromosome ends also exhibit elevated substitution rates (ICSAC 2005). Adding to the complexity, in dog, syntenic breakpoints display a significantly increased GC-content as well as increased divergence (IDGSC 2005).

To better understand the variation in substitution rates across genomes it might be helpful to track how individual substitutions vary with base composition. It has been shown in humans that rates of GC-decreasing substitutions (G:C→A:T, G:C→T:A or CpG→CpA/TpG) decrease with increasing GC-content, while the opposite is observed for rates of GC-increasing substitution (A:T→G:C or A:T→C:G) (IHGSC 2001), (Arndt et al. 2005). Furthermore it seems as if the GC-decreasing substitution rates depend more heavily on GC-content than the opposite types of changes do. There is a dramatic decrease in C:G → X:Y transversion rates (C:G → G:C or C:G → A:T) in AT-rich regions as GC increase (Arndt et al. 2005). A:T → T:A substitutions on the other hand display little rate variation with GC.

Base composition is not at equilibrium in mammals

Perhaps surprisingly it has been found that the total rate at which new G and C nucleotides are added to the genome cannot match the rate at which G and C are removed. Consequently, the human genome is not at compositional equilibrium (IHGSC 2001), instead, the average GC-content is decreasing. This was later found to be true for several other mammalian species (Duret et al. 2002) and generally the shift towards a reduced average GC-content manifest in GC-rich regions becoming GC-poorer (Duret et al. 2002). Arndt and colleagues (Arndt et al. 2003) showed that substitution patterns changed from having a preserving to a homogenising effect on base composition at around 90 Mya. Interestingly, rates of CpG-mutations increased 4- to 8 fold at this time point.

Birds

In **paper II** we analysed the alignments of 155 chicken and turkey genes and observed a positive relationship between GC3 (the GC-level at the third codon position) and synonymous substitution rates (r=0.40). On the other hand, divergence estimated in 67 introns spread across the chicken and turkey genomes displayed no dependence on GC-content (**paper II**). Then again, when analysing substitution patterns in CR1 elements in chicken, we found a strong positive correlation between substitution rate and GC-content (r = 0.832) (**paper III**). In fact there is a more than twofold increase in substitution rates when CR1 elements in GC-poor regions are compared to those in GC-rich.

We also investigated how individual substitution rates vary with base composition in chicken *(Fig. 2)* (**paper III**). Except for substitutions in CpG-sites, we found a negative relationship between rates of GC-decreasing substitutions (G:C→A:T or G:C→T:A) and base composition in chicken, thus mimicking the situation in human. CpG substitutions on the other hand increase with GC-content in chicken. This is contrary to what is seen in human. Turning to GC-increasing substitutions (A:T→G:C or A:T→C:G), we find, as in human, an increase of rates in GC-rich regions *(Fig. 2)*. The strong increase in C:G → X:Y transversion frequencies (C:G → G:C or C:G → A:T) in AT-rich areas found in human is not observed in chicken.

CpG substitutions hence seem to represent a change in pattern between human and chicken.

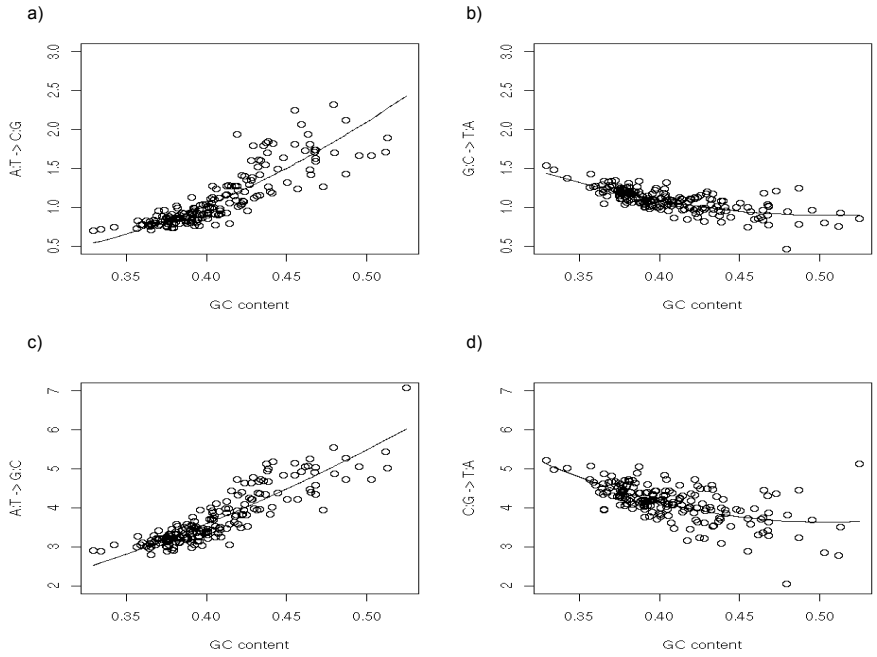

Figure 2. Substitution frequencies inferred from CR1 elements in the chicken genome plotted against GC-content for: a) GC-increasing transversions; b) GC-decreasing transversions; c) GC-increasing transitions; c) GC-decreasing transitions (**paper III**)

Base composition is close to equilibrium in birds

We use substitution rates inferred from the analysis of chicken repeat sequences to estimate GC*, the equilibrium GC content, of different genomic

regions (**paper III**). When comparing GC* with present day GC we note that the heterogeneity in base composition is being reinforced, instead of homogenised as seen in several mammalian genomes. We also estimated GC* using substitution rates inferred from chicken, turkey and quail intron sequence alignments. Once again we see no sign of decaying GC-rich regions but instead our analysis suggests that variation in base composition is being preserved.

In conclusion, base composition and mutation rate are in some way related phenomena, both in birds and mammals. It is possible to explain a reasonably large proportion of substitution rate variation, perhaps especially in birds, from this relationship. To further understand the causes of mutation rate variation I will therefore now turn to discuss factors that potentially influence base composition.

Recombination and biased gene conversion

Many studies have reported a positive correlation between recombination rate and GC-content (Eyre-Walker 1993), (Lercher and Hurst 2002), (Kong et al. 2002), (Hellmann et al. 2003), (Spencer et al. 2006). In addition, genes residing in the frequently recombining human psudoautosomal region (Filatov 2004) and genes found on the microchromosomes of chickens are unusually GC-rich (**paper II**). These observations suggest a link between recombination and base composition. Given a causative relationship between these two factors, which is the determining one? The fate of the mouse F_{XY} gene serves as a natural experiment to help answering this question; this gene underwent a rapid increase in GC-content following translocation from an autosomal location to the highly recombining PAR (Montoya-Burgos et al. 2003). Furthermore, recombination correlates better with GC* than with the present day GC, an observation that also indicates that recombination drives base composition, not the other way around (Meunier and Duret 2004).

In theory, recombination could affect base composition in two different manners. First, recombination requires the formation of DSB which in turn has to be repaired. If the repair mechanism is biased towards using either G or C nucleotides as templates in the repair process, recombination will effect nucleotide composition. Evidence for biased repair processes has been observed in both mammals (Brown and Jiricny 1988) and birds (Kunkel and Alexander 1986). However, considering that few mutations in mammals seem to arise due to recombination this effect might be insignificant.

Secondly, recombination could influence the fate of existing mutations. When DSB are repaired, sequence from the homologous chromosome is used as template. Consequently, in the case of allelic differences between the two chromosomes heteroduplexes will be formed as the two sequences are paired. These heteroduplexes may in turn be repaired in a biased manner (Strathern et al. 1995), leading to a process which has been referred to as

biased gene conversion (BGC). BGC does not influence the rate of mutation, but rather the rate at which mutations are becoming fixed. (In doing so, this process resembles selection and should perhaps not be included in a discussion of mutation rate variation. On the other hand BGC is a neutral process with potential relevance to the origin of a heterogeneous base composition.)

Individual substitution rates do change with GC-content in a way that to some extent fit with predictions from the BGC-hypothesis. In human there is tendency for rates of A:T→G:C or A:T→C:G substitutions to increase in areas of high GC, while the opposite is true for G:C→A:T or G:C→T:A substitutions (Arndt et al. 2005). This could indicate that GC-increasing mutations go to fixation at higher rates in regions with high recombination rates. The same, and in fact even stronger, pattern is also observed in chicken (**paper III**). Considering the large chromosomal differences in recombination rate in chicken (ICGSC 2004), patterns of substitution might be expected to change more clearly with GC-content in birds than in mammals if BGC is important. Moreover it is known that telomeric regions of the human genome experience high rates of recombination (Kong et al. 2002). It is therefore interesting to note that in human there is a change of substitutions patterns in subtelomeric DNA, relative to other genomic regions, that is consistent with an increased action of BGC (Arndt et al. 2005). A similar change in the chicken genome is reflected in increased levels of G and C nucleotiedes in telomeres of macrochromosomes (ICGSC 2004) as well as in an increased rate of substitution in these regions (**paper III**).

Furthermore, since the consequence of BGC is equivalent to the action of selection, it is expected that polymorphic sites segregating with either a G or C and an A or T should display allele frequency shifts away from what is expected under neutrality. Such shifts have been observed in several human SNP studies. For instance Lercher and colleagues (Lercher et al. 2002a) could not reject a model incorporating BGC when describing allele frequency distributions in GC-rich regions of the human genome. In another study it was noted that AT→GC polymorphisms segregate at higher frequencies in GC-rich regions (Webster et al. 2003). Finally, elevated allele frequencies for AT→GC relative to GC→AT mutations have been observed in regions of both low and, in particular, high recombination (Spencer et al. 2006).

There is thus much evidence to support the existence of BGC. However, arguments have been lifted that BGC only rarely overcomes the power exerted by stochastic events on fixation probabilities (see section on fixation). Also, the individual substitution patterns observed in human are probably better explained by mutation biases, with a possible additional effect of BGC in highly recombining areas such as subtelomeric regions (Arndt et al. 2005). Furthermore, indirect evidence in the form of correlations between recombination and GC-content could be explained trough recombination and GC covarying with a third factor. Finally, if BGC creates GC-rich regions in verte-

brate genomes (that is, there is no mutation bias) then the overall rate of substitution is expected to correlate negatively with GC-content (Piganeau et al. 2002). This is not observed in either mammals, or birds.

Repair

There are many different DNA polymerases present in vertebrate cells (Kunkel 2004). The function of all has not yet been determined but it is clear that there is a large variation in accuracy among these polymerases (Kunkel 2004). As a consequence mutation rates and base composition might be expected to vary if different polymerases operate in different regions of the genome. For instance, Filipski (Filipski 1987) noted that both DNA polymerase β and DNA polymerase α are involved in the repair of DNA lesions. However, *in vitro*, DNA polymerase β appears to be more error prone than DNA polymerase α. It has furthermore been hypothesised that repair of DNA lesions in relaxed, GC-rich and transcriptionally active chromatin is performed by DNA polymerase β, while lesions in the condensed, AT-rich regions to a higher degree are repaired by DNA polymerase α. We should in other words expect to see an increase in mutation rates as sequences become more GC-rich.

This hypothesis is not entirely supported by observations from studies of mammalian genomes, although it is interesting to note that substitution rates at four fold degenerate sites exhibit a less complex, positive relationship with GC-content, whereas divergence in untranscribed regions display a U-shaped dependency on GC. The situation in birds is as mentioned different, in that both types of sequence (with the exception of introns) show a similar and rather strong positive correlation with levels of GC. In relation to this it is very interesting to note that observations in mouse show that mistakes made by DNA polymerase β indicate no bias in the probability of inserting A or T, in relation to C or G nucleotides, while in chicken there seems to be a 3:2 bias in favour of G and C (Kunkel and Alexander 1986). Therefore this theory offers a potential explanation not only to the differences in compositional status between birds and mammals, but as well to the different relationships between divergence and GC-content in the two classes.

Another example of a regional variation in repair mechanisms is transcription coupled repair. This repair mechanism is believed to be activated as RNA polymerase II stales when trying to read through a mismatched base pair. The nucleotides of the transcribed strand is then excised and replaced using the coding strand as template. This means misinsertions introduced by the DNA polymerase will not be corrected for if they happen on the coding strand. In turn, any strand associated biases in errors made by the DNA polymerase should be evident from differences in base composition when transcribed and non-transcribed regions are compared. Most mammalian genes show a slight increase in G and T on the coding strand thus supporting the existence of transcription coupled repair (Green et al. 2003). Furthermore a

negative correlation between substitution rate and expression breadth in introns has been reported (Webster et al. 2004). Although it is hard to directly connect this process to any large scale differences in substitution rates and base composition it functions as an example of the complexity of these questions. The status of transcription coupled repair and possible associated strand asymmetries is unknown in birds. It would be interesting to investigate this further.

Replication timing

In vitro experiments have shown that mutagenesis can increase as a consequence of unbalanced dNTP pools (Kunz and Kohalmi 1991). Replication errors occur either directly through misinsertion due to excess or deficiency of a dNTP, or through inhibition of proofreading in which excess of a dNTP forces extension past a mismatch (next nucleotide effect) (Mathews 2006). As a consequence, over time it might also be expected that dNTP concentrations during replication should be reflected in the genomic base composition (Mathews 2006). Almost all organisms analysed so far display asymmetries in the nucleotide pool at replication, and generally dGTP is clearly under represented in comprising only between 5 and 10% of the dNTP pool (Mathews 2006). *In vivo* replication reactions run at dGTP levels observed in vivo do not show clearly increased mutation rates when compared to reactions run at equal nucleotide concentrations (Martomo and Mathews 2002). On the other hand there is evidence that too high *in vivo* nucleotide concentrations cause mutations and that accumulation of especially dGTP can increase mutation rates (Martomo and Mathews 2002).

Adding to this, there is also evidence to suggest that levels of nucleotides change during the long time required for a complete genome to replicate (Wolfe 1991) and that the time point for the onset of replication varies across different regions of genomes. In human it has been shown that gene rich and above all GC-rich regions are replicated early during S-phase of the cell-cycle (Woodfine et al. 2004). Transcriptional activity is an even better predictor of replication timing. When comparing groups of genes with different replication timing there is a clear tendency for early replicating genes to be expressed during replication (Woodfine et al. 2005). As origins of replication are spaced at 100-150 kb intervals in the human genome (Donaldson 2005) it is interesting to note that the correlation between transcriptional activity and the timing of replication is improved at a ~100 Kb, compared to a 1Mb scale, while the opposite is observed for GC and replication timing (Woodfine et al. 2005).

Considering observations of regional differences in replication timing, possible mutagenic effects of dNTP pool concentrations and suggestions that concentrations of dNTPs change during the course of replication – can these circumstances explain regional mutation rate variation? Interestingly, a U-shaped relationship between mutation rates and GC-content is expected if

the next nucleotide effect is strong (Wolfe 1991), (Hurst and Williams 2000). This does in fact agree with observations in mammals. On the other hand we do not observe this in birds (**paper III**) and given that a U-shaped relationship is expected even for sequences at equilibrium with the dNTP pool this could be problematic for this model.

Non equilibrium

When the present pattern of substitution does not match that which shaped the base composition in a DNA sequence, the sequence is at a state of compositional non equilibrium. As mentioned, this seems to be the case for many regions in several mammalian genomes. Given that regions vary in pattern of substitution, as described above in both mammals and birds, a shift from equilibrium to non-equilibrium could be imagined if a sequence is moved from one chromosomal location to another. Moreover, regardless of the exact nature of mechanisms underlying mutation or fixation, it is also possible that chromosomal rearrangements could induce changes in pattern of substitution. If regional differences in GC-content are caused by biases in mutation rates, a consequence of compositional non equilibrium would be an increase in the overall rate of substitution (Piganeau et al. 2002). Furthermore, the greater the difference between the equilibrium GC-content and the present day GC-content, the greater the increase should be.

Kumar and Subramanian (Kumar and Subramanian 2002) argued that the above reasoning potentially explain large proportions of the observed regional mammalian substitution rate differences. Testing for this they compared the rate of substitutions at four fold degenerate sites in two different groups of genes; one group contained genes for which nucleotide frequencies (or GC content) differed more than expected when orthologues from different mammals were compared, while the other contained genes of more similar base composition. When estimating the average evolutionary distance for the two sets of genes using cow and pig, it was found that the group of compositionally different genes had evolved at a rate 46% faster than the other group. A similar result was observed when comparing the dog, human and mouse genomes, where the ratio of lineage specific divergences in orthologous regions is positively correlated with the ratios of current GC-content in the same regions (IDGSC 2005).

However, since base composition is at equilibrium in chicken (**paper III**) the above explanation fails to explain the positive correlation between GC and divergence there. It would anyhow be interesting to test this hypothesis further by seeing if substitution rates vary to a less extent in birds than in mammals.

Conclusions on mutation

The many examples of male biased mutation (e.g. **paper I**) and the observation of a generation time effect probably mean most mutations arise during replication. There is little convincing evidence of a specific reduction in the mutation rates of X- and Z-chromosomes (**paper I**). However, the substitution rate of neutral, autosomal sequence varies (e.g. **paper II**) and there are many potential explanations to this variation. Up till now the best way to predict the substitution rate of a neutral sequence is GC-content (e.g. **paper III**) but this relationship is not trivial as it explains far from all variation in substitution rates (Hardison et al. 2003), (ICSAC 2005), (IDGSC 2005); seems to vary depending on what kind of sequence is used for substitution rate estimation (IMGSC 2002) ; varies between different regions of the genome (ICSAC 2005); and differs between different organisms (Arndt et al. 2005), (**paper III**). In addition, we lack a firm understanding of why base composition display regional variation in the first place (e.g. **paper III**). A large difficulty in understanding how both mutation rates and base composition varies in a genome is that many factors with the potential to explain the observations are interrelated.

In summary, mutation rates can vary due to differences in CpG-densities (**paper II**) and the level of methylation of these sites. Furthermore, recombination is mutagenic, but to what extent is a debated issue as is the strength of the effect of BGC. The roles of heterogeneous repair processes and different timings of replication are very interesting but more needs to be known about the specificities of the large variety of DNA polymerases before any conclusions can be drawn. Whether a sequence is in compositional equilibrium or not may affect mutation rates, but this cannot explain mutation rate variation in chicken (**paper II** and **paper III**).

We are thus left with speculations as to why mutation rates and base composition vary. In mammals it has been noted that the decay of GC-rich regions started at a time point that also saw an increase in rates of chromosomal rearrangements (Arndt 2003). Birds in contrast, have experienced relatively few changes in karyotype (Griffin et al. unpublished) and likely also in substitution pattern as is indicated by the comparatively stable base composition of the chicken genome (**paper III**). Furthermore, due to the difference in rate of rearrangements the karyotype ancestral to all amniotes was more likely bird like than mammal like, which means many chromosome fusions probably occurred in the mammalian lineage. As one chiasmata per chromosome arm is required at meioses (Pardo-Manuel de Villena and Sapienza 2001), a consequence of these fusions may have been a reduced rate of recombination in mammals. We have argued that differences in the rate of recombination between mammals and birds (ICGSC 2004), via the effect of BGC, have the potential to explain the difference in base composition between birds and mammals (**paper III**). However, this hypothesis

has its problems and if chromosomal rearrangements are related to shifts in patterns of substitution it is equally well possible that other mechanisms, apart from recombination, change so that mutation rates and base composition are affected.

As it stands now, it thus seems hard to exclude any of the above potential contributors to mutation rate variation. It might be that all influence mutation rates to varying degrees in different parts of genomes. Furthermore, due to the difficulties of controlling for different factors in comparative analyses, experimental biology could be the way forward in this field.

Even though it is difficult to fully understand the causes of mutation rate variation is still very important to understand how mutation rates vary, not the least when searching for functionally relevant changes or regions, in genomes.

Fixation

Once a new mutation has arisen it will represent a new variant, or allele, in a pool of 2*N* (where *N* is the population size of a diploid organism) copies of a particular site in the genome. This allele will eventually come to meet one of two possible fates; either it is lost, or it spreads through the population and finally completely replaces all other alleles. In the latter case the mutation has gone to fixation.

Evolution could thus be described as a process in which the raw material – mutation – is shaped by factors that influence the probability of fixation. Inferring patterns of fixed mutations between different populations forms the foundation of comparative analyses of DNA sequences and offers the possibility to understand and quantify forces responsible for the fixation of alleles. Then, which are these forces? Is evolution mostly a random process or is it driven by directional change in the form of selection?

Here I will first describe the factors believed to influence fixation probabilities and how they are expected to affect rates of evolution. This will be followed by a section on how the action of these factors are detected or measured. Then, observations from sequence comparisons, mostly in birds and mammals, will be discussed and contrasted to theoretical expectations in order to better understand how mutations become fixed. In relation to this I will also look at how genome organisation and protein function affects rates of evolution in birds.

Drift

Random events, or genetic drift as it is termed in population genetics, affect the fate of mutations. Stochasticity plays a role during meioses when chromosomes segregate into gametes and also adds variation in the number of offspring between individuals. As result of genetic drift the amount of variation, or evolutionary raw material, available in the population will be reduced. This holds for all populations, but the rate at which this reduction proceeds is inversely proportional to the number of individuals in a population. Large populations will therefore suffer less from random events, while small populations are more sensitive to drift. Ultimately the genetic variation at a certain site will be lost (only in very large populations like that of present day humans can mutation balance the pace of drift so that any site al-

most always is represented by more than one allele) leaving one allele as the sole survivor. The chance that a particular allele drifts to fixation is equal to its currant frequency in the population.

For a new allele this is only $1/2N$ meaning that the chances that it will go to fixation are usually small. In other words drift may play a small role for the fixation of existing variation in a large population, but the destiny of new mutations will be heavily influenced by drift in all populations. Still on rare occasions new mutations will make it all the way to fixation and the rate at which this is expected to take place if the mutation lacks fitness effects, k, is equal to the product of the number of mutations added to a population each generation, $2Nu$. That is $k = u$. Consequently, if evolution is only due to the interaction between mutation and drift, then the rate of fixation should be independent of population size.

Selection

When a mutation affects an individual's capacity to survive or reproduce its frequency in the population, as well as the probability of fixation, may come to change due to selection. The effect of selection is usually measured in s, the selection coefficient, which is a fitness measure of a homozygote of the allele of interest relative to that of an alternative allele. Alleles that confer an advantage to the individual might increase in frequency as they improve the bearer's ability to reproduce successfully. This process has been termed positive selection. The opposite, negative or purifying selection reduces fixation probabilities when mutations affect fitness negatively.

When the fate of a mutation is solely determined by drift, the probability of fixation is equal to the allele frequency. However the rate at which drift affects the change in allele frequency from generation to generation is independent of allele frequency. In this respect selection differs from drift. The strength of selection is depending on the frequency of the allele. For new and rare alleles selection is a very weak force. The strength of selection acting on a new mutation is roughly s multiplied by $1/2N$, which is less than the strength of drift. Only after the allele has increased in frequency, will selection become important, and if $s >> 1/2N$ selection will dominate drift for common alleles. When $s \approx 1/2N$ selection and drift have similar abilities to influence the destiny of alleles. This is true also for common alleles. What determines the outcome of the interaction between selection and drift?

Positive selection and drift

The probability of fixation for a new mutation is as mentioned small due to random events in nature and as selection is a weak force for rare alleles most

beneficial mutations will be lost. For relatively small values of s and in sufficiently large populations, the probability of going to fixation for a beneficial new mutation is roughly equal to $2hs$, where h is a measure of dominance (if $h = 0$ then the new mutation is completely recessive, while if $h = 1$ it is completely dominant). As before, multiplying this probability with the rate at which new mutations are added gives the expected rate of adaptive evolution: $2Nu$ x $2hs = 4Nhus$. From this expression it can be seen that, in contrast to for neutral evolution, adaptive evolution depends on the size of the population. Evolution should, if positive selection is common, be faster in large populations. Furthermore, it is clear that dominant mutations should go to fixation faster than recessive.

Negative selection and drift

It has already been mentioned that most mutations probably are harmful. Therefore, due to negative selection, they are less likely to contribute to evolution. However under certain circumstances drift has the ability to push also deleterious alleles to fixation. The probability of fixation for a new mutation with a negative effect on fitness, given a relatively small s, is:

$$P(\frac{1}{2N}) \approx \frac{s}{e^{2Ns} - 1}$$

Here it is seen that for large populations, such that $2Ns >> 1$, it is very unlikely that a harmful mutation would become fixed. On the other hand, if $2Ns \approx 1$ then fixation is not unlikely. The expected rate of deleterious evolution, k, can be obtained through multiplication with the mutation rate:

$$k \approx \frac{2Nus}{e^{2Ns} - 1}$$

That is, the rate of fixation of deleterious mutations is expected to be slower in large as compared to small populations.

Tests for selection

There are numerous ways of testing if, or to what degree, selection has influenced the evolution of a DNA-sequence. A common theme for most of these tests is that they use Kimura's deduction of expectations under neutrality as a null model. I will here briefly describe the tests for selection used in this thesis.

d_N vs. d_S

Due to the redundancy of the genetic code a mutation in a coding region may either cause a change in amino acid, a so called *nonsynonymous* mutation, or leave the amino acid as is, in what is termed a *synonymous* mutation. A simple and widely used measure of selection is the comparison of the relative amount of nonsynonymous and synonymous substitutions in a gene of interest. Here the number of nonsynonymous changes per nonsynonymous site (d_N *or* K_A) should be equal to the number of synonymous changes per synonymous site (d_S *or* K_S) if no selection is acting on the gene. A reduced value of d_N relative to d_S is indicative of negative selection while the opposite scenario suggests positive selection.

In chicken and turkey, for 155 genes analyzed in this manner, we found the average d_N/d_S – ratio to be equal to 0.133 **(paper II)**. This reflects what has been pointed out several times; most mutations that cause change in the amino acid sequence are harmful and will be selected against. This is a general finding echoed in many organisms (see table 4).

The d_N/d_S – ratio is an average measure of the forces that influence fixation probabilities across all sites of a gene. In the comparison of chicken and turkey genes mentioned above, this measure can thus be used to say that at least 77% of nonsynonymous mutations are deleterious, while the 13% of mutations that became fixed could be deleterious, neutral or beneficial. This is a limitation in discerning exactly how genes evolve as different amino acids might be affected by varying degrees of selection or even by differentially oriented selection.

McDonald-Kreitman tests

In the McDonald-Kreitman test (MK-test) (McDonald and Kreitman 1991), patterns of polymorphisms within a species are contrasted with patterns of fixed differences between species. If mutations contributing to polymorphism and divergence, both at synonymous and nonsynonymous sites in a gene, are neutral, then there should be no difference between the ratio of nonsynonymous to synonymous changes within as compared to between species. Deviations from this can be tested in a 2x2 contingency table, where fixed synonymous and non synonymous differences are tabled in one row while polymorphic synonymous and non synonymous sites are tabled in the other (table 3). Significant deviations can then be interpreted as evidence of non neutrality. For example, as advantageous mutations only remain polymorphic for a relatively short period of time in a population once they have reached an appreciable frequency, they will contribute little to the within species variation but add to the between species divergence. If adaptive evolution is common, an excess of fixed non synonymous differences relative to non synonymous polymorphisms will be detected in the MK-test. The oppo-

site pattern, a deficiency of amino acid replacements relative to polymorphisms might indicate that harmful mutations are segregating in the populations. This test has the advantage that it can detect the existence of positive selection even when negative selection is dominant.

However one should be cautions when drawing conclusions from a test like this. Still segregating polymorphisms represent the near past of the history of many species, while differences between species might have been accumulating for long periods of time. Therefore, as the interaction between drift and selection is influenced by the population size, the MK-test will be sensitive to fluctuations in N (Eyre-Walker 2002). In addition it should be mentioned that since the power of the test is relatively low sufficient amounts of both within and between species data is required for a reliable result.

Table 3. 2x2 contingency table showing fixed as well as polymorphic synonymous and nonsynonymous differences for the EF-hand containing 2 gene. Divergence data is from chicken and zebra finch and polymorphisms are from chicken. This gene displays a significant excess of fixed nonsynonymous differences based on neutral expectations.

	Nonsynonymous	Synonymous
Polymorphic	2	6
Fixed	53	31

Multi-species analysis using maximum likelihood

Maximum likelihood is a statistical method for parameter estimation. Given a certain data set, which in the case of molecular evolution often is a set of DNA-sequences, this method estimates one or several parameters, like e.g. the d_N/d_S – ratio, that maximize the likelihood of observing the data. The methodology has the potential to overcome difficulties in parameter estimation due to, for instance, multiple substitutions at nucleotide sites, which are known to cause problems for parsimony based methods (Eyre-Walker 1998). Of particular use, independent of what statistical methodology is used, are inferences of evolutionary parameters from multispecies comparisons. This allows for the possibility of letting parameters adopt different values in dif-

ferent parts of a phylogenetic tree. In this way comparisons of the evolutionary forces that have affected separate lineages can be performed and in turn analyzed with respect to known biological differences between the lineages. Furthermore it is possible to test for differences in the evolutionary history of separate amino acid sites of a sequence, or even of the same site when located in different branches. In deciding which model of evolution, or in other words which parameter settings that best explains the data, a likelihood ratio test (LRT) can then be performed. The test statistics of a LRT, which is equal to 2 times the log likelihood difference between the two tested models, is χ^2 - distributed and the number of degrees of freedom is determined by the difference in the number of parameters estimated in each of the models. Apart from providing a means to statistically test for differences in rates of evolution in different lineages, this methodology in part overcomes the limitations with pair wise d_N/d_S – ratio analyses as different sites can be tested separately.

Neutral, nearly neutral or adaptive evolution?

The probability of fixation for new mutations is small due to drift; so small that a majority of all mutations are lost. From average d_N/d_S – ratios it is also evident that additional mutations are removed due to their deleterious fitness effects. Still a fraction of all amino acid mutations contribute to evolution as they become fixed. An important question that remains to be answered is through which process these mutations go to fixation. Are they driven to fixation by positive selection or by drift? Are they perhaps deleterious, but still able to replace other alleles? In other words is evolution adaptive, neutral or nearly neutral?

Neutral theory of molecular evolution

In the 1960s, access to the amino acid sequence of a few proteins in a number of different species made it possible to estimate the rate at which proteins evolve. From these estimates it was observed that the rate of change is fairly constant in different lineages (when measured per year)(Zuckerkandl and Pauling 1965). Kimura (Kimura 1968), and Jukes and King (King and Jukes 1969) argued that if selection is the main force responsible for evolutionary change, such as lineage specific adaptations, a more erratic pattern should be observed. They meant that the rate constancy was in good agreement with expectations from the balance between mutation and drift. Consequently they argued that protein evolution to a large extent is due to the fixation, through drift, of neutral mutations, with only a small contribution of selectively driven changes. This view of evolution has been

referred to as the neutral theory of molecular evolution. Already at this time point it was suggested that their might exist rate differences between different proteins, or regions of proteins, as certain sequences could be more sensitive to change than others (King and Jukes 1969). In relation to this, observed d_N/d_S – ratios of less than 1 were argued to fit the neutral theory well as this could mean that deleterious mutations are removed, some neutral mutations go to fixation, while positive selection contributes very little to evolution (Kimura 1977).

Adaptive evolution

Considering the many obvious adaptations of various organisms to their environment it might be expected that directional changes, by which mutations contributing to new adaptations are selected for, should be plentiful. It has however been difficult to observe the effects of positive selection at the DNA-sequence level. This could perhaps be explained by the limitations of the tests for selection described above, but it has also been argued, as in the neutral theory of molecular evolution, that adaptive evolution is a minor determinant of overall rates of evolution. For instance, the efficiency of adaptive evolution is expected to depend on N (4Nush). Large populations should thus evolve faster than small populations, a prediction for which there is conflicting evidence (table 4). It is only relatively recently that signs of positive selection at the molecular level have started to accumulate.

In a comparison of 8079 human and chimpanzee orthologous gene pairs d_N/d_S was found to be larger than 1 for 733 genes. However, for only 35 of these could the ratio be said to be larger than 1 with statistical certainty (Nielsen et al. 2005). Bustamente and colleagues (Bustamante et al. 2005) included human polymorphism data in their analyses and where thereby able to perform MK-tests for a large number of human and chimpanzee genes. Evidence for positive selection was here found for 304 out of 3277 genes. In addition to providing proof of positive selection these studies also exemplifies the use of polymorphism data in search for evidence of adaptive evolution. In rodents it was noted that in codons with two or three substitutions between mouse and rat both or all substitutions occurred in the same lineage unexpectedly often (Bazykin et al. 2004). This was interpreted as evidence for positive selection and it was estimated that at least 0.5% of all amino acid replacements between mouse and rat were driven to fixation for adaptive reasons.

In order to test for selection in birds we aligned and estimated the d_N/d_S of 5542 chicken – zebra finch orthologous gene pairs (**paper V**). Of these only 16 displayed a ratio larger than 1 (whether the ratio was statistically larger than 1 or not was not tested). This is a much smaller fraction than the 733 genes out of a total 8079, observed in the study including human (Nielsen et al. 2005). We also mapped chicken single nucleotide polymor-

phisms (SNP) to a subset of these genes and for genes with at least 4 SNPs associated to them we divided the ratio of non synonymous to synonymous fixed differences by the ratio of non synonymous to synonymous polymorphisms in chicken. 167 out of 815 genes displayed a ratio above 1, however only a fraction (n=13), had significant deviations from expectations when MK-tests were performed (**paper V**). These analyses thus provide little evidence for positive selection in chicken and zebra finch. This in turn is probably to some extent explained by the fact that the zebra finch sequences were obtained from cDNA libraries originating in brain. Mammal brain genes also evolve slowly (Kuma et al. 1995) (Duret and Mouchiroud 2000). An overrepresentation of conserved genes in this data set is seen when the average d_N/d_S – ratio (0.085) is compared to that of the 155 chicken – turkey gene pairs (0.133) **(paper I** and **paper V)**.

If adaptive evolution is frequent enough, the MK-test can also be used to estimate the proportion of nonsynonymous substitutions that were driven to fixation by positive selection. Assuming that polymorphisms segregating in a population are neutral, the ratio of nonsynonymous to synonymous polymorphisms, and the number of synonymous substitutions between two species, can be used to estimate the number of nonsynonymous substitutions expected under neutrality. This can then be compared to the observed value. In this way Smith and Eyre-Walker (Smith and Eyre-Walker 2002), based on some 30 genes in D. simulans and D. melanogaster, estimated that ~ 45 % of amino acid replacements were driven to fixation by selection. Similar conclusions have come from other analyses in Drosophila (Fay et al. 2002), (Bierne and Eyre-Walker 2004). Studies including human genes have produced varying results, with one study, based on 182 pairs of orthologous genes in human and an old world monkey, arguing that 35% of non synonymous substitutions are due to selection in these lineages (Fay et al. 2001). Other studies did however not find any evidence at all for adaptive evolution in human and chimpanzee gene pairs (ICSAC 2005; Zhang and Li 2005). The result of the former study is likely due to a gene set biased towards rapidly evolving genes.

In conclusion, fixations due to selection could be frequent in large populations like those of Drosophila, while for mammals it seems they are relatively rare. A general conclusion on the situation in birds will have to await the analysis of a less biased data set than the one used here. Still, as indicated by some of the human studies mentioned here, and based on theoretical expectations, even in mammals and birds there are probably certain classes of genes that evolve under a large influence of positive selection. I will return to these shortly.

The nearly neutral theory

One of the observations from which the neutral theory came to light was the rather constant rate of protein evolution measured per year, i.e. the molecular clock. A problem with using this finding as argument for the neutral theory is that mutation rates of different lineages are similar when measured per generation, not per year (table 1). In order to accommodate theory to this small but important difference, Ohta (Ohta 1973) made an addition to the neutral theory. Instead of proposing that all fixed amino acid replacements are neutral, a proportion might have slightly deleterious fitness effects, she argued. This new version of the neutral theory was termed the nearly neutral theory. I previously described that when $2Ns \approx 1$, fixation is not unlikely. Mutations with small s, or slightly deleterious mutations, will thus be removed rather efficiently when N is large, while in small populations they may contribute substantially to protein evolution. The nearly neutral theory explains why a generation time effect exists for synonymous changes (table 1) while a clock like ticking, as observed by Kimura, holds for non synonymous changes; in the latter case the increase in the rate of nearly neutral substitutions caused by the generation time effect is cancelled by the increased efficiency of purifying selection in large populations.

The result of this effect can also be seen as a negative correlation between population size and the d_N/d_S – ratio (table 4). In table 4 I have compared the d_N/d_S – ratios from a few analyses on a genomic scale in different organisms to that of our chicken and turkey study including 155 genes (**paper II**). From this comparison it can be seen that the bird genes appear to be evolving under similar constraint to those in rodents. This in turn could be seen as an indicator of similar historical population sizes in birds and rodents. In support of this it can be noted that an estimation of $N \approx 5 \times 10^5$, based on the level of genetic variation in chicken, agrees well with expectations from the d_N/d_S – ratio and the nearly neutral theory. Furthermore comparing the mean d_N/d_S – ratio between chicken and zebra finch (0.085, **paperV**) to those estimated in the lineages leading from a common ancestor to the mouse (0.088), dog (0.095) and human (0.112), respectively (IDGSC 2005), supports a similar conclusion. However these comparisons may be halting due to a small or biased data set in birds. In a more fair comparison we contrasted the evolution of orthologous gene pairs in mammals and birds. The observation of a smaller d_N/d_S – ratio in chicken and zebra finch (0.084) compared to that of human and dog (0.095) again suggests that the population size of birds generally has been larger than for many mammalian populations (**paper V**), perhaps with the exception of rodents.

More support for the nearly neutral theory comes from comparisons of d_N/d_S – ratios for polymorphisms and for fixed differences respectively. Some slightly deleterious mutations might not make it all the way to fixation but could well contribute to the level of polymorphisms (Eyre-Walker et al.

2002), an expectation that should manifest in higher d_N/d_S – ratios for polymorphisms than for fixed differences. A genome wide observation in accordance with this prediction has been made in human (ICSAC 2005). A similar study awaits to be done in birds, although there is a beautiful demonstration of the effect of population size on the rate of non synonymous change in island as compared to mainland bird populations (Johnson and Seger 2001).

Table 4. d_N/d_S - ratios for organisms with different population sizes. The bird analysis is based on 155 chicken and turkey orthologous gene pairs (**paper II**). The population size in chicken is estimated from θ (ICPMC 2004).

Species	d_N/d_S	N
H. sapiens – P. troglodytes	0.23	$\sim 10^4$
M. domesticus – R. norvegicus	0.12 – 0.13	$\sim 10^5$
D. pseudoobscura – D. melanogaster	0.08	$\sim 10^6$
G. gallus – M. gallopavo	0.13	$\sim 10^5$

Divergences from (ICSAC 2005), (IMGSC 2002), (Richards et al. 2005). Population size estimates from (Eyre-Walker et al. 2002).

Why are proteins evolving at different rates?

So far in this text, the discussion on evolutionary rate differences has focused on variation among lineages. In this context the importance of population size in increasing the efficiency of selection against slightly deleterious mutations was emphasized. As is obvious from the analysis of d_N/d_S – ratios for large sets of genes, there is large variation in the rate at which proteins of the same genome evolve. What is the cause of such rate variation? One obvious answer is that the proportion of amino acids under functional constraint varies among proteins. It may also be that organisms tolerate slight changes to the function of some proteins, while, on the other hand, altering other proteins would result in severe fitness reduction. Moreover, no matter how important a protein is, it usually needs a three dimensional structure to function. Rate variation could result from the fact that some structures may be built with various amino acid sequences while others require a specific sequence. Furthermore proteins may be specific in their function or take part in several different processes and it is easy to imagine that pleiotrophic proteins should be more constrained than others. Finally, the rate at which both transcription and translation take place varies across proteins, which in turn could impact on rates of protein evolution.

In addition to factors, like those mentioned above, that influence the strength of purifying selection, some functional classes of proteins are strongly affected by positive selection which leads to increased evolutionary rates.

Here I will briefly discuss the importance of some of these factors in birds using gene ontology terms and measurements of expression as starting points. In relation to this I will also discuss the effect of chromosomal location on rate of change.

Expression pattern and expression level

In yeast a negative correlation between expression level and the rate of protein evolution has been observed (Drummond et al. 2006). It was suggested that this correlation could account for almost half of the variation seen in divergence. The reason for this relationship could in turn be several. The function of a highly expressed gene could simply be more important than that of others, resulting in stronger purifying selection (Rocha 2006). Alternatively, highly expressed genes could have evolved to tolerate errors during translation (Drummond et al. 2006). This last hypothesis is motivated by the rather frequent occurrence of mistranslations and the sometimes harmful effects of misfolded protein aggregations in the cell. A negative correlation between expression levels and divergence have been reported in mammals (Subramanian and Kumar 2004), but there are conflicting results (Liao et al. 2006), perhaps indicating that this relationship is more important in yeast than in mammals (Pal et al. 2006).

In **paper V** we used the relative contribution of a particular transcript to the overall number of ESTs sequenced from different cDNA libraries as a measure of expression level in chicken. Expression data for a total of 4468 genes for which we estimated the divergence between chicken and zebra finch was available. There is a significant negative correlation between expression level and rate of protein evolution in these data ($r = -0.130$) (**paper V**).

As mentioned above, a protein might have to fulfill several functions or interact with many different proteins. This is likely the case for proteins expressed in several different tissues. As a consequence of such a pleiotrophy of functions, a stronger negative selection might be expected. A negative correlation between breadth of expression, as measured in the number of tissues in which a protein is expressed, and divergence has been noted in a diverse array of organisms (Zhang and Li 2004), (Pal et al. 2006), including chicken (ICGSC 2004). In **paper V** we too analysed this relationship using divergence data from chicken and zebra finch, and EST counts from various cDNA libraries. More specifically the EST data originate from 10 different chicken tissues and based on the relative abundance of gene specific ESTs in these tissues we were able to calculate an index of tissue specificity, τ (where $\tau = 1$ means a gene is expressed in only one tissue and $\tau = 0$ indicates that a gene is expressed in all sampled tissues), for a large set of chicken genes. In concordance with earlier studies we detected a positive relationship when analysing how d_N/d_S varies as a function of τ ($r=0.105$) (**paper V**).

In mammals a stronger relationship between expression breadth, rather than level, and divergence have been observed (Liao et al. 2006), (Pal et al. 2006). This is thus not confirmed in chicken (**paper V**). However, EST data represent a crude measure of both levels and breadth of expression, why these results may come to change when better methods are used.

Functional classes

Even if positive selection probably affects the overall rates of protein evolution little in most mammals (Bazykin et al. 2004) (ICSAC 2005), there are some types of genes that are expected to be more strongly affected. Genes that are involved in resistance to disease and/or in sexual reproduction are such examples. As pathogens rapidly develop new means of getting past the immune system of their hosts, there is a continual room for improvement of function in immune related genes. As for reproductive genes, intense sexual selection, perhaps in the form of male-male competition, is expected to lead to increased rates of evolution due to selection. Evidence for fast rates of amino acid replacements in male reproductive genes has been observed in several species including rodents and human (Wyckoff et al. 2000). In addition to genes involved in immune defence and gametogenesis, many genes belonging to functional classes such as sensory perception, apoptosis and transcription factors are evolving rapidly in human (Bustamante et al. 2005; Nielsen et al. 2005). As a consequence of this, in the human genome genes expressed exclusively in testis, thyroid and thymus show most evidence of positive selection (Nielsen et al. 2005).

Echoing some of these observations we find that chicken genes with maximum expression level in digestive and genitourinary tissues have higher average d_N/d_S – ratio than genes expressed elsewhere (**paper V**). To be able to functionally characterize rapidly evolving chicken and zebra finch genes in more detail, we studied gene ontology terms (GO-terms) associated to these genes. We find that genes with high d_N/d_S – ratios are overrepresented among GO-terms related to response to endogenous stimuli, DNA-repair and apoptosis (**paper V**), thus again, partly mirroring results from studies in human.

Avian adaptations

A long term goal of comparative genome analyses in birds is to understand avian adaptations at the molecular level. In **paper V** we present a first attempt at approaching this goal. Here we aligned 2686 1:1:1:1:1 orthologous gene quintets from chicken, zebra finch, opossum, mouse and human in order to compare rates of protein evolution in the avian and mammalian lineages. In total we find 186 genes that evolve at a significantly faster rate in birds than in the other lineages and 204 genes that evolve at a significantly slower rate in mammals as compared to the other lineages (**paper V**). Although it should be kept in mind that rapidly evolving bird genes may have

experienced periods of relaxed constraint instead of adaptive evolution; genes in these groups do represent candidate targets for positive selection during avian evolution. GO-terms overrepresented among these rapidly evolving bird genes include for instance skeletal development, lipid binding, pericentric region and microtubule (**paper V**). The molecular evolution of these genes should be investigated further using sequence data from several avian lineages, preferably in association with expression and functional analyses, before any conclusions can be drawn as to why these genes have evolved rapidly and what function they may have. Still we note that birds have light and hollow bone structures to allow for flight and a special lipid metabolism which facilitates energy demanding long-distance flights. In addition, the centromere, which belongs to the pericentric region, is believed to be involved in an arms race with the meiotic spindle, which in turn is built from microtubule (Henikoff and Malik 2002).

Chromosomal location and rate of evolution

Autosomal location

In **paper II** we studied the effect of chromosomal location on the rate of protein evolution, as measured in d_N/d_S. From sequence comparisons in 155 chicken and turkey orthologues gene pairs we found a significantly smaller d_N/d_S – ratio for genes located on microchromosomes (d_N/d_S = 0.094) as compared to genes on macrochromosomes (d_N/d_S = 0.185). Genes on the intermediately sized chromosomes displayed a mean d_N/d_S ratio (d_N/d_S = 0.105) close to that of microchromosomes but did not differ significantly from the average of machrochromosomal genes. These findings have been confirmed in a genome wide comparison of chicken and human (ICGSC 2004). Furthermore, in **paper V** 5542 genes evenly spread across the chicken genome were compared to their zebra finch orthologues and a similar pattern was observed. In this study the average d_N/d_S – ratios for genes of all three chromosome classes were found to differ significantly (macrochromosomes, d_N/d_S = 0.096; intermediate chromosomes, d_N/d_S = 0.090; microchromosomes = 0.071) (**paper V**). It is thus clear that the average level of constraint among avian autosomes varies so that genes on the frequently recombining and GC-rich microchromosomes on average evolve slowly. This finding is echoed by observations in mouse and rat where the rate of protein evolution was found to be negatively correlated with local GC-content (Williams and Hurst 2000). What causes these differences in rates of protein evolution?

It has been shown that the efficiency of selection is expected to vary among genomic regions as a function of the local recombination rate (Nordborg et al. 1996) (Hill and Robertson 1966). When two loci are physically linked selection acting at one locus will reduce the effective population size of the other, in turn leading to a reduction in the efficiency of selection

(Birky and Walsh 1988). Recombination has the potential to break up such linkage. In regions of low recombination linkage can thus lead to both an increased rate of fixation of deleterious alleles as well as a decreased rate of fixation of advantageous alleles (Birky and Walsh 1988). However, as most amino acid mutations are slightly deleterious (Li 1997), recombination might be expected to reduce the overall rate of protein evolution. In agreement with this we observe a negative correlation between rates of recombination and d_N/d_S (r= -0.162) (**paper II**). Consequently, the low average d_N/d_S – ratio of microchromosomal genes could perhaps be explained by the relatively high recombination rates of microchromosomes.

However, the efficiency of selection may vary little due to recombination in organisms with as high average recombination rates as birds and mammals (Nordborg et al. 1996). Furthermore genes on microchromosomes are separated by shorter intergenic distances and have shorter introns than genes on macrochromosomes (ICGSC 2004) meaning a potential effect of recombination would first have to overcome this tighter linkage in order to facilitate selection. Finally the negative correlation between recombination and d_N/d_S could be due to some other effects. In yeast it has been shown that genes in highly recombining regions often are expressed at high levels and after controlling for this effect, recombination and rate of evolution are no longer related (Pal et al. 2001).

Another reason for the difference in average d_N/d_S – ratios between avian chromosome classes could be a distinct distribution of genes with regard to function and/or expression pattern. In order to test this hypothesis we compared the tissue specificity, as measured in τ, of macro- and microchromosomal genes. We found that genes on microchromosomes on average are expressed in more tissues ($\tau = 0.537$) than genes on macrochromosomes ($\tau = 0.552$) (**paper V**). While two earlier studies failed to observe this difference (Hubbard et al. 2005; ICGSC 2004) we believe this could be due to differences in methodology or to limitations in size of previous data sets. Our finding thus indicates that microchromosomes are enriched for slowly evolving housekeeping genes which in turn offers a potential explanation to the difference in rate of evolution between micro- and macrochromosomes (**paper V**). Interestingly broadly expressed genes have been observed to cluster in the human genome (Lercher et al. 2002b).

Aiming at categorising genes of the two chromosomal classes further we mapped polymorphisms to a subset of the chicken – zebra finch orthologues. For each of these genes we then used an approach that resembles that of a MK-test and calculated the ratio $(D_N/D_S)/(P_N+1)/(P_S))$ where D_N is the number of non-synonymous substitutions, D_S is the number of synonymous substitutions, P_N is the number of non-synonymous polymorphisms and P_S is the number of synonymous polymorphisms. Here a ratio exceeding one may indicate that positive selection has contributed to the evolution of the gene. We analysed the chromosomal location of genes showing this sign of posi-

tive selection and found that a larger proportion resides on the macrochromosomes than on the microchromosomes (23,9% vs 19.4%) (**paper V**).

Finally, genes expressed specifically in certain chicken tissues do not seem to be preferentially located on micro- or macrochromosomes (Hubbard et al. 2005), an observation that is also mirrored in human (Lercher et al. 2002b). Moreover, when we compare the average level of expression of genes on macrochromosomes and microchromosomes we see no difference (**paper V**).

Differences in the rates of protein evolution between chromosomal classes of birds can thus potentially be explained by a combination of at least two factors: microchromosomes harbour many house-keeping genes while macrochromosomes have a higher proportion of genes that evolve under positive selection (**paper V**).

Chromosome Z

Recessivity is generally expected to lead to a relatively slow rate of adaptive protein evolution as beneficial effects of mutations cannot be expressed until frequencies are high enough for homozygotes to form. However, if recessive advantageous mutations arise on the X or Z chromosome their fitness effects will be fully exposed to selection in the hemizygous sex. As a consequence, if beneficial mutations are sufficiently common and recessive (h<0.5) to their nature, genes located on the hemizygote sex chromosome, like the Z chromosome of birds, might be expected to evolve faster than autosomal genes (Charlesworth and Coyne 1987). This hypothesis has been termed the fast-X theory. Evidence to support the predictions of this hypothesis has been observed in human and chimpanzee, where the average d_N/d_S – ratio of genes on chromosome X is higher than that of autosomal genes (Lu and Wu 2005), (ICSAC 2005). This observation still holds after taking a more efficient negative selection on synonymous sites on X into consideration (Lu and Wu 2005). However, several studies in Drosophila have so far failed to observe a fast-X effect (Betancourt et al. 2002) (Counterman et al. 2004) (Thornton et al. 2006). In order to test for a fast-Z effect, we compared the evolutionary rates of Z-linked and autosomal genes in chicken and zebra finch (**paper IV**). In total we gathered and estimated the d_N/d_S – ratios of 4874 autosomal and 176 Z chromosomal orthologous gene pairs in zebra finch and chicken. In this comparison we observed significantly elevated d_N/d_S – ratios for Z genes as compared to autosomal genes, a result in agreement with the fast-Z theory (**paper IV**). This finding could however also be due to a reduced efficiency of negative selection on Z as the population size of this chromosome is only ¾ of that of autosomes, and since recombination rates are reduced on Z (ICGSC 2004) (Birky and Walsh 1988). In order to test this possibility we compiled chicken polymorphism data from the two types of chromosomes (ICPMC 2004) with the intent to also look for an elevated ratio of d_N/d_S when measured for polymor-

phisms, on Z relative to autosomes. Finding this would indicate that slightly deleterious mutations make up a larger proportion of polymorphisms segregating on Z as compared to autosomes. This, in turn, would mean that our observation of high d_N/d_S – ratios for fixed differences on Z would be due to poor purging of deleterious alleles. We could however not see a difference in the d_N/d_S – ratio of polymorphic relative to fixed differences suggesting that an increased rate of adaptive evolution on Z explains our observations (**paper IV**).

Why should the signature of fast-X be observable in mammals and birds, but not in Drosophila? It has been suggested that the X-chromosomes may accumulate genes with sex-biased fitness effects (Rice 1987). This idea is based on the occurrence of antagonistic alleles, mutations with opposing effects on fitness in males and females. When dominant, a female beneficial mutation is here expected to go to fixation easier when X-linked, as compared to autosomal, since it will be subject to negative selection only one third, in contrast to half, of the time. If mutations have dominant fitness effects it might thus be expected that female biased genes should accumulate on X. On the other hand, when mutations are recessive, an accumulation of male biased genes on X is predicted as deleterious effects are masked at low frequencies in females, while at the same time the advantage is fully expressed in males. In human, genes expressed exclusively in males seem to be overrepresented on the X chromosome (Lercher et al. 2003), a finding hence in line with a recessive nature of many mutations. The same pattern holds also for mouse (Khil et al. 2004), while in D. simulans, an excess of genes with a female biased expression pattern is instead observed on the X chromosome (Ranz et al. 2003). Are dominance effects different between mammals and Drosophila? Birds show a pattern similar to that of D. simulans as chromosome Z has a deficit of female brain and ovary genes, and an excess of male brain genes (Kaiser and Ellegren 2006). It is unclear as to why these differences are observed. Nevertheless, in addition to the enhanced selection for recessive beneficial mutations on Z and X, a bias towards harbouring male specific genes could also add to the high average d_N/d_S – ratios observed on these chromosomes.

Conclusions

The forces responsible for evolution leave their marks in DNA sequences. These are marks, expressed as patterns of substituted nucleotides that can be observed when comparing orthologous sequences of different species. By contrasting observed patterns from different genomic regions or from different species it is possible to understand which factors that shapes evolution and their relative importance. The access to near complete genome sequences of many organisms makes this a powerful approach, but so far relatively few studies have been conducted in birds. This thesis brings new knowledge to the field of avian molecular evolution but as well to more general questions in molecular evolution.

In order to test for selection it is of big importance to know how rates of mutation vary across the genome and what neutral processes that affect the probabilities of selection. Unless this is clear, it is hard to argue that regions with unusual numbers of change are selectively important. In birds we show that GC-content predicts mutation rates very well and that this phenomenon explains why microchromosomes have accumulated more substitutions than macrochromosomes. The high numbers of hyper mutable CpG – dinucleotides in turn partly explain why GC-rich areas mutate often. This thesis thus represents a step forward in the effort to understand how mutation rates vary, especially in birds.

Rates of protein evolution differ among evolutionary lineages. As the fixation of slightly deleterious mutations is effected by population size, and since the level of constraint in avian lineages is relatively high, we draw the conclusion that historical population sizes of birds may have been larger than those of the mammalian lineages leading to dog and human respectively.

Except for in organisms with very large population sizes, like many species of the Drosophila genera, most amino acid changes are likely neutral or slightly deleterious. This probably holds for birds as well. We find few genes that show direct evidence for positive selection in birds although it must be kept in mind that our data set is biased towards slowly evolving genes. However, there are certain classes of genes that evolve fast. We characterise some of these classes in birds and find that genes expressed in digestive or genitourinary tissue and genes related to the biological functions: apoptosis, response to stress and response to endogenous stimuli are evolving fast.

In a first large scale attempt to classify adaptations that distinguish birds from mammals we characterize genes that evolve more rapidly in birds than

in other organisms. Among these genes we see that genes related to, among others, the gene ontology terms: skeletal development, lipid binding, microtubule and pericentric region are overrepresented.

Furthermore, genes located on the avian Z-chromosome on average evolve faster than autosomal genes. This is probably explained by the exposure of recessive beneficial alleles on this chromosome when in female birds.

Finally this thesis shows that genes expressed at high levels or in many chicken tissues evolve slower than other genes. The broadly expressed genes likely code for conserved housekeeping proteins and we find that microchromosomes are enriched for this type of genes, which in turn provides an explanation to why the average rate of protein evolution is slower on micro than macrochromosomes.

Future prospects

We are soon at a point were it is possible to understand how adaptations that distinguish birds from mammals evolved. With help of comparative analyses of genes from many bird species we may be able to understand how some birds have acquired the ability to learn and remember songs or to accurately navigate across continents and oceans. The evolution of flight will most likely have left molecular traces possible to track in a similar fashion. Furthermore, through analyses of genes responsible for the large variety of colour patterns in birds it will be interesting to study the role of sexual selection in avian evolution.

These interesting aspects of bird evolution can already be addressed to some degree but the sequencing of two genomes, those of the zebra finch and the anoles lizard will expand these possibilities enormously in the near future.

Access to a better data set of polymorphisms for at least one bird species would benefit to our understanding of the relative contributions of drift and selection to avian evolution. In addition such a data set would also make it possible to study the role of biased gene conversion in greater detail.
A second bird genome would also cast more light on the evolution of the genomic base composition in birds.

Svensk sammanfattning

Genom att jämföra ortologa DNA-sekvenser från olika arter kan vi lära oss om de krafter som formar evolutionen. Eftersom i stort sett kompletta DNA-sekvenser från ett flertal däggdjur, fiskar, insekter, svampar och växter har kartlagts är det nu möjligt att göra sådana jämförande analyser i stor skala. Än så länge har dock relativt få studier av fåglars arvsanlag presenterats. Fågelstudier är viktiga eftersom de ger en bild av hur evolutionen fortgår i en annan organism grupp än de vanligen undersökta däggdjuren och bananflugorna (*Drosophila* ssp.). Genom att jämföra observationer gjorda i fåglar med observationer från andra organismgrupper kan allmängiltigheten av tidigare observationer undersökas. I detta avseende bidrar fåglar med en intressant kontrast till andra utvecklingslinjer i form av en annorlunda kromosomuppsättning i allmänhet, och en omvänd könskromosomsuppsättning i synnerhet. Dessutom gör studier av fåglars DNA-sekvens det möjligt att undersöka frågor som rör fåglars biologi specifikt; vilka gener har t.ex. betytt mest för utvecklingen av de egenskaper som är karakteristiska för fåglar? I den här avhandlingen presenteras framsteg inom området fåglars molekylära evolution.

Grunden för all evolution är den variation som skapas i samband med mutationer. Samtidigt verkar paradoxalt nog de flesta mutationer som leder till en förändrad proteinsekvens vara skadliga. En del av den här avhandlingen behandlar frågor som rör mutationers uppkomst. Hur ofta sker mutationer och sker dom lika ofta över hela arvsanlaget? Genom att jämföra ackumuleringen av substitutioner på könskromosomerna Z och W har det tidigare visats att fler mutationer har sitt ursprung i hanar än i honor. Anledningen till detta skeva mönster är förmodligen att tillverkningen av spermier kräver betydligt fler celldelningar än tillverkningen av ägg och att mutationer framförallt uppstår när DNA replikeras, vilket sker i samband med just celldelning. Om DNA muterar ofta när det befinner sig i hanlig könscellslinje borde Z-kromosomen, som under två tredjedelar av tiden är i hanar, förändras fortare än autosomer eftersom dessa finns i honor och hanar lika ofta. Teoretiska argument har dock framställts som förutspår att mutationshastigheten på just Z-kromosomen borde vara låg. Anledningen skulle vara att recessiva, skadliga mutationer på denna kromosom uttrycks när dom befinner sig i honliga fåglar, något som inte sker med autosomala mutationer. Vi kan dock inte finna bevis för en reduktion av mutationshastigheten på Z. Genom den-

na studie understryks därför betydelsen av replikering för uppkomsten av nya mutationer.

I tillägg till de skilda mutationshastigheter som tidigare har uppmätts mellan könskromosomer och mellan könskromosomer och autosomer ser vi att olika autosomala DNA-sekvenser också uppvisar olika mutationshastigheter. Vi visar att denna variation till stor del kan förutsägas med hjälp av sekvensers bassammansättning då ett högt GC-innehåll sammanfaller med en hög mutationshastighet. Detta fenomen kan i sin tur till viss del tillskrivas den relativt höga närvaron av snabbt muterande CpG-nukleotider i områden som är rika på C och G. Vidare förklarar denna observation dessutom varför små, s.k. mikrokromosomer, har en högre mutationshastighet än de större makrokromosomerna.

Vidare ser vi att lokala substitutionsmönster i kycklinggenomet bidrar till att bevara, eller kanske till och med förstärka, den regionala variationen i bassammansättning. Detta är tvärt emot observationer i många däggdjur där GC-rika regioner sakta blir mer AT-rika så att variationen i bassammansättningen där förväntas suddas ut. Vi spekulerar i om detta skulle kunna förklaras av en långsammare förändring av kromosomstrukturen i fåglar än i däggdjur.

För att mutationer ska leda till evolutionär förändring krävs normalt att de ersätter andra alleler i populationen. De ska med andra ord fixeras. Detta kan ske med hjälp av slumpen, så kallad genetisk drift, eller genom selektion. Ytterliggare fokus för den här avhandlingen är dels vilka faktorer som styr fixeringssannolikheter, dels vilka grupper av gener som varit särskilt viktiga för evolutionen av fåglar.

Eftersom det antas att måttligt skadliga mutationer selekteras bort mer effektivt i stora än små populationer har vi undersökt hur fort proteiner förändras i kyckling och zebrafink jämfört med i hund och människa. Baserat på denna jämförelse drar vi slutsatsen att fåglars populationsstorlekar historiskt sett kan ha varit större än för de av däggdjurens utvecklingslinjer som leder till hund respektive människa. Vidare finner vi relativt få tecken på adaptiv evolution i ~5000 gener som är uttryckta i zebrafinkens hjärna. Men eftersom gener uttryckta i hjärna utvecklas långsammare än många andra gener i andra utvecklingslinjer är det svårt att dra några generella slutsatser om betydelsen av positiv selektion för fåglars evolution baserat på dessa resultat. Däremot kan vi se att vissa grupper av gener utvecklas fortare än andra, vilket kan tolkas som att adaptiv evolution har spelat en stor roll för dessa geners utveckling. Till de grupper av gener som utvecklas fort hör de som är uttryckta i vävnader med matsmältnings- samt reproduktionsrelaterad funktion. Dessutom visar vi att gener som kontrollerar celldöd samt gener som styr svar på stress och endogena stimuli (termern stress och endogena stimuli verkar i det här fallet syfta på gener som aktiveras av skador på DNA-sekvensen) har förändrats snabbt under fåglarnas evolution. Vidare utvecklas gener som är Z-bundna fortare än autosomala. Vi tror att den sistnämnda

observationen beror på att recessiva, fördelaktiga mutationer uttrycks i honor om dom finns på Z, men inte på autosomer.

I ett första storskaligt försök att på molekylär nivå kartlägga adaptiva förändringar som har bidragit till fenotypiska drag som är karakteristiska för fåglar har vi jämfört evolutionen av ca 2400 ortologa gener i kyckling, zebrafink, opossum, mus och människa. Gener som har utvecklats fortare i fåglar än i andra utvecklingslinjer representerar sekvenser som kan ha påverkats av positiv selektion i fåglar men ej i däggdjur. Bland dessa gener märker vi bland annat en överrepresentation av gener som styr skelettets utveckling, lipidbindning samt gener som är involverade i centromerernas funktion.

Tillsist ser vi att den genomsnittliga hastigheten av aminosyraförändringar, relaterat till den lokala mutationshastigheten, är lägre på mikro- än makrokromosomer. Vi visar att detta delvis kan förklaras genom att gener på mikrokromosomer generellt uttrycks i fler vävnader än gener på makrokromosomer.

Acknowledgements

Tack:

Hans Ellegren för fin handledning, bra idéer, hög ambitionsnivå, full fart framåt och för att jag fick jobba här. Det har varit mycket givande!

Matthew Webster för alla diskussioner, mycket hjälp, inspiration, kunnande och för att du är en god vän.

Nick Smith för kunnande och inspiration.

Datagurun Mikael Brandström för mycket värdefull programmeringshjälp. Inte ett felmeddelande är okänt för dig vid det här laget.

Nicklas, Sofia, Lina, Lujiang, Martin, Judith, Mikael och alla andra och tidigare medlemmar av gruppen för molekylär evolution. Det är kul att jobba med er! Ett extra tack till Johanna för hjälp i labbet, samt till Lina för dina insatser mot slutet av den här tiden.

Cia, Frank och Mikael för att det har gått så bra att dela rum med er trots att jag förlorade lottdragningarna.

Norska konsulatet – Jon, Tomas, Glenn-Peter, Henrik, Ståle, Stein-Are, Annika (även om du är svensk) och Öystein för allt skoj! Heja Norge!

Lunchklubben – Markus, Fredrik, Frank, Öystein och Stein-Are för kulinariska utflykter från trädgårdsland i Uppsala till norska fjällfår, pölsa och överkokt pasta med tomatsås.

Alla andra trevliga människor på avdelningen och hela EBC!

Kameleonterna för att inget hände under långa stunder.

Martin C. för viktiga fikapauser.

Adriana för den supersnabba framsidesbilden

Sven och Lilly Lawskis fond, Helge Ax:son-Johnsons stiftelse, KVA, Liljewalch och Wilhelm Zethelius för finansiellt stöd.

EBC-iskalla för motionen. Det viktigaste är inte att vinna utan att spela så jämt så möjligt med så få avbytare som möjligt.

Alla fredagsbandyspelare för att ni har bidragit till det perfekta avslutet på en arbetsvecka.

Gröna Volvon med delägarna Jon och Matthew för kalla mek-helger, p-böter och utflykter med oviss utgång.

Johan C., Johan F., Erik, Tomas, Petra, Anna, Anna-Maria, Mattias, Cissi, Torbjörn, Linda, Patrik, Sandro, Jenny, Gustav, Jan samt allt Karlsöfolk för skidexpeditioner, kajakturer, ö-utflykter, resor, vintillverkning, spelkvällar och all annat som är kul!

Weronika och Pissen för att ni har gjort denna långa mörka höst riktigt trevlig och för många goda råd som gjort arbetet mindre svårt.

Min familj; Göran, Vivan, Lars, Pernilla och Ivan för massor av ovärderligt stöd och avkoppling under alla år.

References

Alberts, B., D. Bray, J. Lewis, M. Raff, K. Roberts, and J.D. Watson. 1994. *The Cell*. Garland Publishing, New York.

Arndt, P.F., T. Hwa, and D.A. Petrov. 2005. Substantial regional variation in substitution rates in the human genome: importance of GC content, gene density, and telomere-specific effects. *J Mol Evol* **60:** 748-763.

Arndt, P.F., D.A. Petrov, and T. Hwa. 2003. Distinct changes of genomic biases in nucleotide substitution at the time of Mammalian radiation. *Mol Biol Evol* **20:** 1887-1896.

Bartosch-Harlid, A., S. Berlin, N.G. Smith, A.P. Moller, and H. Ellegren. 2003. Life history and the male mutation bias. *Evolution Int J Org Evolution* **57:** 2398-2406.

Bauer, V.L. and C.F. Aquadro. 1997. Rates of DNA sequence evolution are not sex-biased in Drosophila melanogaster and D. simulans. *Mol Biol Evol* **14:** 1252-1257.

Bazykin, G.A., F.A. Kondrashov, A.Y. Ogurtsov, S. Sunyaev, and A.S. Kondrashov. 2004. Positive selection at sites of multiple amino acid replacements since rat-mouse divergence. *Nature* **429:** 558-562.

Bernardi, G., B. Olofsson, J. Filipski, M. Zerial, J. Salinas, G. Cuny, M. Meunier-Rotival, and F. Rodier. 1985. The mosaic genome of warm-blooded vertebrates. *Science* **228:** 953-958.

Betancourt, A.J., D.C. Presgraves, and W.J. Swanson. 2002. A test for faster X evolution in Drosophila. *Molecular Biology and Evolution* **19:** 1816-1819.

Bierne, N. and A. Eyre-Walker. 2004. The genomic rate of adaptive amino acid substitution in Drosophila. *Molecular Biology and Evolution* **21:** 1350-1360.

Birky, C.W. and J.B. Walsh. 1988. Effects of Linkage on Rates of Molecular Evolution. *Proceedings of the National Academy of Sciences of the United States of America* **85:** 6414-6418.

Bohossian, H.B., H. Skaletsky, and D.C. Page. 2000. Unexpectedly similar rates of nucleotide substitution found in male and female hominids. *Nature* **406:** 622-625.

Brown, T.C. and J. Jiricny. 1988. Different base/base mispairs are corrected with different efficiencies and specificities in monkey kidney cells. *Cell* **54:** 705-711.

Bustamante, C.D., A. Fledel-Alon, S. Williamson, R. Nielsen, M.T. Hubisz, S. Glanowski, D.M. Tanenbaum, T.J. White, J.J. Sninsky, R.D. Hernandez, D. Civello, M.D. Adams, M. Cargill, and A.G. Clark. 2005. Natural selection on protein-coding genes in the human genome. *Nature* **437:** 1153-1157.

Charlesworth, B. and J.A. Coyne. 1987. The relative rates of evolution of sex chromosomes and autosomes. *Am. Nat.* **130:** 113-146.

Counterman, B.A., D. Ortiz-Barrientos, and M.A.F. Noor. 2004. Using comparative genomic data to test for fast-x evolution. *Evolution* **58:** 656-660.

Donaldson, A.D. 2005. Shaping time: chromatin structure and the DNA replication programme. *Trends Genet* **21:** 444-449.
Drummond, D.A., A. Raval, and C.O. Wilke. 2006. A single determinant dominates the rate of yeast protein evolution. *Mol Biol Evol* **23:** 327-337.
Duret, L. and D. Mouchiroud. 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Mol Biol Evol* **17:** 68-74.
Duret, L., M. Semon, G. Piganeau, D. Mouchiroud, and N. Galtier. 2002. Vanishing GC-rich isochores in mammalian genomes. *Genetics* **162:** 1837-1847.
Ellegren, H. and A.K. Fridolfsson. 1997. Male-driven evolution of DNA sequences in birds. *Nat Genet* **17:** 182-184.
Eyre-Walker, A. 1993. Recombination and mammalian genome evolution. *Proc Biol Sci* **252:** 237-243.
Eyre-Walker, A. 1998. Problems with parsimony in sequences of biased base composition. *J Mol Evol* **47:** 686-690.
Eyre-Walker, A. 2002. Changing effective population size and the McDonald-Kreitman test. *Genetics* **162:** 2017-2024.
Eyre-Walker, A. and L.D. Hurst. 2001. The evolution of isochores. *Nat Rev Genet* **2:** 549-555.
Eyre-Walker, A., P.D. Keightley, N.G. Smith, and D. Gaffney. 2002. Quantifying the slightly deleterious mutation model of molecular evolution. *Mol Biol Evol* **19:** 2142-2149.
Fay, J.C., G.J. Wyckoff, and C.I. Wu. 2001. Positive and negative selection on the human genome. *Genetics* **158:** 1227-1234.
Fay, J.C., G.J. Wyckoff, and C.I. Wu. 2002. Testing the neutral theory of molecular evolution with genomic data from Drosophila. *Nature* **415:** 1024-1026.
Filatov, D.A. 2004. A gradient of silent substitution rate in the human pseudoautosomal region. *Mol Biol Evol* **21:** 410-417.
Filatov, D.A. and D. Charlesworth. 2002. Substitution rates in the X- and Y-linked genes of the plants, Silene latifolia and S. dioica. *Mol Biol Evol* **19:** 898-907.
Filipski, J. 1987. Correlation between molecular clock ticking, codon usage fidelity of DNA repair, chromosome banding and chromatin compactness in germline cells. *FEBS Lett* **217:** 184-186.
Gaffney, D.J. and P.D. Keightley. 2005. The scale of mutational variation in the murid genome. *Genome Res* **15:** 1086-1094.
Gilbert, N., S. Boyle, H. Fiegler, K. Woodfine, N.P. Carter, and W.A. Bickmore. 2004. Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers. *Cell* **118:** 555-566.
Goetting-Minesky, M.P. and K.D. Makova. 2006. Mammalian male mutation bias: impacts of generation time and regional variation in substitution rates. *J Mol Evol* **63:** 537-544.
Green, P., B. Ewing, W. Miller, P.J. Thomas, and E.D. Green. 2003. Transcription-associated mutational asymmetry in mammalian evolution. *Nat Genet* **33:** 514-517.
Griffin, D.K., L.B.W. Robertson, H.G. Tempest, and B.M. Skinner. unpublished. The evolution of the avian genome as revealed by comparative molecular cytogenetics.
Haldane, B.S. 1947. The mutation rate of the gene haemophilia, and its segregation ratios in males and females. *Annual Eugenetices* **13:** 262-271.
Hardison, R.C., K.M. Roskin, S. Yang, M. Diekhans, W.J. Kent, R. Weber, L. Elnitski, J. Li, M. O'Connor, D. Kolbe, S. Schwartz, T.S. Furey, S. Whelan, N. Goldman, A. Smit, W. Miller, F. Chiaromonte, and D. Haussler. 2003. Covaria-

tion in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res* **13:** 13-26.
Hellmann, I., I. Ebersberger, S.E. Ptak, S. Paabo, and M. Przeworski. 2003. A neutral explanation for the correlation of diversity with recombination rates in humans. *Am J Hum Genet* **72:** 1527-1535.
Henikoff, S. and H.S. Malik. 2002. Centromeres: selfish drivers. *Nature* **417:** 227.
Hill, W.G. and A. Robertson. 1966. The effect of linkage on limits to artificial selection. *Genet Res* **8:** 269-294.
Hubbard, S.J., D.V. Grafham, K.J. Beattie, I.M. Overton, S.R. McLaren, M.D. Croning, P.E. Boardman, J.K. Bonfield, J. Burnside, R.M. Davies, E.R. Farrell, M.D. Francis, S. Griffiths-Jones, S.J. Humphray, C. Hyland, C.E. Scott, H. Tang, R.G. Taylor, C. Tickle, W.R. Brown, E. Birney, J. Rogers, and S.A. Wilson. 2005. Transcriptome analysis for the chicken based on 19,626 finished cDNA sequences and 485,337 expressed sequence tags. *Genome Res* **15:** 174-183.
Hughes, S., D. Zelus, and D. Mouchiroud. 1999. Warm-blooded isochore structure in Nile crocodile and turtle. *Mol Biol Evol* **16:** 1521-1527.
Hurst, L.D. and H. Ellegren. 1998. Sex biases in the mutation rate. *Trends Genet* **14:** 446-452.
Hurst, L.D. and E.J. Williams. 2000. Covariation of GC content and the silent site substitution rate in rodents: implications for methodology and for the evolution of isochores. *Gene* **261:** 107-114.
Hwang, D.G. and P. Green. 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc Natl Acad Sci U S A* **101:** 13994-14001.
ICGSC. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432:** 695-716.
ICPMC. 2004. A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. *Nature* **432:** 717-722.
ICSAC. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437:** 69-87.
IDGSC. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438:** 803-819.
IHGSC. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.
IMGSC. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520-562.
IRGSC. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428:** 493-521.
Johnson, K.P. and J. Seger. 2001. Elevated rates of nonsynonymous substitution in island birds. *Molecular Biology and Evolution* **18:** 874-881.
Kaiser, V.B. and H. Ellegren. 2006. Nonrandom distribution of genes with sex-biased expression in the chicken genome. *Evolution* **60:** 1945-1951.
Khil, P.P., N.A. Smirnova, P.J. Romanienko, and R.D. Camerini-Otero. 2004. The mouse X chromosome is enriched for sex-biased genes not subject to selection by meiotic sex chromosome inactivation. *Nature Genetics* **36:** 642-646.
Kimura, M. 1968. Evolutionary rate at the molecular level. *Nature* **217:** 624-626.
Kimura, M. 1977. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* **267:** 275-276.
King, J.L. and T.H. Jukes. 1969. Non-Darwinian evolution. *Science* **164:** 788-798.
Kondrashov, A.S. 2003. Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Hum Mutat* **21:** 12-27.

Kong, A., D.F. Gudbjartsson, J. Sainz, G.M. Jonsdottir, S.A. Gudjonsson, B. Richardsson, S. Sigurdardottir, J. Barnard, B. Hallbeck, G. Masson, A. Shlien, S.T. Palsson, M.L. Frigge, T.E. Thorgeirsson, J.R. Gulcher, and K. Stefansson. 2002. A high-resolution recombination map of the human genome. *Nat Genet* **31:** 241-247.

Kuma, K., N. Iwabe, and T. Miyata. 1995. Functional constraints against variations on molecules from the tissue level: slowly evolving brain-specific genes demonstrated by protein kinase and immunoglobulin supergene families. *Mol Biol Evol* **12:** 123-130.

Kumar, S. and S. Subramanian. 2002. Mutation rates in mammalian genomes. *Proc Natl Acad Sci U S A* **99:** 803-808.

Kunkel, T.A. 2004. DNA replication fidelity. *J Biol Chem* **279:** 16895-16898.

Kunkel, T.A. and P.S. Alexander. 1986. The base substitution fidelity of eucaryotic DNA polymerases. Mispairing frequencies, site preferences, insertion preferences, and base substitution by dislocation. *J Biol Chem* **261:** 160-166.

Kunz, B.A. and S.E. Kohalmi. 1991. Modulation of mutagenesis by deoxyribonucleotide levels. *Annu Rev Genet* **25:** 339-359.

Lawson, L.J. and G.M. Hewitt. 2002. Comparison of substitution rates in ZFX and ZFY introns of sheep and goat related species supports the hypothesis of male-biased mutation rates. *J Mol Evol* **54:** 54-61.

LeClerc, J.E., B. Li, W.L. Payne, and T.A. Cebula. 1996. High mutation frequencies among Escherichia coli and Salmonella pathogens. *Science* **274:** 1208-1211.

Leigh, E.G. 1970. Natural selection and mutability. *The american naturalist* **104:** 301-305.

Lercher, M.J. and L.D. Hurst. 2002. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet* **18:** 337-340.

Lercher, M.J., N.G. Smith, A. Eyre-Walker, and L.D. Hurst. 2002a. The evolution of isochores: evidence from SNP frequency distributions. *Genetics* **162:** 1805-1810.

Lercher, M.J., A.O. Urrutia, and L.D. Hurst. 2002b. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat Genet* **31:** 180-183.

Lercher, M.J., A.O. Urrutia, and L.D. Hurst. 2003. Evidence that the human X chromosome is enriched for male-specific but not female-specific genes. *Molecular Biology and Evolution* **20:** 1113-1116.

Lercher, M.J., E.J. Williams, and L.D. Hurst. 2001. Local similarity in evolutionary rates extends over whole chromosomes in human-rodent and mouse-rat comparisons: implications for understanding the mechanistic basis of the male mutation bias. *Mol Biol Evol* **18:** 2032-2039.

Li, W.-H. 1997. *Molecular Evolution*. Sinauer Associates, Inc., Sunderland.

Liao, B.Y., N.M. Scott, and J. Zhang. 2006. Impacts of gene essentiality, expression pattern, and gene compactness on the evolutionary rate of mammalian proteins. *Mol Biol Evol* **23:** 2072-2080.

Lu, J. and C.I. Wu. 2005. Weak selection revealed by the whole-genome comparison of the X chromosome and autosomes of human and chimpanzee. *Proc Natl Acad Sci U S A* **102:** 4063-4067.

Makova, K.D. and W.H. Li. 2002. Strong male-driven evolution of DNA sequences in humans and apes. *Nature* **416:** 624-626.

Martomo, S.A. and C.K. Mathews. 2002. Effects of biological DNA precursor pool asymmetry upon accuracy of DNA replication in vitro. *Mutat Res* **499:** 197-211.

Matassi, G., P.M. Sharp, and C. Gautier. 1999. Chromosomal location effects on gene sequence evolution in mammals. *Curr Biol* **9:** 786-791.

Mathews, C.K. 2006. DNA precursor metabolism and genomic stability. *Faseb J* **20:** 1300-1314.
McDonald, J.H. and M. Kreitman. 1991. Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351:** 652-654.
McVean, G.T. and L.D. Hurst. 1997. Evidence for a selectively favourable reduction in the mutation rate of the X chromosome. *Nature* **386:** 388-392.
Meunier, J. and L. Duret. 2004. Recombination Drives the Evolution of GC-Content in the Human Genome. *Mol Biol Evol* **21:** 984-990.
Miyata, T., H. Hayashida, K. Kuma, K. Mitsuyasu, and T. Yasunaga. 1987. Male-driven molecular evolution: a model and nucleotide sequence analysis. *Cold Spring Harb Symp Quant Biol* **52:** 863-867.
Montoya-Burgos, J.I., P. Boursot, and N. Galtier. 2003. Recombination explains isochores in mammalian genomes. *Trends Genet* **19:** 128-130.
Nachman, M.W. and S.L. Crowell. 2000. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156:** 297-304.
Nekrutenko, A. and W.H. Li. 2000. Assessment of compositional heterogeneity within and between eukaryotic genomes. *Genome Res* **10:** 1986-1995.
Nielsen, R., C. Bustamante, A.G. Clark, S. Glanowski, T.B. Sackton, M.J. Hubisz, A. Fledel-Alon, D.M. Tanenbaum, D. Civello, T.J. White, J.S. J, M.D. Adams, and M. Cargill. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol* **3:** e170.
Nilsson, M. and N. Snoad. 2002. Optimal mutation rates in dynamic environments. *Bull Math Biol* **64:** 1033-1043.
Nordborg, M., B. Charlesworth, and D. Charlesworth. 1996. The effect of recombination on background selection. *Genet Res* **67:** 159-174.
Ohta, T. 1973. Slightly deleterious mutant substitutions in evolution. *Nature* **246:** 96-98.
Pal, C., B. Papp, and L.D. Hurst. 2001. Does the recombination rate affect the efficiency of purifying selection? The yeast genome provides a partial answer. *Mol Biol Evol* **18:** 2323-2326.
Pal, C., B. Papp, and M.J. Lercher. 2006. An integrated view of protein evolution. *Nat Rev Genet* **7:** 337-348.
Pardo-Manuel de Villena, F. and C. Sapienza. 2001. Recombination is proportional to the number of chromosome arms in mammals. *Mamm Genome* **12:** 318-322.
Pecon Slattery, J. and S.J. O'Brien. 1998. Patterns of Y and X chromosome DNA sequence divergence during the Felidae radiation. *Genetics* **148:** 1245-1255.
Piganeau, G., D. Mouchiroud, L. Duret, and C. Gautier. 2002. Expected relationship between the silent substitution rate and the GC content: implications for the evolution of isochores. *J Mol Evol* **54:** 129-133.
Ranz, J.M., C.I. Castillo-Davis, C.D. Meiklejohn, and D.L. Hartl. 2003. Sex-dependent gene expression and evolution of the Drosophila transcriptome. *Science* **300:** 1742-1745.
Rice, W.R. 1987. The Accumulation of Sexually Antagonistic Genes as a Selective Agent Promoting the Evolution of Reduced Recombination between Primitive Sex-Chromosomes. *Evolution* **41:** 911-914.
Richards, S., Y. Liu, B.R. Bettencourt, P. Hradecky, S. Letovsky, R. Nielsen, K. Thornton, M.J. Hubisz, R. Chen, R.P. Meisel, O. Couronne, S. Hua, M.A. Smith, P. Zhang, J. Liu, H.J. Bussemaker, M.F. van Batenburg, S.L. Howells, S.E. Scherer, E. Sodergren, B.B. Matthews, M.A. Crosby, A.J. Schroeder, D. Ortiz-Barrientos, C.M. Rives, M.L. Metzker, D.M. Muzny, G. Scott, D. Steffen, D.A. Wheeler, K.C. Worley, P. Havlak, K.J. Durbin, A. Egan, R. Gill, J. Hume, M.B. Morgan, G. Miner, C. Hamilton, Y. Huang, L. Waldron, D. Verduzco,

K.P. Clerc-Blankenburg, I. Dubchak, M.A. Noor, W. Anderson, K.P. White, A.G. Clark, S.W. Schaeffer, W. Gelbart, G.M. Weinstock, and R.A. Gibbs. 2005. Comparative genome sequencing of Drosophila pseudoobscura: chromosomal, gene, and cis-element evolution. *Genome Res* **15:** 1-18.

Rocha, E.P. 2006. The quest for the universals of protein evolution. *Trends Genet* **22:** 412-416.

Silva, J.C. and A.S. Kondrashov. 2002. Patterns in spontaneous mutation revealed by human-baboon sequence comparison. *Trends Genet* **18:** 544-547.

Smith, N.G. and A. Eyre-Walker. 2002. Adaptive protein evolution in Drosophila. *Nature* **415:** 1022-1024.

Smith, N.G., M.T. Webster, and H. Ellegren. 2002. Deterministic mutation rate variation in the human genome. *Genome Res* **12:** 1350-1356.

Sniegowski, P.D., P.J. Gerrish, T. Johnson, and A. Shaver. 2000. The evolution of mutation rates: separating causes from consequences. *Bioessays* **22:** 1057-1066.

Sniegowski, P.D., P.J. Gerrish, and R.E. Lenski. 1997. Evolution of high mutation rates in experimental populations of E. coli. *Nature* **387:** 703-705.

Spencer, C.C., P. Deloukas, S. Hunt, J. Mullikin, S. Myers, B. Silverman, P. Donnelly, D. Bentley, and G. McVean. 2006. The influence of recombination on human genetic diversity. *PLoS Genet* **2:** e148.

Strathern, J.N., B.K. Shafer, and C.B. McGill. 1995. DNA synthesis errors associated with double-strand-break repair. *Genetics* **140:** 965-972.

Subramanian, S. and S. Kumar. 2003. Neutral substitutions occur at a faster rate in exons than in noncoding DNA in primate genomes. *Genome Res* **13:** 838-844.

Subramanian, S. and S. Kumar. 2004. Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. *Genetics* **168:** 373-381.

Taylor, J., S. Tyekucheva, M. Zody, F. Chiaromonte, and K.D. Makova. 2006. Strong and weak male mutation bias at different sites in the primate genomes: insights from the human-chimpanzee comparison. *Mol Biol Evol* **23:** 565-573.

Thornton, K., D. Bachtrog, and P. Andolfatto. 2006. X chromosomes and autosomes evolve at similar rates in Drosophila: No evidence for faster-X protein evolution. *Genome Research* **16:** 498-504.

Trasler, J.M. 1998. Origin and roles of genomic methylation patterns in male germ cells. *Semin Cell Dev Biol* **9:** 467-474.

Webster, M.T., N.G. Smith, and H. Ellegren. 2003. Compositional evolution of noncoding DNA in the human and chimpanzee genomes. *Mol Biol Evol* **20:** 278-286.

Webster, M.T., N.G. Smith, M.J. Lercher, and H. Ellegren. 2004. Gene expression, synteny, and local similarity in human noncoding mutation rates. *Mol Biol Evol* **21:** 1820-1830.

Williams, E.J. and L.D. Hurst. 2000. The proteins of linked genes evolve at similar rates. *Nature* **407:** 900-903.

Vinogradov, A.E. 2005. Noncoding DNA, isochores and gene expression: nucleosome formation potential. *Nucleic Acids Res* **33:** 559-563.

Wolfe, K.H. 1991. Mammalian DNA replication: mutation biases and the mutation rate. *J Theor Biol* **149:** 441-451.

Wolfe, K.H., P.M. Sharp, and W.H. Li. 1989. Mutation rates differ among regions of the mammalian genome. *Nature* **337:** 283-285.

Woodfine, K., D.M. Beare, K. Ichimura, S. Debernardi, A.J. Mungall, H. Fiegler, V.P. Collins, N.P. Carter, and I. Dunham. 2005. Replication timing of human chromosome 6. *Cell Cycle* **4:** 172-176.

Woodfine, K., H. Fiegler, D.M. Beare, J.E. Collins, O.T. McCann, B.D. Young, S. Debernardi, R. Mott, I. Dunham, and N.P. Carter. 2004. Replication timing of the human genome. *Hum Mol Genet* **13:** 191-202.

Wyckoff, G.J., W. Wang, and C.I. Wu. 2000. Rapid evolution of male reproductive genes in the descent of man. *Nature* **403:** 304-309.

Zhang, L. and W.H. Li. 2005. Human SNPs reveal no evidence of frequent positive selection. *Mol Biol Evol* **22:** 2504-2507.

Zhang, L.Q. and W.H. Li. 2004. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Molecular Biology and Evolution* **21:** 236-239.

Zhao, Z. and E. Boerwinkle. 2002. Neighboring-nucleotide effects on single nucleotide polymorphisms: a study of 2.6 million polymorphisms across the human genome. *Genome Res* **12:** 1679-1686.

Zuckerkandl, E. and L. Pauling. 1965. Molecules as documents of evolutionary history. *J Theor Biol* **8:** 357-366.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 261

Editor: The Dean of the Faculty of Science and Technology

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. (Prior to January, 2005, the series was published under the title "Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology".)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-7432

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2007