Bioinformatic Analysis of Mutation and Selection in the Vertebrate Non-coding Genome

MIKAEL BRANDSTRÖM
Dissertation presented at Uppsala University to be publicly examined in Zootissalen, EBC, Villavägen 9, Uppsala, Wednesday, October 31, 2007 at 09:00 for the degree of Doctor of Philosophy. The examination will be conducted in English.

Abstract

The majority of the vertebrate genome sequence is not coding for proteins. In recent years, the evolution of this noncoding fraction of the genome has gained interest. These studies have been greatly facilitated by the availability of full genome sequences. The aim of this thesis is to study evolution of the noncoding vertebrate genome through bioinformatic analysis of large-scale genomic datasets.

In a first analysis we addressed the use of conservation of sequence between highly diverged genomes to infer function. We provided evidence for a turnover of the patterns of negative selection. Hence, measures of constraint based on comparisons of diverged genomes might underestimate the functional proportion of the genome.

In the following analyses we focused on length variation as found in small-scale insertion and deletion (indel) polymorphisms and microsatellites. For indels in chicken, replication slippage is a likely mutation mechanism, as a large proportion of the indels are parts of tandem-duplicates. Using a set of microsatellite polymorphisms in chicken, where we avoid ascertainment bias, we showed that polymorphism is positively correlated with microsatellite length and AT-content. Furthermore, interruptions in the microsatellite sequence decrease the levels of polymorphism.

We also analysed the association between microsatellite polymorphism and recombination in the human genome. Here we found increased levels of microsatellite polymorphism in human recombination hotspots and also similar increases in the frequencies of single nucleotide polymorphisms (SNPs) and indels. This points towards natural selection shaping the levels of variation. Alternatively, recombination is mutagenic for all three kinds of polymorphisms.

Finally, I present the program ILAPlot. It is a tool for visualisation, exploration and data extraction based on BLAST.

Our combined results highlight the intricate connections between evolutionary phenomena. It also emphasises the importance of length variability in genome evolution, as well as the gradual difference between indels and microsatellites.

Keywords: non-coding genome, selection, mutation, chicken, human, indel, microsatellite, sequence alignment, comparative genomics, bioinformatics, vertebrate

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

© Mikael Brandström 2007

ISSN 1651-6214
ISBN 978-91-554-6981-8
urn:nbn:se:uu:diva-8240 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-8240)
List of papers

The thesis is based on the following papers, hereafter referred to by their roman numerals:


Paper I and II are reproduced with permission from the publisher.
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>9</td>
</tr>
<tr>
<td>Comparative genomics</td>
<td>11</td>
</tr>
<tr>
<td>Comparing the appropriate sequences</td>
<td>11</td>
</tr>
<tr>
<td>Bioinformatics</td>
<td>13</td>
</tr>
<tr>
<td>The non-coding genome</td>
<td>15</td>
</tr>
<tr>
<td>Single copy DNA</td>
<td>15</td>
</tr>
<tr>
<td>Selection on the non-coding genome</td>
<td>16</td>
</tr>
<tr>
<td>Positive selection on non-coding sequences</td>
<td>17</td>
</tr>
<tr>
<td>Function of non-coding sequences</td>
<td>17</td>
</tr>
<tr>
<td>Repetitive sequences</td>
<td>18</td>
</tr>
<tr>
<td>Microsatellites</td>
<td>19</td>
</tr>
<tr>
<td>Sequence polymorphism</td>
<td>21</td>
</tr>
<tr>
<td>Small scale sequence length polymorphisms</td>
<td>22</td>
</tr>
<tr>
<td>Short insertion and deletion polymorphisms</td>
<td>23</td>
</tr>
<tr>
<td>Microsatellite polymorphisms</td>
<td>24</td>
</tr>
<tr>
<td>Research aims</td>
<td>25</td>
</tr>
<tr>
<td>Summaries of papers</td>
<td>26</td>
</tr>
<tr>
<td>Paper I: Evidence for turnover of functional noncoding DNA in mammalian genome evolution</td>
<td>26</td>
</tr>
<tr>
<td>Paper II: The genomic landscape of short insertion and deletion polymorphisms in the chicken (Gallus gallus) genome: a high frequency of deletions in tandem duplicates</td>
<td>29</td>
</tr>
<tr>
<td>Paper III: Microsatellite polymorphism is elevated in human recombination hotspots</td>
<td>32</td>
</tr>
<tr>
<td>Paper IV: A genome-wide analysis of microsatellite polymorphism circumventing the ascertainment bias</td>
<td>34</td>
</tr>
<tr>
<td>Paper V: ILAPlot: An Interactive Local Alignment Plotting Tool</td>
<td>37</td>
</tr>
<tr>
<td>Summary in Swedish</td>
<td>39</td>
</tr>
<tr>
<td>Acknowledgement</td>
<td>44</td>
</tr>
<tr>
<td>References</td>
<td>45</td>
</tr>
</tbody>
</table>
### Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>BAC</td>
<td>Bacterial artificial chromosome</td>
</tr>
<tr>
<td>bp</td>
<td>Basepair(s)</td>
</tr>
<tr>
<td>C.I.</td>
<td>Confidence Interval</td>
</tr>
<tr>
<td>Gb</td>
<td>Gigabasepair</td>
</tr>
<tr>
<td>Indel</td>
<td>Insertion/Deletion</td>
</tr>
<tr>
<td>kb</td>
<td>Kilobasepair</td>
</tr>
<tr>
<td>LINE</td>
<td>Long Interspersed Nuclear Element</td>
</tr>
<tr>
<td>LTR</td>
<td>Long Terminal Repeat</td>
</tr>
<tr>
<td>Mb</td>
<td>Megabasepair</td>
</tr>
<tr>
<td>SINE</td>
<td>Short Interspersed Nuclear Element</td>
</tr>
<tr>
<td>SNP</td>
<td>Single Nucleotide Polymorphism</td>
</tr>
</tbody>
</table>
Introduction

Seeing all the biodiversity surrounding us, it is hard to refrain from asking what processes are involved in forming it. We see differences between species, but also obvious differences within species. Ultimately, it is the genome, and interactions between the genome and the environment, that will determine the phenotype. The importance of the genome has directed research interest towards understanding evolution on the molecular level. What parts are important? How does DNA sequence mutate? What forces shape the patterns of variation we see? Reading and decoding the genome reveals answers to these questions.

Since the advent of DNA sequencing the amount of available sequence data from various organisms has skyrocketed. For example, at the time of publication of the human genome (International Human Genome Sequencing Consortium 2001), there were roughly 10 gigabasepairs (Gb) of sequence deposited in the traditional divisions of the public databases GenBank, EMBL and DDBJ. Today, the same database divisions contain more than 100 Gb of sequence data (http://www.ncbi.nlm.nih.gov/Genbank/). In addition to this, there are 125 vertebrate genomes that, at time of writing this, have been sequenced to at least the state of draft assembly (http://www.ncbi.nlm.nih.gov/sites/entrez?db=genomeprj).

In addition to the genome projects several related projects produce vast amounts of data in the quest to answer various medical or evolutionary questions. A few examples of these include the HapMap project aimed at developing a detailed haplotype map of the human genome (International HapMap Consortium 2005); the NISC comparative vertebrate sequencing initiative aimed at sequencing a set of orthologous target regions in a wide range of vertebrates (Thomas et al. 2003); and the ENCODE project aimed at providing an encyclopaedia of all functional DNA elements in the human genome (The ENCODE Project Consortium 2007). The data sets produced by projects such as these provide an important resource for the evolutionary biologist, as many questions, other than the initial questions of the data producers, can be addressed. Furthermore, the availability of these data sets has prompted the development of new computerised methods for the analysis of data.

The work presented in this thesis is based on data obtained from a number of sources of large genomic data sets. I have applied different in silico methods to address evolutionary aspects of the non-coding vertebrate genome.
The organisms of interest have either been human, placental mammals or chicken. Much of the discussion in this introduction is human-centric by virtue of the human genome being one of the most well studied. However, most of the described phenomena and patterns seem to be shared among vertebrates. I will also cover bird-specific aspects, in some detail, given that special emphasis has been placed on the analysis of the non-coding chicken genome.
Comparative genomics

Vertebrate comparative genomics, as known today, came to a start when the mouse genome was sequenced, and the two genomes of human and mouse were available (International Human Genome Sequencing Consortium 2001; Mouse Genome Sequencing Consortium 2002). The term “comparative genomics” is commonly used to denote studies where full genome sequences, full gene sets or at least significant parts thereof are used (Filipski and Kumar 2005; Hardison 2003). The addition of new genomes continuously sheds light on new aspects on the evolution of the genome.

Evolutionary biologists have since Darwin used comparative methods to understand trait evolution. The application of the comparative method to whole genomes can be traced back to the advent of DNA-binding dyes, which made studies of karyotypes and chromosome banding possible. With these methods, it could be seen that, although the number and layout of chromosomes differed, the overall variation in genetic content among mammals was not as pronounced as initially hypothesised (Filipski and Kumar 2005). It became apparent that it was more a matter of reshuffling of regions within the different genomes. The term comparative genomics seems to have appeared in the early 1990’s (Charlebois and St Jean 1995; Lyons et al. 1994; Womack and Kata 1995). At that time it was mostly used in conjunction to comparative mapping of genes, and also high level comparative cytological studies like chromosome painting (Wienberg and Stanyon 1995; Womack and Kata 1995).

Comparing the appropriate sequences

An underlying assumption of comparative genomics is the homology of studied sequences. In other words, that the sequences studied have a common ancestry. More specifically, the sequences must have diverged as a result of speciation, they must be orthologous. If a sequence is duplicated within a genome, before speciation, then the two copies are said to be paralogous (Figure 1). The use of paralogous sequences in a comparative study would void any inferences on species-specific traits, as one cannot distinguish between what has occurred before or after the species-split.
During evolution, genomes undergo rearrangements like inversions, duplications and deletions on various scales (Kent et al. 2003; Ma et al. 2006). All this shuffling of genetic material increases the difficulties of determining orthology of regions in the genome. Much research and development has been spent on developing methods to address the problem of asserting orthology. I will just briefly mention two approaches to it that are important for this thesis.

The first approach is when large genomic sequences are used in comparative studies, such as in Paper I and the chicken-turkey comparisons of Paper II. Here pairwise or multiple alignments are created through the use of locally anchored alignment procedures like BLASTZ (Schwartz et al. 2003), MAVID (Bray and Pachter 2003; Bray and Pachter 2004) or MLAGAN (Brudno et al. 2003). Whole genome alignments can be done with BLASTZ (Schwartz et al. 2003) or the chaining and netting method presented by Kent et al. (2003). However, when shorter sequences, like fully sequenced BACs, are to be aligned, one needs to have some knowledge of homology, which often can be obtained through homology searches like BLAST (Altschul et al. 1997). When searching with a long query sequence against a genomic database, homologous regions can often be identified as a large number of consecutive search hits. The program presented in Paper V aids this semi-manual determination of homology by parsing and visualising the BLAST search results, and cutting out sequences of interest.

The other approach is when using large sets of short sequences, like reads from sparse shotgun sequencing for polymorphism discovery or EST sequences. The most common method is to use a fast homology search method, like BLAST/MegaBLAST (Altschul et al. 1997; Zhang et al. 2000).
or SSAHA (Ning et al. 2001) to place the sequences onto a reference genome. The reference genome may be the same species or a related species, depending on the question of interest. The obvious problem with this method is placement of sequences matching to repeated sequences like repeat elements or genes within gene families. There are two major ways of handling this problem. The first is to require the hits to be of at least a defined quality or statistical significance level. However, with only this criterion, some sequences may match to many places with similar scores, and to avoid using the incorrect placement, a second criterion is often applied. This criterion is that the best hit must have at least a predefined advantage in score to the second best hit, or that there is only a single hit. If this separation is not fulfilled, the sequence is discarded (e.g. International Chicken Polymorphism Map Consortium 2004). The other common way of handling the risk of paralogy is to use reciprocal best matches between two sets of sequences (e.g. Mouse Genome Sequencing Consortium 2002). This method ensures that the two sequences grouped as orthologs are the most similar sequences. This is a reasonable assumption, given that both sets of sequences are complete. However, if one of the sets is incomplete, there is a risk of incorrectly grouping sequences as orthologs.

Bioinformatics

Bioinformatics, the collection, management and analysis of biological data using computers, has seen a rapid development in the last decade. Two major forces have been driving this: the rapid increase in the availability of biological data, as seen from the many large scale genomics projects, and the rapid development of computer power, with roughly doubling in capacity every two years (known as Moore's law. Moore 1965). In fact, current day’s genome sequencing projects would not have been possible without computer aided data management.

The large datasets also pose computational challenges in the analysis of data. Classical computer programs for the study of molecular evolution are now mature. They often provide easy to use graphical “point and click” interfaces (eg. Rozas et al. 2003; Tamura et al. 2007). However, for comparative genomic analysis, we have not yet seen the move to desktop programs, although this will likely be the case once full genome sequencing becomes affordable for a larger number of research groups. There are some exceptions to this, notably programs to perform full genome analysis of specific phenomena, for example the microsatellite screening and analysis program SciRoKo (Kofler et al. 2007).

On the other hand, there are nowadays a large number of web-based resources for the analysis of genomic data. For genomic sequence and gene data, as used in this thesis, GenBank (http://www.ncbi.nlm.nih.gov/), En-
sembl (http://www.ensembl.org/) and University of California at Santa Cruz Genome Browser (http://genome.ucsc.edu) have been the major source for data. These web sites offer powerful ways of exporting data sets, based on various selection criteria including homology between species. They also provide ways of overlaying local data onto graphical presentations of the genome. There are also numerous other web resources focused on different aspects of comparative genomics, enabling researchers to do quite elaborate studies without having to resort to in-house programming solutions (e.g. the Vista tool set, http://genome.lbl.gov/vista/).

When there are no off-the-shelf computer programs or web-applications available for the analysis to be done, the only solution is to set up a framework or pipeline locally. It is common practice to use available programs for the various sub-tasks within an analysis. For instance, to use the PAML package (Yang 1997) to estimate divergence and testing various models, or to use various alignment or homology search programs. For other parts of the analysis custom computer programs have to be written. This is often also the case for data flow between the different parts of the pipeline, as well as data management, extraction and statistical testing.

There are many different programming languages available and suitable for bioinformatic projects. Among these, the scripting language Perl is likely the most commonly used language for programs written with the sole intent of performing a single task within a project. There are, however, several other languages available with similar properties, e.g. powerful string handling and easy writing. Python and Java are good examples. For all these languages there are extensions available to simplify many tasks, like data input and output, and parsing of similarity search results to mention a few. In the Perl case the module set is called BioPerl (Stajich et al. 2002), and for Python and Java there are BioPython (http://www.biopython.org) and BioJava (http://www.biojava.org/), respectively.

Even though most parts of a comparative genomics analysis can be performed using ordinary desktop computers, there are still many tasks that are limited by the extreme computational times required. Typically these tasks are homology searches or alignments, which have to be run on super computers or large clusters of computers.
The non-coding genome

One of the many discoveries coming through the sequencing of the human genome was the small number of genes actually found. The initial predictions, based on the sequence, pointed towards somewhere around 30,000 protein-coding genes, although the total number of gene products is considerably higher due to alternative splicing (International Human Genome Sequencing Consortium 2001). This figure reappeared at the sequencing of other vertebrate genomes, including the mouse and the chicken, both with similar number of protein-coding genes (International Chicken Genome Sequencing Consortium 2004; Mouse Genome Sequencing Consortium 2002). It seems that the gene set is rather constant among most vertebrates, although the genome size varies significantly (Gregory 2005a). The set of protein-coding genes makes up only a few percent of the genome. The rest can be divided into a number of different categories, all of which I refer to as non-coding DNA, on the basis of them not coding for proteins (Table 1) (Filipski and Kumar 2005; Gregory 2005a; Lynch 2007). I will discuss some of these DNA classes in more detail below.

Table 1. The different major classes of DNA in the human genome.

<table>
<thead>
<tr>
<th>Structure of DNA</th>
<th>DNA Class</th>
<th>Proportion of genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-repetitive</td>
<td>Protein coding sequences</td>
<td>2%</td>
</tr>
<tr>
<td></td>
<td>&quot;Conserved non-coding sequences&quot;</td>
<td>3%</td>
</tr>
<tr>
<td></td>
<td>Other unique DNA</td>
<td>46%</td>
</tr>
<tr>
<td>Repetitive</td>
<td>Interspersed repeats</td>
<td>4%</td>
</tr>
<tr>
<td></td>
<td>Tandem repeats</td>
<td>5%</td>
</tr>
</tbody>
</table>

Single copy DNA

Roughly half of the human genome can be classified as single copy non-coding DNA, i.e. unique sequence occurring only at one place in the genome (Filipski and Kumar 2005; International Human Genome Sequencing Consortium 2001). More than half of this sequence is introns (Lynch 2007). Since the discovery of introns, and even more so, when it was realised that
significant parts of the genome were intergenic sequences, there has been ongoing efforts to try to understand the role of the non-coding genome. Much of these efforts have focused on evolutionary conserved regions, as sequence conservation is often regarded an indicator of functional constraint. The recently published, and to date most elaborate study of the non-coding genome is the ENCODE project (The ENCODE Project Consortium 2007). The ENCODE project confirmed earlier results suggesting that evolutionary conserved regions provide important functions in the genome (Drake et al. 2006; Margulies et al. 2003; Woolfe et al. 2005). However, a surprising result of ENCODE is that many seemingly unconstrained elements, show evidence of being biochemically active (The ENCODE Project Consortium 2007).

Selection on the non-coding genome

Early this decade, when enough sequence data had accumulated through genome sequencing projects, evidence for widespread negative selection on the non-coding genome was presented from different interspecies comparisons (Bergman and Kreitman 2001; Loots et al. 2000; Shabalina et al. 2001). In pairwise or multispecies comparisons, regions under negative selection are expected to show a lower degree of sequence divergence, as fewer new mutations will reach fixation due to them being (slightly) deleterious. The simplicity of the method, to just scan genomic alignments for conservation, has probably contributed to its appeal.

Based on alignments of human and mouse, the proportion of the genome conserved by negative selection was estimated to about 5%. Excluding the 1.5% of protein coding sequence leaves 3.5% of the genome as non-coding sequences conserved through negative selection. Similar figures are obtained when more genomes have added to the alignments (e.g. Thomas et al. 2003).

The use of sequence conservation as criterion for detection of functional regions is not without limitations. For example, the results are much affected by the alignment methods used. This becomes especially important with sequences where the divergence approaches saturation of substitution (Bergman and Kreitman 2001; Clark 2001). Furthermore, in a two-species comparison it is not possible to distinguish if conserved regions are the result of constrained evolution, or simply natural low mutation rate. Adding a third species can to some extent resolve the issue, as it is less likely that regions in three divergent species are conserved by chance (Dubchak et al. 2000). However, this requires divergent species as variation in mutation rate correlates between species (Smith et al. 2002). Interspecies comparisons tend to produce quite noisy results, and several methods have been proposed to make the analysis more robust (Elnitski et al. 2003; Siepel et al. 2005). One has also to be cautious when choosing species to ensure sufficient diver-
gence to gain power, although there is a trade-off between power in detection and reliability of the alignments.

Later, when the chicken genome was added to the multispecies mammalian comparisons, a core set of ultra conserved regions among vertebrates could be defined. Excluding protein coding genes, this set covers about 1.25% of the human genome (International Chicken Genome Sequencing Consortium 2004).

Drake et al. (2006) used polymorphism information from the International HapMap Consortium (2005) to study polymorphism levels in conserved non-coding sequences. Derived alleles were significantly less common in these regions, indicating selective constraints, rather than reduced mutation rate (Drake et al. 2006). This and other studies indicating conservation by selection has spurred research in what the functions of these regions are.

Positive selection on non-coding sequences

With such clear evidence of negative selection in non-coding sequence, an obvious question is whether there are regions under positive selection as well? In recent years a number of studies have provided evidence for positive selection. For example Andolfatto (2005) showed that positive selection is acting on non-coding regions of in *Drosophila*. However, for hominids evidence for positive selection is still lacking, although it is possible that this could be due to lack of power in the tests for selection (Eyre-Walker 2006). Andolfatto (2005) contrasted polymorphism levels to divergence in the search for positive selection. Pollard et al. (2006), on the other hand, contrasted levels of divergence within a phylogeny to find regions conserved among mouse, rat and chimpanzee, but with accelerated divergence after the human-chimpanzee split. Their argument is that these human accelerated regions are regions of adaptive evolution in humans. However, adaptive evolution is not the only possible explanation of the human acceleration, as it has been suggested that the rapid divergence in humans might be due to biased gene conversion in recombination hot-spots (Galtier and Duret 2007).

Function of non-coding sequences

Already before the sequencing of the human genome, there was extensive knowledge of many kinds of RNA genes where the final product is not a protein. These include the many RNAs with enzymatic functions, and also involved in the translational machinery like tRNA, ribosomal RNA, snRNA (small nuclear RNA, part of the spliceosome) (reviewed in Eddy 1999). However, most of these non-coding RNA genes have been discovered through experimental methods.

The newly discovered conserved non-coding sequences has spurred research into what their potential function may be. Early studies showed that
some of the sequences had cis-regulatory functions (Bergman and Kreitman 2001; Loots et al. 2000). Other functions that has become apparent is transcription factor binding (Margulies et al. 2003) and developmental regulation (Woolfe et al. 2005). However, the ENCODE project (2007) found that 40% of the inter-species constrained bases were still un-annotated, and function not yet determined.

Repetitive sequences

About half the human genome consists of non-unique DNA sequences that are repeated in one way or the other (International Human Genome Sequencing Consortium 2001). The amount of repetitive sequence varies between genomes, for example the more compact chicken genome contains only about 15% repetitive sequence (International Chicken Genome Sequencing Consortium 2004). The repetitive sequences can be divided into different classes, which I will briefly describe below (Table 2). Particular emphasis will be devoted to microsatellites, as this specific class of sequences is one of the main topics of interest in this thesis.

Table 2. Different classes of repetitive sequences in the human genome

<table>
<thead>
<tr>
<th>Sequence type</th>
<th>Proportion of genome</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Interspersed repeats</strong></td>
<td></td>
</tr>
<tr>
<td>Long Interspersed Nuclear Elements (LINEs)</td>
<td>20%</td>
</tr>
<tr>
<td>Short Interspersed Nuclear Elements (SINEs)</td>
<td>13%</td>
</tr>
<tr>
<td>LTR retrotransposons</td>
<td>8%</td>
</tr>
<tr>
<td>DNA transposons</td>
<td>3%</td>
</tr>
<tr>
<td><strong>Tandem repeats</strong></td>
<td></td>
</tr>
<tr>
<td>Satellites</td>
<td>&lt; 1%</td>
</tr>
<tr>
<td>Minisatellites</td>
<td>&lt; 1%</td>
</tr>
<tr>
<td>Microsatellites</td>
<td>4%</td>
</tr>
</tbody>
</table>

The major dichotomy when classifying repetitive sequences is between interspersed repeats and tandem repeats. Interspersed repeats are sequences repeated on different places within the genome. The majority of the repetitive sequences in the human genome (almost 50% of the genome) are interspersed repeats (International Human Genome Sequencing Consortium 2001). Mechanisms exist for most of these sequences that, in one way or the other, can replicate the sequences within the genome. Some of the motifs can be seen as selfish DNA elements, as they, themselves, contain all functions necessary for replication, e.g. LINEs. The transposable elements can be further divided into two classes based on if the transposition includes transcription to RNA followed by reverse transcription to DNA before insertion, or if
it is purely DNA based (Finnegan 1989). The first class is the most abundant in human, but also in chicken (International Chicken Genome Sequencing Consortium 2004; International Human Genome Sequencing Consortium 2001). In the human genome, L1 LINEs (Long Interspersed Nuclear Elements) are the most common with regard to total sequence content. The L1 elements account for about 16% of the total genome sequence, whereas LINEs together account for roughly 20% of the genome. By numbers SINEs (Short Interspersed Nuclear Elements) are the most abundant, with Alu elements being most common (International Human Genome Sequencing Consortium 2001). Endogenous retroviruses also belong to this first class of transposable elements. The second class, DNA-elements, is much smaller and accounts for only about 3% of the genome (International Human Genome Sequencing Consortium 2001).

Interspersed repeats can pose great problems in homology searches in comparative analyses. Specifically, the high similarity between copies makes it impossible to assess orthology. The most common way of handling this is to mask the repetitive sequences before homology searches using programs such as RepeatMasker (Smit, A., unpublished, http://www.repeatmasker.org/). On the other hand, interspersed repeats can provide a good neutrally evolving marker for the study of regional variation in mutation patterns, where changes from the ancestral sequence of the repeat can be assessed (International Human Genome Sequencing Consortium 2001; Webster et al. 2006; Webster et al. 2005).

Tandem repeats, the other part of the dichotomy mentioned above, are sequences repeated in tandem head to tail, or head to head (inverted repeats). These repeats are mostly classified by their size, with the long (> 100 bp per unit) repeats referred to as satellite DNA (Corneo et al. 1967), a shorter group of repeats between 10 and 30 bp repeat units called minisatellites (Jeffreys et al. 1985), and the shortest repeat units, up to five or six bp called microsatellites (Litt and Luty 1989; Weber and May 1989). Satellite DNA is mostly found as tandem-repeated about 170 bp long monomers known as alpha-satellite repeats in the centromeres. Other tandem repeats are spread throughout the genome, with the telomeres as a noticeable other region rich in repeats (International Human Genome Sequencing Consortium 2001).

Microsatellites

It was early found that the occurrence of microsatellites in the genome is much higher than would be expected by pure chance (Hamada and Kakunaga 1982; Hamada et al. 1982; Tautz and Renz 1984; Tautz et al. 1986). For example, the Human Genome Sequencing Consortium (2001) estimated that 3% of the human genome comprise microsatellite sequence. However, the microsatellite content varies between species, with a positive correlation
between genome size and microsatellite number (Dieringer and Schlötterer 2003; Ellegren 2004; Toth et al. 2000).

An intriguing aspect of microsatellite occurrence is the differences in prevalence of different motifs, including differences within repeat unit length classes. For example, \( CA_n \) is the most common dinucleotide repeat motif in the human genome, accounting for roughly 50% of the dinucleotide repeats found, while GC repeats account for less than a percent (International Human Genome Sequencing Consortium 2001). Similar skews are seen for other repeat unit length classes (International Human Genome Sequencing Consortium 2001). The skewness tends to be different among different species (Dieringer and Schlötterer 2003; Toth et al. 2000).

The distinction between microsatellites and minisatellites can seem arbitrary. Satellite DNA got its name from the observation that it formed a satellite band in caesium-chloride density gradient separation experiments (Corneo et al. 1967). When a smaller class of repeats was found, it was dubbed minisatellites (Jeffreys et al. 1985), and consequently the smallest sized repeats were called microsatellites (Litt and Luty 1989; Weber and May 1989). The mutation process is distinctly different between minisatellites and microsatellites, although it for both types of repeats involves the addition or loss of repeat units. However, whereas minisatellites mutate through unequal crossing over or gene conversion at recombination (Berg et al. 2003), microsatellites mutate through replication slippage (Levinson and Gutman 1987; Schlötterer and Tautz 1992). Replication slippage involves the dissociation and shifted reassociation of the DNA strands during replication and leads to the addition or loss of repeat units.

As for many of the classes of non-coding DNA, the question whether some function can be attributed to microsatellites has received quite some attention. For instance, microsatellites within genes have been well studied, including examples of microsatellite associated disease involving trinucleotide expansions (Li et al. 2004; Lindblad and Schalling 1999). There are also examples from \textit{in vitro} studies of microsatellites being involved in gene regulation (Contente et al. 2002; Struhl 1985). Furthermore, there is an example of a microsatellite affecting social behaviour in a mole species (Hammock and Young 2004). It has also been debated whether microsatellites are involved in recombination, either as recombination signals, or with recombination being mutagenic to microsatellites. There is some experimental evidence for the latter (Hile and Eckert 2004). However, most current evidence points towards this not being the case (Morral et al. 1991), with the strongest evidence from the non-recombining Y chromosomes, which has microsatellite mutation rates similar to autosomes (Kayser et al. 2000).
Polymorphisms are generally considered to be naturally occurring variation between individuals within a population, or between populations within a species. Some polymorphisms are seen as phenotypic differences between individuals, but most are just manifested as differences in the primary DNA sequence, without a noticeable phenotypic effect, or even being selectively neutral. Intraspecific polymorphism is also denoted diversity, as opposed to interspecific divergence. The differences in the DNA sequence between individuals can occur on many scales, ranging from large-scale segmental duplications down to single bp differences.

The amount of genetic diversity seen varies on many different levels. For instance there are interspecies differences, where some species have high levels of genetic diversity while other species have low, e.g. higher in chicken than in humans (International Chicken Polymorphism Map Consortium 2004). There is also within genome variation between chromosomes (International Chicken Polymorphism Map Consortium 2004) or even at a finer levels (Berlin et al. 2006; Gaffney and Keightley 2005; International Chicken Polymorphism Map Consortium 2004; Spencer et al. 2006). There are two basic forces governing the levels of genetic diversity seen: mutation and selection. Mutation governs the emergence of new polymorphisms, whereas selection governs the fate of mutations once they occur.

New mutations occur continuously, for instance as failed repair of errors at replication of the DNA, at recombination or as lesions due to extrinsic factors. If a new mutation is neutral only genetic drift will govern its fate, and the fixation rate is determined by the effective population size. In this case, given that the effective population size is constant, the variation in the levels of genetic diversity will reflect the underlying mutation rate variation, (Kimura 1968; Li 1997). On the other hand, if there is selection on a mutation, either positive or negative, the new mutation will be more rapidly lost or fixed than under neutrality. However, in this process, linked loci will also be affected, and as a consequence the levels of polymorphism around a selected site will be reduced. In the case of a locus with a deleterious allele, selection will reduce the frequency of the allele, and as a consequence the frequency of neutral alleles at linked loci will be reduced. This process is called background selection (Charlesworth et al. 1993). In the case of positive selection on an advantageous allele, the allele will rapidly increase in frequency through positive selection. In this case neutral alleles on linked
loci will be dragged along and also increase in frequency through genetic hitch-hiking (Kaplan et al. 1989; Maynard Smith and Haigh 1974). Thus, in both cases, the levels of recombination will also affect the levels of genetic diversity seen in a population, as recombination rate determines how far from the selected site effects of background selection and hitch-hiking will be seen.

What are the consequences of the above on the levels of genetic diversity within a genome? To mention a few: Firstly, if there is variation in mutation rate we can expect to see variation in the levels of genetic diversity. This has been shown in e.g. birds and rodents (Berlin et al. 2006; Gaffney and Keightley 2005). Secondly, we can expect to see a positive correlation between levels of polymorphism and recombination rate (e.g. Hudson 1994; Nachman 2001; Spencer et al. 2006). Thirdly, effects of selection and drift should affect different types of polymorphisms similarly, i.e. the levels are expected to correlate, as shown by Mills et al. (2006).

An obvious question in extension to this is how can we distinguish between selection and mutation rate in a region of low polymorphism? The probably most common approaches are to use the Hudson-Kreitman-Aguade test of neutrality (Hudson et al. 1987) or the McDonald-Kreitman test (McDonald and Kreitman 1991). The latter was initially devised to contrast synonymous polymorphism and divergence to non-synonymous polymorphism and divergence. The latter was recently used in a more generalised framework where a neutral region is compared to a region where one would like to test if selection is acting, to infer positive selection in non-coding regions of the *Drosophila* genome (Andolfatto 2005). The problem of this approach is that one needs *a priori* knowledge of what to use as neutral reference. This is not always evident (e.g. Andolfatto 2005).

**Small scale sequence length polymorphisms**

Polymorphisms in sequence length, i.e. the insertion or removal of sequence, occur on various scales. The most commonly occurring variations are in the size-range of one to fifty bp, with a negative exponential relationship between length and occurrence (Mills et al. 2006). However, longer sequences are also involved in length variation, with lengths up to several hundred thousand bp (Bailey et al. 2002). It is likely that the longer the insertions and deletions are, the greater the risk that they are lethal or strongly deleterious, simply because they affect more DNA. However, to what extent the seen size-distribution is caused by selection or if long mutations happen more seldom, is not known. Furthermore, the mutation rate, and thus to some extent the occurrence, of length polymorphisms, is dependent on sequence context. The most obvious example of this is tandem repeated regions, like microsatellites or minisatellites. Such regions have a higher probability of
mutating than unique sequence, although the mechanism varies (Bois 2003; Ellegren 2004; Mills et al. 2006).

Short insertion and deletion polymorphisms

Short non-repetitive insertions and deletions (indels) have to date not been studied to the same detail as base-substitution. A number of studies have used comparative genomic approaches to study divergence between species through indels (Chen et al. 2007; Makova et al. 2004; Ogurtsov et al. 2004; Taylor et al. 2004; Yang et al. 2004). However, indels in alignments have mostly been regarded as a nuisance as they cause sparse columns in alignments that cannot be used in divergence estimates with methods aimed at studying substitution patterns. This can probably be traced to the fact that the patterns of indels seen in alignments are to large extent dependent on the alignment parameters (e.g. Holmes 2005). Thus, the ideal solution would be to study indels segregating in a population, where the alignments will be mostly unambiguous.

Most studies where polymorphism screening has been performed have been aimed at SNPs or microsatellites. Recently, however, two large polymorphism screenings have been performed where indels has been characterised: one in chicken (International Chicken Polymorphism Map Consortium 2004) and one in human (Mills et al. 2006). These studies have shed new light on the distribution of indels in the genome, with the perhaps most important findings being that indels are in the order of one tenth as common as SNPs and that the genomic density of indels and of SNPs correlate strongly within the human genome.

The strong correlation between the occurrence of indels and SNPs points towards that the same evolutionary forces affect them. There is evidence from in vitro studies that indels, similarly to substitutional mutations, are caused by erroneous replications of DNA (Bierne et al. 1997; Bierne and Michel 1994). However, unless the fidelity of replication with regard to indels and SNPs is affected by the same factors to create such a correlation, a more likely scenario is that population genetic effects, other than mutation rate variation, cause the correlation. This could be variation in the effects of background selection or hitchhiking due to variation in recombination rate (Charlesworth et al. 1993). Another possible explanation could be that recombination is mutagenic (Hellmann et al. 2003; Lercher and Hurst 2002).

A common result from many studies of indels is the presence of a deletion bias of up to five deletions per insertion event (Comeron and Kreitman 2000; Cooper et al. 2004; Mills et al. 2006; Neafsey and Palumbi 2003; Ophir and Graur 1997; Petrov et al. 2000; Vinogradov 2002; Zhang and Gerstein 2003). This bias has been argued to be an important aspect in the evolution of genome size, as it keeps the genome size down (Kozlowski et al. 2003; Vinogradov 1997; Vinogradov and Anatskaya 2006). However, others have
argued that such a small bias would not be capable of counteracting large-scale events such as insertions of transposable elements or segmental duplication (e.g. Gregory 2003; Gregory 2005a). Moreover, it is unlikely that the fitness effects of individual small deletions would be sufficient for natural selection to favour a mechanism that affects the indel bias.

Microsatellite polymorphisms

Microsatellites are probably the most frequently used polymorphic marker of today. This owes to a number of important properties making them attractive to work with: as said earlier, microsatellites are amply available in the genomes of most eukaryotes; it is comparably easy to design new markers without the need for full genomic sequence, e.g. through enrichment libraries; microsatellites are highly informative as a single locus can have many alleles, not only two as for SNPs; and, typing of microsatellites can be done through simple fragment length separation (see e.g. Buschiazzo and Gemmell 2006; Ellegren 2004).

To use microsatellites in population genetic contexts to infer, for instance, population separation, mutational models describing mutational patterns are required. Many models for microsatellite mutation have been proposed to match the underlying mutational mechanism including the commonly used stepwise mutation model and derivates of it (reviewed in Buschiazzo and Gemmell 2006). However, to study microsatellite mutation using available microsatellite markers has been difficult. The major reason for this is that most markers available from population genetic studies have passed through a screening process where high heterozygosity has been a selection criterion. Thus there is likely to be an ascertainment bias introduced. Any description of variation within and between genomes using such microsatellites will be inherently biased. While keeping this in mind, these markers have shown evidence of very high mutation rates, in the order of one mutation per 1000 generations, although this varies between loci (Ellegren 2000).

Another approach to study microsatellite mutations is to analyse de novo mutations seen in pedigrees (Primmer et al. 1996; Weber and Wong 1993), which has showed that longer microsatellites are more mutable than shorter. This finding agrees with the general practice of choosing long microsatellites when screening microsatellites for polymorphic loci to be used as markers. Furthermore, there seems to be a bias towards mutations causing increases in size rather than decreases (Brinkmann et al. 1998; Kayser et al. 2000; Wierdl et al. 1997). The consequences of such a continuous growth are quite obvious. There seems, however, to be a limit on microsatellite length, as most microsatellites do not exceed a certain length. It is likely that this is due to the accumulation of point mutations which stabilises the microsatellite by dividing it into two smaller microsatellites and thereby stops its growth (Calabrese et al. 2001; Kruglyak et al. 1998).
Research aims

The general aim of this thesis has been to study the evolution of the non-coding genome. More specifically to use bioinformatic approaches to explore:

- the extent of negative selection on vertebrate non-coding sequences.
- properties and evolution of short insertion and deletion mutations in the chicken genome.
- effects of recombination on microsatellite polymorphisms.
- properties and evolution of microsatellites in the chicken genome.
Summaries of papers

**Paper I: Evidence for turnover of functional noncoding DNA in mammalian genome evolution**

One of the most important questions regarding the evolution of the non-coding genome is to what extent it is functional and constrained by negative selection. The most common approach to this has been to use multiple species comparisons and infer function from the conservation of sequences (Elnitski et al. 2003; Glazko et al. 2003; Ludwig 2002). However, there are aspects that potentially can render the proportion conserved sequence (PC) an inaccurate estimate of the proportion under negative selection (PN). It is well known that there is strong regional variation in mutation rates within genomes (Ellegren et al. 2003; Silva and Kondrashov 2002; Smith et al. 2002; Wolfe et al. 1989). This variation will lead to that some regions found to be conserved are just a result of low mutation rate (Ellegren et al. 2003; Hare and Palumbi 2003). Thus one can divide PC into two fractions: sequence conserved through negative selection (PCN) and sequence conserved through variation in mutation rate (PCM).

The use of PCN is problematic for other reasons too. Using PCN in comparisons of distantly related species to estimate PN will not estimate the current day PN. Rather, it will estimate the shared proportion of the genome that has been under negative selection among the studied species. There is evidence for a high rate of turnover of regulatory elements (Dermitzakis and Clark 2002; Ludwig 2002; Ludwig et al. 2000; Ohta 2002), and in light of this, it is unlikely that the patterns of negative selection on non-coding sequences have been constant over such time spans. Thus, PCN will underestimate PN, if there is turnover. For instance, based on whole genome alignments of humans and rodents, the estimated proportion of the genome being conserved and non-coding is 3.5% (Gibbs et al. 2004; Mouse Genome Sequencing Consortium 2002). However, if there is an independent turnover of 50% of the constrained non-coding sequences in the branches to the common ancestor of rodents and humans, then PN could be as much as 14%.
In this paper we investigated the turnover of sequences under selective constraint, using a dataset of sequences from homologous regions of eight mammals.

Results and discussion

We used sequence data from eight mammalian species previously reported by Thomas et al. (2003). The sequences were aligned with either MAVID (Bray and Pachter 2003) or MLAGAN (Brudno et al. 2003), as different alignment methods affect the patterns of sequence divergence and conservation seen. In general both alignment programs produce well-behaved alignments, although differences in gap-placement yields slightly different divergence estimates.

To estimate PCM we took a simulation approach similar to (Hare and Palumbi 2003). Simulation of mutation rate variation is complicated by the fact that mutation rates varies at different levels, with most variation found on the bp and kb levels (Silva and Kondrashov 2002). To account for this we used a model with mutation rate variation at these two levels. We used the gamma distribution to model the different parameters, as a single shape parameter, commonly called alpha, allows the distribution to take different forms. The small scale rate variation was modelled as an among-site gamma distribution, while the regional variation was modelled with an among region gamma distribution. We estimated the alpha parameter of the among-site variation from the genomic alignments, and the among-region alpha parameter from human-baboon alignments. By simulating multiple alignments, the alpha-parameters were tuned to 6 and 25, for among-site and among-region variation, respectively. As this simulation is tuned to real sequences, it will rather overestimate PCM than underestimate it, as conservation seen in the real sequences is the sum of PCM and PCN.

A simple method to define conserved regions in a pairwise alignment is to identify regions with a given minimum size and level of conservation. This approach is, obviously, highly dependent on the parameters chosen. However, as our primary intent was to study the turnover of functional noncoding DNA, we chose to tune our detection to a given benchmark of PCN = 1% between mouse and human. Using a 50 bp long sliding window, where we varied the threshold of conservation required, and also simulations of PCM, we found 90% conservation to be the threshold where PCN was 1%.

The basic idea of the test for turnover of functional elements is that PCN will be negatively correlated to pairwise divergence, in the presence of turnover. This can be tested using all the 21 pairwise comparisons among the 8 species in our alignment. We acknowledge that all pairs are not fully independent as they share some ancestry. This should not, however, be critical to the test of turnover. Coding sequences was removed from the alignments, and PCN was estimated as described, as well as divergence, K, which was estimated using PAML (Yang 1997). If we assume that the turnover of func-
The proportion of noncoding elements is proportional to the divergence, then we expect a simple negative exponential model to describe the relationship of PCN and K,

$$PCN = PNe^{-BK}$$

where B is the proportionality constant. Taking the natural logarithm of both sides gives

$$\ln(PCN) = \ln(PN) - BK,$$

which can be used to fit a linear regression, as \(\ln(PCN)\) and K are available from the pairwise comparisons (Figure 2). Based on this model \(\ln(PN)\) is -2.3, corresponding to a PN of 10%. This figure should be treated with caution, though, as the y-axis intercept is uncertain. However, the qualitative finding that there is turnover is rather more certain. The finding of turnover is robust to choice of alignment program, as indicated when we used MLAGAN instead of MAVID to align the sequences.

The results presented here unveil a potential problem of current comparative methods of finding functionally constrained sites. Cooper et al. (2003) suggest that a broad mammalian sample would suffice to detect selection at the single bp level. Their method relies on the assumption of no turnover, and will never be capable of detecting individual bases under constraint, if there is turnover. However, sequences under extreme constraints will likely still be detectable. A more viable alternative might be to use many closely related species instead (Boffelli et al. 2003).
We emphasise the need for future work to confirm the results presented here, especially in terms of a wider sample of genomic regions. Furthermore, the addition of more species could provide the possibility of selecting fully independent species pairs, which would benefit the statistical analysis.

**Paper II: The genomic landscape of short insertion and deletion polymorphisms in the chicken (*Gallus gallus*) genome: a high frequency of deletions in tandem duplicates**

Despite the fact that indels contribute significantly to the divergence between species (Britten 2002; Britten et al. 2003; Chimpanzee Sequencing and Analysis Consortium 2005), they have received relatively little attention. One reason for this could potentially be that indels are a rather heterogeneous class of mutations, spanning from small events encompassing one or a few bp, up to large segmental duplications and deletions. Some of these types have been investigated to varying detail, such as retrotransposons (Price et al. 2004), minisatellites (Bois 2003), microsatellites (Ellegren 2004) and segmental duplications (Samonte and Eichler 2002).

Small mutation events, encompassing a few bp of non-repetitive nature have been studied using comparative genomics approaches (Makova et al. 2004; Ogurtsov et al. 2004; Taylor et al. 2004; Yang et al. 2004). However, methodological problems of aligning divergent sequences might produce alignments with ambiguous indel placements and properties (Holmes 2005). A preferable solution to this is to use alignments of closely related species, or even better, intraspecific detection of polymorphism (e.g. Mills et al. 2006).

There are a number of reasons to gather more information on the evolution of indels: 1) Indels are likely to represent an important source of phenotypic variation (Chen et al. 2005a; Chen et al. 2005b). 2) Indels have been recognised as an important determinator of genome size (Gregory 2005b). 3) Analysis of indels might reveal constraints in e.g. regulatory regions with, at least in part, length dependence rather than sequence dependence (Ometto et al. 2005). 4) Indels might be used as unique event markers for phylogenetic reconstruction, thus avoiding some of the problems of homoplasy and convergence (Fain and Houde 2004; Hamilton et al. 2003; Kawakita et al. 2003; Müller 2006).

**Results and discussion**

We used a set of 274,000 indel polymorphisms provided by the International Chicken Polymorphism Map Consortium (2004). These length polymorphisms consist of both length variants in unique sequence, as well as in re-
petitive sequence, i.e. microsatellites. The former group was the focus of the study, and thus we removed all indels in repetitive sequence, where the longer allele was repeated three or more times. The resulting dataset used in all analysis consisted of 140,484 indels.

The size distribution of indels follows a negative exponential distribution, with 1bp events being by far most common and with the average length 3.6 bp. We found a total indel density of roughly one indel per 5 kb, which is 5% of the SNP rate (International Chicken Polymorphism Map Consortium 2004). In general the patterns of indels are similar to what has been seen in other species (Bhangale et al. 2005; Chimpanzee Sequencing and Analysis Consortium 2005; Mills et al. 2006; Petrov et al. 2000; Zhang and Gerstein 2003). For 2 bp to 5 bp indel words, we found significant deviations from the expected frequencies. For instance, AT and AG were overrepresented, while GC was underrepresented. In general, AT-rich motifs were overrepresented.

We made several observations indicating that replication slippage is an important mechanism for indel formation. Indels within duplicated words were significantly overrepresented among the polymorphic indels. Moreover, using turkey as outgroup, we were able to infer the direction of mutation for 569 indel events. Among these, we saw a strong overrepresentation of deletions in duplicate words, whereas this overrepresentation was not nearly as pronounced for insertions. Furthermore, the above mentioned overrepresentation of AT-rich motifs is compatible with the weaker association between the DNA strands, which possibly can promote temporary dissociation, i.e. the first step in replication slippage (Levinson and Gutman 1987).

It has been debated whether small indels provide a significant contribution to the evolution of genome size (reviewed in e.g. Gregory 2004; Petrov 2002a ). Using the set of indels where we could infer direction of the mutation (insertion or deletion), we found a bias of 1.4 towards deletions. This is in the lower end of what has been seen in other species (Comeron and Kreitman 2000; Cooper et al. 2004; Neafsey and Palumbi 2003; Ophir and Graur 1997; Petrov et al. 2000; Vinogradov 2002; Zhang and Gerstein 2003). It is unlikely, however, that there is selection on a modified insertion-deletion ratio, as each event involves a limited number of bp (Petrov 2002b).

We divided the chicken genome into 1 megabasepair (Mb) non-overlapping windows. The indel distribution within the chicken genome shows a significant heterogeneity, with trend of lower densities in the small micro chromosomes. The density of indels was strongly correlated to the SNP density in the windows (Figure 3). Local genomic context has been shown to have an influence on the rates of evolution and intraspecific polymorphism. This often results in correlations between several genomic parameters, including indels (Hardison et al. 2003). For instance, GC content has been shown to correlate with nucleotide substitution rate (reviewed in Ellegren et al. 2003), possibly due to a combination of biased gene conver-
We found that both indel and SNP density correlates with GC content. This could indicate that rates of both types of polymorphism are affected by a similar mechanism, manifested in GC content. However, correcting both indel density and SNP density for GC does not remove the correlation. Indeed, it could be the case that the same factors, for instance repair, replication or recombination, affect both indels and SNPs. Another possible explanation is that the correlation is due to varying levels of selection, governing the overall levels of variability. To fully disentangle the relationships, further studies are required.

Figure 3. Correlates of indel and SNP density in 1 Mb windows. A significant correlation between indel density and SNP density is seen (A). Both indel density (B) and SNP density (C) are correlated with GC level. However, correcting indel density and SNP density for GC does not remove the correlation (D).
**Paper III: Microsatellite polymorphism is elevated in human recombination hotspots**

It has been a long-standing issue whether microsatellites simply represent “junk” DNA or if some function can be attributed to them (Kashi et al. 1997; Kashi and King 2006; Li et al. 2002; Li et al. 2004). Although most researches have regarded them as neutrally evolving markers, there has recently been frequent reports on association between microsatellite alleles and various phenotypic traits (Contente et al. 2002; Curi et al. 2005; Hammock and Young 2005; Uhlemann et al. 2004), supporting the idea that microsatellites are involved in gene regulation (Hamada et al. 1984; Li et al. 2002; Naylor and Clark 1990; Santoro et al. 1984; Struhl 1985).

Another aspect of the potential function of microsatellites is evidence for an association of microsatellites and recombination. For instance, in mammals, an association between microsatellite occurrence and regions of high recombination rate has been seen (Isobe et al. 2002; Jensen-Seaman et al. 2004; Kong et al. 2002; Myers et al. 2005; Schultes and Szostak 1991; Templeton et al. 2000). However, microsatellite variation has not been found to be associated with recombination rates in humans (Huang et al. 2002; Payseur and Nachman 2000).

Recombination is concentrated to narrow hotspots of a few kb in length (Jeffreys et al. 2001; Jeffreys et al. 1998; McVean et al. 2004; Myers et al. 2005). If there is an association between microsatellites and recombination, we might expect to see the strongest signals of such an association within the hotspots. In this study we investigate the relationship between recombination and microsatellites by using data on the location of recombination hotspots and genome wide information on microsatellite occurrence and polymorphism.

**Results and discussion**

If recombination promotes microsatellite growth, or if microsatellites stimulate recombination, we should expect to see an increased microsatellite density in recombination hotspots. We mined the human genome for perfect microsatellites using a modified version of sputnik (Morgante et al. 2002) and found, indeed, that microsatellites are both longer and more common in recombination hotspots. However, the biological relevance of these differences is not obvious, as the differences are very small. In contrast to this result, microsatellites are more than twice as frequent in *Saccharomyces cerevisiae* recombination hotspots compared to non-hot regions (Bagshaw and Gemmel, unpublished). Moreover, on broad scales, the association between simple sequences and recombination is quite marked (Jensen-Seaman et al. 2004; Kong et al. 2002). A possible explanation to the contrasting results on different scales is the ephemeral nature of human hotspots (Ptak et al. 2005; Winckler et al. 2005), i.e. that they do not last long enough to exert
a strong effect on local microsatellite occurrence. On a broad scale, recombination is more constant and an association is more likely to be seen.

To analyse the relationship between recombination and microsatellite mutation we turned to the Allele FREquency Database (ALFRED) (Rajeevan et al. 2003). We obtained three mutability estimates for the 282 microsatellites we found in the database: allele span, as the difference between longest and shortest allele; number of alleles; and heterozygosity. None of these estimates deviated significantly between microsatellites in hotspots and the genomewide expectation. However, this result could be a methodological artefact. Most markers found in allele frequency databases are likely to have been developed for population screening or linkage mapping, and thus initially chosen on basis of known high heterozygosity (Selkoe and Toonen 2006). Such an ascertainment bias (see Ellegren et al. 1997; Ellegren et al. 1995) could mask an underlying difference between regions of the genome.

To overcome the problem of possibly biased markers, we turned to a set of about 44,000 length polymorphisms within tandem repeats (Mills et al. 2006). These were identified through the reanalysis of shotgun sequencing initiatives. Using this data set we saw a significant increase of microsatellite polymorphism of 14% (Table 3). There are at least three possible explanations to this increase in microsatellite polymorphism: polymorphic microsatellites promote recombination; recombination is mutagenic to microsatellites; and/or recombination reduces the effect of selection at linked sites.

Table 3. Polymorphism data for microsatellites, SNPs and indels in recombination hotspots compared to genomic expectations. Ninety-five percent confidence intervals were estimated using resampling test with 1,000 replicates.

<table>
<thead>
<tr>
<th>Genomic expectation</th>
<th>Hotspots</th>
<th>Lower 95% CI</th>
<th>Resampling median</th>
<th>Upper 95% CI</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of polymorphic microsatellites</td>
<td>15,059</td>
<td>12,172</td>
<td>12,288</td>
<td>12,363</td>
<td>p &lt; 0.001</td>
</tr>
<tr>
<td>Proportion polymorphic microsatellites</td>
<td>0.0136</td>
<td>0.0116</td>
<td>0.0119</td>
<td>0.0122</td>
<td>p &lt; 0.001</td>
</tr>
<tr>
<td>Number of SNPs</td>
<td>442,015</td>
<td>346,640</td>
<td>355,125</td>
<td>384,772</td>
<td>p &lt; 0.001</td>
</tr>
<tr>
<td>Number of indels</td>
<td>22,218</td>
<td>18,653</td>
<td>19,051</td>
<td>19,477</td>
<td>p &lt; 0.001</td>
</tr>
</tbody>
</table>

There are several examples of long microsatellites that have been identified within recombination hotspots (Cullen et al. 2002; Majewski and Ott 2000; Rana et al. 2004), and also examples of microsatellites affecting hotspot activity (Schultes and Szostak 1991). However, only a handful of hotspots have yet been thoroughly examined with regard to mechanism, and thus the involvement of microsatellites is still an open question. An interesting, yet speculative, possibility is that length variation as such promotes recombination through the decreased stability between sister chromatids (Kayser et al. 2006). On the other hand, there are several other forms of
length differences which ought to have greater impact in this respect, e.g. minisatellites and segmental duplications (Jeffreys et al. 1999; Sharp et al. 2005).

There is some experimental evidence supporting that recombination is mutagenic to microsatellites in *Escherichia coli* and *S. cerevisiae* (Gendrel et al. 2000; Hashem et al. 2004). However, there is conflicting evidence showing that recombination is not important for microsatellite mutation (Henderson and Petes 1992; Levinson and Gutman 1987). Most notable are the studies showing similar mutation rates on the non-recombining Y-chromosome as on autosomes (Ellegren 2004; Gusmao et al. 2005; Kayser et al. 2000).

The third explanation to increased levels of microsatellite polymorphisms within recombination hotspots is the reduced effects of selection on linked sites. Correlations of genetic diversity and recombination rate have been seen in many species (Andolfatto and Przeworski 2001; Aquadro et al. 1994; Begun and Aquadro 1992; Nachman 1997; Nachman et al. 1998). Such correlation could be due to background selection (Charlesworth 1994; Charlesworth et al. 1995; Hudson 1994; Hudson and Kaplan 1995; Nordborg et al. 1996) and/or genetic hitchhiking (Kaplan et al. 1989; Maynard Smith and Haigh 1974). However, previous studies have not found any association between local recombination rate and microsatellite diversity (Huang et al. 2002; Payseur and Nachman 2000; Thomas et al. 2005; Yu et al. 2001). In contrast to this, we found a rather strong association between microsatellite diversity and recombination.

If we assume that this selectionist explanation is correct, we would expect other kinds of polymorphism to follow the same patterns. To explore this, the same data set was used to look at non-microsatellite indels as well as SNPs. Also for these kinds of polymorphisms we saw significant increases in polymorphisms, with similar magnitudes as for microsatellites (Table 3). This result supports an explanation where the increased levels of recombination in hotspots decouples selection, thus allowing increased levels of polymorphism. However, it does not rule out the alternative explanation of recombination being mutagenic to all three kinds of polymorphisms, SNPs, indels and microsatellites.

**Paper IV:** A genome-wide analysis of microsatellite polymorphism circumventing the ascertainment bias

A number of models have been suggested to describe the mutational patterns of microsatellites. These range from the application of the simple stepwise mutation model to much more complex models involving various parame-
ters, including point mutations and length dependent mutation patterns (Bell 1996; Bell and Jurka 1997; Cornuet et al. 2006; Di Rienzo et al. 1994; Dieringer and Schlötterer 2003; Kruglyak et al. 1998). To test the validity of these models a number of empirical approaches have been taken. For instance, direct observations of de novo mutations in pedigrees (Ellegren 2000; Huang et al. 2002; Weber and Wong 1993). However, any approaches based on already characterised microsatellite markers will suffer from ascertainment bias. This is due to the way researchers have initially selected the markers, a process where several screening steps are involved and aimed at finding markers with high levels of heterozygosity and many alleles. Such markers are likely biased towards long microsatellites with high mutation rates (Pardi et al. 2005). Whole genome sequence surveys provide a means of testing microsatellite distributions against theoretical models (Dieringer and Schlötterer 2003). This approach will, however, provide a snapshot of the current microsatellite distribution, rather than capturing the ongoing processes of microsatellite evolution. This can to some extent be remedied by comparing two or more genomes (Sainudiin et al. 2004). However, for an accurate description of the ongoing processes, unbiased mutation data is needed.

Some of the genome sequencing projects have been complemented with separate screenings for polymorphisms, through re-sequencing or light shotgun sequencing of a number of unrelated individuals (International HapMap Consortium 2005; Lindblad-Toh et al. 2005). One of the largest efforts of this kind is the light shotgun sequencing of three unrelated domestic chicken strains (0.25 X coverage each), done in conjunction to the sequencing of the chicken genome (International Chicken Genome Sequencing Consortium 2004). The traces generated from these projects provide an opportunity to study microsatellite polymorphism without ascertainment bias.

In this study we use the traces from the chicken polymorphism project to obtain an unbiased set of microsatellites. An important aspect of the data is that in addition to polymorphism information, we also get information on monomorphic loci.

**Results and discussion**

We identified 390,000 perfect di- through pentanucleotide repeats of at least 3 repeat-units, where two chromosomes were sampled. Among these, 1.8% were polymorphic. Microsatellite mutation has previously been shown to be length dependent, with higher heterozygosity seen for longer microsatellites (Brinkmann et al. 1998; Ellegren 2000). We found a clear pattern of longer microsatellites being more polymorphic (Figure 4). To examine what factors affect the mutability of microsatellites in more detail we fitted a logistic regression model with presence of polymorphism as a function of microsatellite length, length of repeat unit, repeat unit motif, repeat unit GC-content.
We found a significant effect of all parameters. In general, longer microsatellites are more polymorphic, and so are AT-rich microsatellites.

Figure 4. Microsatellite heterozygosity for different repeat lengths. A) all motifs and B) dinucleotide microsatellites.

When microsatellites are typed using fragment length separation methods, there is a risk for size homoplasy, i.e. two alleles have the same size, but different lengths. We noted that 6% of the microsatellites showed no length difference, but a sequence difference. Furthermore, adding presence of imperfections to the logistic regression model showed that microsatellites with imperfections need to be longer to be polymorphic. This fits well with models of microsatellite mutation suggesting that point mutations in microsatellites will have a stabilising effect (Bell 1996; Bell and Jurka 1997; Kruglyak et al. 1998).

The patterns of polymorphism are heterogeneous among chromosomes, and also among microsatellite motifs. As we have previously shown a strong correlation between indel polymorphisms and SNPs (Paper II), we wanted to see if microsatellites follows the same patterns. We studied the variation in microsatellite variability across the chicken genome in 1 Mb windows, in relation to a number of contextual features of the windows. Microsatellite polymorphism was found to correlate strongly with both SNP and indel density. The reasons for such a correlation can be extensively discussed. Either similar mechanisms affect the mutation rates for all three kinds of polymorphisms, or the correlation may be a result of varying levels of selection. The latter would be mediated by background selection (Charlesworth 1994; Charlesworth et al. 1995) or selective sweeps (Maynard Smith and Haigh 1974).

We studied the patterns of microsatellite occurrence in 1 Mb windows in the genome. We found significant variation in microsatellite occurrence among chromosomes. This is expected as G+C-content varies among the chromosomes (International Chicken Genome Sequencing Consortium...
2004) and base composition is important in determining microsatellite occurrence (Dieringer and Schlötterer 2003; Toth et al. 2000). A more interesting result was that microsatellite occurrence is negatively correlated with SNP density. Although the explanatory value was rather low, this gives an important clue to the mutational mechanisms of microsatellites. Within-genome variation in mutation rates has been seen in a number of organisms (Berlin et al. 2006; Gaffney and Keightley 2005). Furthermore, point mutations have been suggested to be important in determining the fate of microsatellites (Bell 1996; Bell and Jurka 1997; Kruglyak et al. 1998). Thus, our observation supports the importance of point mutations stopping microsatellite growth, if some of the variation in SNP density can be explained by mutation rate variation.

**Paper V: ILAPlot: An Interactive Local Alignment Plotting Tool**

There are a number of computer programs and web services available for plotting and exploration of results from homology search programs. A few examples are NIH BLAST viewer (www.ncbi.nlm.nih.gov/blast), PipMaker (Schwartz et al. 2000), VISTA (Dubchak et al. 2000; Mayor et al. 2000), Theatre (Edwards et al. 2003), SimiTri (Parkinson and Blaxter 2003), DiagHunter/GenoPix2D (Cannon et al. 2003), Dotter (Sonnhammer and Durbin 1995), Cross (http://comb5-156.umbi.umd.edu/poster/Cross/) and Laj (Wilson et al. 2001).

In the initial steps of comparative studies when, for instance, a fully sequenced BAC is to be aligned to a genome of another organism, BLAST (Altschul et al. 1997) can be used to find the correct placement. When plotting the BLAST result like a dot-plot, a diagonal will be formed by hits spanning the region where the BAC aligns. However, few programs allow the user to export data aided by the plot. Laj (Wilson et al. 2001) provides a means of exporting sequence data, but only the alignments of the individual hits.

**Results and discussion**

To overcome the limitations of the available programs, ILAPlot was written. ILAPlot parses BLAST results and plots them similarly to a dot-plot (Figure 5). The program then lets the user explore the hits by zooming and selecting hits of interest. The most important functions of the program are provided by the different ways of exporting data aided by the selection of hits. There are four ways of exporting data provided by the program: 1) Export a list of start and end coordinates of all selected hits. 2) Export FASTA-formatted alignments of all selected hits. 3) Export coordinates of the region spanned by
two selected hits. 4) Export FASTA formatted sequence files with the sub-
selections of the two sequences spanned by the selected hits. The fourth
method allows for an easy way to select the part of a long sequence (e.g. a
fully sequenced BAC) that matches a specific part of another long sequence
(e.g. a chromosome) by just selecting the first and last hit of a diagonal in the
plot. The sequences can later be rigorously aligned using a suitable align-
ment program.

ILAPlot was used successfully for this purpose in Paper II in this thesis,
where it reduced the time required for preparation of data drastically com-
pared to manual parsing of BLAST results to find regions of orthology.

Figure 5. The plotting window of ILAPlot showing the diagonal formed by BLAST
hits between two long sequences.

Möjligheten att kunna avläsa DNA-sekvensen i en organism var ett viktigt steg för att öka förståelsen av hur arvsmassan evolverar. Tekniken för detta har rusat framåt. För sex år sedan hade man avläst människans arvsmassa. Detta gav en mängd nya insikter i hur informationen i arvsmassan är uppbyggd. En av de viktigaste upptäckterna var att endast omkring 2,5% av arvsmassan utgörs av proteinkodande gener, och att dessa var endast drygt 20 000 till antalet. Tidigare hade man antagit att stora delar av arvsmassan bestod av proteinkodande gener och att det kunde vara uppåt 100 000 gener. Idag är ett flertal arters arvsmassa avläst. Vi kan därmed jämföra deras DNA-sekvenser och utifrån skillnader och likheter dra slutsatser om vilka krafter som genom den evolutionära historien varit med och format genomen som vi ser idag.

När mängden data växer behövs nya effektiva verktyg och metoder för att hantera den. Dessa metoder och verktyg finns i nya och kraftfulla datorer och program, tekniker som även de utvecklas i ett högt tempo. Tillämpningen av datorer för att hantera och analysera biologiska data brukar kallas bioinformatik och är den teknik som används i de olika delarbetena som ingår i den här avhandlingen.

När bara 2,5% av människans arvsmassa är proteinkodande gener, vad är då resten, den ickekodande delen av arvsmassan? Efterhand som mer DNA-sekvens blev avläst visade det sig att större delen av arvsmassan bestod av olika former av upprepade sekvenser. Den största gruppen är upprepade, men utspridda DNA-sekvenser som på ett eller annat sätt kan kopiera in sig själva på nya platser i arvsmassan. Ett exempel på sådana sekvenser är de så
kallade LINE-elementen. De består av en uppsättning gener med allt som behövs för att kopiera sig självt till en ny plats.

En annan kategori av det ickekodande genomet är mikrosatelliterna. De är tandem-upprepade sekvenser av 1-5 baser långa sekvensmotiv, som till exempel ATATATATAT. Det intressanta med den här gruppen sekvenser är deras mutationsprocess. Den vanligaste mutationen är att det tillkommer eller försvinner en eller flera upprepningar och det händer kanske tusen gånger oftare än mutationer som orsakar enbasparsskillnader. Mikrosatelliter utgör i storleksordningen några procent av de flesta högre ryggradsdjurs arvsmassa.


Det finns en risk att siffran 3,5% är en underskattning av vilken andel av arvsmassen som utsätts för negativ selektion. Om det förekommer omsättning av regioner som utsätts för negativ selektion, kan det i själva verket vara en betydligt större andel. Låt oss anta att 50% av de sekvenser som utsätts för negativ selektion i mus har bytts ut sedan den gemensamma anfådern med människa. Anta också att samma sak skett i grenen till människa. Detta skulle innebära att hela 14% av arvsmassan utsätts för negativ selektion. I så fall är 3,5% en underskattning.

I delarbete I i den här avhandlingen studerade vi DNA-sekvensen i en inpassning (alignment) av nästan 2 megabaser sekvens från åtta däggdjur. Om det förekommer omsättning av sekvenser utsatta för negativ selektion skulle det kunna beskrivas med en modell där andelen bevarat ickekodande DNA avtar exponentiellt över tid. Vi tog alla parvisa jämförelser i inpassningen och avsatte andelen bevarat ickekodande DNA mot tid sedan arterna skiljdes åt. Vi fann ett signifikant samband och extrapolera vi till idag ger det att drygt 10% av arvsmassan utsätts för negativ selektion. Eftersom det är en extrapolation på en logaritmisk skala är uppskatningen av andelen DNA som utsätts för negativ selektion osäker. Däremot är det med detta resultat tämligen klart att det förekommer omsättning av vilka sekvenser som utsätts för negativ selektion.


mikrosatelliter finns det en sådan möjlighet. Det är främst i in vitro-försök man sett tecken på det. Å andra sidan finns det andra försök som visat att mikrosatelliter kan vara en rekombinationssignal. I människa sker de flesta rekombinationshändelserna i korta regioner, ofta några enstaka kilobaser långa som åtskiljs av långa regioner med låg rekombination. Vi jämförde dessa aktiva regioner med övriga genomet och fann en viss ökning i mikrosatellittäthet, men även en 14%-ig ökning av antalet mikrosatellitpolymorfi-
er. Vi noterade också en motsvarande ökning av polymorfigraden för längd-
variationer som inte är del av mikrosatelliter och enbasparspolymorfier. Det
senare talar för att effekter av selektion i kombination med rekombination
svarar för den variation vi ser, snarare än att rekombination är mutagent. Den
kanske främsta orsaken är att den ökade rekombinationen minskar effekten
av genetisk koppling, så att man ser en ökad variation. En annan möjlighet är
att någon gemensam faktor påverkar mutationstakten för alla tre typer av
polymorfiir i mikrosatelliter, längdvariationer och enbasparspolymorfier.

I delarbete IV har vi undersökt förekomsten och variationsgraden för mik-
rosatelliter i kyckling-genomet. Även för den här studien använde vi rådata
från samma glesa slumpmässiga sekvensering som för de icke-repetitiva
längdvariationerna. Den hittills vanligaste metoden för att studera mikrosa-
tellitevolution har varit att se på mutationsmönster i redan kända mikrosatel-
litmarkörer. Dessa har ofta en gång i tiden karaktäriserats för att användas i
kopplingsanalyser eller populationsgenetiska studier. Ett viktigt kriterium för
dessa användningsområden är att markörer har ett högt informationsinnehåll,
något de har om de har en hög mutationshastighet och därmed många varian-
ter. Detta medför en uppenbar risk att de slutsatser man drar om mikrosatel-
litmutation och evolution endast beskriver en liten delmängd mikrosatelliter
med hög mutationshastighet och inte mikrosatelliter i allmänhet. Vårt tillvä-
gångssätt medför att vi minimerar risken för felrepresentation i dataupp-
sättningen. Vi fann att längden av en mikrosatellit starkt påverkar dess po-
lymorfgrad. Långa mikrosatelliter är mer polymorfa än korta. Det tycks vara
antalet upprepningar i en mikrosatellit snarare än längden i baspar som är
avgörande för mutationsbenägenheten. Dessutom är AT-rika mikrosatelliter
mer polymorfa än GC-rika. Dessa resultat stämmer väl överens med teorin
om att mikrosatelliter muterar genom att polymeraset halkar framåt eller
bakåt längs med den DNA-sträng som replikeras och därmed lägger till eller
tar bort en enhet. I kyckling fann vi också ett starkt samband mellan poly-
morfigraden för mikrosatelliter, längdvariationer och enbasparspolymorfier.
Vi såg att polymorfityperna samvarierade i megabaspar-stora avsnitt av ge-
nomet. Detta stärker resultatet från delarbete III, att en stor del av variatio-
nen i polymorfigrad beror på effekten av selektion och rekombination i
kombination.

Det sista delarbetet i denna avhandling (delarbete V) beskriver ett pro-
gram, ILAPlot, som visualiserar resultaten från homologisökningsprogram-
met BLAST. BLAST används ofta för att hitta regioner som är snarliga ge-
nom sin härstamning. ILAPlot markerar alla träffar mellan två sekvenser på en tvådimensionell rityta där $x$ och $y$-axlarna är koordinater inom de två jämförda sekvenserna. Om de två sekvenserna har stora likheter kommer det att visa sig som en diagonal på ritytan. Programmet har funktioner för att klippa ut delsekvenser ur de ursprungliga sekvenserna baserat på vilka träffar man markerat på ritytan. Denna möjlighet underlättar studier där delsekvenser ur till exempel hela kromosomer ska klippas ut.

Resultaten som presenterats i den här avhandlingen bidrar till en ökad förståelse för vilka processer som är med och formar arvsmassan.
Acknowledgement

I know most of you will read this part of my thesis among the first things you read… Feel free to do so.

There are a number of people to whom I would like to express my gratitude:

Hans Ellegren, for accepting me as a Ph D student and letting me be part of your research group. It is a very inspiring environment with ambitious projects. I have learnt a lot during these years.

Erik Axelsson, for many interesting discussions on research (and other things). The best way to learn is to discuss and test new ideas.

Sofia Berlin, for good discussions on various topics, and also for good comments on parts of this thesis.

Nick Smith, who left us too early, for asking me what my null hypothesis is. I always keep that in mind now.

Niclas Backström – no, I have not caught any pike this weekend.

Matthew Webster, Judith Mank, Niclas Backström, Erik Axelsson, Sofia Berlin, Lina Hultin-Rosenberg, and other (former) members of the molecular evolution research group. You have provided an inspiring environment for me to be a part of with many exciting discussions of new and ongoing projects. It is a shame we couldn’t clone ourselves to pursue some more of the new ideas that always pop up.

My former room-mates, Cia Anderung, Erik Axelsson and Frank Hailer. Odd that the box for bad jokes never was filled. It must imply the jokes were good…

All other people that are part of the department of Evolutionary Biology. It has been a great environment to work in.

Louise Durling, for good comments on my texts, but also for telling me that I have to do other things than work in my spare time.

My parents, for always believing in me. I have a feeling that you were fairly sure I would end up like this. Hopefully I will have some more spare time now to fix some of the stuff I have intended to.

The Sven and Lilly Lawskis fond and Stiftelsen B. von Beskows fond are gratefully acknowledged for financial support.

Now when you have read this, there are a lot of more interesting things to be read in this book…
References


Berlin D. Bessette B. Blitshteyn T. Bloom J. Blye L. Boguslavskiy
C. Bonnet B. Boughalter A. Brown P. Cahiil N. Calixte J.
Camarata Y. Cheshatsang J. Chu M. Citroen A. Collymple P. Cooke
Dooley P. Dorje K. Dorjee L. Dorris N. Duffey A. Dupes O. Egbi-
remolen R. Elong J. Falk A. Farina S. Faro D. Ferguson P. Ferreira
S. Fisher M. FitzGerald K. Foley C. Foley A. Franke D. Friedrich D.
Gage M. Garber G. Gearin G. Giannoukos T. Goode A. Goyette J.
Graham E. Grandbois K. Gyaltsen N. Hafez D. Hagopian B. Hagos
J. Hall C. Healy R. Hegarty T. Honan A. Horn N. Houd E. Hughes
L. Hunnicutt M. Husby B. Jester C. Jones A. Kamat B. Kanga C.
Kells D. Khazanovich A.C. Kieu P. Kisner M. Kumar K. Lance T.
Landers M. Lara W. Lee J.P. Leger N. Lennon L. Leuper S. LeVine
J. Liu X. Liu Y. Lokyitsang T. Lokyitsang A. Lui J. Macdonald J.
Major R. Marabella K. Maru C. Matthews S. McDonough T. Mehta
J. Meldrim A. Melnikov L. Meneus A. Mihalev T. Mihova K. Miller
R. Mittelman V. Mlenga L. Mulrain G. Munson A. Navidi J. Naylor
Norbu N. Novod T. Nyima P. Olandt B. O'Neill K. O'Neill S. Osman
L. Oyono C. Patti D. Perrin P. Phunkhang F. Pierre M. Priest A.
Rachupka S. Raghuraman R. Rameau V. Ray C. Raymond F. Rege
M. Sheehan N. Sherpa J. Shi D. Shih J. Sloan C. Smith T. Sparrow
J. Stalker N. Stange-Thomann S. Stavropoulos C. Stone S. Stone S.
Sykes P. Tchuinga P. Tenzing S. Tesfaye D. Thoultutsang Y. Thou-
lutsang K. Topham I. Topping T. Tsamla H. Vassiliev V. Venkata-
raman A. Vo T. Wangchuk T. Wangdi M. Weiand J. Wilkinson A.
Wilson S. Yadav S. Yang X. Yang G. Young Q. Yu J. Zainoun L.
Zembek A. Zimmer and E.S. Lander. 2005. Genome sequence, com-
parative analysis and haplotype structure of the domestic dog. 

vitro amplification of a dinucleotide repeat within the cardiac muscle

Loots, G.G., R.M. Locksley, C.M. Blankespoor, Z.E. Wang, W. Miller, E.M.
Rubin, and K.A. Frazer. 2000. Identification of a coordinate regula-
tor of interleukins 4, 13, and 5 by cross-species sequence compari-
sions. _Science_ 288: 136-140.

Genet Dev_ 12: 634-639.

for stabilizing selection in a eukaryotic enhancer element. _Nature_
403: 564-567.

MA.

nomics: The next generation. _Animal Biotechnology_ 5: 103-111.


Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 348

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-8240